Universiteit Gent
Faculteit Ingenieurswetenschappen en Architectuur
Vakgroep Telecommunicatie en Informatieverwerking

Numerieke methoden voor wachtrijen met gezamenlijke bediening

Numerical Methods for Queues with Shared Service

Eline De Cuypere

Proefschrift tot het bekomen van de graad van
Doctor in de Ingenieurswetenschappen
Academiejaar 2014-2015

Universiteit Gent
Faculteit Ingenieurswetenschappen en Architectuur
Vakgroep Telecommunicatie en Informatieverwerking

Promotoren: Prof. dr. ing. Dieter Fiems
Dr. ir. Koen De Turck

Universiteit Gent
Faculteit Ingenieurswetenschappen en Architectuur

Vakgroep Telecommunicatie en Informatieverwerking
Sint-Pietersnieuwstraat 41, B-9000 Gent, België

Proefschrift tot het behalen van de graad van
Doctor in de Ingenieurswetenschappen
Academiejaar 2014-2015

Dankwoord

Hierbij wens ik mijn dank te betuigen aan iedereen die rechstreeks of onrechstreeks bijgedragen heeft bij het verwezenlijken van mijn doctoraat. Eerst en vooral wil ik mijn promotoren Dieter Fiems en Koen De Turck van harte bedanken voor hun uitstekende begeleiding, enthusiasme, inspiratie, steun en vertrouwen. Ik had me geen betere begeleiding kunnen voorstellen. Bedankt aan beide voor deze onvergetelijke samenwerking! Mijn dank gaat ook uit naar Herwig Bruneel om me de kans te geven om onderzoek te verrichten in de vakgroep SMACS.

Ik kijk met veel plezier terug naar mijn doctoraatsperiode, wat niet in de laatste plaats te danken is aan de goede sfeer op het bureau 1.18. In het bijzonder wil ik hiervoor mijn mede bureau-genoten Katia, Marco, Marijn, Patrick, Sofian en Thijs en ex-bureaugenoten Alessandro, Frederick, Koen, Luigi en Wouter bedanken. Ik hou mooie herinneringen over aan onze 'choco loco events', comedy avonden en glühweinsamenkomsten. Ook wens ik mijn SMACS collega's te bedanken voor de leerrijke en aangename momenten zowel op het werk als op conferenties.

Lieve familie, schoonfamilie en vrienden, jullie mogen zeker niet in dit dankwoord ontbreken. Bedankt voor jullie steun, vriendschap en liefde. Graag een speciale dank aan mijn ouders voor hun onvoorwaardelijke liefde en steun de afgelopen jaren. Bedankt mama om er altijd te zijn voor mij en me de nodige morele en emotionele steun te geven. Bedankt papa om me goede adviezen te geven op belangrijke momenten in mijn leven. Ik dank ook graag mijn broer Vincent en zijn vriendin Pauline voor de leuke momenten samen en de lekkere kookavonden, tante Dodo voor de onvergetelijke woensdag middagpauzes, mamie en tonton voor de gezellige kaartavonden, mamie voor de spannende voetbalwedstrijden en tante Ann voor de aanmoediging en hulp. Ook wil ik mijn fantastische vrienden zowel uit Vlaanderen als uit Wallonië bedanken voor alle steun, vriendschap, gezelligheid en betrokkenheid. Ik hoop in de toekomst nog veel lunch-, cinéma-, loop-, sauna-, reis-, praat- en dansplezier met jullie te hebben. Tot slot wil ik me richten tot mijn schat Lode. Bedankt om er altijd te zijn voor mij en voor je grenzeloze steun, vertrouwen en liefde.

Gent, 2014
Eline De Cuypere

Table of Contents

Nederlandse samenvatting –Dutch Summary–

Een wachtrijsysteem is een wiskundige abstractie van een situatie waarin elementen, de zogenaamde klanten, in een systeem aankomen en wachten totdat ze bediend worden. Wachtrijsystemen zijn alomtegenwoordig. Voorbeelden bij uitstek zijn mensen die aan een loket wachten om bediend te worden, vliegtuigen die wachten om op te stijgen, files tijdens de spits etc. Wachtrijtheorie is de wiskundige studie van wachtfenomenen. Aangezien vaak noch de aankomstmomenten van de klanten, noch hun bedieningstijden vooraf bekend zijn, gaat wachtrijtheorie er meestal van uit dat die stochastische variabelen zijn. Het wachtrijproces zelf is dan een stochastisch proces en meestal ook een Markovproces, gegeven dat een gepaste toestandsbeschrijving van het wachtrijproces geïntroduceerd wordt. Dit proefschrift onderzoekt numerieke oplossingsmethoden voor een bijzonder type Markoviaanse wachtrijsystemen, namelijk voor wachtrijsystemen met gezamenlijke bediening. Deze wachtrijsystemen zijn erg anders dan traditionele wachtrijsystemen doordat de klanten vooraan in de verschillende wachtrijen gelijktijdig bediend worden. Bovendien is er slechts bediening als er klanten in alle wachtrijen aanwezig zijn. De afwezigheid van bediening wanneer één van de wachtrijen leeg is veroorzaakt een bijzondere dynamiek die niet te vinden is bij traditionele wachtrijsystemen.

Deze wachtrijen met gezamenlijke bediening vormen niet alleen een mooi wiskundig studieonderwerp, maar worden ook gemotiveerd door een brede waaier aan toepassingen. De motivatie om wachtrijsystemen met gezamenlijke bediening te bestuderen kwam oorspronkelijk uit een bepaald proces in vooraadbeheer, met name kitting. Een kittingproces verzamelt de benodigde onderdelen voor een eindproduct in een speciaal ontworpen kist vooraleer het toe te leveren aan de assemblagelijn. De onderdelen en hun voorraden, zijnde de klanten en wachtrijen, hebben een “gezamenlijke bediening” daar kitting niet kan doorgaan als sommige onderdelen ontbreken. Nog steeds in het gebied van voorraadbeheer, vertoont de ontkoppelingsvoorraad van een hybride make-to-stock/make-to-order systeem een gezamenlijke bediening. Het productieproces voorafgaande aan de ontkoppelingsvoorraad is make-to-stock en wordt gedreven door vraagvoorspellingen. In tegenstelling, het productieproces na de ontkoppelingsvooraad is make-to-order en wordt gedreven door de werkelijke vraag daar de onderdelen van de ontkoppelingsvooraad samengesteld worden volgens de specificaties van de klanten. Aan het ontkoppelingspunt, is er naast de ontkoppelingsvoorraad ook een wachtrij van

openstaande orders. Daar maatwerk enkel begint wanneer de ontkoppelingsvoorraad niet leeg is en er minstens een order aanwezig is, is er opnieuw sprake van gezamenlijke bediening. Ook in telecommunicatie is er sprake van gezamenlijke bediening, in het bijzonder bij de studie van energie-oogstende sensoren. Dergelijke sensoren halen energie uit de omgeving om hun volledige energieverbruik te dekken of om hun levensduur te verlengen. Een oplaadbare batterij werkt heel gelijkaardig aan een wachtrij waarbij klanten gediscretiseerde energiepaketten zijn. Daar een sensor zowel waargenomen data als energie nodig heeft voor transmissie, kan gezamenlijke bediening opnieuw geïdentificeerd worden.

In het Markoviaanse kader, komt het "oplossen" van een wachtrijsysteem ongeveer overeen met het bepalen van de evenwichtsdistributie van het Markovproces dat het wachtrijsysteem beschrijft. Inderdaad, de belangrijkste prestatiematen kunnen uitgedrukt worden in termen van de evenwichtsdistributie van het onderliggende Markovproces. Voor een eindig ergodisch Markovproces, is de evenwichtsdistributie de unieke oplossing van $N-1$ balansvergelijkingen aangevuld met de normalisatievoorwaarde, waarbij N de grootte van de toestandsruimte is. Voor wachtrijsystemen met gezamenlijke bediening groeit de grootte van de toestandsruimte van de Markovprocessen exponentieel met het aantal wachtrijen. Vandaar dat, zelfs indien slechts enkele wachtrijen beschouwd worden, de omvang van de toestandsruimte enorm is. Dit is het fenomeen van toestandsruimte-explosie. Daar directe oplossingsmethoden van de Markovprocessen rekenkundig onmogelijk zijn, richt dit proefschrift zich op het uitbuiten van bepaalde structurele eigenschappen van de bestudeerde Markovprocessen om de berekening van de evenwichtsdistributie te versnellen.

De eerste eigenschap die gebruikt kan worden is de ijlheid van de generatormatrix van het Markovproces. Inderdaad, het aantal gebeurtenissen dat zich kan voordoen in eender welke toestand — of het aantal transities dat zich kan voordoen naar andere toestanden — is veel kleiner dan de grootte van de toestandsruimte. Dit betekent dat de generatormatrix van het Markovproces voornamelijk met nullen gevuld is. Iteratieve methoden voor ijle lineaire systemen — in het bijzonder de Krylov-deelruimtesolver GMRES — bleken computationeel efficiënt te zijn voor het bestuderen van kittingprocessen met een beperkt aantal wachtrijen. Voor meerdere wachtrijen (of een grotere toestandsruimte) kan deze methode de evenwichtsdistributie echter onvoldoende snel berekenen. De toepassingen gerelateerd aan de ontkoppelingsvoorraad en de energie-oogstende sensor hebben slechts twee wachtrijen. In dit geval vertoont de generatormatrix een homogene bloktridiagonale structuur. Dergelijke Markovprocessen kunnen efficiënt worden opgelost door middel van matrix-geometrische methoden, zowel in het geval dat de toestandsruimte een eindige grootte heeft als — en nog efficiënter — in het geval dat deze oneindig is met eindige blokgrootte. Geen van de voormalige exacte oplossingsmethoden laat toe om systemen met vele wachtrijen te bestuderen. Daarom ontwikkelen we een benaderende numerieke oplossingsmethode, gebaseerd op Maclaurin-reeksontwikkelingen. In plaats van zich toe te leggen op de structurele eigenschappen van de Markovprocessen voor eender welke parameterinstelling, maakt de reeksontwikkelingstechniek gebruik van structurele

eigenschappen van het Markovproces wanneer een bepaalde parameter naar nul gaat. Voor de wachtrijen met een exponentiële gezamenlijke bediening en met een bedieningsintensiteit gaande naar nul heeft het resulterende proces een absorberende toestand en kunnen de toestanden geordend worden zodat de generatormatrix boven-diagonaal is. In dit geval is de oplossing wanneer de bedieningsintensiteit gelijk is aan nul triviaal en heeft het berekenen van de hogere-ordetermen in de reeksontwikkeling rond nul een computationele complexiteit evenredig met de grootte van de toestandsruimte. Dit is een geval van reguliere verstoring van de parameter. Daartegenover staat singuliere verstoring die hier wordt toegepast op het kittingproces met phase-type-verdeelde bedieningstijden. Bij singuliere verstoring heeft het Markovproces geen unieke evenwichtsdistributie wanneer de parameter naar nul gaat. Vergelijkbare technieken zijn nog steeds van toepassing, hoewel deze een ietwat grotere rekentijd vergen.

Tot slot merken we op dat de numerieke reeksontwikkelingstechnieken zich niet beperken tot het analyseren van wachtrijen met gezamenlijke bediening. Door een Markovproces met een multidimensionale toestandsruimte gelijkaardig aan een wachtrijsysteem met gezamenlijke bediening te beschouwen tonen we aan dat de reguliere reeksontwikkelingstechniek toegepast kan worden op een epidemiologisch model voor de studie van opinieverspreiding in een sociaal netwerk. Hierbij is het interessant te bemerken dat de reeksontwikkelingstechniek complementair is aan de gebruikelijke vloeistof-aanpak uit de epidemiologische literatuur.

English Summary

A queueing system is a mathematical abstraction of a situation where elements, called customers, arrive in a system and wait until they receive some kind of service. Queueing systems are omnipresent in real life. Prime examples include people waiting at a counter to be served, airplanes waiting to take off, traffic jams during rush hour etc. Queueing theory is the mathematical study of queueing phenomena. As often neither the arrival instants of the customers nor their service times are known in advance, queueing theory most often assumes that these processes are random variables. The queueing process itself is then a stochastic process and most often also a Markov process, provided a proper description of the state of the queueing process is introduced. This dissertation investigates numerical methods for a particular type of Markovian queueing systems, namely queueing systems with shared service. These queueing systems differ from traditional queueing systems in that there is simultaneous service of the head-of-line customers of all queues and in that there is no service if there are no customers in one of the queues. The absence of service whenever one of the queues is empty yields particular dynamics which are not found in traditional queueing systems.

These queueing systems with shared service are not only beautiful mathematical objects in their own right, but are also motivated by an extensive range of applications. The original motivation for studying queueing systems with shared service came from a particular process in inventory management called kitting. A kitting process collects the necessary parts for an end product in a box prior to sending it to the assembly area. The parts and their inventories being the customers and queues, we get "shared service" as kitting cannot proceed if some parts are absent. Still in the area of inventory management, the decoupling inventory of a hybrid make-to-stock/make-to-order system exhibits shared service. The production process prior to the decoupling inventory is make-to-stock and driven by demand forecasts. In contrast, the production process after the decoupling inventory is make-to-order and driven by actual demand as items from the decoupling inventory are customised according to customer specifications. At the decoupling point, the decoupling inventory is complemented with a queue of outstanding orders. As customisation only starts when the decoupling inventory is nonempty and there is at least one order, there is again shared service. Moving to applications in telecommunications, shared service applies to energy harvesting sensor nodes. Such a sensor node scavenges energy from its environment to meet its energy expenditure or to prolong its lifetime. A rechargeable battery operates very much like a queue, customers being discretised as chunks of energy. As a sensor node

requires both sensed data and energy for transmission, shared service can again be identified.

In the Markovian framework, "solving" a queueing system corresponds to finding the steady-state solution of the Markov process that describes the queueing system at hand. Indeed, most performance measures of interest of the queueing system can be expressed in terms of the steady-state solution of the underlying Markov process. For a finite ergodic Markov process, the steady-state solution is the unique solution of $N-1$ balance equations complemented with the normalisation condition, N being the size of the state space. For the queueing systems with shared service, the size of the state space of the Markov processes grows exponentially with the number of queues involved. Hence, even if only a moderate number of queues are considered, the size of the state space is huge. This is the state-space explosion problem. As direct solution methods for such Markov processes are computationally infeasible, this dissertation aims at exploiting structural properties of the Markov processes, as to speed up computation of the steady-state solution.

The first property that can be exploited is sparsity of the generator matrix of the Markov process. Indeed, the number of events that can occur in any state — or equivalently, the number of transitions to other states — is far smaller than the size of the state space. This means that the generator matrix of the Markov process is mainly filled with zeroes. Iterative methods for sparse linear systems — in particular the Krylov subspace solver GMRES — were found to be computationally efficient for studying kitting processes only if the number of queues is limited. For more queues (or a larger state space), the methods cannot calculate the steady-state performance measures sufficiently fast. The applications related to the decoupling inventory and the energy harvesting sensor node involve only two queues. In this case, the generator matrix exhibits a homogene block-tridiagonal structure. Such Markov processes can be solved efficiently by means of matrix-geometric methods, both in the case that the process has finite size and — even more efficiently — in the case that it has an infinite size and a finite block size. Neither of the former exact solution methods allows for investigating systems with many queues. Therefore we developed an approximate numerical solution method, based on Maclaurin series expansions. Rather than focussing on structural properties of the Markov process for any parameter setting, the series expansion technique exploits structural properties of the Markov process when some parameter is sent to zero. For the queues with shared exponential service and the service rate sent to zero, the resulting process has a single absorbing state and the states can be ordered such that the generator matrix is upper-diagonal. In this case, the solution at zero is trivial and the calculation of the higher order terms in the series expansion around zero has a computational complexity proportional to the size of the state space. This is a case of regular perturbation of the parameter and contrasts to singular perturbation which is applied when the service times of the kitting process are phase-type distributed. For singular perturbation, the Markov process has no unique steady-state solution when the parameter is sent to zero. However, similar techniques still apply, albeit at a higher computational cost.

Finally we note that the numerical series expansion technique is not limited to evaluating queues with shared service. Resembling shared queueing systems in that a Markov process with multidimensional state space is considered, it is shown that the regular series expansion technique can be applied on an epidemic model for opinion propagation in a social network. Interestingly, we find that the series expansion technique complements the usual fluid approach of the epidemic literature.

1
Introduction

This dissertation investigates queuing systems with multiple queues that are jointly served. Joint service not only means that there is a departure in every queue upon service completion, but also that service is only possible when each queue is nonempty. Models with two queues and more than two queues are here referred to as *paired queueing models* and *coupled queueing models*, respectively. We aim to gain insights into the dynamics of such systems under uncertainty. Accurate closed-form expressions of performance measures cannot be expected given the complexity of these queueing systems. Therefore, we mainly rely on numerical analysis techniques to accurately evaluate system performance with reasonable computational effort. As will become clear in the following chapters, coupled queueing systems find applications in diverse areas including inventory management and telecommunications. Nevertheless, at the onset of this dissertation we also would like to mention that these queueing systems are beautiful mathematical objects in their own right, with particularly interesting dynamics.

This introductory chapter provides some background on as well as an outline of the subject of the dissertation and is organised as follows. In Section 1.1, we give a brief overview of the various applications of coupled queueing systems that have motivated our investigations in these systems. In Section 1.2, we shortly describe the nature and utility of stochastic modelling, as well as survey the numerical techniques used throughout the dissertation. In Section 1.3, an outline of the later chapters is provided, with special attention to the main modelling assumptions of and differences between these chapters, both in terms of applications under study and methodology. Finally, an overview of the publications on which this

dissertation is based is given in Section 1.4.

1.1 Applications of paired and coupled queues

1.1.1 Kitting system

Many manufacturing systems pursue a high product variety strategy to gain a competitive advantage. Indeed, companies try to differentiate themselves from their competitors by supplying a wide assortment of assembled products. However, this strategy is likely to increase the total material handling time and required storage space at the assembly line [40]. To cope with this tendency, a kitting process can be introduced in the production and assembly process. Kitting is a strategy for supplying parts to an assembly line. More specifically, kitting collects the necessary parts for a given end product into a container, referred to as a kit, prior to arriving at the assembly line [7, 9, 41, 50, 58]. The overall material handling time is reduced as activities like selecting and gripping parts are performed more efficiently [40, 52]. Moreover, kitting mitigates storage space requirements at the assembly station since no part inventories need to be kept there. Additional benefits include reduced assembly times when parts are placed in proper positions in the container and reduced operator walking times since kits are brought as a whole to the assembly station.

The introduction of kitting obviously does not come for free as an additional stage is introduced prior to assembly. Assessing whether or not it is beneficial to introduce a kitting operation requires a deep understanding of its dynamics. Most literature study kitting performance in a deterministic production environment [7, 12] but this may lead to an unrealistic optimal solution with high stock and idle times. As to get a more detailed cost/benefit assessment of the kitting operation, we study the performance of kitting operations under uncertainty in demand and production times. Kitting involves multiple part inventories and kits can only be compiled if all parts are available. Hence, the kitting process is modelled as a paired or coupled queueing system if there are respectively two or more than two queues. Figure 1.1 shows an abstract representation of a kitting process with two types of parts. These parts arrive at their respective part inventories, the queues, and 'wait' there until they are collected into a kit.

1.1.2 Hybrid MTS/MTO system

In supply chain management, well-known production strategies are make-to-stock (MTS) and make-to-order (MTO). Under pure MTS management, the activities are forecast-driven. Indeed, the end products are manufactured independently of any customer requirements and are stocked in advance. In contrast, in a pure MTO strategy, the activities are demand-driven. The manufacturing of a product

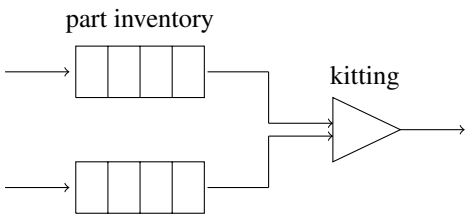


Figure 1.1: Kitting process with two parts.

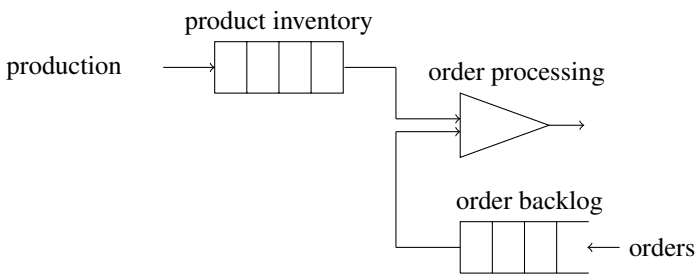


Figure 1.2: Generic model for the decoupling point in an MTS/MTO system.

is triggered only when a customer order is placed. To benefit from both systems, production is gradually moving towards a hybrid MTS/MTO strategy [59]. In such systems, the decoupling point — the boundary between forecast-driven and demand-driven activities — is positioned in the middle of the production process. This leads to shorter delivery times than in a pure MTO system (which corresponds to a decoupling point prior to delivery) and a higher production flexibility and responsiveness to demand than in a pure MTS system (which corresponds to a decoupling point prior to production) [23, 36, 49].

As firms suffer from increased demand fluctuations, inventory replenishment issues and variable order processing times, we describe hybrid MTS/MTO systems as a stochastic inventory model with two queues: the order backlog, which tracks the production orders that have not yet been processed, and the decoupling inventory of semi-finished products. Production prior to the product inventory here corresponds to the MTS stage of the hybrid MTS/MTO system. The MTO stage is captured by the order processing times. When there are orders in the order backlog, products in the product inventory, and previous orders have been completed, a semi-finished product is processed into a finished product according to the order instructions. Hence, the queues are paired as shown in Figure 1.2. The dynamics of the system however differs considerably from the dynamics of the kitting process, as the order queue has infinite capacity. Moreover, a threshold-based control policy can be implemented: production of semi-finished products starts when the

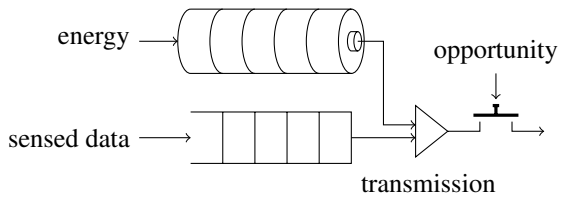


Figure 1.3: Stochastic model of energy harvesting sensor nodes.

inventory level goes below a certain value, referred to as the threshold value, and stops when the inventory level reaches maximum capacity. Such a control is sometimes referred to as an (s,S) policy and reduces the number of times production starts and stops due to the finite capacity of the decoupling inventory.

1.1.3 Energy harvesting sensor node

Sensor networks, formed by collections of intercommunicating sensor nodes, are used to collect and monitor spatially distributed data like temperature, humidity, movement, noise etc [2, 3]. As it is often not convenient to recharge or replace the batteries, the lifetime of sensor nodes is largely determined by the energy of on-board batteries. Applications for which this is mostly difficult are found in hard-to-reach locations or locations where power lines do not exist such as for volcano monitoring [64], habitat monitoring [39] and vehicle tracking [33]. To mitigate or overcome this dependency, the necessary energy can be scavenged from the sensor node's environment. This alternative technique is called energy harvesting [31, 35].

In this dissertation, we evaluate the performance of energy harvesting sensor nodes under uncertainty in energy harvesting, energy expenditure, data acquisition and data transmission. To this end, the sensor node is described as a system with two queues: the accumulated harvested energy and the data packet buffer, as shown in Figure 1.3. Indeed, a rechargeable battery operates very much like a queue: charging the battery corresponds to arrivals of "energy customers", and depletion corresponds to the departure of these customers. Again, pairing of both buffers is natural as any data transmission requires both the availability of data as well as energy. While the buffer battery equivalence comes natural, it should be noted that the operational point of a typical data buffer and a battery can differ considerably. For the data buffer, one aims for small buffer content: the fewer packets one has to store the better the performance. In contrast, for the battery it is beneficial to have a lot of energy present.

Independently of the availability of sufficient battery power, data cannot always be transmitted. The introduction of these limited transmission opportunities is motivated by but not restricted to scenarios where a mobile sink is responsible

for data collection. A mobile sink moves towards the energy harvesting sensor node and gathers the sensed data when it is located in the transmission range of the sensor node. This means that data transmission is only possible when there is sufficient energy, a data packet available and a transmission opportunity.

1.1.4 Opinion propagation in a social network

Given the rapid growth of companies in the internet sector and the ascent of social networks in particular (e.g. Facebook, LinkedIn, Twitter etc.), there is a very strong interest in understanding how new opinions spread through a community [29]. Indeed, the analysis of opinion propagation can improve our comprehension of social relations among individuals online as well as offline.

In this dissertation, opinion spreading is described as a Markovian non-standard Susceptible-Infected-Recovered (SIR) epidemic model [4, 27]. Indeed, opinion propagation holds many qualitative similarities with infectious diseases spread: if an individual without a specific opinion about a topic (susceptible) encounters an opinioned individual (infected), this individual may form an opinion with some probability, and therefore also get opinioned. Afterwards, opinioned individuals (infected) may, with some probability, stop transmitting their opinion to other individuals. Indeed, these individuals may become neutral to the topic (recovered). We extend the standard stochastic model in two ways. We account for the situation where a non-opinioned (susceptible) individual becomes neutral (recovered) directly and we allow for state-dependent infection and recovery rates.

If there are many individuals partaking in the spreading of the opinion, epidemic theory learns that deterministic approximations of the evolution of the number of opinioned persons are appropriate. In this case, the evolution of the state of the individuals can be described by a set of differential equations. However, when the population size is limited, deterministic epidemic theory does not apply. In this case, opinion spreading can be captured by a multidimensional Markov process which very much resembles the Markov processes of coupled queueing systems, although individuals can now leave the system one by one (there is no coupling). If the population size is small, the Markov process can be analysed by standard solution methods. However, when the population size increases this is no longer the case. We show that the methodology that was developed to study coupled queueing systems, also applies to the opinion spreading model. This observation enables us to study epidemics in medium-sized populations.

1.2 Methodology

Often, neither the arrival nor the departure process of the systems under study are fully known. To cope with this inherent uncertainty, these are preferably modelled as stochastic processes. In this section, we aim to convey the general ideas of stochastic modelling and the numerical techniques used throughout the dissertation.

To this end, we first introduce the basic notions of Markov processes below. Given the complexity of the dynamics of our Markov processes, accurate closed-form solutions of the performance measures of interest cannot be expected. Therefore, this research mainly relies on numerical techniques. Three techniques are used and discussed in this section: (i) the iterative method for sparse linear systems GMRES, (ii) the matrix-geometric method, and (iii) the Maclaurin-series expansion approach. We conclude this section by comparing speed and accuracy of the different numerical techniques in Section 1.2.5.

1.2.1 Markov processes

A more specific and for us relevant stochastic process is the Markov process. In fact, in this dissertation, all studied systems are described as continuous-time Markov processes with a discrete state space. This means that the Markov process is defined for a continuous set of times and only adopts values in a countable collection. The vast majority distinguishes Markov processes from Markov chains based on the time parameter: chains proceed in discrete time, processes in continuous time [42]. A smaller number of authors however make the distinction between the two based on the state space in which they are operating: if it is finite or countable, then it is a Markov chain, else it is a Markov process [45]. Throughout this dissertation, we describe our continuous-time stochastic processes with a discrete state space as Markov processes. For such a process $\{X_t, t \in \mathbb{R}\}$, the future outcome depends only on the current value or state of the process and not on its past. At any point in time t and for any positive integer n, and for points $t_1, t_2, \ldots, t_n < t_0 \leq t$ the random variable X_t has the property,

$$\mathbb{P}[X_t = x_t | X_{t_0} = x_{t_0}, X_{t_1} = x_{t_1}, \ldots, X_{t_n} = x_{t_n}] = \mathbb{P}[X_t = x_t | X_{t_0} = x_{t_0}].$$

This specific kind of 'memorylessness' is called the Markov property: the probability distribution of the future value of the process X_t is independent of the past values of X_{t_1} to X_{t_n} given the current value X_{t_0}.

A Markov process X_t with a countable state space $\mathcal{X}$ can be completely characterised by the so-called generator matrix $\mathbf{Q}$ with elements

$$q_{ij} = \lim_{\triangle t \to 0} \frac{\mathbb{P}[X_{t+\triangle t} = j | X_t = i]}{\triangle t}, \quad i \neq j, i, j \in \mathcal{X}. \tag{1.1}$$

The parameter q_{ij} is referred to as the transition rate from state i to state j ($j \neq i$); there is a transition from state i to state j in an interval of length dt with probability $q_{ij}dt + o(dt)$. The diagonal elements of $\mathbf{Q}$ are defined as

$$q_{ii} = -\sum_{j \neq i} q_{ij}\,. \tag{1.2}$$

Note that the definition of q_{ii} implies that the row sums of $\mathbf{Q}$ are 0. Moreover, assuming that the Markov process is in state i, it remains in state i for an exponentially distributed time with rate $-q_{ii}$. We also use the term generator matrix when the diagonal elements are equal to zero ($q_{ii} = 0$), which sometimes is notationally more convenient. Throughout the dissertation, the diagonal elements of the generator matrices are explicitly defined.

The basic quantitative result that can be obtained from our Markov processes is the stationary distribution. This distribution tells us how the system behaves after a long period of time, when the effects of the initial state have faded out. In this dissertation, we consider time-homogeneous Markov processes with a finite and an infinite number of states. Time-homogeneity means that transition rates are time-independent. For Markov processes with a finite state space, a unique solution of the stationary distribution can be found if the Markov process has a single communicating class. In the infinite case, it is additionally required that the states in the communicating class are positive recurrent. This means that the expected return time to each state of this class is finite. Under these conditions, the stationary distribution $\boldsymbol{\pi}$ can be found and is the unique solution of the following equations:

$$\sum_{i \in \mathcal{X}} \pi_i q_{ij} = 0 \quad \text{and} \quad \sum_{i \in \mathcal{X}} \pi_i = 1.$$

It is now useful to write the above equations in matrix form. Let $\boldsymbol{\pi}$ be the row vector with elements π_i. We then have,

$$\boldsymbol{\pi}\mathbf{Q} = \mathbf{0}, \tag{1.3}$$

and

$$\boldsymbol{\pi}\mathbf{1} = 1, \tag{1.4}$$

where $\mathbf{1}$ denotes a column vector of appropriate size. Equation (1.4) denotes the so-called normalisation condition and states that the sum of the probabilities in steady state must be equal to one. This condition is necessary to find the unique solution of the stationary vector $\boldsymbol{\pi}$ together with the system of equations (1.3). Throughout the dissertation, the stationary distribution of a queueing system is also referred to as the steady-state probability vector.

In the remainder, a number of Markov processes are used as building blocks for the paired and coupled queueing models. The most important processes are the Markovian arrival process and the renewal process with phase-type renewal-time distribution discussed below.

Markovian arrival process A Markovian arrival process (MAP) is a continuous-time Markov process $\{(N(t), J(t)), t \geq 0\}$, with $N(0) = 0$, on the state space $\mathcal{S} = \mathbb{N} \times \{1, \ldots, m\}$, with $m \geq 1$. The function $N(t)$ counts the number of arrivals in interval $[0,t]$ and $J(t)$ represents the phase of the arrival process at time t. Assuming that the states are ordered according to count first, the generator matrix $\mathbf{Q}$ of the MAP has the following block matrix representation,

$$\mathbf{Q} = \begin{bmatrix} \mathbf{D}_0 & \mathbf{D}_1 & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{D}_0 & \mathbf{D}_1 & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{D}_0 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}. \tag{1.5}$$

where $\mathbf{D}_1$ and $\mathbf{D}_0$ are matrices of size $m \times m$. The elements of $\mathbf{D}_1$ and $\mathbf{D}_0$ represent transitions which are and are not accompanied with arrivals, respectively.

Many familiar arrival processes represent special cases of MAPs. The simplest example is the Poisson process where $m = 1$, $\mathbf{D}_1 = \lambda$ and $\mathbf{D}_0 = -\lambda$ where λ is the arrival rate of the Poisson process. Another specific case of MAP is the interrupted Poisson process (IPP). An IPP is a two-state MAP in which arrivals occur only in one of the states, denoted as the active state, and state jumps do not cause arrivals. In this case, $m = 2$ and the two matrices $\mathbf{D}_1$ and $\mathbf{D}_0$ are defined as follows:

$$\mathbf{D_1} = \begin{bmatrix} 0 & 0 \\ 0 & \lambda \end{bmatrix}, \quad \mathbf{D_0} = \begin{bmatrix} -\beta & \beta \\ \alpha & -(\alpha+\lambda) \end{bmatrix}.$$

where α (β) is the rate at which the system goes from an active (inactive) to an inactive (active) state in an infinitesimal time interval.

Phase-type distribution The phase-type distribution is a probability distribution constructed by a mixture of exponential distributions occurring in phases [44]. The sequence in which the phases occur may be a stochastic process in itself. More precisely, a phase-type distributed random variable describes the time until absorption of a Markov process with one absorbing state. Special cases of interest for this research are the exponential distribution, the Erlang distribution (with two or more identical phases in sequence), the hyperexponential distribution (with two or more non-identical phases and where each phase has a probability of occurring in a parallel manner), and the hypoexponential distribution (with two or more phases in sequence, that can be non-identical or a mixture of identical and non-identical phases).

1.2.2 Iterative methods

All numerical methods in this dissertation exploit structural properties of the Markov processes, as to find the stationary probability vector more efficiently. A first structural property that can be exploited is the sparsity of the generator matrices. Most processes in this dissertation possess the property that the number of states that can be directly reached from a certain state is far smaller than the size of the state space. This implies that the generator matrices are sparse.

The first methods developed for solving sparse linear systems of equations are the so-called sparse direct solvers, which are clever implementations of Gaussian elimination. These methods produce the result in a prescribed, finite number of steps. However, these methods are computationally often too expensive for large systems, even on today's fastest supercomputers. To cope with this shortcoming, researchers developed iterative methods. These methods are an attempt to solve a system of equations by finding successive approximations to the solution starting from an initial guess.

In this section, a brief introduction into the world of iterative techniques, and in particular of the Krylov subspace method GMRES, is given. In Chapter 2, the GMRES technique is implemented to find numerical solutions of a two-part kitting process. Compared to other numerical techniques, the use of this technique in this research is rather limited. This is because of the remaining difficulty to solve large state-space systems (see Figure 1.5 in Section 1.2.5) and the keen interest we had in other numerical techniques and their applications.

We now give a short introduction of the basic steps of an iterative method. Consider a large sparse linear system of equations of the form

$$\mathbf{A}\mathbf{x} = \mathbf{b}. \tag{1.6}$$

where $\mathbf{A}$ is the generator matrix of size $n \times n$, $\mathbf{b} \in \mathbb{R}^n$ is a known column vector, and $\mathbf{x} \in \mathbb{R}^n$ is the vector of unknowns. The first step is to define an initial guess $\mathbf{x}_0 \in \mathbb{R}^n$ that approximates the exact solution $\mathbf{x}$. Once we have $\mathbf{x}_0$, we use it to generate a new guess $\mathbf{x}_1$ which is used to guess $\mathbf{x}_2$ and so on. Each guess is improved by reducing the error with a convenient and cheap approximation.

In the next part, we discuss two main types of iterative methods: stationary and nonstationary methods. Nonstationary methods differ from stationary methods in that the computations involve information that changes at each iteration. Stationary methods are older, less complex but usually also less effective. On the other hand, nonstationary methods, also called Krylov subspace methods, are more recent — the GMRES method discussed below dates back to 1986 — and their analysis is more complex but they can be highly effective. Below is a brief overview of the most important stationary iterative methods given.

1.2.2.1 Stationary iterative methods

Stationary iterative methods, also called relaxation methods, involve passing from one iterate to the next by modifying one or a few components of an approximate vector solution at a time. The criteria used at each iteration is the minimisation of the residual vector $|\mathbf{b}-\mathbf{A}\mathbf{x}|$ of equation (1.6). Let the matrix $\mathbf{A}$ be decomposed as $\mathbf{A}=\mathbf{D}+\mathbf{L}+\mathbf{U}$, where $\mathbf{D}$ is the diagonal component of $\mathbf{A}$ and $\mathbf{L}$ and $\mathbf{U}$ are respectively the strictly lower and strictly upper-triangular components of $\mathbf{A}$. This matrix splitting allows us to describe properly the Jacobi, Gauss-Seidel and the Successive Over-Relaxation method (SOR) [57], [60] (p. 125–138).

The simplest iterative method is the Jacobi iteration. The solution is obtained iteratively via

$$\mathbf{x}^{(k+1)}=\mathbf{D}^{-1}(\mathbf{b}-\mathbf{R}\mathbf{x}^{(k)}) \tag{1.7}$$

where $\mathbf{R}=\mathbf{U}+\mathbf{L}$. Note that the inverse of a diagonal matrix is trivial. The element-based formula is then

$$x_i^{(k+1)}=\frac{1}{a_{ii}}\left(\mathbf{b}_i-\sum_{j\neq i}a_{ij}\mathbf{x}_j^{(k)}\right). \tag{1.8}$$

with $i\neq j$ and $x_j^{(k)}\in\mathbf{x}_k$. The computation of $x_i^{(k+1)}$ requires each element in $\mathbf{x}^{(k)}$ except itself. This means that we cannot overwrite $x_i^{(k)}$ with $x_i^{(k+1)}$, as that value will be needed in the remainder of the computation. The minimum amount of storage comprises thus two vectors of size n. However, the new expression of the solution vector $\mathbf{x}^{(k+1)}$ is calculated by using only the old approximation $\mathbf{x}^{(k)}$. So the computed elements of $\mathbf{x}^{(k+1)}$ are not already used. Therefore, this method possesses a high degree of natural parallelism and is thus very convenient to vectorise and to parallelise [11].

The Gauss-Seidel iteration is defined as

$$\mathbf{x}^{(k+1)}=\mathbf{S}^{-1}\left(\mathbf{b}-\mathbf{U}\mathbf{x}^{(k)}\right)$$

where $\mathbf{S}=\mathbf{D}+\mathbf{L}$. Equivalently

$$\mathbf{D}\mathbf{x}^{(k+1)}=\mathbf{b}-\mathbf{L}\mathbf{x}^{(k+1)}-\mathbf{U}\mathbf{x}^{(k)}$$

By taking advantage of the triangular form of $\mathbf{S}$, the elements of $\mathbf{x}^{(k+1)}$ can be computed sequentially using forward substitution:

$$x_i^{(k+1)}=\frac{1}{a_{ii}}\left(\mathbf{b}_i-\sum_{j<i}a_{ij}\mathbf{x}_j^{(k+1)}-\sum_{j>i}a_{ij}\mathbf{x}_j^{(k)}\right). \tag{1.9}$$

Unlike the Jacobi method, the approximate solution is updated immediately after the new component is determined. The computation of $x_i^{(k+1)}$ uses only the

elements of $\mathbf{x}^{(k+1)}$ that have already been computed, and only the elements of $\mathbf{x}^{(k)}$ that have not yet advanced to iteration $k+1$. This means that only one storage vector is required, which can be advantageous for very large problems. The method of Gauss-Seidel is therefore not convenient for vectorisation or parallelisation, but typically converges faster than the Jacobi iteration.

The Successive Over-Relaxation method is a variant of the Gauss-Seidel and the Jacobi method. The equation may be written as

$$\omega\mathbf{A} = (\mathbf{D}+\omega\mathbf{L}) + (\omega\mathbf{U} - (1-\omega)\mathbf{D}). \tag{1.10}$$

This may be rewritten as:

$$(\mathbf{D}+\omega\mathbf{L})\mathbf{x}^{(k+1)} = \omega\mathbf{b} - (\omega\mathbf{U} - (1-\omega)\mathbf{D})\mathbf{x}^{(k)}, \tag{1.11}$$

where the constant ω is the relaxation factor. Note that if $\omega = 1$, we have the Gauss-Seidel method. Note also that the SOR method converges only if $0 < \omega < 2$ [65]. Analytically, this may be written as:

$$\mathbf{x}^{(k+1)} = (\mathbf{D}+\omega\mathbf{L})^{-1}\Big(\omega\mathbf{b} - (\omega\mathbf{U} + (\omega-1)\mathbf{D})\mathbf{x}^{(k)}\Big). \tag{1.12}$$

However, by taking advantage of the triangular form of $(\mathbf{D}+\omega\mathbf{L})$, the elements of $\mathbf{x}^{(k+1)}$ can be computed sequentially using forward substitution:

$$x_i^{(k+1)} = (1-\omega)x_i^{(k)} + \frac{\omega}{a_{ii}}\left(b_i - \sum_{j<i} a_{ij}x_j^{(k+1)} - \sum_{j>i} a_{ij}x_j^{(k)}\right) \tag{1.13}$$

where $i = 1,2,\ldots,n$. A last important stationary method is the Symmetric Successive Over-Relaxation (SSOR). This method combines two successive Over-Relaxation methods (SOR),

$$\begin{aligned}(\mathbf{D}+\omega\mathbf{L})\mathbf{x}^{(k+\frac{1}{2})} &= \omega\mathbf{b} - (\omega\mathbf{U} - (1-\omega)\mathbf{D})\mathbf{x}^{(k)},\\ (\mathbf{D}+\omega\mathbf{U})\mathbf{x}^{(k+1)} &= \omega\mathbf{b} - (\omega\mathbf{L} - (1-\omega)\mathbf{D})\mathbf{x}^{(k+\frac{1}{2})}.\end{aligned}$$

The main advantage of SSOR schemes is that the iteration matrix is similar to a symmetric matrix when the original matrix is symmetric.

1.2.2.2 Krylov subspace solvers

The first step is to construct a sequence of approximations of the solution $\mathbf{x}$ of equation (1.6) as

$$\mathbf{x}_m \in \mathbf{x}_0 + \mathcal{K}_m(\mathbf{A},\mathbf{r}_0). \tag{1.14}$$

where $\mathbf{x}_0$ is the initial guess and

$$\mathbf{r}_i := \mathbf{b} - \mathbf{A}\mathbf{x}_i \tag{1.15}$$

is the ith residual. The order-m Krylov subspace $\mathcal{K}_m(\mathbf{A},\mathbf{r}_0)$ is defined as the linear subspace spanned by the images of $\mathbf{r}_0$ under the first $m-1$ powers of $\mathbf{A}$,

$$\mathcal{K}_m(\mathbf{A},\mathbf{r}_0) = \text{span}\{\mathbf{r}_0, \mathbf{A}\mathbf{r}_0, \mathbf{A}^2\mathbf{r}_0, \ldots, \mathbf{A}^{m-1}\mathbf{r}_0\}. \tag{1.16}$$

Suppose that after exactly m iterations the solution is contained in the current affine Krylov subspace, i.e.

$$\mathbf{x} \in \mathbf{x}_0 + \mathcal{K}_m(\mathbf{A},\mathbf{r}_0). \tag{1.17}$$

This means that m is the smallest index for which the following formula holds

$$\dim[\mathcal{K}_m(\mathbf{A},\mathbf{r}_0)] = \dim[\mathcal{K}_{m+1}(\mathbf{A},\mathbf{r}_0)]. \tag{1.18}$$

such that we have found an invariant subspace where $\mathbf{r}_m = \mathbf{0}$ and $\mathbf{x}_m = \mathbf{x}$. Hence, the Krylov procedure stops.

To find this approximate solution $\mathbf{x}_m$, an orthonormal basis $\{\mathbf{v}_1,\ \mathbf{v}_2,\ \ldots,\ \mathbf{v}_m\}$ of the Krylov subspace $\mathcal{K}_m(\mathbf{A},\mathbf{r}_0)$ is generated. Let $\mathbf{V}_m = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m]$. The next step is to seek an approximation of the solution of $\mathbf{A}\mathbf{x} = \mathbf{b}$ in the set of $\mathbf{x}_0 + \mathcal{K}_m(\mathbf{A},\mathbf{r}_0)$ and of the form $\mathbf{x}_m = \mathbf{x}_0 + \mathbf{V}_m\mathbf{y}_m$ for some $\mathbf{y}_m \in \mathbb{R}_m$.

The difference between the Krylov methods arise in the choices of $\mathbf{V}_m$ and $\mathbf{y}_m$. The generation of the orthonormal basis and the steps to determine $\mathbf{y}_m$ for the GMRES method are discussed next.

GMRES The generalized minimum residual method, is designed to solve non-symmetric linear systems [60, 47, 10, 56]. The orthonormal basis of the GMRES method is defined by using Arnoldi's method. Hence, we first explain the Arnoldi algorithm applied on the GMRES method [63].

Given the generator matrix $\mathbf{A}$ and the initial residual $\mathbf{r}_0$ as defined in equation (1.6) and (1.15) respectively, the Arnoldi process begins with

$$\mathbf{v}_1 := \frac{\mathbf{r}_0}{\beta}. \tag{1.19}$$

where $\beta := \|\mathbf{r}_0\|$. Suppose now that we have generated $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_j\}$ as a basis for $\mathcal{K}_j\ (\mathbf{A},\ \mathbf{v}_1)$. We now wish to find a vector $\mathbf{v}_{j+1}$ such that $\mathcal{K}_{j+1}(\mathbf{A},\mathbf{v}_1) = \text{span}\{\mathbf{v}_1,\ \mathbf{v}_2, \ldots, \mathbf{v}_{j+1}\}$. The Arnoldi iteration uses the stabilised Gram-Schmidt process to produce $\mathbf{v}_{j+1}$ as follows

$$h_{i,j} = < \mathbf{A}\mathbf{v}_j, \mathbf{v}_1 >, \quad i = 1, \ldots, j, \tag{1.20}$$

$$\bar{\mathbf{v}}_{j+1} = \mathbf{A}\mathbf{v}_j - \sum_{i=1}^{j} \mathbf{v}_i h_{ij}, \tag{1.21}$$

$$h_{j+1,j} = \|\bar{\mathbf{v}}_{j+1}\|, \tag{1.22}$$

and if $h_{j+1,j} \neq 0$,

$$\mathbf{v}_{j+1} = \frac{\bar{\mathbf{v}}_{j+1}}{h_{j+1,j}} \tag{1.23}$$

From the above equations, we get

$$\mathbf{A}\mathbf{v}_j = \sum_{i=1}^{j} \mathbf{v}_i h_{ij} + \mathbf{v}_{j+1} h_{j+1,j} \Rightarrow \mathbf{A}\mathbf{v}_j = \sum_{i=1}^{j+1} \mathbf{v}_j h_{i,j} \tag{1.24}$$

If $\bar{\mathbf{v}}_{\mathbf{j+1}} = 0$, then $h_{j+1,j} = 0$ and $\text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_j\} = \mathcal{K}_j(\mathbf{A}, \mathbf{v}_1)$ is invariant under $\mathbf{A}$ such that the Arnoldi process terminates.

This process runs for $M = \{1, 2, \ldots, m\}$ so that

$$\mathbf{A}\mathbf{V}_m = \mathbf{V}_{m+1}\mathbf{H}_{m+1,m}. \tag{1.25}$$

where $\mathbf{V}_m = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m]$ as defined previously and the matrix $\mathbf{H}_{m+1,m}$ denotes the $(m+1) \times m$ upper Hessenberg matrix whose (i, j) entry h_{ij} is such that

$$\mathbf{A}\mathbf{v}_j = \sum_{i=1}^{j} \mathbf{v}_i h_{ij}. \tag{1.26}$$

if $i \leq j+1$ and $h_{ij} = 0$ if $i > j+1$.

Based on the mth step of Arnoldi's method given by equation (1.25), the iterate $\mathbf{x}_m$ in the mth step of GMRES (satisfying expression (1.14)) can be written as

$$\mathbf{x}_m = \mathbf{x}_0 + \mathbf{V}_m \mathbf{y}_m \tag{1.27}$$

and the vector $\mathbf{y}_m$ is chosen to minimise the norm of the residual $\mathbf{r}_m$ such that

$$\begin{aligned}
\min_{y \in \mathbb{R}_m} \|\mathbf{b} - \mathbf{A}\mathbf{x}_m\| &= \min_{y \in \mathbb{R}_m} \|\mathbf{b} - \mathbf{A}(\mathbf{x}_0 + \mathbf{V}_m \mathbf{y})\| \\
&= \min_{y \in \mathbb{R}_m} \|\mathbf{b} - \mathbf{A}\mathbf{x}_0 - \mathbf{A}\mathbf{V}_m \mathbf{y}\| \\
&= \min_{y \in \mathbb{R}_m} \|\mathbf{r}_0 - \mathbf{V}_{m+1}\mathbf{H}_{m+1,m}\mathbf{y}\| \\
&= \min_{y \in \mathbb{R}_m} \|\beta \mathbf{v}_1 - \mathbf{V}_{m+1}\mathbf{H}_{m+1,m}\mathbf{y}\| \\
&= \min_{y \in \mathbb{R}_m} \|\mathbf{V}_{m+1}(\beta \mathbf{e}_1 - \mathbf{V}_{m+1}\mathbf{H}_{m+1,m}\mathbf{y})\| \\
&= \min_{y \in \mathbb{R}_m} \|\beta \mathbf{e}_1 - \mathbf{V}_{m+1}\mathbf{H}_{m+1,m}\mathbf{y}\|.
\end{aligned}$$

where $\mathbf{e}_1$ is the first Euclidian vector [55] (p.25–27).

To sum up, the GMRES approximation is the unique vector of $\mathbf{x}_0 + \mathcal{K}_m(\mathbf{A}, \mathbf{r}_0)$ which minimises the residual $\|\mathbf{b} - \mathbf{A}\mathbf{x}_m\|$ where

$$\mathbf{x}_m = \mathbf{x}_0 + \mathbf{V}_m \mathbf{y}_m,$$

and

$$\mathbf{y}_m = \min_y \|\beta \mathbf{e}_1 - \mathbf{V}_{m+1}\mathbf{H}_{m+1,m}\mathbf{y}\|$$

Note. In Chapter 2, the GMRES method is compared with the LU decomposition method [61]. This direct method decomposes the generator matrix $\mathbf{Q}$ of equation (1.3) into a lower and upper-triangular matrix such that their product is equal to this matrix. The solution of this equation is then obtained by solving two triangular systems.

1.2.3 Matrix-geometric methods

A common characteristic of the studied paired and coupled queueing systems is the structure of the state space; the generator matrix has a tridiagonal block structure with level-independent transition rates (see Chapters 3 – 6). Such a Markov process is called a homogeneous quasi-birth-death (QBD) process. To exploit this structure, efficient numerical solution techniques such as the matrix-geometric method have been developed [37, 43]. This section aims to convey a general idea of the homogeneous QBD process and the matrix-geometric method.

The state of the Markov process can be written as (n,m), where $n \geq 0$ and $0 \leq m \leq M$. The first coordinate n indicates the block-row number, called the level of the Markov process, and the second coordinate m indicates the index within a block element of size $M+1$, called the phase of the QBD. This means that states belonging to a similar block matrix have the same level but a different phase.

The main property of QBD processes is the constraint of allowing only 'neighbouring' transitions. In particular, the one-step transitions are restricted to states in the same level (from state $(n,*)$ to state $(n,*)$) or in two adjacent levels (from state $(n,*)$ to state $(n+1,*)$ or $(n-1,*)$).

Hence, we have a continuous-time Markov process with the following generator matrix $\mathbf{Q}$:

$$\mathbf{Q} = \begin{bmatrix} \mathbf{B} & \mathbf{A}_2 & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{A}_0 & \mathbf{A}_1 & \mathbf{A}_2 & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{A}_0 & \mathbf{A}_1 & \mathbf{A}_2 & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_0 & \mathbf{A}_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \tag{1.28}$$

where the block matrices $\mathbf{A}_0$ and $\mathbf{A}_2$ have nonnegative values and the block matrices $\mathbf{A}_1$ and $\mathbf{B}$ have nonnegative off-diagonal elements and strictly negative diagonal elements. The row sums of $\mathbf{Q}$ are equal to zero, so that we have $(\mathbf{B}+\mathbf{A}_2)\mathbf{1} = \mathbf{0}$ and $(\mathbf{A}_0+\mathbf{A}_1+\mathbf{A}_2)\mathbf{1} = \mathbf{0}$, where $\mathbf{1}$ and $\mathbf{0}$ are column vectors of adequate size with all elements equal to zero and one, respectively.

Next, let the steady-state probability vector $\boldsymbol{\pi}$ of the above defined homogeneous QBD process with equation (1.3) be partitioned conformally with the levels of $\mathbf{Q}$, i.e.

$$\boldsymbol{\pi} = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots)$$

where

$$\boldsymbol{\pi}_i = (\pi(i,0), \pi(i,1), \pi(i,2), \ldots, \pi(i,M)).$$

This gives the following balance equations

$$\begin{aligned} \boldsymbol{\pi}_0\mathbf{B} + \boldsymbol{\pi}_1\mathbf{A}_0 &= 0 \\ \boldsymbol{\pi}_0\mathbf{A}_2 + \boldsymbol{\pi}_1\mathbf{A}_1 + \boldsymbol{\pi}_2\mathbf{A}_0 &= 0 \\ \boldsymbol{\pi}_1\mathbf{A}_2 + \boldsymbol{\pi}_2\mathbf{A}_1 + \boldsymbol{\pi}_3\mathbf{A}_0 &= 0 \\ &\vdots \\ \boldsymbol{\pi}_{i-1}\mathbf{A}_2 + \boldsymbol{\pi}_i\mathbf{A}_1 + \boldsymbol{\pi}_{i+1}\mathbf{A}_0 &= 0 \end{aligned} \tag{1.29}$$

for $i = 1, 2, \ldots$

Stability condition According to theorem 7.2.3 of [37], if (i) the QBD is irreducible, (ii) the number of phases is finite and (iii) the generator matrix $\mathbf{A} = \mathbf{A}_0 + \mathbf{A}_1 + \mathbf{A}_2$ is irreducible, then the QBD is positive recurrent if and only if,

$$\boldsymbol{\pi}_A\mathbf{A}_0\mathbf{1} < \boldsymbol{\pi}_A\mathbf{A}_2\mathbf{1}$$

where $\mathbf{1}$ is the unit vector and $\boldsymbol{\pi}_A$ is the stationary distribution of the generator matrix $\mathbf{A}$. That is, $\boldsymbol{\pi}_A$ is the normalised solution of $\boldsymbol{\pi}_A\mathbf{A} = \mathbf{0}$. When this equation is satisfied, the stationary distribution of the QBD process exists. Intuitively, elements of $\mathbf{A}_0$ move the process up a level while elements of $\mathbf{A}_2$ move the process down a level. Hence, the so-called drift to higher numbered levels must be strictly less than the drift to lower numbered levels. Note that, assuming all states to be positive recurrent, and thus not only one class of states, this stability condition is stronger than the one given in Section 1.2 for time-homogeneous infinite Markov processes.

To solve such systems, we introduce the rate matrix $\mathbf{R}$ as the minimal nonnegative solution of the nonlinear matrix equation

$$\mathbf{A}_2 + \mathbf{R}\mathbf{A}_1 + \mathbf{R}^2\mathbf{A}_0 = \mathbf{0}. \tag{1.30}$$

Now assume that the equilibrium probabilities satisfy

$$\boldsymbol{\pi}_i = \boldsymbol{\pi}_{i-1}\mathbf{R}, \quad \text{for } i = 1, 2, \ldots.$$

which can be rewritten as

$$\boldsymbol{\pi}_i = \boldsymbol{\pi}_0 \mathbf{R}^i. \quad \text{for } i = 0, 1, \ldots \tag{1.31}$$

Provided that such an $\mathbf{R}$ can be found, it is easy to show that the solution above solves the balance equations. Indeed, if we substitute equation (1.31) in equation (1.29), we have

$$\boldsymbol{\pi}_0 \mathbf{R}^i \mathbf{A}_2 + \boldsymbol{\pi}_0 \mathbf{R}^{i+1} \mathbf{A}_1 + \boldsymbol{\pi}_0 \mathbf{R}^{i+2} \mathbf{A}_0 = \boldsymbol{\pi}_0 \mathbf{R}^i (\mathbf{A}_2 + \mathbf{R}\mathbf{A}_1 + \mathbf{R}^2 \mathbf{A}_0) = \mathbf{0},$$

for $i = 0, 1, \ldots$, where the last equality follows from (1.30). Hence, if $\mathbf{R}$ satisfies equation (1.30), then the vectors $\boldsymbol{\pi}_i$ are geometrically related to each other. The remaining unknown vector $\boldsymbol{\pi}_0$ satisfies,

$$\boldsymbol{\pi}_0 \mathbf{B} + \boldsymbol{\pi}_0 \mathbf{R} \mathbf{A}_0 = \mathbf{0}. \tag{1.32}$$

Moreover, the normalisation condition yields,

$$\sum_{i=0}^{\infty} \boldsymbol{\pi}_i \mathbf{1} = \boldsymbol{\pi}_0 \sum_{i=0}^{\infty} \mathbf{R}^i \mathbf{1} = \boldsymbol{\pi}_0 (\mathbf{I} - \mathbf{R})^{-1} \mathbf{1} = 1 \tag{1.33}$$

where $\mathbf{I}$ is the identity matrix and $\mathbf{1}$ is a column vector of ones. Equations (1.32) and (1.33) uniquely determine $\boldsymbol{\pi}_0$.

The former argument shows that the solution of the QBD reduces to finding the solution of (1.30). Several iterative procedures exist for solving $\mathbf{R}$ in equation (1.30). For example, starting with $\mathbf{R}$ equal to a matrix of zeros-entries, Gun [22] uses the following simple recursion

$$\mathbf{R} \leftarrow -(\mathbf{A}_2 + \mathbf{R}^2 \mathbf{A}_0) \mathbf{A}_1^{-1}. \tag{1.34}$$

In this dissertation the rate matrix $\mathbf{R}$ is computed by implementing the improved iterative algorithm of [37, Chapter 8, p.179-187].

1.2.4 Series expansion

The queueing systems under study cannot always be analysed as QBD processes. If we identify the level of the queue content of one queue and the phase with the queue content of the other queue of a paired queueing system, the resulting Markov process is a QBD. If we move beyond two queues, the phase can still describe the queue content of all but one queue and we still obtain a QBD. However, now the state-space explosion problem translates into increasing block sizes and matrix-geometric methods are no longer computationally efficient.

Nevertheless, these systems can sometimes be easily solved when some parameter is sent to zero. Assuming the resulting generator matrix to have an upper-triangular structure, the solution at zero is trivial as there is only one final state.

Also, as we will see further, the derivatives with respect to that parameter can easily be found. Hence, we use a series expansion in that parameter to approximate the steady-state probability vector of the Markov process (see Chapter 7 – 9). More specifically, we characterise the steady-state behaviour of the Markov processes that describe kitting systems and opinion propagation by means of a Maclaurin-series expansion in the departure rate μ. The general ideas behind the methods are sketched below.

Consider a 'perturbed' Markov process with the following generator matrix

$$\mathbf{Q}_\mu = \mathbf{Q}^{(\mathbf{0})} + \mu \mathbf{Q}^{(\mathbf{1})}. \tag{1.35}$$

In perturbation theory, $\mathbf{Q}^{(\mathbf{0})}$ represents the unperturbed part, $\mathbf{Q}^{(\mathbf{1})}$ the perturbed part of $\mathbf{Q}_\mu$ and μ is the perturbation parameter. We now aim to find the steady-state probability vector $\boldsymbol{\pi}_\mu$ of the Markov process,

$$\boldsymbol{\pi}_\mu \mathbf{Q}_\mu = \mathbf{0}. \tag{1.36}$$

Now, assume that $\boldsymbol{\pi}_\mu$ is an analytic function of μ. That is, $\boldsymbol{\pi}_\mu$ satisfies the following series expansion representation,

$$\boldsymbol{\pi}_\mu = \sum_{n=0}^{\infty} \mu^n \boldsymbol{\pi}^{(n)} = \boldsymbol{\pi}^{(0)} + \mu \boldsymbol{\pi}^{(1)} + \mu^2 \boldsymbol{\pi}^{(2)} + \ldots \tag{1.37}$$

To determine the different terms of $\boldsymbol{\pi}_\mu$, it is important to make a distinction between a regular and singular perturbation problem. In the former case, the Markov process is irreducible in $\mu = 0$ which means that we can find one unique steady-state solution for $\boldsymbol{\pi}^{(0)}$ by means of the equation:

$$\boldsymbol{\pi}^{(0)} \mathbf{Q}^{(\mathbf{0})} = \mathbf{0}. \tag{1.38}$$

In the latter case, the Markov process is reducible in $\mu = 0$ (but typically irreducible in the neighbourhood of $\mu = 0$) which means that we cannot find a unique steady-state solution for $\boldsymbol{\pi}^{(0)}$ by means of equation (1.38). In this dissertation, we study both types of perturbation.

Regular perturbation First we consider the studied regular perturbation problem. If we substitute (1.35) and (1.37) into (1.36) and take (1.38) into account, we have

$$\boldsymbol{\pi}_\mu \mathbf{Q}_\mu = \boldsymbol{\pi}^{(0)} \mathbf{Q}^{(\mathbf{0})} + \mu(\boldsymbol{\pi}^{(1)} \mathbf{Q}^{(\mathbf{0})} + \boldsymbol{\pi}^{(0)} \mathbf{Q}^{(\mathbf{1})}) + \mu^2 (\boldsymbol{\pi}^{(2)} \mathbf{Q}^{(\mathbf{0})} + \boldsymbol{\pi}^{(1)} \mathbf{Q}^{(\mathbf{1})}) + \ldots = \mathbf{0}$$

such that,

$$\boldsymbol{\pi}^{(n+1)} \mathbf{Q}^{(\mathbf{0})} = -\boldsymbol{\pi}^{(n)} \mathbf{Q}^{(\mathbf{1})} \tag{1.39}$$

for $n \geq 0$. Complementing equations (1.38) and (1.39) with the normalisation conditions

$$\boldsymbol{\pi}^{(0)}\mathbf{1} = 1, \quad \boldsymbol{\pi}^{(n)}\mathbf{1} = 0, \tag{1.40}$$

allows for recursive calculation of all $\boldsymbol{\pi}^{(n)}$ for $n \geq 1$.

The perturbation technique is efficient if the equations (1.38) and (1.39) can be solved efficiently. This is the case for the problems in Chapter 7 and 9, as $\mathbf{Q}^{(\mathbf{0})}$ is triangular. This regular perturbation technique is applied in Chapter 7 on kitting processes with more than two parts. Indeed, the generator matrix of a kitting system with Poisson arrivals and exponential service times has only positive elements above its main diagonal when the service rate μ equals zero. In other words, $\mathbf{Q}^{(\mathbf{0})}$ has an upper-triangular structure when $\mu \to 0$ is the perturbation parameter. This triangular structure implies that the process is transient and that there is only one final state. All queues will eventually fill up completely in the absence of service. Hence, the Markov process for $\mu = 0$ has a unique (and trivial) stationary distribution and the perturbation is regular. This regular perturbation technique is also applied in Chapter 9 on opinion propagation in a social network.

Singular perturbation In contrast to regular perturbation, the unperturbed generator matrix $\mathbf{Q}^{(\mathbf{0})}$ of singular perturbation problems has more than one possible final state when the parameter is sent to zero. In Chapter 8, a kitting system with phase-type service times is analysed. When there is no service, all queues will eventually be full but the system will remain in one of the phases of the service process. Hence, the perturbation is singular as there is an absorbing state for each phase of the phase-type distribution.

Nevertheless, the stationary distribution is analytic in a deleted neighbourhood of 0 and there exists a unique analytic continuation for 0. Practically, the singular perturbation reflects in not having enough equations to solve term by term in the expansion, by consecutively equating terms in $\boldsymbol{\pi}^{(n)}$. That is, $\mathbf{Q}^{(\mathbf{0})}$ in equation (1.39) has a rank lower than its size minus one. It is however possible to find the terms of the expansion by combining the equations one gets for $\boldsymbol{\pi}^{(n)}$ till $\boldsymbol{\pi}^{(n+k)}$ for some integer k. The value k is equal to the order of the Laurent series expansion of the deviation matrix of the Markov process and can be determined by solving a combinatorial problem.

For the singular perturbation problem of Chapter 8, we show that we only have to combine two pairs of equations ($k = 1$). Moreover, we are only "missing" $P - 1$ equations, for solving it directly, where P denotes the number of phases of the phase-type distribution. In this case, the additional computational effort induced by the singular perturbation is limited.

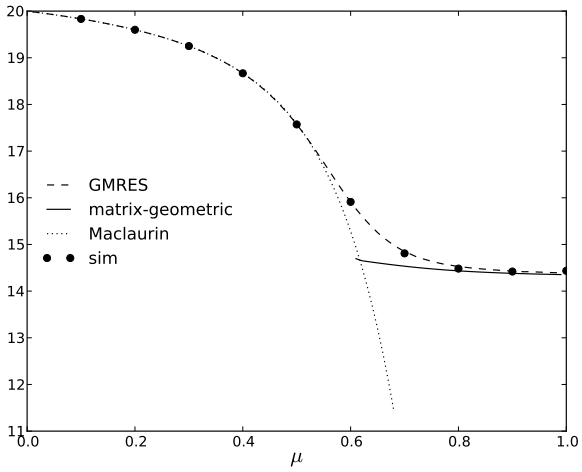


Figure 1.4: The mean number of parts of type 2 *of a two-part kitting process versus the service rate* μ *for the different solution methods.*

1.2.5 Performance analysis of the numerical solution methods

In this section, the proposed numerical techniques are compared in terms of speed and accuracy. To this end, we study a two-part kitting process which can be solved numerically by the different methods introduced above. Parts arrive according to a Poisson process with rate λ_i at their respective part inventories with capacity C_i, $i = \{1,2\}$ and service times are exponentially distributed with rate μ. We choose different λ_i for the queues, to ensure that there exists a stable QBD. Without loss of generality, assume $\lambda_1 < \lambda_2$.

Before discussing the results found in Figures 1.4 and 1.5, the values of the parameters that affect speed and accuracy are given for each solution method. Concerning the GMRES method, the maximum total number of iterations is defined as the product of inner and outer iterations. In both figures, the number of inner iterations equals 20 and the number of outer iterations equals 1. Concerning the matrix-geometric method, the error term to calculate the rate matrix R iteratively equals 10^{-10} and the maximum number of iterations is equal to 10^6. Finally, the number of terms of the developed Maclaurin-series expansion equals 10.

Figure 1.4 depicts the mean number of parts of type 2 of a two-part kitting process calculated with the different solution methods, as well as simulation results which allow for assessing the accuracy of the solution methods. In this figure, the service rate μ varies from 0 to 1 and the inventory capacity $C_1 = C_2$ equals 20 and we assume that $\lambda_1 = 0.6$ and $\lambda_2 = 0.8$. For the QBD process, we make the infi-

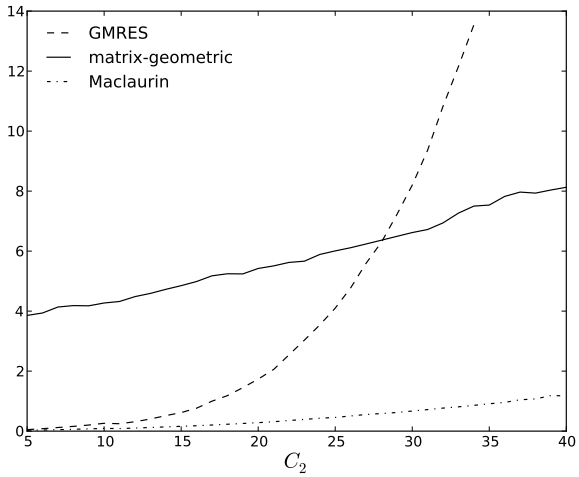


Figure 1.5: The CPU time (in seconds) to calculate the steady-state probability vector of a two-part kitting system by the different solution methods.

nite buffer approximation for queue 1. As the figure shows, the level of accuracy of the Maclaurin-series expansion and the matrix-geometric method vary according to the service rate μ. As expected, the proposed Maclaurin-series expansion is accurate for low values of μ while the matrix-geometric method is accurate for high values of μ. Note that the stability condition of the QBD process (given in Section 1.2.2) requires that $\mu > \lambda_1 (= 0.6)$. Concerning the Maclaurin-series expansion method, as the perturbation parameter μ is sent to zero, the kitting process is obviously well approximated when the service rate has a value in the neighbourhood of zero. Concerning the GMRES method, the level of accuracy does not vary significantly according to the value of the service rate μ. Indeed, the GMRES method is accurate for any value of μ varying from 0 and 1. To summarise, the regions in which the Maclaurin-series expansion and the matrix-geometric method are accurate are complementary for this set of parameter values whereas GMRES is accurate for all μ.

However, the accuracy of GMRES does not come cheap in terms of computational effort. Figure 1.5 depicts the CPU time (averaged over 40 runs) needed by the GMRES, matrix-geometric and Maclaurin-series expansion method to calculate the steady-state probability vector versus the inventory capacity. In this figure, C_1 varies together with C_2 for the GMRES and the Maclaurin-series expansion method and is equal to infinity for the matrix-geometric method, as explained above. The CPU time of the GMRES method includes the time to generate the

sparse generator matrix and to calculate the steady-state probability vector. The CPU time of the matrix-geometric method includes the time to generate the generator matrix of the QBD process and to calculate the rate matrix R and the steady-state probability vector using the rate matrix R. The CPU time of the Maclaurin-series expansion approach includes the time to develop the series expansion of the steady-state probability vector. As the figure shows, the CPU time clearly varies according to the value of the inventory capacity. Indeed, the computational complexity increases as the state space size increases. When comparing the CPU time of the different solution methods, the Maclaurin-series expansion method clearly outperforms the other methods. The matrix-geometric method is shown to have the highest CPU time for an inventory capacity of C_2 varying from 5 to 28 and the GMRES method is shown to have the highest CPU time for an inventory capacity $C_1 = C_2$ varying from 28 to 40. To summarise, the Maclaurin-series expansion method is clearly preferred for low values of μ. For high values of μ, the GMRES and matrix-geometric method are here preferred when the inventory capacity is respectively smaller and larger than 28.

1.3 Dissertation outline

In this section we explain how the chapters of this dissertation are interconnected. As previously mentioned, we investigate three numerical methods applied on four main applications arising in the context of inventory management and telecommunications (see Figure 1.6).

In Chapter 2, we investigate the performance of kitting processes with two parts. In particular, the impact of uncertainty in part arrivals and kit assembly times on the behaviour of the part inventories is assessed. To this end, the kitting system is modelled as a paired queueing system. Methodologically, the sparse matrix technique GMRES is applied and compared with the LU decomposition method in terms of speed. A cost-profit analysis is conducted with the aim of determining the optimal inventory capacity.

The use of matrix-geometric methods starts in Chapter 3 with the performance analysis of a hybrid MTS/MTO system. To account for uncertain demand, inventory replenishment and order processing, these systems are described as stochastic inventory models with two queues: the inventory of semi-finished products and the order backlog. The main modelling differences with Chapter 2 are the assumption of an infinite capacity of one of the queues (instead of finite) and a threshold-based control policy. This means that production of semi-finished products starts when the inventory level drops to a threshold value and stops when the capacity is reached. Methodologically, the studied queueing system is analysed as a homogeneous quasi-birth-death (QBD) process and solved by matrix-geometric methods.

Chapter 4 extends Chapter 3 by increasing the versatility of the developed

queueing model. In particular, we allow for a controlled and uncontrolled replenishment of the semi-finished product inventory. Furthermore, a cost analysis is conducted with the aim of determining the optimal value of the inventory capacity and threshold.

In Chapter 5, we evaluate the performance of energy harvesting sensor nodes under uncertainty in energy capture, energy expenditure, data acquisition and data transmission by means of numerical examples. To this end, the energy harvesting sensor node is again modelled as a paired queueing system. One queue has infinite capacity and keeps track of not yet transmitted data packets. The other queue has finite capacity and represents the energy level of the battery. As in Chapters 3 and 4, this system is modelled as a QBD and solved by matrix-geometric methods. The main difference with previous models is the introduction of limited time periods in which a service can occur. This assumption is motivated by but not limited to scenarios where a mobile sink is responsible for data collection. A mobile sink moves towards the energy harvesting sensor node and gathers the sensed data when it is located in the transmission range of the sensor node. Hence, data transmission is only possible when there is sufficient energy, a data packet available and a transmission opportunity.

Chapter 6 extends Chapter 5 by increasing the versatility of the developed queueing model. Indeed, we allow for simultaneous and non-simultaneous departures from the data packet buffer and the battery level and for non-zero transmission times. As in the previous chapter, the performance of energy harvesting sensor nodes is evaluated under uncertainty in energy capture, energy expenditure, data acquisition and data transmission by means of numerical examples.

Lastly, we consider an alternative methodology to assess the performance of coupled queueing systems with finite capacity. Although the sparse matrix techniques, as elaborated in Chapter 2, lead to quite efficient results compared to the LU decomposition method, there is room for improvement. Indeed, scenarios with more than two finite queues and a reasonable capacity require so much additional calculations that the utility of the analysis as compared to simulations seems limited or are even out of reach due to memory consumption. Hence, mathematical techniques based on series expansions are considered.

In Chapter 7, a numerical algorithm is presented which calculates the Maclaurin-series expansion of the steady-state probability vector, under the condition that the generator matrix reduces to a triangular matrix when a certain rate is sent to zero. The proposed algorithm is illustrated by a kitting process with more than two parts and exponentially distributed service times. In this case, the solution when the service rate is sent to zero is unique; the perturbation is regular. Furthermore, a proof of convergence of the series expansion and a lower bound on the convergence radius are provided. The convergence domain is illustrated by a numerical example.

Chapter 8 extends the results of Chapter 7 by also considering the singular perturbation case. In particular, kitting systems with phase-type distributed service times are assumed. In contrast with the regular case, the unperturbed generator matrix has more than one recurrent class. A numerical method of analysis to cope with singular perturbation is given which keeps the additional computational effort limited.

Chapter 9 is devoted to the study of opinion propagation in a social network. To model such dynamics, we use the contagion approach which is based on the spreading of diseases. Specifically, opinion propagation is modelled as a Markovian non-standard Susceptible-Infected-Recovered (SIR) epidemic model. As in Chapter 7, we develop a Maclaurin-series expansion of the steady-state probability vector in the departure rate. Assuming exponential departure times, we can find a unique steady-state solution when the departure rate is sent to zero. This is again a case of regular perturbation. Also, the fluid limit of the Markov process is derived. By means of numerical examples, we show that the series expansion approximation and fluid limit are complementary such that by combining them we get accurate estimates of performance values for all parameter values.

1.4 Publications

The research performed during the doctoral research led to a number of publications in and submissions to recognised international research fora – journals and conferences.

1.4.1 Papers in international journals

1. E. DE CUYPERE, K. DE TURCK, D. FIEMS, Performance analysis of a kitting process as a paired queue, Hindawi Publishing Corporation Mathematical Problems in Engineering, 2013, vol. 2013, Article ID 843184, 10 pages, http://dx.doi.org/10.1155/2013/843184.

2. E. DE CUYPERE, K. DE TURCK, D. FIEMS, A Maclaurin-series expansion approach to multiple paired queues, 2014, Operations Research Letters, 2014, 42(3), pp.203-207, http://dx.doi.org/10.1016/j.orl.2014.02.003.

3. E. DE CUYPERE, K. DE TURCK, D. FIEMS, Performance analysis of hybrid MTS/MTO systems with stochastic demand, production and service times, submitted to the International Journal of Production Economics, June 2014.

4. E. DE CUYPERE, K. DE TURCK, D. FIEMS, Performance evaluation of an energy harvesting sensor node, submitted to Ad Hoc Networks Journal, June 2014.

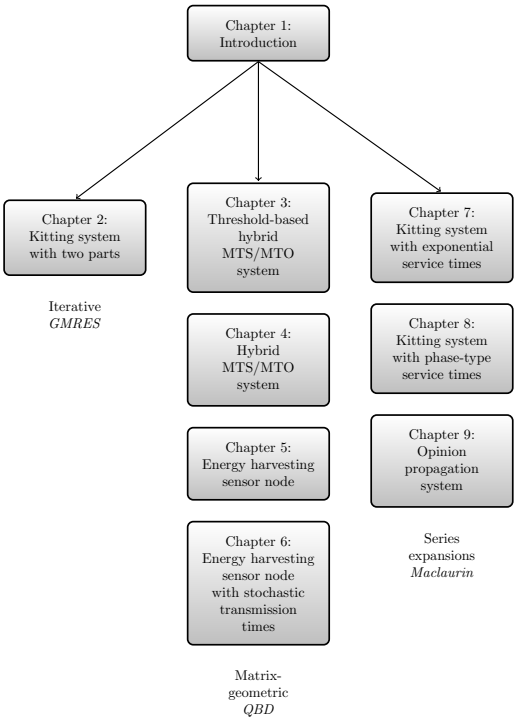


Figure 1.6: Overview of this dissertation.

5. E. DE CUYPERE, K. DE TURCK, S. WITTEVRONGEL, D. FIEMS, A Maclaurin-series expansion approach to coupled queues with phase-type distributed service times, submitted to the European Journal of Operations Research, June 2014.

6. E. DE CUYPERE, K. DE TURCK, D. FIEMS, Opinion propagation in medium-sized populations, submitted to Performance Evaluation, June 2014.

1.4.2 Papers in proceedings of international conferences

1. E. DE CUYPERE, D. FIEMS, Performance evaluation of a kitting process, Lecture Notes in Computer Science, 2011, vol. 6571, pp. 175-188, http://dx.doi.org/10.1007/978-3-642-21713-5_13.

2. E. DE CUYPERE, K. DE TURCK, D. FIEMS, Performance analysis of a decoupling stock in a make-to-order system, IFAC Proceedings Volumes, INCOM-2012 (Bucharest 23-25 May 2012), 14(1), pp. 1493-1498.

3. E. DE CUYPERE, K. DE TURCK, D. FIEMS, A queueing theoretic approach to decoupling inventory, Lecture Notes in Computer Science, 2012, vol. 7314, pp. 150-164, http://dx.doi.org/10.1007/978-3-642-30782-9_11.

4. E. DE CUYPERE, K. DE TURCK, D. FIEMS, Stochastic modelling for low power sensor nodes, 7th International Conference on Queueing Theory and Network Applications, QTNA 2012 (Kyoto, 1-3 August 2012).

5. E. DE CUYPERE, K. DE TURCK, S. WITTEVRONGEL, D. FIEMS, Algorithmic approach to series expansions around transient Markov chains with applications to paired queueing systems, Proceedings of the 6th International Conference on Performance Evaluation Methodologies and Tools, Valuetools 2012 (Cargèse, 9-12 October 2012), pp. 38-44, http://dx.doi.org/10.4108/valuetools.2012. 250292.

6. E. DE CUYPERE, K. DE TURCK, D. FIEMS, H. BRUNEEL, Optimal inventory management in a fluctuating market, Lecture Notes in Computer Science, 2013, vol. 7984, pp. 158-170, http://dx.doi.org/10.1007/978-3-642-39408-9_12.

7. E. DE CUYPERE, K. DE TURCK, S. WITTEVRONGEL, D. FIEMS, Markovian SIR model for opinion propagation, Proceedings of the 25th International Teletraffic Congress, ITC 2013 (Shanghai, 10-12 September 2013), http://dx.doi.org/10. 1109/ITC.2013.6662953.

1.4.3 Abstracts and other presentations

1. E. DE CUYPERE, D. FIEMS, The impact of production interruptions on kitting, an analytical study, Book of Abstracts of the 12th FirW PhD Symposium (Ghent University, 1 December 2010), p. 72.

2. E. DE CUYPERE, D. FIEMS, The impact of production interruptions on kitting, an analytical study, Book of Abstracts of the 22nd Conference on Quantitative Methods for Decision Making, abstract, ORBEL 25 (Ghent, 10-11 February 2011).

3. E. DE CUYPERE, D. FIEMS, Analyse numérique des processus kitting, 12e congrès annuel de la société française de recherche opérationnelle et d'aide à la decision, ROADEF 2011 (Saint-Etienne, 2-4 March 2011), p.899-900.

4. E. DE CUYPERE, D. FIEMS, Economic order quantity of a kitting process with stochastic demand and reordering times, abstract, International Conference on Operations Research, OR 2011 (Zurich, 30 August-2 September 2011).

5. E. DE CUYPERE, K. DE TURCK, D. FIEMS, Performance evaluation of a decoupling inventory for hybrid push-pull systems, abstract, ORBEL 26 (Brussels, 2-3 February 2012).

6. E. DE CUYPERE, K. DE TURCK, D. FIEMS, Economic order quantity of an inventory control system with order backlog and stochastic set-up times, abstract, 25th European Conference on Operational Research, EURO 2012 (Vilnius, 8-11 July 2012).

7. E. DE CUYPERE, K. DE TURCK, D. FIEMS, Analysis of coupled queues (abstract). In session: Analytic methods in queueing systems, sixth Young European Queueing Theorists (YEQT-VI) workshop, Eurandom (Eindhoven, 1-3 November 2012).

8. E. DE CUYPERE, K. DE TURCK, H. BRUNEEL, D. FIEMS, Optimal inventory management in a fluctuating market, abstract, International Conference on Frontiers of Statistics and Its Applications, ICONFROST 2012 (Puducherry, 21-23 December 2012).

9. E. DE CUYPERE, K. DE TURCK, D. FIEMS, Analysis and application of coupled queues, abstract, In session: Management and control of queues, EURO& INFORMS 2013 (Rome 1-4 July 2013).

10. E. DE CUYPERE, K. DE TURCK, D. FIEMS, Opinion spreading of a tourism-related topic in an online travel forum, ttra 45th Annual Interna-

tional Conference Tourism and the New Global Economy, (Brugge, 18-20 June 2014).

11. E. DE CUYPERE, K. DE TURCK, D. FIEMS, Coupled inventory management systems with perishable goods, abstract, First European Conference on Queueing Theory, ECQT 2014 (Ghent 20-22 August 2014).

1.4.4 Award

Best paper Award at the Nineteenth International Conference on Analytical and Stochastic Modelling Techniques and Applications, ASMTA 2012, for the paper: A queueing theoretic approach to decoupling inventory (with K. De Turck and D. Fiems).

References

[1] R. Ackerley. *Telecommunications Performance Engineering, Classical Microscopic Theory, Markov Modulated Poisson Processes*, The Institution of Electrical Engineers, London, U.K., p.17, 2003.

[2] I.F. Akyildiz, W. Su, Y. Sankarasubramaniam and E. Cayirci. *Wireless sensor networks: a survey*, Computer Networks, 38(4), p.393–422, 2002.

[3] I.F. Akyildiz and M.C. Vuran. *Wireless sensor networks*, Wiley, 2010.

[4] H. Andersson and R.M. May. *Infectious Diseases of Humans*, Oxford University Press, Oxford, UK, 1992.

[5] P. Bell. *A Decoupling Inventory Problem with Storage Capacity Constraints*, Operations Research, 28, p.476–488, 1980.

[6] D.A. Bini and B. Meini. *Solving block banded block Toeplitz systems with structured blocks: algorithms and applications*, Advances In Computation: Theory And Practice. Structured matrices: recent developments in theory and computation, Nova Science Publishers, Inc., Commack, NY, USA, p.21–41, 2001.

[7] Y. Bozer and L. McGinnis. *Kitting versus line stocking: A conceptual framework and a descriptive model*, International Journal of Production Economics 28, p.1–19, 1992.

[8] I.N. Bronshtein, K.A. Semendyayev, G. Musiol and H. Muehling. *Power series*, Handbook of Mathematics, Fifth edition, Springer, p.416–418, 2013.

[9] H. Bryznr and M. Johansson. *Design and performance of kitting and order picking systems*, International Journal of Production Economics 41, p.115, 1995.

[10] P. Buchholz. *Structured analysis approaches for large Markov chains*, Applied Numerical Mathematics, 31(4), p.375–404, 1999.

[11] J. Bylina and B. Bylina. *Merging Jacobi and Gauss-Seidel methods for solving Markov chains on computer clusters*, Proceedings of the International multiconference on Computer Science and Information Technology, 20-22 October 2008, p.263–268.

[12] O. Carlsson and B. Hensvold. *Kitting in a high variation assembly line: a case study at caterpillar BCP-E, theoretical framework*, Master thesis, Lule University of Technology, p.6–21, 2008.

[13] K. Chang and Y. Lu. *Queueing analysis on a single-station make-to-stock/make-to-order inventory-production system*, Applied Mathematical Modelling, 34, p.978–991, 2010.

[14] D. Beyer, F. Cheng, S.P. Sethi and M. Taksar. *Markovian Demand Inventory Models*, Series: International Series in Operations Research & Management Science, volume 108, 2010.

[15] A. Chesnokov and M. Van Barel. *A direct method to solve block banded block Toeplitz systems with non-banded Toeplitz blocks*, KU Leuven, Department of Computer Science, report TW 531, September 2008.

[16] J. Cohen. *The single server queue*, North-Holland Pub. Co., Amsterdam, 1969.

[17] E. De Cuypere and D. Fiems. *Performance evaluation of a kitting process*, Proceedings of the 17th International Conference on analytical and stochastic modelling techniques and applications, Lecture Notes in Computer Science, volume 6751, Venice, Italy, 2011.

[18] F.Y. Edgeworth. *The Mathematical Theory of Banking*. Journal of the Royal Statistical Society, 51(1), p.113–127, JSTOR 2979084, 1888.

[19] S.N. Ethier and T.G. Kurtz. *Markov Processes: Characterization and Convergence*, Wiley 2005.

[20] W. Feller. *An Introduction to Probability Theory and Its Applications*, New York: Wiley, 1957.

[21] M. Gribaudo, C.F. Chiasserini, R. Gaeta, M. Garetto, D. Manini and M. Sereno. *A spatial fluid-based framework to analyse large-scale wireless sensor networks*, Proceedings of IEEE International Conference on Dependable Systems and Networks. 2005.

[22] L. Gun. *Experimental results on matrix-analytical solutions techniques - extensions and comparisons*, Stochastic models, 5(4), p.669–682, 1989.

[23] D. Gupta and S. Benjaafar. *Make-to-order, make-to-stock, or delay product differentiation? A common framework for modeling and analysis*, IIE Transactions, 36(6), p.529–546, 2004.

[24] M.H. Gutknecht. *A Brief Introduction to Krylov Space Methods for Solving Linear Systems*, Frontiers of Computational Science, p.53–62, 2007.

[25] F.W. Harris. *How many parts to make at once*, Factory, the magazine of Management, 10(2), p.135–136, 1913.

[26] R. Hassin and M. Haviv. *Mean passage times and nearly uncoupled Markov chains*, SIAM Journal on Discrete Mathematics, 5(3), p.386–397, 1992.

[27] H.W. Hethcote. *The mathematical of infectious diseases*, SIAM Review, 42, p.599–653, 2000.

[28] D.P. Heyman and M. J. Sobel. *Stochastic models in Operation Research: Stochastic processes and Operating Characteristics*, Mc Graw-Hill Book Company, p.112, 1982.

[29] A.L. Hill, D.G. Rand, M.A. Nowak and N.A. Christakis. *Infectious Disease Modeling of Social Contagion in Networks*, Plos computational Biology, 6, 11, 2010.

[30] W.J. Hopp and J. T. Simon, *Bounds and Heuristics for assembly-like queues*, Queueing Systems, 4, p.137–156, 1989.

[31] F.B. James and M. Gilbert. *Comparison of energy harvesting systems for wireless sensor networks*, International Journal of Automation and Computing, 5(4), p.334, 2008.

[32] P. Kaminsky and O. Kaya. *Combined make-to-order/make-to-stock supply chains*, IIE Transactions, 41, p.103–119, 2009.

[33] M. Karpiriski, A. Senart and V. Cahill, *Sensor Networks for Smart Roads*, Fourth Annual IEEE International Conference on Pervasive Computing and Communications Workshops, p.310–314, 2006.

[34] J. Kelif and E. Altman. *Downlink fluid model of cdma networks*, Proceedings of IEEE 61th Vehicular Technology Conference, 2005.

[35] H.S. Kim, J.-H. Kim and J. Kim. *A Review of Piezoelectric Energy Harvesting Based on Vibration*, International Journal of Precision Engineering and Manufacturing, 12(6), p.1129–1141, 2011.

[36] J. Köber and G. Heinecke. *Hybrid Production Strategy between Make-to-order and Make-to-stock - A Case Study at a Manufacturer of Agricultural Machinery with Volative and Seasonal Demand*, 45th CIRP Conference on Manufacturing Systems, 3, p.453–458, 2012.

[37] G. Latouche and V. Ramaswami. *Introduction to Matrix Analytic Methods in Stochastic Modeling*, 1999.

[38] F. Machihara. *A fractal Poisson process and its input queue*, International Journal Computers & Mathematics with Applications, 51, p.199–200, 2006.

[39] A. Mainwaring, D. Culler, J. Polastre, R. Szewczyk and J. Anderson. *Wireless Sensor Networks for Habitat Monitoring*, Proceedings 1st ACM International Workshop on Wireless Sensor Networks and Applications, p.88–97, 2002.

[40] J. Mapes, C. New and M. Szwejczewski. *Trade-offs in manufacturing plants*, International Journal of Operations and Production Management, 17, p.1020–1033.

[41] L. Medbo. *Assembly work execution and materials kit functionality in parallel flow assembly systems*, International Journal of Industrial Ergonomics, 31, p.263–281, 2003.

[42] S.P. Meyn and R.L. Tweedy. *Markov chains and stochastic stability*, Springer-Verlag, London, 1997.

[43] M.F. Neuts. *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*, The Johns Hopkins Press, 1981.

[44] B.F. Nielsen. *Lecture notes on phase-type distributions for 02407, Stochastic Processes*, DTU Informatics, Technical University of Denmark, October 2012.

[45] J.R. Norris. *Markov chains*, Cambridge University Press, 1997.

[46] H. Ohta, T. Hirota and A. Rahim. *Optimal production-inventory policy for make-to-order versus make-to-stock based on the M/Er/1 queuing model*, International Journal of Advanced Manufacturing Technologies, 33, p.36–41, 2007.

[47] B. Philippe, Y. Saad, and W. Stewart. *Numerical methods in Markov chain modelling*, Operations Research, 40(6), p.1156–1179, 1992.

[48] A. Popescu. *Tools for modern Techniques of Networking: Markov - Modulated Poisson Processes*, Department of Telecommunication Systems, p.1, 2000.

[49] H. Rafiei and M. Rabbani. *Capacity coordination in hybrid make-to-stock/make-to-order production environments*, International Journal of Production Research, 50(3), p.773–789, 2012.

[50] S. Ramachandran and D. Delen. *Performance analysis of a Kitting process in stochastic assembly systems*, Computers & Operations Research, 32(3), p.449–463, 2005.

[51] K. Ramachandran, L. Whitman and A. Ramachandran. *Criteria for determining the push-pull boundary*, Industrial Engineering Research Conference. Orlando, FL, USA, 2002.

[52] R. Ramakrishnan and A. Krishnamurthy. *Analytical approximations for Kitting systems with multiple inputs*, Asia-Paific Journal of Operations Research 25(2), p.187–216, 2008.

[53] X. Ren and W. Liang. *Delay-Tolerant Data Gathering in Energy Harvesting Sensor Networks With a Mobile Sink*, Proceedings of GLOBECOM, IEEE, 2012.

[54] W. Liang and J. Luo. *Network lifetime maximization in sensor networks with multiple mobile sinks*, Proceedings of LCN, IEEE, 2011.

[55] N.R. Roshyara, *Krylov Subspace Iteration*, 2005.

[56] Y. Saad and M. Schultz. *GMRES: A Generalized Minimal Residual Algorithm for Solving Non symmetric Linear Systems*, SIAM Journal on Scientific and Statistical Computing, 7, p.586–869, 1986.

[57] Y. Saad. *Iterative methods for sparse linear systems, second edition*, Chapter 4: Basic iterative methods, SIAM, p.105–112, 2003.

[58] P. Som, W. Wilhelm and R. Disney. *Kitting process in a stochastic assembly system*, Queueing Systems, 17, p.471–490, 1994.

[59] C. Soman, D. van Donk and G. Gaalman. *Combined make-to-order and make-to-stock in a food production system*, International Journal of Production Economics, 90, p.223–235, 2004.

[60] W. Stewart. *Introduction to the Numerical Solution of Markov Chains*, Princeton University Press, p.197-205, 1994.

[61] G. Strang. *Linear Algebra and Its Applications*, 2nd Ed., Orlando, FL, Academic Press, Inc., 22, 1980.

[62] D. Turgut and L. Bölöni. *Heuristic approaches for transmission scheduling in sensor networks with multiple mobile sinks*, The Computer Journal, 54(3), p.332–344, 2009.

[63] D.S. Watkins. *The Matrix Eigenvalue Problem: GR and Krylov Subspace Methods*, Chapter 9: Krylov subspace method, 2007.

[64] G. Werner-Allen, K. Lorincz, M. Ruiz, O. Marcillo, J. Johnson, J. Lees and M. Welsh. *Deploying a Wireless Sensor Network on an Active Volcano*, IEEE Internet Computing, 10(2), p. 18–25, 2006.

[65] D.M. Young. *Iterative solutions of large linear systems*, Academic Press, New York, 1971.

[66] Y. Yun and Y. Xia. *Maximizing the lifetime of wireless sensor networks with mobile sink in delay-tolerant applications*, IEEE Trans. Mobile Computing, 9, p.1308–1318, 2010.

2

Performance analysis of a kitting process as a paired queue

Eline De Cuypere, Koen De Turck and Dieter Fiems

published in Hindawi Publishing Corporation Mathematical Problems in Engineering, 2013, vol. 2013, Article ID 843184, 10 pages.

Abstract. Nowadays, customers request more variation in a company's product assortment leading to an increased amount of parts moving around on the shop floor. To cope with this tendency, a kitting process can be implemented. Kitting is the operation of collecting the necessary parts for a given end product in a specific container, called a kit, prior to arriving at an assembly unit. As kitting performance is critical to the overall cost and performance of the manufacturing system, this paper analyses a two-part kitting process as a Markovian model. In particular, kitting is studied as a paired queue, thereby accounting for stochastic part arrivals and kit assembly times. Using sparse matrix techniques, we assess the impact of kitting interruptions, bursty part arrivals and phase-type distributed kit assembly times on the behaviour of the part buffers. Finally, a cost-profit analysis of kitting processes is conducted.

2.1 Introduction

Nowadays manufacturing systems are often composed of multiple in-house fabrication units [12]. The semi-finished products stemming from these units are the input materials for other fabrication units or for assembly lines. Hence, efficient transport of materials between the different stages of the production process is key for overall production cost minimisation. Kitting is a particular strategy for supplying materials to an assembly line. Instead of delivering in containers, each holding a single type part and all holding the same number of parts, kitting collects the necessary set of parts for an individual end product in a specific container, referred to as kit, prior to arriving at an assembly unit [1, 2, 12, 14, 15, 17].

Kitting mitigates storage space requirements at the assembly station since no part inventories need to be kept there. Moreover, parts are placed in proper positions in the container such that assembly time reductions can be realised. Additional benefits include reduced learning time of the workers at the assembly stations and increased quality of the product. Although kitting is a non-value adding activity, its application can reduce the overall materials handling time [15]. Indeed activities such as selecting and gripping parts are performed more efficiently. Furthermore, the whole operator walking time is drastically reduced or even eliminated since kits, each containing a complete set of components, are brought to the assembly station [9]. The advantages mentioned above do not come for free since the kitting operation itself incurs additional costs such as the time and effort for planning the allocation of the parts into kits and the kit preparation itself. Moreover, the introduction of a kitting operation in a production process involves a major investment and the effect on efficiency are uncertain. Therefore, it is important to analyse the performance of kitting in a production environment prior to its actual introduction. This is the subject of the present paper.

In literature, most authors consider a kitting process as a queueing system with stochastic part arrivals and kit assembly times. Hopp and Simon [8] developed a model for a kitting process with exponentially distributed processing times for kits and Poisson arrivals. They found accurate bounds for the required buffer capacity of kitting processes with two parts. Explicitly accounting for finite buffer capacities, Som et al. [17] further refined the results of Hopp and Simon.

Of course real buffers always have a finite capacity, the capacity being constrained by the storage room. However, if the capacity is large enough, we can have a good approximation of a process with a finite capacity on the basis of a model with unlimited capacity. This means that there is always enough space for upcoming parts, which simplifies the analysis. Unfortunately, the assumption of an infinite buffer is not valid for kitting processes. If the capacity is assumed to be infinite, then the model will degrade to an unstable stochastic system. Harrison [7] showed for a multiple input generalisation of the GI/G/1 queue that it is

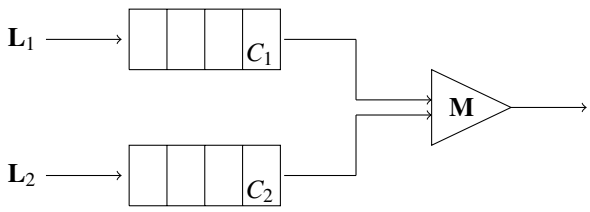


Figure 2.1: Kitting process: the buffers are on the left, and the triangularly shaped kitting process is on the right.

necessary to impose a restriction on the size of the buffer to ensure stability in the operations of a kitting process. Under this assumption, the probability to have a certain long-term stock position is equal and independent of the current stock position. This was also demonstrated by Latouche [10] who studied waiting lines with paired customers. We can consider this analysis as an abstraction of a kitting process with two types of parts.

In this work, we focus on a kitting process modulated by a Markovian environment. The introduction of this environment allows us to study kitting under more realistic stochastic assumptions: kitting interruptions, bursty part arrivals, phase-type distributed kit assembly times etc. Our paper extends the results on kitting in a Markovian environment [5].

The remainder of this paper is organised as follows. Section 2.2 describes the kitting process at hand. In Section 2.3, Chapman-Kolmogorov equations are derived and their numerical solution is discussed. To illustrate our approach, Section 2.4 considers a number of numerical examples. In particular, we assess the impact of kitting interruptions, bursty part arrivals and phase-type distributed kit assembly times on the behaviour of the part buffers. Then, a cost-profit analysis of kitting processes is conducted. Finally, conclusions are drawn in Section 2.5.

2.2 Model description

In this paper, we study a two-queue kitting process, as depicted in Figure 2.1. Each queue has a finite capacity — let C_ℓ denote the capacity of buffer ℓ, $\ell = 1,2$ — and models the inventory of parts of a single type. New parts arrive at the buffers and, if both buffers are nonempty, a kit is assembled by collecting a part from each buffer. Hence, departures from the buffers are synchronised, the buffers are paired. Operation of part buffers therefore considerably differs from other queueing systems.

Arrivals at both buffers are modelled by a Markovian arrival process and kit assembly is not instantaneous. For ease of modelling, it is assumed that there is a modulating Markov process, arrival and service rates depending on the state of

this process. To be more precise, the kitting process is modelled as a continuous-time Markov process with state space $\mathcal{C}_1 \times \mathcal{C}_2 \times \mathcal{K}$, whereby $\mathcal{C}_\ell = \{0, \ldots, C_\ell\}$ for $\ell = 1,2$ and with $\mathcal{K} = \{1,2,\ldots,K\}$ being the state space of the modulating process. At any time, the state of the kitting process is described by the triplet $[m,n,i]$, m and n being the number of parts in the first and second buffer respectively, and i being the state of the modulating process. We now describe the state transitions.

- The state of the modulating process can change when there are neither arrivals nor departures. Let α_{ij} denote the transition rate from state i to state j ($i,j \in \mathcal{K}$, $i \neq j$) and let $\mathbf{A}$ denote the corresponding generator matrix.
- The state of the modulating process may remain the same or may change when there is an arrival. Let $\lambda_{ij}^{(\ell)}$ denote the (marked) transition rate from state i to state j when there is an arrival at buffer ℓ, $\ell = 1,2$. Moreover, let $\mathbf{L}_\ell$ denote the corresponding generator matrix. Note that such marked transitions from state i to state i are allowed.
- Analogously, the state of the modulating process may remain the same or may change when there is a departure (in each buffer). Let μ_{ij} and $\mathbf{M}$ denote the corresponding transition rate and generator matrix respectively.

Summarising, arrivals at and departures from the buffers are described by the generator matrices $\mathbf{A}$, $\mathbf{L}_1$, $\mathbf{L}_2$ and $\mathbf{M}$. So far, no diagonal elements of $\mathbf{A}$ have been defined. To simplify notation, it will be further assumed that the diagonal elements are chosen such that the row sums of $\mathbf{A}+\mathbf{L}_1+\mathbf{L}_2+\mathbf{M}$ are zero.

The computational method employed here does not require any homogeneity of the generator matrices. When required by the applications at hand, intensities may depend on the buffer content. In this case, we introduce superscripts to make this dependence explicit. For example, $\mathbf{M}^{(m,n)}$ denotes the generator matrix of state transitions with departure when there are m parts in buffer 1 and n parts in buffer 2. In addition, we use arguments for rates as we already used superscripts and subscripts to distinguish the arrival rates at the different queues. For example, $\lambda_{ij}^{(\ell)}(m,n)$ denotes the arrival rate at buffer $\ell = 1,2$ from state i to state j when there are m parts in buffer 1 and n parts in buffer 2.

Example 1. In the most basic setting, parts arrive at the buffers in accordance with an independent Poisson process with rate λ_1 and λ_2 and kit assembly times are exponentially distributed with parameter μ. In this case, there is no need to have a modulating Markov process, the state is completely described by the number of parts in each buffer, (m,n). We have,

$$\mathbf{M} = \begin{bmatrix}\mu\end{bmatrix}, \quad \mathbf{L}_1 = \begin{bmatrix}\lambda_1\end{bmatrix}, \quad \mathbf{L}_2 = \begin{bmatrix}\lambda_2\end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix}-\lambda_1-\lambda_2-\mu\end{bmatrix}.$$

Example 2. To account for burstiness in the arrival process of the parts at the different buffers, the modulating process allows the mitigation of the Poissonian arrival assumptions: We can replace the Poisson processes by a two-class Markovian arrival processes. Multi-class Markovian arrival processes allow for intricate correlation and can be efficiently characterised from trace data [4, 6]. As we have two types of arrivals, the Markovian arrival process is described by the generator matrix $\mathbf{\Lambda}_1$ of transitions with arrivals at buffer 1, the generator matrix $\mathbf{\Lambda}_2$ with arrivals at buffer 2 and the generator matrix $\mathbf{\Lambda}_0$ without arrivals. As usual, the diagonal elements of $\mathbf{\Lambda}_0$ are negative and ensure that the row sums of $\mathbf{\Lambda}_0+\mathbf{\Lambda}_1+\mathbf{\Lambda}_2$ are zero. Retaining exponentially distributed kit assembly times, we have,

$$\mathbf{L}_1 = \mathbf{\Lambda}_1\,, \quad \mathbf{L}_2 = \mathbf{\Lambda}_2\,, \quad \mathbf{A} = \mathbf{\Lambda}_0 - \mu\mathbf{I}\,, \quad \mathbf{M} = \mu\mathbf{I}\,.$$

Here $\mathbf{I}$ denotes the identity matrix.

Example 3. As for the arrival processes, the model at hand is sufficiently flexible to include phase-type distributed kit assembly times. The phase-type distribution is completely characterised by an initial probability vector $\boldsymbol{\tau}$ and the matrix $\mathbf{T}$ which corresponds to non-absorbing transitions [11]. Let $\mathbf{t}' = -\mathbf{T1}$ be the column vector with the rates to the absorbing state and let $\mathbf{f}$ be a row vector with zero-elements except the first one. Assuming Poisson arrivals in both buffers (with rate λ_1 and λ_2, respectively), we get the following matrices,

$$\begin{aligned}
\mathbf{L}_1^{(m,n)} &= \lambda_1\mathbf{I}\left(1-\mathbb{1}_{\{m=0,n>0\}}\right)+\lambda_1\mathbf{1}\boldsymbol{\tau}\mathbb{1}_{\{m=0,n>0\}}\\
\mathbf{L}_2^{(m,n)} &= \lambda_2\mathbf{I}\left(1-\mathbb{1}_{\{m>0,n=0\}}\right)+\lambda_2\mathbf{1}\boldsymbol{\tau}\mathbb{1}_{\{m>0,n=0\}}\\
\mathbf{A}^{(m,n)} &= \mathbf{T}\mathbb{1}_{\{m>0,n>0\}}-\lambda_1\mathbf{I}-\lambda_2\mathbf{I}\\
\mathbf{M}^{(m,n)} &= \mathbf{t}'\boldsymbol{\tau}\mathbb{1}_{\{m>1,n>1\}}+\mathbf{t}'\mathbf{f}(1-\mathbb{1}_{\{m>1,n>1\}})
\end{aligned}$$

Here, it is assumed that the background state equals 1 if one of the buffers is empty. When service starts again, the background state is chosen in accordance with the probability vector $\boldsymbol{\tau}$.

2.3 Analysis

Having established the modelling assumptions and settled our notation, we now focus on the analysis of the kitting process.

2.3.1 Balance equations

We aim to define a set of equations for the steady-state probability vector for the Markov process $[Q_1(t), Q_2(t), S(t)]$, $Q_\ell(t)$ being the number of parts in buffer ℓ at time t and $S(t)$ being the state of the background process at time t.

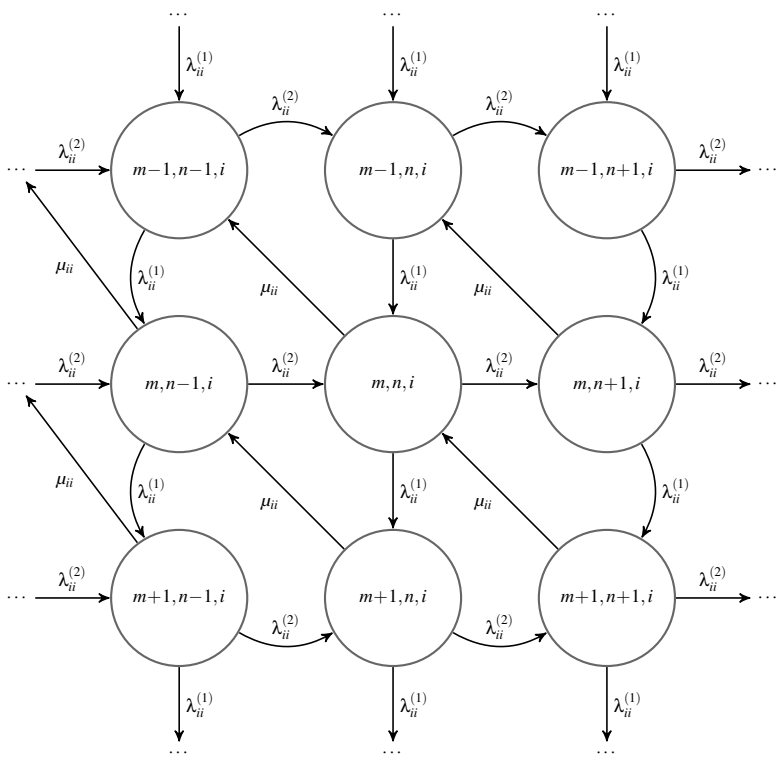


Figure 2.2: Fragment of the transition rate diagram for state (m,n,i)

Let $\pi_i(m,n) = \lim_{t\to\infty} \mathbb{P}[Q_1(t) = m, Q_2(t) = n, S(t) = i]$ be the steady-state probability to be in state $[m,n,i]$ and let $\boldsymbol{\pi}(m,n)$ be the vector with elements $\pi_i(m,n)$, for $i \in \mathcal{K}$. Figure 2.2 shows a fragment of the transition rate diagram of the kitting model in state $[m,n,i]$. As mentioned above, two independent input streams arrive at the buffers with intensity $\lambda_{ij}^{(\ell)}$ and are processed into kits with intensity μ_{ij}. Upon completion of a kit, the content of both buffers is decreased by 1. Note that we only show the transitions whereby the modulating Markov process remains in state i. Moreover, possible dependence of the transition rates on the buffer sizes is not indicated.

Based on the transition rate diagram, we now derive the balance equations of the kitting process at hand.

- First, consider the case where both buffers are neither empty nor full ($0 < n < C_1$ and $0 < m < C_2$). We have,

$$\begin{aligned}
&\pi_i(m,n)\left(\sum_{j=1}^{K}\lambda_{ij}^{(1)}(m,n)+\sum_{j=1}^{K}\lambda_{ij}^{(2)}(m,n)+\mu_{ij}(m,n)+\sum_{j=1,j\neq i}^{K}\alpha_{ij}(m,n)\right)\\
&\quad=\sum_{j=1}^{K}\pi_j(m-1,n)\lambda_{ji}^{(1)}(m-1,n)+\sum_{j=1}^{K}\pi_j(m,n-1)\lambda_{ji}^{(2)}(m,n-1)\\
&\qquad+\sum_{j=1}^{K}\pi_j(m+1,n+1)\mu_{ji}(m+1,n+1)\\
&\qquad+\sum_{j=1,j\neq i}^{K}\pi_j(m+1,n+1)\alpha_{ji}(m,n)\,,
\end{aligned}$$

  for $i,j \in \mathcal{K}$ or equivalently,

$$\begin{aligned}
\boldsymbol{\pi}(m-1,n)\mathbf{L}_1^{(m-1,n)}&+\boldsymbol{\pi}(m,n-1)\mathbf{L}_2^{(m,n-1)}\\
&+\boldsymbol{\pi}(m+1,n+1)\mathbf{M}^{(m+1,n+1)}+\boldsymbol{\pi}(m,n)\mathbf{A}^{(m,n)}=\mathbf{0}\,.
\end{aligned}$$

- If buffer 1 is empty and buffer 2 neither empty nor full ($m = 0$ and $0 < n < C_2$), we have,

$$\begin{aligned}
&\pi_i(0,n)\left(\sum_{j=1}^{K}\lambda_{ij}^{(1)}(0,n)+\sum_{j=1}^{K}\lambda_{ij}^{(2)}(0,n)+\sum_{j=1,j\neq i}^{K}\alpha_{ij}(0,n)\right)\\
&\quad=\sum_{j=1}^{K}\pi_j(0,n-1)\lambda_{ji}^{(2)}(0,n-1)+\sum_{j=1}^{K}\pi_j(1,n+1)\mu_{ji}(1,n+1)\\
&\qquad\qquad+\sum_{j=1,j\neq i}^{K}\pi_j(0,n)\alpha_{ji}(0,n)\,,
\end{aligned}$$

for $i, j \in \mathcal{K}$ or equivalently,

$$\boldsymbol{\pi}(0,n-1)\mathbf{L}_2^{(0,n-1)} + \boldsymbol{\pi}(1,n+1)\mathbf{M}^{(1,n+1)} + \boldsymbol{\pi}(0,n)(\mathbf{A} + \mathrm{diag}(\mathbf{M}^{(0,n)}\mathbf{1}) = \mathbf{0}.$$

- Similarly, if buffer 2 is empty and buffer 1 neither empty nor full ($n = 0$ and $0 < m < C_1$), we have,

$$\boldsymbol{\pi}(m-1,0)\mathbf{L}_1^{(m-1,0)} + \boldsymbol{\pi}(m+1,1)\mathbf{M}^{(m+1,1)} + \boldsymbol{\pi}(m,0)(\mathbf{A} + \mathrm{diag}(\mathbf{M}^{(m,0)}\mathbf{1}) = \mathbf{0}.$$

- If both buffers are empty ($m = 0$ and $n = 0$), we have,

$$\boldsymbol{\pi}(1,1)\mathbf{M}^{(1,1)} + \boldsymbol{\pi}(0,0)(\mathbf{A}^{(0,0)} + \mathrm{diag}(\mathbf{M}^{(0,0)}\mathbf{1})) = \mathbf{0}.$$

- If buffer 1 is empty and buffer 2 is full ($m = 0$ and $n = C_2$), we get,

$$\boldsymbol{\pi}(0,C_2-1)\mathbf{L}_2^{(0,C_2-1)} + \boldsymbol{\pi}(0,C_2)(\mathbf{A}^{(0,C_2)} + \mathrm{diag}(\mathbf{M}^{(0,C_2)}\mathbf{1} + \mathbf{L}_2^{(0,C_2)}\mathbf{1})) = \mathbf{0}.$$

- Similarly, if buffer 1 is full and buffer 2 is empty ($m = C_1$ and $n = 0$), we have,

$$\boldsymbol{\pi}(C_1-1,0)\mathbf{L}_1^{(C_1-1,0)} + \boldsymbol{\pi}(C_1,0)(\mathbf{A}^{(C_1,0)} + \mathrm{diag}(\mathbf{M}^{(C_1,0)}\mathbf{1} + \mathbf{L}_1^{(C_1,0)}\mathbf{1})) = \mathbf{0}.$$

- Finally, if both buffers are full ($m = C_1$ and $n = C_2$), we find,

$$\begin{aligned}\boldsymbol{\pi}(C_1-1,C_2)\mathbf{L}_1^{(C_1-1,C_2)} &+ \boldsymbol{\pi}(C_1,C_2-1)\mathbf{L}_2^{(C_1,C_2-1)} \\ &+ \boldsymbol{\pi}(C_1,C_2)(\mathbf{A}^{(C_1,C_2)} + \mathrm{diag}(\mathbf{L}_1^{(C_1,C_2)}\mathbf{1} + \mathbf{L}_2^{(C_1,C_2)}\mathbf{1})) = \mathbf{0}.\end{aligned}$$

2.3.2 Performance measures

Given the steady-state vectors $\boldsymbol{\pi}(m,n)$, we can now obtain a number of interesting performance measures for the kitting system. For ease of notation, let $\pi^{(t)}(m,n) = \boldsymbol{\pi}(m,n)\mathbf{1}$ denote the probability to have m parts in buffer 1 and n parts in buffer 2. Moreover let $\pi^{(1)}(m) = \sum_n \pi^{(t)}(m,n)$ and $\pi^{(2)}(n) = \sum_m \pi^{(t)}(m,n)$ denote the marginal probability mass functions of the queue content of buffer 1 and buffer 2, respectively.

The following performance measures are of interest

- The mean buffer 1 and 2 content: $\mathrm{E}\,Q_1$ and $\mathrm{E}\,Q_2$ respectively,

$$\mathrm{E}\,Q_1 = \sum_m^{C_1} \pi^{(1)}(m)m, \quad \mathrm{E}\,Q_2 = \sum_n^{C_2} \pi^{(2)}(n)n.$$

- The variance of the buffer 1 and 2: $\operatorname{Var} Q_1$ and $\operatorname{Var} Q_2$ respectively,

$$\operatorname{Var} Q_1 = \sum_{m}^{C_1} \pi^{(1)}(m) m^2 - (\mathrm{E}\, Q_1)^2, \quad \operatorname{Var} Q_2 = \sum_{n}^{C_2} \pi^{(2)}(n) n^2 - (\mathrm{E}\, Q_2)^2.$$

- The effective load of the system ρ_{eff} is the fraction of time that kitting is ongoing. As kitting is only ongoing when none of the buffers is empty, we have,

$$\rho_{\text{eff}} = 1 - \pi^{(1)}(0) - \pi^{(2)}(0) + \pi^{(t)}(0,0).$$

- Let the throughput η be defined as the number of kits departing from the system per time unit. Taking into account all possible states from which we can have a departure, we find,

$$\eta = \sum_{m=1}^{C_1} \sum_{n=1}^{C_2} \boldsymbol{\pi}(m,n) \mathbf{M}^{(m,n)} \mathbf{1}.$$

- The shortage probability K is the probability that one of the buffers is empty,

$$K = \pi^{(1)}(0) + \pi^{(2)}(0) - \pi^{(t)}(0,0).$$

- The loss probability b_i in buffer i is the probability that an arriving part cannot be stored in buffer i, $i = 1,2$. By noting that the accepted arrival load equals the departure rate, we find,

$$b_i = \frac{\rho_i - \eta}{\rho_i}.$$

  Where ρ_i is the arrival load of part $i = \{1,2\}$. If the arrival load in both buffers are the same, then the loss probability is also the same: $b_1 = b_2$. If the arrival load is not the same, then the excess load in the most loaded buffer is lost as well.

2.3.3 Methodology: the sparse matrix technique

Queueing models for kitting processes are rather complicated. Indeed, the modelled kitting process has a multidimensional state space. Even for relative moderate buffer capacity, the multidimensionality leads to huge state spaces; this is the so-called state-space explosion problem. For many queueing systems, infinite-buffer assumptions may mitigate this problem. Given some buffer system with finite capacity, more efficient numerical routines can be constructed for the corresponding queueing system with infinite capacity. Unfortunately, as mentioned above, the infinite-buffer capacity assumption is not applicable for kitting processes and therefore cannot simplify the analysis. Recall that the infinite-capacity model is

always unstable. For all input parameters except trivial ones (no arrivals), some or all of the queues grow unbounded with positive probability.

Consequently, the multidimensionality of the state space and the inapplicability of the infinite-buffer assumption yield Markov processes with a finite but very large state space. However, the number of possible state transitions from any specific state is limited. This means that most of the entries in the generator matrix are zero; the matrix is sparse. As illustrated by the numerical examples, using sparse matrices and their associated specialised algorithms results in manageable memory consumption and processing times, compared to standard algorithms.

The method used here to solve the sparse matrix equation is the projection method GMRES (Generalized Minimum Residual). Details can be found in Stewart [18] (p.197–205), Philippe et al. [13], Buchholz [3] and Saad and Schultz [16]. To solve the linear equation $Ax = b$, the GMRES algorithm approximates x by the vector $x_n \in K_n$ in a Krylov subspace $K_n = \text{span}\{b, Ab, ..., A^{n-1}b\}$ which minimises the norm of the residual $\|Ax_n - b\|$. Since the residual norm is minimised at every step of the method, it is clear that it is non-increasing. However, the work and storage requirement per iteration increases linearly with the iteration count. Hence, the cost of n iterations grows by $O(n^2)$ which is a major drawback of GMRES. This limitation is usually overcome by restarting the algorithm. After a chosen number of iterations m, the accumulated data is cleared and the intermediate results are used as the initial data for the next m iterations [16].

To ensure fast convergence, it is also key to properly choose the initial vector that is passed on to the algorithm. We rely on MATLAB's build in GMRES algorithm and assume a uniform initial vector if no additional information about the solution is known. However, it is often the case that there is additional information about the solution. Indeed, calculation speed can be further improved as, in practice, performance measures are not calculated for an isolated set of parameters. E.g., when a plot is created, a parameter is varied over a range of values. In this case, a previously calculated steady-state vector for some set of parameters can be used as a first estimate of the steady-state vector for a new "perturbed" set of parameters. Using previously calculated steady-state vectors is trivial if the state spaces corresponding to the parameter sets are equal. In this case, the previously calculated steady-state vector can be passed on unmodified. If the state space changes, the steady-state vector must be rescaled to the new state space. In general, adding zero-probability states if the state space increases or removing states if the state space decreases, turns out to be ineffective. This is easily explained by a simple example. Assume that we increase the queue capacity of one of the part buffers. Typically, even for moderate load, a considerable amount of probability mass can be found for queue size equal to capacity. Increasing the queue size and assigning zero probability to the new states is not a good estimate for the new steady-state vector. Also for the system with higher capacity, a considerable

amount of probability mass can be found when the queue size equals the capacity (while zero probabilities were assigned).

2.4 Numerical results

With the balance equations at hand, we now illustrate our numerical approach by means of some examples.

2.4.1 Bursty part arrivals

As a first example, we quantify the impact of production inefficiency on the performance of a kitting process. To this end, we compare part buffers with Poisson arrivals to corresponding kitting systems with interrupted Poisson arrivals. The arrival interruptions account for inefficiency in the production process. Kit assembly times are assumed to be exponentially distributed with service rate equal to one, this value being independent of the number of parts in the different buffers. This is a kitting process with Markovian arrivals as described in Example 2 in Section 2.2.

The interrupted Poisson process considered here is a two-state Markovian process. In the active state, new parts arrive in accordance with a Poisson process with rate λ whereas no new parts arrive in the inactive state. Let α and β denote the rate from the active to the inactive state and vice versa, respectively. We then use the following parameters to characterise the interrupted Poisson process,

$$\sigma = \frac{\beta}{\alpha+\beta}, \quad \kappa = \frac{1}{\alpha} + \frac{1}{\beta}, \quad \rho = \lambda\sigma.$$

Note that σ is the fraction of time that the interrupted Poisson process is active, the absolute time parameter κ is the average duration of an active and an inactive period, and ρ is the arrival load of the parts.

Figure 2.3 shows the mean number of parts in buffer 1 versus the arrival load, for various values of the buffer capacities C_1 and C_2 for Poisson arrivals (for both buffers) as well as for interrupted Poisson arrivals (again for both buffers). We set $\sigma = 0.4$ and $\kappa = 10$ for the interrupted Poisson processes. Clearly, the mean buffer content increases as the arrival load increases as expected. Moreover, if more buffer capacity is available, it will also be used: the mean buffer content increases for increasing values of $C_1 = C_2$. Comparing interrupted Poisson and Poisson processes, burstiness in the production process has a negative impact on performance — more buffering is required — if the queues are not fully loaded ($\rho < 1$). As for ordinary queues, the opposite can be observed for overloaded buffers. In this case, the effect of a large burst is mainly reflected in loss and not in additional queue content. Burstiness also yields larger periods without arrivals during which the buffer size decreases.

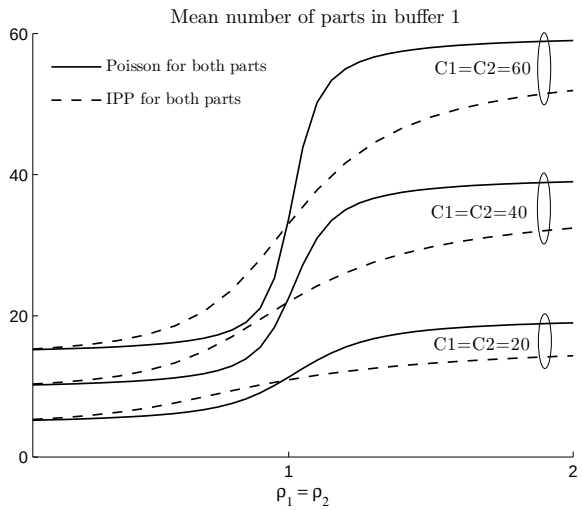


Figure 2.3: When the queues are not fully loaded, burstiness in the production process has a negative impact on the performance.

By numerical examples, we can quantify the expected buffer behaviour - e.g. more production yields higher buffer content, higher buffer capacity mitigates the loss probability etc. However, less trivial behaviour can be observed as well. Figure 2.4 depicts the probability that the buffer is full versus the buffer capacity $C_1 = C_2$. We compare performance of kitting with Poisson arrivals to kitting with interrupted Poisson arrivals at one buffer and at both buffers. As in the preceding figure, the interrupted Poisson processes are characterised by $\sigma = 0.4$ and $\kappa = 10$. The arrival load of both buffers ρ_1 and ρ_2 is equal to 0.8. As expected, the probability that the buffer is full decreases for increasing values of the buffer capacities. Moreover, to reduce this probability, more buffer capacity is required for the case of two interrupted Poisson processes than for the case of two Poisson processes. For the kitting process with one Poisson and one interrupted Poisson process, non-trivial performance results can be observed. Namely, interruptions in the production of a part more negatively affect buffer performance of the other part. Indeed, buffer 1 is full with higher probability if the arrivals at buffer 2 are interrupted than if the arrivals at buffer 1 are interrupted. First note that the loss probabilities in both buffers are the same. For the Poisson buffer, parts are rejected at the arrival rate ρ_1 whenever the buffer is full. For the IPP buffer, parts are rejected when the buffer is full and the arrival process is in the on-state at rate ρ_2/σ. We define S as the background state of the arrival process. This state equals 0 when the arrival process is in an off-state and equals 1 when the arrival process is in an on-state. As the loss probability in both buffers is

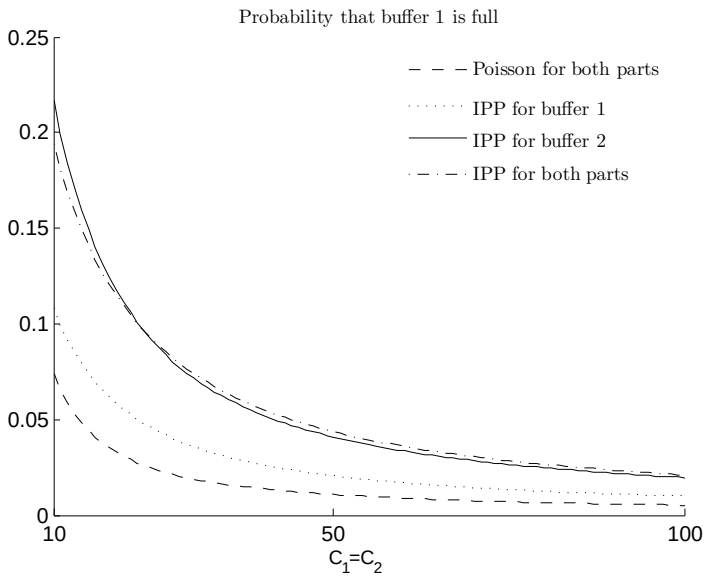


Figure 2.4: Irregularity in the production of a part leads to a higher probability to have a full buffer for the other part.

the same, we have then $\mathbb{P}[Q_1 = C_1]\rho_1 = \mathbb{P}[Q_2 = C_2, S = 1]\rho_2/\sigma$, or equivalently, $\mathbb{P}[Q_1 = C_1] = \mathbb{P}[Q_2 = C_2|S = 1]$. As the second queue is more likely to be full when there are arrivals we have $\mathbb{P}[Q_2 = C_2|S = 1] \geq \mathbb{P}[Q_2 = C_2]$ which shows $\mathbb{P}[Q_1 = C_1] \geq \mathbb{P}[Q_2 = C_2]$.

2.4.2 Phase-type distributed kit assembly times

The second numerical example quantifies the impact of the distribution of the kit assembly times on kitting performance. In particular, we here study Erlang-distributed kit assembly times. Limiting ourselves to Poisson arrivals to both buffers, this numerical example fits Example 3 of Section 2.2.

Figures 2.5 and 2.6 depict the mean number of parts in buffer 1 and the loss probability in buffer 1 for the kitting process and, as a reference point, for the M/E/1/n queue as well. In both figures, the arrival load is varied and different values of the variance of the kitting time distribution are assumed as indicated. The mean kitting time is equal to 1 for all curves and the capacity of both buffers is equal to 20. In underload ($\rho < 1$), kitting performs worse than the M/E/1/n queue: the mean buffer content and the loss probability have a higher value. This follows from the fact that kitting stops when one of the buffers is empty. By increasing the load, it is obvious that the buffer content converges to the capacity and the loss

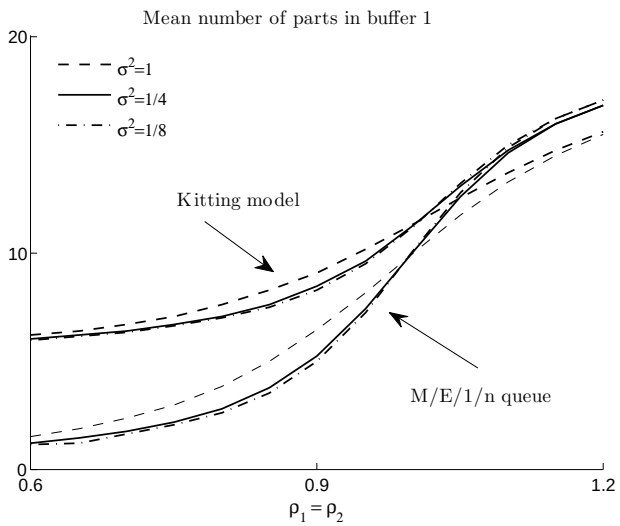


Figure 2.5: Given the mean kitting time, the corresponding kitting time distribution has only a limited impact on the mean number of parts in buffer 1 *in this case.*

probability to one. It is most interesting to observe that the shape of the service time distribution only has a small effect on these performance measures. Indeed, there is no significant performance difference when σ^2 equals $1/4$ and when it equals $1/8$.

2.4.3 Cost and profit analysis

We now add a cost structure to the kitting process under study. In particular, cost and profit for the kitting systems of Section 2.4.1 and 2.4.2 are analysed.

The proposed cost function is,

$$c_1(\mathrm{E}Q_1 + \mathrm{E}Q_2) + c_2K + c_3(b_1 + b_2)$$

where c_1 is the holding cost of a part in the buffer, c_2 is the shortage cost in one or in both of the buffers and c_3 is the loss cost. Note that for all figures, the input parameters are symmetric for both parts such that $\mathrm{E}Q_1 = \mathrm{E}Q_2$ and $b_1 = b_2$.

Poisson arrivals

In Figure 2.7 the total cost of applying kitting systems versus the buffer capacity (varying from 1 to 30) is depicted. Limiting ourselves to Poisson arrivals for both buffers, $\rho_i = \lambda_i = 0.8$ for part $i = \{1,2\}$. We compare kitting performance for different costs as depicted. As expected, higher cost values lead to higher total

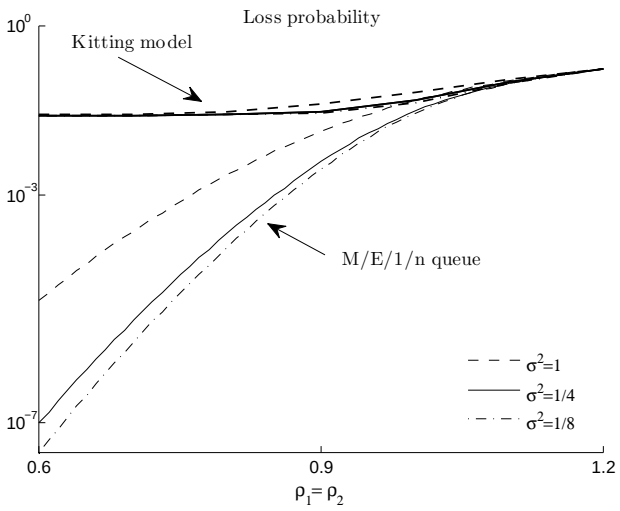


Figure 2.6: Given the mean kitting time, the corresponding kitting time distribution has only a limited impact on the loss probability in this case.

cost. The cost structure also affects the optimal buffer capacity. When looking at the different costs separately, we observe that a higher holding cost decreases the optimal buffer capacity (from 15 to 12), a higher loss cost increases the optimal capacity (from 12 to 14), whereas the optimal buffer capacity remains the same (12) for a higher shortage cost. Obviously, buffering is interesting when the storage cost is low and when the cost of rejecting parts (because of a full buffer) is high. It is most interesting to observe that, as the capacity increases, the cost models converge when the sum of the holding and the shortage cost $(c_1 + c_2)$ is equal. Indeed, as the loss tends to zero, the loss cost tends to zero as well. Furthermore, when the capacity equals one, the state space of the kitting model has size 4 and hence the cost is easily calculated explicitly,

$$\text{total cost} = (2(c_1 + c_3)(\mu + 2\lambda) + 3\mu c_2)(3\mu + 2\lambda)^{-1}.$$

Bursty part arrivals

Next, we analyse the cost and the profit of kitting systems with different Markovian arrivals. We use the same values to model the arrivals as in Figure 2.4. In Figure 2.8(a) and 2.8(b) the total cost (left) and the profit (right) versus the buffer capacity are depicted. We consider a holding cost c_1 equal to 2, a shortage cost c_2 equal to 55 and a loss cost c_3 equal to 40. On the left figure, the optimal buffer capacity for Poisson arrivals is 12, for an interrupted Poisson arrival at one buffer

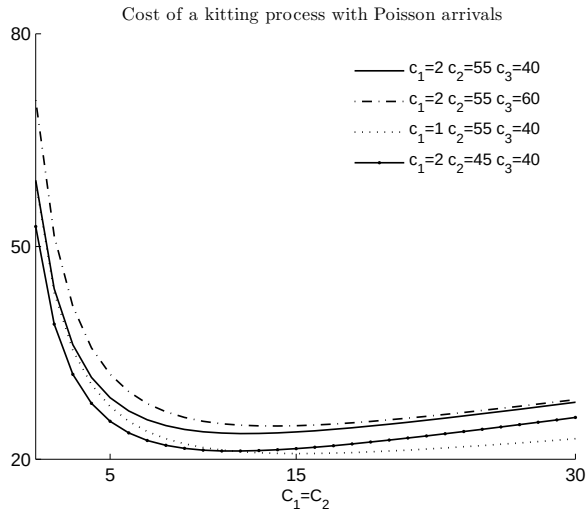
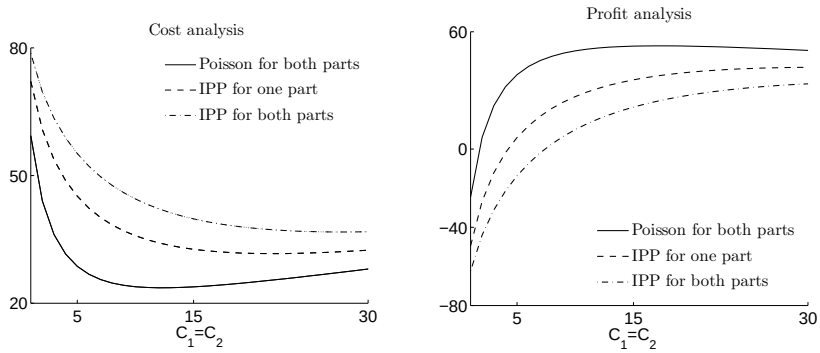


Figure 2.7: Each type of cost has a different impact on the value of the optimal buffer capacity.

it equals 22 and for an interrupted Poisson process at both buffers the optimal capacity is 28. As expected, the optimal buffer capacity increases as the burstiness in the production process increases. Concerning the profit analysis, we assume that the yield equals the throughput multiplied by a sale unit equal to 100. Assuming a maximum storage room of 30 parts, we observe that the optimal buffer capacity is 5, 11 and 12 for the depicted arrival processes respectively. Consequently, kitting systems under production inefficiency require much higher storage space, especially when profit is applied as the parameter to determine the optimal buffer capacity. Moreover, the optimal capacity is very sensitive to the burstiness parameters σ and κ. In the example at hand, the plots suggest that the cost function is a convex function of the buffer capacity. However, one can also easily choose the cost parameters to obtain non-convex cost functions.

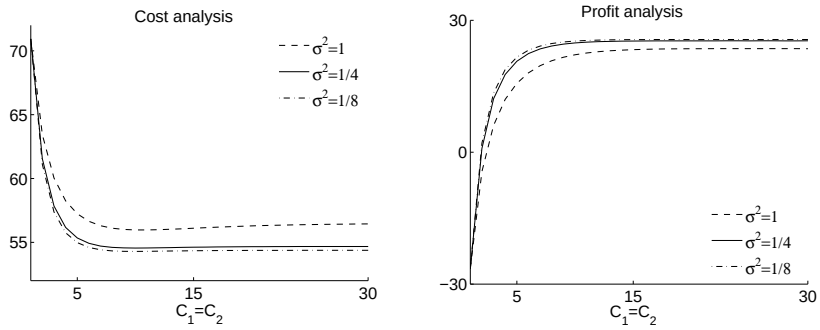
Phase-type kit assembly times

Finally, Figure 2.9(a) and 2.9(b) depict the cost (left) and the profit (right) for a kitting system with Erlang-distributed kit assembly times versus the buffer capacity. As in the preceding figure, the profit equals the difference between the yield (equal to the throughput multiplied by 100) and the cost (defined by the parameters $c_1 = 2$, $c_2 = 55$ and $c_3 = 40$). Out of the numerical examples, we observe that the shape of the service time distribution has a very limited impact on profit and cost. These results confirm those found in Section 2.4.2.



(a) (b)

Figure 2.8: Production inefficiency results in higher required storage space.



(a) (b)

Figure 2.9: Given the mean kitting time, the corresponding kitting time distribution has only a limited impact on the value of the optimal buffer capacity in this case.

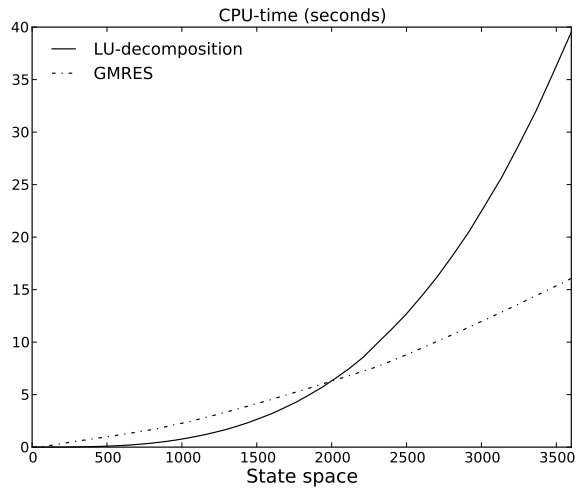


Figure 2.10: The GMRES method performs well in terms of speed.

2.4.4 Performance analysis of solution methods

We compare the performance of (sparse) GMRES and solving the Markov process by standard LU decomposition [19] on a kitting process with Poisson arrivals with rate $\lambda_i = 0.8$ for part $i = \{1, \ldots, 2\}$ and with exponentially distributed kit assembly times with rate μ. Figure 2.10 depicts both methods in terms of speed versus the state space for a kitting process starting from a symmetric buffer capacity $C_1 = C_2 = 1$ to $C_1 = C_2 = 60$. While LU performance is better than GMRES when the capacity (and state space) is small, the figure clearly shows that GMRES performs considerably better than LU decomposition for a symmetric buffer capacity equal or larger than 44.

2.5 Conclusion

In this paper, we evaluate the performance of two-part kitting processes in a Markovian setting. Furthermore, a cost-profit analysis is conducted. Note that the particularity of the studied kitting systems is that the part buffers are paired. This means that each demand requires both parts such that a kit can only be assembled if both inventories are nonempty. Methodologically, we have applied sparse matrix techniques (e.g. GMRES) as most of the entries in the generator matrix have a value equal to zero. The solution is not exact but performs well in terms of solution speed and accuracy.

As our numerical results show, the interplay between the different queues leads to complex performance behaviour. For example, interruptions in the production of a part more negatively affect buffer performance of the other part. Overall, we observe extreme sensitivity of kitting performance on arrival process parameters while performance is reasonable insensitive to variation of the kitting time distribution. Finally, we determine the optimal buffer capacity based on a cost-profit analysis.

References

[1] Y. Bozer and L. McGinnis. *Kitting versus line stocking: A conceptual framework and a descriptive model*, International Journal of Production Economics 28, p.1–19, 1992.

[2] H. Bryznr and M. Johansson. *Design and performance of kitting and order picking systems*, International Journal of Production Economics 41, p.115–125, 1995.

[3] P. Buchholz. *Structured analysis approaches for large Markov chains*, Applied Numerical Mathematics, 31(4), p.375–404, 1999.

[4] P. Buchholz, P. Kemper and J. Kriege. *Multi-class Markovian arrival processes and their parameter fitting*, Performance Evaluation, 67(11), p.1092–1106, 2010.

[5] E. De Cuypere and D. Fiems. *Performance evaluation of a kitting process*, Proceedings of the 17th International Conference on analytical and stochastic modelling techniques and applications, Lecture Notes in Computer Science, vol. 6751, Venice, Italy, 2011.

[6] D. Fiems, B. Steyaert and H. Bruneel. *A genetic approach to Markovian characterisation of H.264 scalable video*, Multimedia Tools and Applications, p.1–22, 2011.

[7] J.M. Harrison. *Assembly-Like Queues*, Journal of Applied Probability, 10(2), p.354–367, 1973.

[8] W. J. Hopp and J. T. Simon. *Bounds and Heuristics for assembly-like queues*, Queueing Systems, 4, p.137–156, 1989.

[9] B. Johansson and M.I. Johansson. *High Automated Kitting System for Small Parts: a Case Study from the Volvo Uddevalla Plant*, Proceedings of the 23rd International Symposium on Automotive Technology and Automation, Vienna, Austria, p.75–82, 1990.

[10] G. Latouche. *Queues with paired customers*, Journal of Applied Probability, 18, p.684–696, 1981.

[11] G. Latouche and V. Ramaswami. *Introduction to Matrix Analytic Methods in Stochastic Modeling*, 1999.

[12] L. Medbo. *Assembly work execution and materials kit functionality in parallel flow assembly systems*, International Journal of Industrial Ergonomics, 31, p.263–281, 2003.

[13] B. Philippe, Y. Saad and W. Stewart. *Numerical methods in Markov chain modeling*, Operations Research, 40(6), p.1156–1179, 1992.

[14] S. Ramachandran and D. Delen. *Performance analysis of a Kitting process in stochastic assembly systems*, Computers & Operations Research, 32(3), p.449–463, 2005.

[15] R. Ramakrishnan and A. Krishnamurthy. *Analytical approximations for Kitting systems with multiple inputs*, Asia-Pacific Journal of Operations Research, 25(2), p.187–216, 2008.

[16] Y. Saad and M. Schultz. *GMRES: A Generalized Minimal Residual Algorithm for Solving Non symmetric Linear Systems*, SIAM Journal on Scientific and Statistical Computing, 7, p.586–869, 1986.

[17] P. Som, W. Wilhelm and R. Disney. *Kitting process in a stochastic assembly system*, Queueing Systems, 17, p.471–490, 1994.

[18] W. Stewart. *Introduction to the Numerical Solution of Markov Chains*, Princeton University Press, 1994.

[19] G. Strang. *Linear Algebra and Its Applications*, 2nd Ed., Orlando, FL, Academic Press, Inc., 22, 1980.

3

A queueing theoretic approach to decoupling inventory

Eline De Cuypere, Koen De Turck and Dieter Fiems

published in Lecture Notes in Computer Science, 2012, vol. 7314, pp.150–164.

Abstract. This paper investigates the performance of different hybrid push-pull systems with a decoupling inventory at the semi-finished products and re-ordering thresholds. Raw materials are 'pushed' into the semi-finished product inventory and customers 'pull' products by placing orders. Furthermore, production of semi-finished products starts when the inventory goes below a certain level, referred to as the threshold value and stops when the inventory attains stock capacity. As performance of the decoupling stock is critical to the overall cost and performance of manufacturing systems, this paper introduces a Markovian model for hybrid push-pull systems. In particular, we focus on a queueing model with two buffers, thereby accounting for both the decoupling stock as well as for possible backlog of orders. By means of numerical examples, we assess the impact of different reordering policies, irregular order arrivals, the setup time distribution and the order processing time distribution on the performance of hybrid push-pull systems.

3.1 Introduction

In a make-to-stock system (push type), products are stocked in advance, while in a make-to-order system (pull type), a product only starts to be manufactured when a customer order is placed, see a.o [26, 18, 14, 27, 9]. Nowadays, as a means to respond quickly to growing variety, shorter product life cycles while keeping inventory costs as low as possible, hybrid push-pull systems are introduced [25]. An important issue in the overall performance of such hybrid systems is the position of the decoupling point [25, 22]. Hoekstra et al. [12] defined the customer order decoupling point (CODP) concept. These authors considered market, product and production related factors as well as the desired service level and associated inventory costs to locate the optimal decoupling point. Under different hybrid push/pull control policies, Pandey and Khokhajaikiat [21] conducted a case study concerning the design and performance evaluation of a multistage production system. Results indicated that the choice of the optimal decoupling positions changes with the extent of raw material constraint operating at the stages and the demand lead time variabilities. To account for a degree of customisation and short delivery times, Blecker and Abdelkafi [2] considered a decoupling point at the inventory of semi-finished products. Here, after an order is received, only the final completion step still needs to be done. A case study at Phoenix showed that, by a hybrid approach, the company would save 20 to 25 percent of the total late costs and inventory costs compared to a pure push approach, which was at that time being used [6]. Research on the performance of the decoupling inventory in a hybrid push-pull system is therefore of main importance. This is the subject of the present paper.

In the present setting, we use a queuing theoretic approach to study the hybrid push-pull system. Queuing theory has already been successfully applied to assess decoupling points. Kaminsky and Kaya [15] considered a variety of combined make-to-order (MTO) and make-to-stock (MTS) supply chains with a single manufacturer and a single supplier in order to minimise a function of the total inventory, lead times and tardiness. The arrival process at the manufacturer is treated as a single facility with multiple classes of Poisson arrivals scheduled FCFS. As in previous research, they concluded that costs can be cut dramatically by using a combined system instead of pure MTO or MTS systems. Ohta and al. [20] analysed a multi-product inventory system where demand for each item arrive according to a Poisson process and the production time has an Erlang distribution. An optimality condition that specifies whether each product should be produced MTS or MTO is proposed. Bell [1] investigated a decoupling inventory between two successive production stages, the demand at stage 2 being independent from production at stage 1. The stages are decoupled by storing intermediate products. Limits on the available storage capacity and the rates of flow production into and out of the decoupling inventory are set, which enables the firm to determine the op-

timum capacities for the storage facility and to determine the value of an additional supply of intermediate product. Chang and Lu [5] studied a one-station production system consistent with MTO and MTS productions and dealing with two types of random demands: ordinary demand and specific demand. In this system, both types of demand arrive according to a Poisson process and production times of the workstation are exponentially distributed. Specific demand has a higher priority with respect to ordinary demand and the performance of this system is studied by means of matrix-geometric methods.

The present study of the decoupling stock closely relates to literature on two-part assembly systems, sometimes termed paired queues or kitting processes. For such systems, there are two queues, each storing a specific part, and production only starts when both part buffers are non-empty. In the current setting, one part-buffer corresponds to the decoupling stock, while the other corresponds to the list of backlogged orders. Also, production only starts when both buffers are non-empty. Indeed, each delivery of a finished product requires both the order specifications and a semi-finished product and can only be satisfied if both are present. If both part-buffers have unlimited capacity, Harrison [11] was the first to prove that, assuming no arrival control strategy, this queueing system is inherently unstable. In particular, he studied the multiple-input extension of the GI/G/1 queue in which arrivals in each stream are described by an independent renewal process and service times are independent and identically distributed. He showed that part waiting times converge to non-defective limiting distributions only if the buffer capacities are bounded. This was also demonstrated by Latouche [16] who termed the two-part assembly system as waiting lines with paired customers. He considered a system of infinite capacity queues with Poisson arrivals for both parts and exponential services. The steady state is attained, i.e. the system is stable, if the arrival rates depend on the difference between queue lengths in a certain way. [3] extended Latouche's research by considering two exponential distributions, one for the part processing distribution, i.e. the synchronisation phase, and the other for the assembly operation distribution. Approximations for the throughput rate and average queue length were given. Lipper and Sengupta [19] is another extension of the work of Latouche. In this paper, multiple Poisson input streams arrive in buffers with finite capacity. A more general structure in which parts are withdrawn from infinite pools and processed prior to assembly has been studied by Hopp et al. [13] and Som et al. [23]. Som and Wilhem [24] studied a two-queue system in which each part is processed according to an exponential distribution and the assembly operation times are generally distributed. They follow a matrix-geometric approach to numerically determine the marginal distributions of both kit and end-product inventory positions. Finally, assuming finite part-buffers, a two-part assembly system in a Markovian environment is studied in [7] by numerically solving the corresponding Markov processes by the generalized minimal residual

method (GMRES).

Furthermore, this article analyses hybrid push-pull systems with a threshold inventory: once the stock of semi-finished products drops below some level, this is either communicated to the production department if the parts are produced in-house or an order is placed with a third-party company if this is not the case. In both cases, it may take some time, the reordering time, before the inventory is replenished. Then, production stops when the semi-finished product inventory level attains stock capacity. The studied inventory control system closely relates to the well-known economic order quantity (EOQ) model [10]. This is a deterministic fluid-model for a single inventory and determines optimal reordering policies which balance the purchase, order and storage costs. While the single-part inventory problem is well understood, both in a deterministic and a stochastic setting, many issues of optimal inventory management in the multi-queue inventory case remain unresolved, most prominently in the stochastic setting.

In contrast to previous research, this paper investigates a two-queue system with one finite and one infinite buffer. Indeed, to limit involved costs, the decoupling stock needs to be sufficiently small. Hence, finite capacity is assumed. However, no such assumption is imposed for the other queue: the order backlog queue has an infinite capacity. Assuming a finite capacity product queue also assures the existence of a steady-state solution, provided that the arrival rate of orders is limited. In particular, this article analyses hybrid push-pull systems under different threshold policies, assuming that production stops when the inventory level reaches maximum capacity. Comparing versatility and numerical tractability, we study the decoupling stock in a Markovian environment as in [7]. This approach allows for assessing the effect of variability in the production process of semi-finished products, the ordering process and the delivery process on the performance of the decoupling stock.

The remainder of this paper is organised as follows. Section 3.2 describes the decoupling stock model at hand. In Section 3.3, the decoupling inventory system is analysed as a quasi-birth-death-process (QBD) and a number of specific application scenarios for the decoupling inventory system are introduced. Also, the numerical solution methodology is discussed and relevant performance measures are determined. To illustrate our approach, Section 3.4 considers some numerical examples. Finally, conclusions are drawn in Section 3.5.

3.2 Model description

The decoupling stock is modelled as a queueing model with two queues, as depicted in Figure 3.1. The product queue has finite capacity C_p and stores the semi-finished products prior to being processed to finished products. Moreover, production of semi-finished products starts when the inventory goes below the

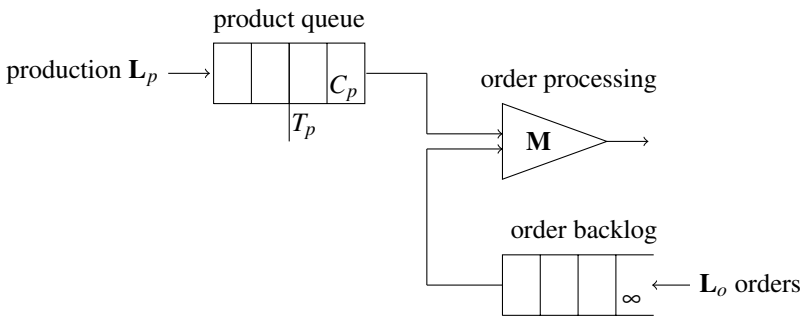


Figure 3.1: Decoupling inventory of semi-finished products in a hybrid push-pull system.

threshold value T_p and stops when the inventory level reaches capacity C_p. The order queue keeps track of the orders that have not yet been delivered and has infinite capacity. Arriving orders are served in accordance with a first-come-first-served queueing discipline. Each order takes a semi-finished product from the product queue and completes the product in accordance with order specifications. Note that the two queues in the model at hand are tightly coupled. Departures from the product queue are only possible when there are orders. Similarly, departures from the order queue are only possible if there are semi-finished products in the product queue.

Arrivals at both queues are modelled according to possibly dependent arrival processes and order completion is not instantaneous. For ease of modelling, it is assumed that there is a modulating Markov process, arrival and service rates depending on the state of this modulating process. To be more precise, the decoupling inventory system is a three-dimensional continuous-time Markov process with infinite state space $\mathbb{N} \times \{0,1,2,\ldots,C_p\} \times \mathcal{K}$, $\mathcal{K} = \{0,1,\ldots,K\}$ being the state space of the modulating process. At any time, the state of the decoupling inventory system is described by the triplet (n,m,i), n being the number of backlogged orders, m being the number of semi-finished products and i being the state of the modulating process. We now describe the state transitions.

- The state of the modulating process can change when there are neither arrivals nor departures. Let α_{ij} denote the transition rate from state i to state j ($i,j \in \mathcal{K}$, $i \neq j$). Further, for ease of notation, let

  $$\alpha_{ii} = -\sum_{j \neq i} \alpha_{ij}.$$

  and let $\mathbf{A} = [\alpha_{ij}]_{i,j \in \mathcal{K}}$ denote the corresponding generator matrix. Further, it is assumed that when either of the queues is empty, different transition rates (when there are neither arrivals nor departures) can be specified: let $\hat{\alpha}_{ij}$ and

$\hat{\mathbf{A}}$ denote the transition rate from state i to state j and the corresponding generator matrix, respectively.

- The state of the modulating process may remain the same or may change when there is an arrival. Let $\lambda_{ij}^{(p)}$ and $\lambda_{ij}^{(o)}$ denote the (marked) transition rate from state i to state j when there is an arrival at the product queue and the order queue, respectively. Moreover, let $\mathbf{\Lambda}_p = [\lambda_{ij}^{(p)}]_{i,j\in\mathcal{K}}$ and $\mathbf{\Lambda}_o = [\lambda_{ij}^{(o)}]_{i,j\in\mathcal{K}}$ denote the corresponding generator matrices. Note that marked self transitions from state i to state i are allowed.

- Analogously, the state of the modulating process may remain the same or may change when there is a departure (in each buffer). Let μ_{ij} and $\mathbf{M}$ denote the corresponding transition rate and generator matrix respectively.

The transition rates are dependent on the product queue size, the state of the modulating process and whether the order queue is empty, e.g. there are no product arrivals when the queue is full, production starts only when the semi-finished product inventory level goes below the threshold value and there are only departures if both queues are non-empty.

3.3 Analysis

3.3.1 Quasi-birth-death process

The studied Markov process is a homogeneous quasi-birth-death process (QBD), see [17]. In the present setting, the so-called level or block-row number, indicates the number of backlogged orders while the phase, i.e. the index within a block element, indicates both the content of the decoupling stock and the state of the Markovian environment. The one-step transitions are restricted to states in the same level (from state $(n,*,*)$ to state $(n,*,*)$) or in two adjacent levels (from state $(n,*,*)$ to state $(n+1,*,*)$ or state $(n-1,*,*)$).

We then find that the generator matrix of the Markov process has the following block matrix representation,

$$\mathbf{Q} = \begin{bmatrix} \mathbf{L}_p' & \mathbf{L}_o & \mathbf{0} & 0 & \cdots \\ \mathbf{W} & \mathbf{L}_p & \mathbf{L}_o & 0 & \cdots \\ 0 & \mathbf{W} & \mathbf{L}_p & \mathbf{L}_o & \cdots \\ 0 & 0 & \mathbf{W} & \mathbf{L}_p & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}. \tag{3.1}$$

The blocks are given by,

$$\mathbf{L}_o = \begin{bmatrix} \mathbf{\Lambda}_o^{(0)} & 0 & 0 & \cdots & 0 \\ 0 & \mathbf{\Lambda}_o^{(1)} & 0 & \cdots & 0 \\ 0 & 0 & \mathbf{\Lambda}_o^{(2)} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \mathbf{\Lambda}_o^{(C_p)} \end{bmatrix}, \tag{3.2}$$

$$\mathbf{L}_p = \begin{bmatrix} \underline{\mathbf{D}}^{(0)} & \mathbf{\Lambda}_p^{(0)} & 0 & \cdots & 0 \\ 0 & \mathbf{D}^{(1)} & \mathbf{\Lambda}_p^{(1)} & \cdots & 0 \\ 0 & 0 & \mathbf{D}^{(2)} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \mathbf{D}^{(C_p)} \end{bmatrix}, \tag{3.3}$$

$$\mathbf{L}_p' = \begin{bmatrix} \underline{\mathbf{D}}^{(0)} & \mathbf{\Lambda}_p^{(0)} & 0 & \cdots & 0 \\ 0 & \underline{\mathbf{D}}^{(1)} & \mathbf{\Lambda}_p^{(1)} & \cdots & 0 \\ 0 & 0 & \underline{\mathbf{D}}^{(2)} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \underline{\mathbf{D}}^{(C_p)} \end{bmatrix}, \tag{3.4}$$

$$\mathbf{W} = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ \mathbf{M}^{(1)} & 0 & \cdots & 0 & 0 \\ 0 & \mathbf{M}^{(2)} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \mathbf{M}^{(C_p)} & 0 \end{bmatrix}. \tag{3.5}$$

with $\mathbf{D}^{(m)} = \mathbf{A}^{(m)} - \partial\mathbf{\Lambda}_o^{(m)} - \partial\mathbf{\Lambda}_p^{(m)} - \partial\mathbf{M}^{(m)}$ and $\underline{\mathbf{D}}^{(m)} = \hat{\mathbf{A}}^{(m)} - \partial\mathbf{\Lambda}_o^{(m)} - \partial\mathbf{\Lambda}_p^{(m)}$ with $m = (0, 1, 2, \ldots, C_p)$ being the number of semi-finished products in the buffer. Note that $\partial\mathbf{X}$ represents a diagonal matrix with diagonal elements equal to the row sums of $\mathbf{X}$. Intensities in the generator matrices Λ_o, Λ_p, $\underline{\mathbf{D}}$, $\mathbf{D}$ and $\mathbf{M}$ are dependent of the product buffer content m. Therefore, we introduce the superscript $^{(m)}$ to make this dependence explicit. Note that if no superscript is indicated, the intensities in the generator matrix are equal for all numbers of semi-finished products in the queue.

To simplify notation, states representing an inactive production and a product queue content equal or less than the threshold value, are taken into account in the generator matrix structure. However, as production is always active when the semi-finished product inventory level is below the threshold value, the next transition changes the given inactive background state to an active one. The matrix structure also considers states where the number of semi-finished products equals

capacity and the background state is active. Again, the next transition changes the background state into an inactive state.

In the general case, arrivals and departures at both queues are modelled according to possibly dependent Markovian arrival processes (MAP) and phase-type distributed order processing times, respectively. The Markovian arrival processes are described by the generator matrices $\mathbf{B}_1^{(m)}$ and $\mathbf{B}_3^{(m)}$ with arrivals of semi-finished products and orders, respectively and the generator matrices $\mathbf{B}_0^{(m)}$ and $\mathbf{B}_2^{(m)}$ without arrivals at the decoupling stock and the queue of backlogged orders, respectively. Let P and O denote the size of the state space of the semi-finished product arrival process $\mathcal{P}$ and the order arrival process O, respectively. The phase-type distribution is completely characterised by an initial probability vector $\boldsymbol{\tau}$, by the matrix $\mathbf{T}$ which corresponds to non-absorbing transitions and by $\mathbf{t}' = -\mathbf{T}\mathbf{1}$ which is the column vector of rates to the absorbing state with $\mathbf{1}$ a column vector of ones [17]. Let T denote the size of the state space $\mathcal{T}$ of the production process.

We use the environment variable to track the states of both arrival processes as well as of the state of the order processing time. It is notationally convenient to define the environment variable as the triplet (x_p, x_o, x_s) with x_p the state of the product arrival process, x_o the state of the order arrival process and x_s the state of the order processing time. As all state variables are finite, we can easily map the triplets on $\mathcal{K}$. Accounting for a threshold-based replenishment policy, we additionally have to specify the state of the product arrival process when it restarts. Let c_i be the probability that the production process restarts in state $i \in \mathcal{P}$ and let $\mathbf{c} = [c_i]_{i \in \mathcal{P}}$. Remark that the inclusion of a start state allows for introducing setup times. For example, a phase-type distributed setup time prior to production is easily introduced as shown in the numerical examples.

When the inventory level m is equal or below T_p, production is always ongoing. Hence for $m \leq T_p$ the environment variable takes values in,

$$\{(x_p, x_o, x_s) : x_p \in \mathcal{P}, x_o \in O, x_s \in \mathcal{T}\}.$$

When the inventory level is at least T_p but not full, the environment variable needs to track whether production is on-going or not. For ease of notation, we assume that the state of the production process (as in the preceding section) is augmented with an additional state θ. When the production process is in this state, there is no production. Hence, for inventory level m and $T_p < m < C_p$, the environment variable takes values in,

$$\{(x_p, x_o, x_s) : x_p \in \mathcal{P} \cup \{\theta\}, x_o \in O, x_s \in \mathcal{T}\}.$$

In contrast to the preceding extension of the state description, we are sure that there is no production when the inventory is full. In this case there is no need to track the state of the production process. Hence for inventory level $m = C_p$, the

environment variable takes values in,

$$\{(x_o, x_s) : x_o \in O, x_s \in \mathcal{T}\}.$$

With the state space defined as above, we now construct the matrices $\mathbf{\Lambda}_o$, $\mathbf{\Lambda}_p$, $\mathbf{A}$, $\hat{\mathbf{A}}$ and $\mathbf{M}$. Introducing the auxiliary matrices and vectors,

$$\widehat{\mathbf{B}}_i = \begin{bmatrix} \mathbf{B}_i & \mathbf{0}_B \\ \mathbf{0}'_B & 0 \end{bmatrix} \quad \widetilde{\mathbf{B}}_i = \begin{bmatrix} \mathbf{B}_i & \mathbf{0}_B \end{bmatrix}$$

$$\mathbf{b}_1 = \begin{bmatrix} \mathbf{B}_1 \mathbf{1}_B \\ 0 \end{bmatrix} \quad \mathbf{C} = \begin{bmatrix} \mathbf{I}_B \\ \mathbf{c} \end{bmatrix} \quad \mathbf{f} = \begin{bmatrix} \mathbf{0}'_B & 1 \end{bmatrix}.$$

The symbols $\mathbf{0}_X$ and $\mathbf{1}_X$ represent column vectors of size X with zeroes and ones, respectively. We get the following generator matrices,

$$\mathbf{\Lambda}_p^{(m)} = \begin{cases} \mathbf{B}_1 \otimes \mathbf{I}_O \otimes \mathbf{I}_T & \text{for } m < T_p \\ \widetilde{\mathbf{B}}_1 \otimes \mathbf{I}_O \otimes \mathbf{I}_T & \text{for } m = T_p \\ \widehat{\mathbf{B}}_1 \otimes \mathbf{I}_O \otimes \mathbf{I}_T & \text{for } T_p < m < C_p - 1 \\ \mathbf{b}_1 \otimes \mathbf{I}_O \otimes \mathbf{I}_T & \text{for } m = C_p - 1 \end{cases}$$

$$\mathbf{\Lambda}_o^{(m)} = \begin{cases} \mathbf{I}_P \otimes \mathbf{B}_3 \otimes \mathbf{I}_T & \text{for } m \leq T_p \\ \mathbf{I}_{P+1} \otimes \mathbf{B}_3 \otimes \mathbf{I}_T & \text{for } T_p < m < C_p \\ \mathbf{B}_3 \otimes \mathbf{I}_T & \text{for } m = C_p \end{cases}$$

$$\mathbf{A}^{(m)} = \begin{cases} \mathbf{B}_0 \otimes \mathbf{I}_O \otimes \mathbf{I}_T & \\ \quad + \mathbf{I}_P \otimes \mathbf{B}_2 \otimes \mathbf{I}_T & \\ \quad + \mathbf{I}_P \otimes \mathbf{I}_O \otimes \mathbf{T} & \text{for } m \leq T_p \\ \widehat{\mathbf{B}}_0 \otimes \mathbf{I}_O \otimes \mathbf{I}_T & \\ \quad + \mathbf{I}_{P+1} \otimes \mathbf{B}_2 \otimes \mathbf{I}_T & \\ \quad + \mathbf{I}_{P+1} \otimes \mathbf{I}_O \otimes \mathbf{T} & \text{for } T_p < m < C_p \\ \mathbf{B}_2 \otimes \mathbf{I}_T + \mathbf{I}_O \otimes \mathbf{I}_T & \text{for } m = C_p \end{cases}$$

$$\hat{\mathbf{A}}^{(m)} = \begin{cases} \mathbf{B}_0 \otimes \mathbf{I}_O \otimes \mathbf{I}_T & \\ \quad + \mathbf{I}_P \otimes \mathbf{B}_2 \otimes \mathbf{I}_T & \text{for } m \leq T_p \\ \widehat{\mathbf{B}}_0 \otimes \mathbf{I}_O \otimes \mathbf{I}_T & \\ \quad + \mathbf{I}_{P+1} \otimes \mathbf{B}_2 \otimes \mathbf{I}_T & \text{for } T_p < m < C_p \\ \mathbf{B}_2 \otimes \mathbf{I}_T & \text{for } m = C_p \end{cases}$$

$$\mathbf{M}^{(m)} = \begin{cases} \mathbf{I}_P \otimes \mathbf{I}_O \otimes \mathbf{t}'\boldsymbol{\tau} & \text{for } 1 < m \leq T_p \\ \mathbf{C} \otimes \mathbf{I}_O \otimes \mathbf{t}'\boldsymbol{\tau} & \text{for } m = T_p + 1 \\ \mathbf{I}_{P+1} \otimes \mathbf{I}_O \otimes \mathbf{t}'\boldsymbol{\tau} & \text{for } T_p + 1 < m < C_p \\ \mathbf{f} \otimes \mathbf{I}_O \otimes \mathbf{t}'\boldsymbol{\tau} & \text{for } m = C_p \end{cases}$$

3.3.2 Methodology: the matrix-geometric technique

Consider the above defined Markov process on the three-dimensional state space $\{(n,m,i) \mid n \geq 0, 0 \leq m \leq C_p, i = 0,1,\ldots,K\}$ where i denotes the state of the modulating process, as the phase set i is defined in the finite state space $\mathcal{K}$ (see Section 3.2). A well-known method for finding the stationary distribution of QBD processes is the matrix-geometric method. With $\pi(n,m,i)$ the stationary probability of the process being in state (n,m,i), and using the vector notation $\boldsymbol{\pi}_k = (\pi(k,0,0), \pi(k,0,1), \ldots, \pi(k,C_p,K))$, the probability vectors can be expressed as,

$$\boldsymbol{\pi}_k = \boldsymbol{\pi}_0 \mathbf{R}^k. \tag{3.6}$$

where the so-called rate matrix $\mathbf{R}$ is the minimal non-negative solution of the non-linear matrix equation $\mathbf{R}^2\mathbf{W} + \mathbf{R}\mathbf{L}_p + \mathbf{L}_o = \mathbf{0}$. Here, we compute the rate matrix by implementing the iterative algorithm of [17, chapter 8].

3.3.3 Performance measures

Once the steady state probabilities have been determined numerically, we can calculate a number of interesting performance measures for the decoupling inventory system. For ease of notation, we introduce the marginal probability mass functions of the queue content of the product queue and the order queue: $\pi^{(p)}(m) = \sum_{i\in\mathcal{K}} \sum_{n=0}^{\infty} \pi(n,m,i)$ and $\pi^{(o)}(m) = \sum_{i\in\mathcal{K}} \sum_{m=0}^{C_p} \pi(n,m,i)$.

Note that as the queue of the backlogged orders is infinite, the throughput of the decoupling inventory system η equals the order arrival rate λ_o. In addition, we have the following performance measures.

- The mean semi-finished product queue and the order backlog content: $\mathrm{E}\,Q_p$ and $\mathrm{E}\,Q_o$ respectively,

$$\mathrm{E}\,Q_p = \sum_{m}^{C_p} \pi^{(p)}(m)m, \quad \mathrm{E}\,Q_o = \sum_{n}^{\infty} \pi^{(o)}(n)n.$$

- The variance of the semi-finished product queue and the order backlog content: $\mathrm{Var}\,Q_p$ and $\mathrm{Var}\,Q_o$ respectively,

$$\mathrm{Var}\,Q_p = \sum_{m}^{C_p} \pi^{(p)}(m)m^2 - (\mathrm{E}\,Q_p)^2,$$

$$\operatorname{Var} Q_o = \sum_{n}^{\infty} \pi^{(o)}(n) n^2 - (\mathrm{E}\, Q_o)^2 .$$

- The mean lead time LT (calculated based on Little's theorem) is the average amount of time between the placement of an order and the completion to a finished product:
$$\mathrm{LT} = \frac{\mathrm{E}\, Q_o}{\lambda_o}$$
- As the product queue has finite capacity, production prior to the decoupling stock may be blocked. This happens when there is a product arrival and the queue is full. Hence, blocking corresponds to the loss probability in the product buffer. The loss probability is most easily expressed in terms of the throughput η. We have,
$$b_p = \frac{\lambda_p - \eta}{\lambda_p} = \frac{\lambda_p - \lambda_o}{\lambda_p} .$$

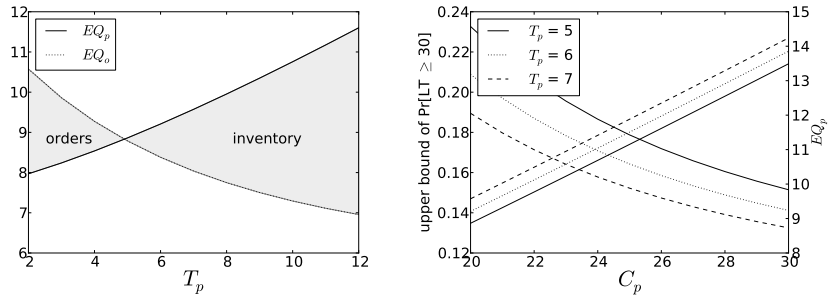
We now illustrate our approach by means of some numerical examples.

3.4 Numerical examples

3.4.1 Poisson arrivals and exponential order processing times

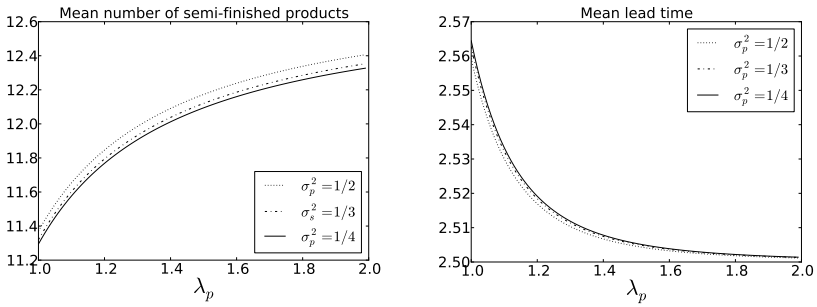
As a first example, the difference between the mean semi-finished product queue and the mean order backlog content versus the threshold value of the semi-finished product inventory is depicted in Figure 3.2(a). We assume that semi-finished products and orders arrive according to a Poisson process with parameter $\lambda_p = 1$ and $\lambda_o = 0.85$, respectively. The inventory capacity C_p equals 20 and order processing times are exponentially distributed with service rate μ equal to 1 for all curves. As the figure shows, the threshold value of 5 results on average in the same amount of backlogged orders and semi-finished products in stock. Under and above this level, products and orders are on average backlogged, respectively. Obviously, there is on average more stock of semi-finished products and less backlog of orders as the threshold value increases.

Figure 3.2(b) represents the trade-off between the upper bound of the probability to have the lead time higher or equal to 30 (left side) and the average stock of the semi-finished products (right side). Note that we calculated the lead time distribution by using the one-sided Chebyshev's inequality. Under the same parameter assumptions of Figure 3.2(a), the upper bound of the probability to have the lead time higher or equal to 30 decreases and the average stock increases as the inventory capacity increases for each threshold value. Indeed, if more buffer capacity is available, it will be used – the mean semi-finished product queue increases such that there is on average less time required to deliver an order. Finally,



(a) (b)

Figure 3.2: There is a trade-off between the average stock of the semi-finished products and the average number of backlogged orders and between the lead time.



(a) (b)

Figure 3.3: Given the mean setup time, the corresponding setup time distribution has a very limited impact on the mean number of semi-finished products and on the mean lead time in this case.

in this numerical example, we observe that the highest threshold value give the average best results: the intersection between the two performance measures and the necessary stock capacity have the lowest value.

3.4.2 Erlang distributed setup times

The second numerical example quantifies the impact of variability in the production process of semi-finished products on the decoupling inventory system. In particular, we here study Erlang-distributed setup times – the setup time starts when the semi-finished product inventory goes below a certain level and stops after some Erlang distributed time. Then, the semi-finished products arrive according to a Poisson process with arrival rate λ_p until the stock capacity is reached.

Figure 3.3(a) 3.3(b) show the mean number of semi-finished products in the buffer and the mean lead time of the system with a buffer capacity equal to 20 and a threshold value equal to 5. In both figures, the arrival rate is varied and different values of the variance of the setup time process are assumed as indicated. The order arrival rate λ_o equals 0.6, order processing times are assumed to be exponentially distributed with service rate μ equal to 1 and the mean setup time equals 1. As expected, the mean number of semi-finished products increases and the mean lead time decreases as the arrival rate of the semi-finished products λ_p increases. Furthermore, the shape of the setup time distribution has a very small effect on both performance measures. In particular, the mean number of semi-finished products and the mean lead time show respectively a slight decrease and increase as the variance of the setup time distribution σ_p^2 increases. This is due to the fact that the more regular the setup time, the less semi-finished products are on average in stock and the more orders are on average backlogged.

3.4.3 Markovian arrival process for orders

We also quantify the impact of irregular order arrivals. To this end, we compare both buffers with Poisson arrivals to corresponding decoupling inventory systems with interrupted Poisson arrivals for the orders and Poisson arrivals for the semi-finished products. The arrival interruptions account for inefficiency in the ordering process. Order processing times are assumed to be exponentially distributed with service rate μ equal to 1, this value being independent of the number of products and orders in the different buffers.

The interrupted Poisson process considered here is a two-state Markovian process. In the active state, new orders arrive in accordance with a Poisson process with rate λ_o whereas no new orders arrive in the inactive state. Let α and β denote the rate from the active to the inactive state and vice versa, respectively. We then use the following parameters to characterise the interrupted Poisson process (IPP),

$$\sigma = \frac{\beta}{\alpha+\beta}, \quad \kappa = \frac{1}{\alpha} + \frac{1}{\beta}, \quad \rho_o = \lambda_o \sigma.$$

Note that σ is the fraction of time that the interrupted Poisson process is active, the absolute time parameter κ is the average duration of an active and an inactive period, and ρ_o is the arrival load of the orders.

Figure 3.4 shows the mean number of backlogged orders versus the arrival rate of semi-finished products with buffer capacity C_p equal to 20 and the threshold value T_p equal to 5 and 7 for Poisson arrivals as well as for interrupted Poisson arrivals of orders. Order processing times are exponentially distributed with service rate μ equal to one for all curves. In addition, we set $\sigma = 0.8$ and $\kappa = 10$ for the interrupted Poisson processes (λ_o equals 0.6 for Poisson arrivals and 0.75 for

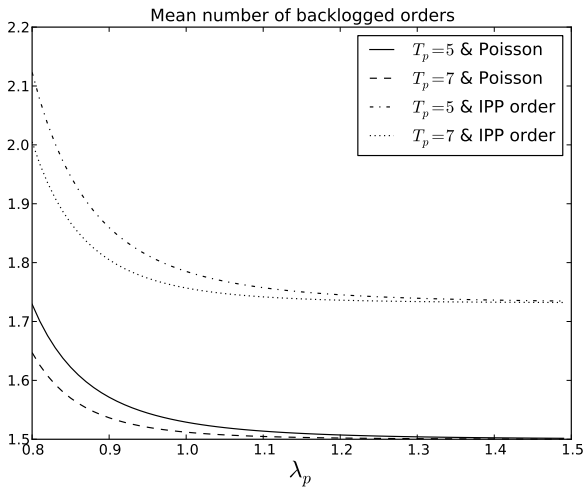


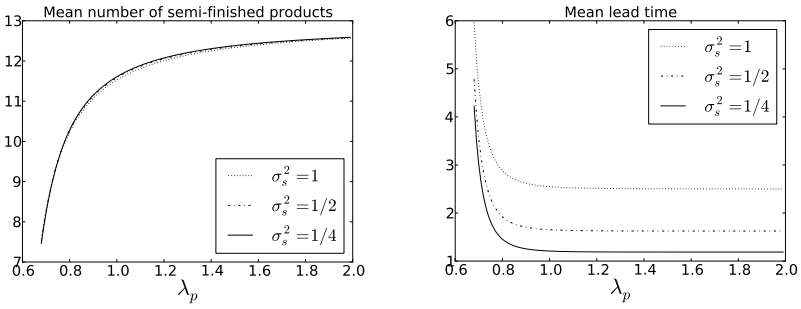
Figure 3.4: Irregular order arrivals result in a higher average number of backlogged orders.

interrupted Poisson arrivals). As expected, the mean number of backlogged orders decreases as the arrival rate of semi-finished products increases. Furthermore, the impact of the threshold value on the average number of backlogged orders decreases as the arrival rate of semi-finished products increases – both Markovian models converge to some value for T_p equal to 5 and 7. Finally, comparing interrupted Poisson and Poisson processes, burstiness in the ordering process has a negative impact on performance – there is on average more time required to deliver an order.

3.4.4 Phase-type distributed order processing times

The last numerical example quantifies the impact of the distribution of the order processing times on the decoupling inventory performance. In particular, we here study Erlang-distributed order processing times.

Figure 3.5(a) and 3.5(b) depict the mean number of semi-finished products in the buffer and the mean lead time of the decoupling inventory system. In both figures, the arrival rate of semi-finished products is varied and different values of the order processing time distribution are assumed as indicated. The service rate μ equals 1 for all curves, the order arrival rate λ_o equals 0.6, the inventory capacity C_p equals 20 and the threshold value T_p is equal to 5. Clearly, Figure 3.5(a) and 3.5(b) show respectively that the buffer content of semi-finished products increases and the lead time decreases until a certain value as the arrival rate of semi-finished



(a) (b)

Figure 3.5: Given the mean order processing time, the corresponding order processing time distribution has no significant impact on the mean number of semi-finished products and has a significant impact on the mean lead time in this case.

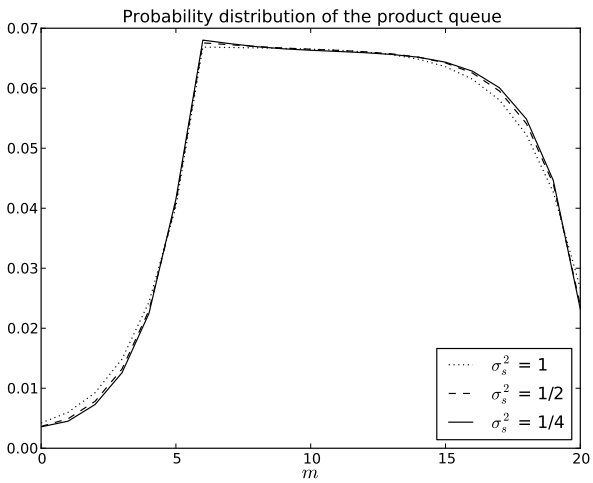


Figure 3.6: The zero probability increases when the variance of the order processing time distribution decreases.

products increases. Concerning the mean number of semi-finished products, we can conclude that the order processing time distribution has no significant effect on this performance measure. Indeed, we observe that the difference is very small and that it decreases as the arrival rate of semi-finished products increases. However, the difference between a variance σ_s^2 equal to 1, $1/2$ and $1/4$ for the mean lead time is significant, especially when the arrival rate λ_p is smaller than 0.7. Furthermore, in this numerical example, the mean number of semi-finished products decreases and the mean lead time increases as the variance increases. Figure 3.6 depicts the probability mass function of the product queue for a decoupling inventory system when λ_p equals one. As the Figure shows, the zero probability increases slightly as the variance of the order processing time distribution increases. As for Erlang distributed setup times in Section 3.4.2, we have a coupling effect between both performance measures – the mean number of semi-finished products increases such that the mean number of backlogged orders (and thus the mean lead time) decreases.

3.5 Conclusion

In this paper, we evaluate the performance of different hybrid push-pull systems with a decoupling inventory at the semi-finished products and reordering thresholds. In particular, we investigate the impact of different reordering policies, irregular order arrivals as well as the setup time distribution and the order processing time distribution on the performance of hybrid push-pull systems. In the studied hybrid push-pull systems, production of semi-finished products starts when the inventory goes below the so-called threshold value and stops when the inventory attains stock capacity. Decoupling means that the completion of a semi-finished product is only possible when there is an order. These orders are backlogged and can be satisfied only when the semi-finished products are available. Therefore, the studied push-pull system is modelled as a homogeneous quasi-birth-death process (QBD) and solved with matrix-analytic methods.

As our numerical examples show, there is trade-off to be made between the inventory cost and the service level, as expected – e.g. a higher threshold value causes on average a higher inventory cost and a smaller lead time. Furthermore, irregular order arrivals have a negative impact on the performance of the hybrid push-pull system. However, system performance is relatively insensitive to variation in the setup time distribution and partially insensitive to variation in the order processing time distribution. Future work will focus on determining the total cost of the studied hybrid push-pull systems.

References

[1] P. Bell. *A Decoupling Inventory Problem with Storage Capacity Constraints*, Operations Research, 28, p.476–488, 1980.

[2] T. Blecker and N. Abdelkafi. *Complexity and variety in mass customization systems: analysis and recommendations*, Management Decision, 44(7), p.908–929, 2006.

[3] F. Bonomi. *An approximate analysis for a class of assembly-like queues*, Queueing Systems Theory and Applications, p.289–309, 1987.

[4] P. Buchholz, P. Kemper and J. Kriege. *Multi-class Markovian arrival processes and their parameter fitting*, Performance Evaluation, 67(11), p.1092–1106, 2010.

[5] K. Chang and Y. Lu. *Queueing analysis on a single-station make-to-stock/make-to-order inventory-production system*, Applied Mathematical Modelling, 34, p.978–991, 2010.

[6] J.K. Cochran and S.S. Kim. *Optimizing a serially combined push and pull manufacturing system by simulated annealing*, Second International Conference on Engineering Design and Automation, 1998.

[7] E. De Cuypere and D. Fiems. *Performance evaluation of a kitting process*, Proceedings of the 17th International Conference on analytical and stochastic modelling techniques and applications, Lecture Notes in Computer Science, vol. 6751, Venice, Italy, 2011.

[8] D. Fiems, B. Steyaert and H. Bruneel. *A genetic approach to Markovian characterisation of H.264 scalable video*, Multimedia Tools and Applications, p.1–22, 2011.

[9] O. Ghrayeb, N. Phojanamongkolkij and B.A. Tan. *A hybrid push/pull system in assemble-to-order manufacturing environment*, Journal of Intelligent Manufacturing, 20, p.379–387, 2009.

[10] F.W. Harris. *How many parts to make at once*, Factory, the magazine of Management, 10(2), p.135–136, 1913.

[11] J.M. Harrison. *Assembly-Like Queues*, Journal of Applied Probability, 10(2), p.354–367, 1973.

[12] S. Hoekstra, J. Romme and S. Argelo. *Integral logistic structures: developing customer-oriented goods flow*, McGraw-Hill, 1992.

[13] W. J. Hopp and J.T. Simon. *Bounds and Heuristics for assembly-like queues*, Queueing Systems, 4, p.137–156, 1989.

[14] W.J. Hopp and M.L. Spearman. *Factory physics: Foundations of manufacturing management*, The McGraw-Hill Companies, Inc., 2000.

[15] P. Kaminsky and O. Kaya. *Combined make-to-order/make-to-stock supply chains*, IIE Transactions, 41, p.103–119, 2009.

[16] G. Latouche. *Queues with paired customers*, Journal of Applied Probability, 18, p.684–696, 1981.

[17] G. Latouche and V. Ramaswami. *Introduction to Matrix Analytic Methods in Stochastic Modeling*, 1999.

[18] C.Y. Lee. *A recent development of the integrated manufacturing system: A hybrid of MRP and JIT*, International Journal of Operations and Production Management, 13(4), p.3–17, 1993.

[19] E.H. Lipper and B. Sengupta. *Assembly-like queues with finite capacity: bounds, asymptotics and approximations*, Queueing Systems: Theory and Applications, 18, p.684, 1986.

[20] H. Ohta, T. Hirota and A. Rahim. *Optimal production-inventory policy for make-to-order versus make-to-stock based on the M/Er/1 queuing model*, International Journal of Advanced Manufacturing Technologies, 33, p.36–41, 2007.

[21] P.C. Pandey and P. Khokhajaikiat. *Performance modeling of multistage production systems operating under hybrid push-pull control*, International Journal Production Economics, 43, p.115–126, 1995.

[22] K. Ramachandran, L. Whitman and A.B. Ramachandran. *Criteria for determining the push-pull boundary*, Industrial Engineering Research Conference, Orlando, FL, USA, 2002.

[23] P. Som, W. Wilhelm and R. Disney. *Kitting process in a stochastic assembly system*, Queueing Systems, 17, p.471–490, 1994.

[24] P. Som and W. Wilhelm. *Analysis of stochastic assembly with GI-distributed assembly time*, INFORMS Journal on Computing, 11, p.104–116, 1999.

[25] CA. Soman, D.P. van Donk and G. Gaalman. *Combined make-to-order and make-to-stock in a food production system*, International Journal of Production Economics, 90, p.223–235, 2004.

[26] M.L. Spearman and M.A. Zazamis. *Push and pull production systems: Issues and comparisons*, Operations Research, 3, p.521–532, 1992.

[27] K. Takahashi and N. Nakamura. *Push pull, or hybrid control in supply chain management*, International Journal of Computer Integrated Manufacturing, 17(2), p.126–140, 2004.

4

Performance analysis of hybrid MTS/MTO systems with stochastic demand, production and service times

Eline De Cuypere, Koen De Turck and Dieter Fiems

submitted to the International Journal of Production Economics.

Abstract. This paper proposes a comprehensive methodology with reasonably light complexity in terms of implementation and computation for evaluating the performance of hybrid make-to-stock (MTS)/make-to-order (MTO) systems. In such systems, semi-finished products are manufactured in advance and processed into finished ones when a customer order is placed. To account for uncertainty in demand, inventory replenishment and order processing, these systems are studied as stochastic inventory models with two "queues": the decoupling inventory and the order backlog. Hence, order processing can only occur when both queues are nonempty. In this work, we exploit structural properties of the stochastic processes at hand by using matrix-analytic methods. This method allows the study of hybrid MTS/MTO systems under non-restrictive stochastic assumptions like arrival correlation and non-exponential order processing times. Furthermore, the hybrid MTS/MTO system can be managed by a continuous review (s,S)-policy in which the replenishment of semi-finished products halts when the inventory is full and restarts when the inventory level drops to a certain threshold value. Furnished

with many numerical examples, we analyse the performance of hybrid MTS/MTO systems by assessing the impact of inventory control, irregular order arrivals, setup and order processing times. Finally, we study the central trade-off that exists in these models, i.e. between inventory and service levels. To this end, a suitable cost structure is introduced, and the optimal inventory capacities and thresholds with respect to this cost structure are determined.

4.1 Introduction

In supply chain management, the position of the customer order decoupling point (CODP) — the boundary between forecast-driven and demand-driven activities — is a key strategic decision. Indeed, the position of the decoupling point has a great impact on market and operational performance [13, 23]. The two most known production strategies are make-to-stock (MTS) and make-to-order (MTO). Under pure make-to stock management, the activities are only forecast driven. Indeed, the end products are manufactured independently of any customer requirements and are stocked in advance. Hence, high holding costs or stock-out costs are unavoidable in highly-fluctuating demand environments [23]. The main performance criteria of such systems are fill rate, demand forecast, average work-in-process etc. [25]. In contrast, the make-to-order system has only order-driven activities. Indeed, the manufacturing of a product is triggered only when a customer order is placed. Hence, response times may become quite long for high demand products. The main criteria of such systems are the average response time, average order delay, manufacturing lead time, due dates etc. [25]. To benefit from both systems, production management is gradually moving toward a hybrid MTS/MTS strategy. In such strategies, products are manufactured and stocked in the first stage of the production (MTS) and are manufactured into end products only after a customer order is received in the second stage (MTO). Ghrayeb et al. [10] investigated a hybrid MTS/MTO system in an assemble-to-order manufacturing environment. In most cases, the hybrid case is shown to inherit the strengths and to conceal the weaknesses of both pure systems. Köber and Heinecke [15] investigated different hybrid production strategies for supply chains with volatile and seasonal demand. Gupta and Benjaafar [9] developed models to compute the costs and the benefits of delaying differentiation in series production systems where the order lead times are load dependent. As expected, a higher load favours later differentiation. Soman et al. [25] proposed Hierarchical Production Planning (HPP) to deal with production management decisions for MTS/MTO situations in food processing. The proposed framework has three levels (strategic, tactical, operational): MTS/MTO decision, capacity coordination and scheduling. These authors [26] also worked on an economic lot scheduling problem of a food production system with different MTS and MTO products. Rafiei and Rabbani [24] focused on the tactical level of the HPP

for hybrid MTS/MTO production systems with pure MTS, pure MTO and hybrid MTS/MTO products. The developed capacity coordination model is shown to be applicable to a real industrial case.

In this work, we describe hybrid MTS/MTO systems as stochastic inventory models to account for uncertainty in demand, inventory replenishment and order processing. Kaminsky and Kaya [14] considered a variety of combined make-to-order (MTO) and make-to-stock (MTS) supply chains in a stochastic environment. As shown by the authors, costs can be cut dramatically by using a combined system instead of pure MTO or MTS systems and that information exchange between the supplier and the manufacturer is critical for effective lead time quotation. Adan and Van der Wal [1] studied a production system with standard and non-standard products as a stochastic process with unit order sizes and exponential production times. They showed that the combination of pure MTS and pure MTO strategies in a production system can significantly reduce lead times. Ohta et al. [21] and Arreola and Decroix [2] proposed optimality conditions for MTO and MTS policies using a base-stock inventory policy. In both works, the production facility is represented as a single server queueing model. In [21], a multi-product inventory system is analysed where demand for each item arrives according to a Poisson process and the production times follow an Erlang distribution. An optimality condition that specifies whether each product should be produced according to a MTS or a MTO strategy is proposed. In continuous make-to-stock operations, the production of semi-finished products is halted when the inventory is full. Assuming that there is a non-negligible setup time when the production of semi-finished products is halted, a restart of the production whenever there is space in the product inventory is inefficient. Hence, a continuous review (s,S)-policy is considered in which replenishment starts when the inventory level drops to the threshold value s and stops when level S is attained. The term 'continuous review' refers to the uninterrupted monitoring of the inventory level to know whether or not level s is reached. Most work conducted on (s,S)-policies assumes that any amount of inventory can be replenished all at once [22, 20, 3, 11, 7]. However, in many real manufacturing systems, the production facility can only produce one item at a time and therefore inventory can only be replenished on an item-by-item basis. Motivated by this fact, various authors have analysed hybrid MTS/MTO systems where inventory is replenished on an item-by-item basis [12, 19, 17, 18].

In contrast to previous work on stochastic hybrid MTS/MTO systems, the present study explicitly accounts for both inventory level and order backlog. This considerably complicates the analysis as the stochastic model now consists of two "queues": a decoupling inventory and an order backlog. The performance of the hybrid MTS/MTO system is assessed when (i) the replenishment of semi-finished products is continuous and (ii) when the inventory is replenished in accordance with a continuous review (s,S)-policy. In the latter case, we also introduce a cost

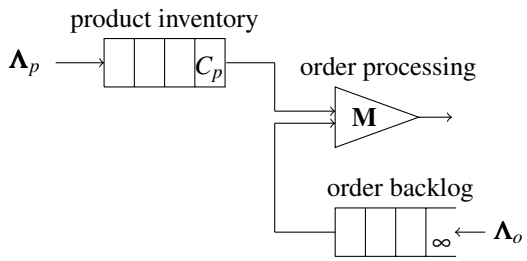


Figure 4.1: Generic inventory model.

function to investigate the known trade-off between inventory and service level and illustrate this trade-off by some numerical experiments. We show that by exploiting structural properties of the stochastic processes involved, matrix-analytic solution techniques can be devised to accurately evaluate system performance with reasonable computational complexity. Furthermore, the approach allows for studying the MTS/MTO systems under non-restrictive stochastic assumptions by introducing a Markovian environment variable [6]: a Markovian arrival process to describe the production of semi-finished products, a Markovian arrival process for demand, phase-type distributed production times to complete semi-finished products etc.

The remainder of this paper is organised as follows. Section 4.2 describes the hybrid MTS/MTO model at hand. In Section 4.3, the generic decoupling inventory system is analysed as a quasi-birth-death-process (QBD), the numerical solution methodology is discussed and relevant performance measures are determined. To illustrate our approach, Section 4.4 considers some numerical examples. In particular, we analyse the cost and performance of hybrid MTS/MTO systems. Finally, conclusions are drawn in Section 4.5.

4.2 Generic inventory model

We consider a generic inventory model, supporting hybrid MTS/MTO operations. The system is depicted in Figure 4.1 and consists of a product inventory, an order backlog and an order processing unit. The product inventory can store up to C_p semi-finished products whereas the order backlog keeps track of the orders that have not yet been delivered and has infinite capacity. Arriving orders are processed in accordance with a first-come-first-served discipline. Each order takes a semi-finished product from the decoupling inventory and sends it to the order processing unit to complete the product in accordance with order specifications. Note that the two "queues" — the product inventory and the order backlog — in the model at hand are tightly coupled. Departures from the inventory are only possible when there are backlogged orders. Similarly, departures from the order backlog are only

possible if there are semi-finished products in the product inventory.

We study the inventory model at hand in a Markovian framework, which combines modelling versatility with computationally efficient analysis techniques. To be more precise, we propose a three-dimensional continuous-time Markov process with infinite state space $\mathbb{N} \times \mathcal{C} \times \mathcal{K}$ with $\mathbb{N} = \{0,1,2,\cdots\}$, $\mathcal{C} = \{0,1,2,\ldots,C_p\}$ and $\mathcal{K} = \{1,\ldots,K\}$. At each point in time, the state of the inventory system is described by the triplet (n,m,i) where $n \in \mathbb{N}$ denotes the number of backlogged orders, $m \in \mathcal{C}$ denotes the number of semi-finished products in the inventory and $i \in \mathcal{K}$ denotes the state of an auxiliary Markovian environment variable. The introduction of such an environment variable provides the necessary versatility in modelling. It allows for introducing order arrival correlation, non-exponential order processing times, threshold-based inventory management, etc (cfr infra). We now describe the (marked) transition rates of the environment variable.

- The environment variable can change when there are neither arrivals nor departures. Let α_{ij} denote the transition rate from state i to state j ($i,j \in \mathcal{K}$, $i \neq j$). Further, let $\mathbf{A} = [\alpha_{ij}]_{i,j\in\mathcal{K}}$ denote the corresponding generator matrix. Note that the diagonal elements of $\mathbf{A}$ equal 0. We allow for different transition rates when either the inventory or the backlog is empty: let $\hat{\alpha}_{ij}$ and $\hat{\mathbf{A}}$ denote the transition rate from state i to state j and the corresponding generator matrix, respectively.

- The environment variable may remain the same or may change when there is an arrival. Let $\lambda_{ij}^{(p)}$ and $\lambda_{ij}^{(o)}$ denote the (marked) transition rate from state i to state j when there is an arrival at the product inventory and the order backlog, respectively ($i,j \in \mathcal{K}$). Moreover, let $\mathbf{\Lambda}_p = [\lambda_{ij}^{(p)}]_{i,j\in\mathcal{K}}$ and $\mathbf{\Lambda}_o = [\lambda_{ij}^{(o)}]_{i,j\in\mathcal{K}}$ denote the corresponding generator matrices. Note that marked self-transitions from state i to state i are allowed.

- Analogously, the environment variable may either remain the same or may change when there is a departure (in inventory and order backlog simultaneously). Let μ_{ij} ($i,j \in \mathcal{K}$) and $\mathbf{M}$ denote the transition rates and the corresponding generator matrix respectively.

When required by the application at hand, the generator matrices, $\mathbf{\Lambda}_o$, $\mathbf{\Lambda}_p$, $\mathbf{A}$, $\hat{\mathbf{A}}$ and $\mathbf{M}$ may depend on the inventory level m. In this case, we use superscripts to make this dependence explicit.

Having introduced the generic inventory model, we now focus on two particular instances. The first instance models a pure decoupling inventory of a hybrid MTS/MTO system. Arrivals in the product inventory correspond to a make-to-stock operation while the final production step is demand-driven or make-to-order. Note that the MTS operation is blocked when the decoupling inventory is full.

Dimensioning of the decoupling inventory aims at avoiding blocking while ensuring product availability. However this is not always possible without further control. Assuming that there is a non-negligible setup-time in the MTS stage, it does not make sense to restart production whenever there is space in the inventory. Therefore, the second instance assumes that the MTS operation is managed by an (S,s)-policy: the MTS operation halts when the inventory is at level S and restarts when the inventory level drops to level s.

4.2.1 Uncontrolled decoupling inventory

In this system, arrivals of products and orders are modelled as independent Markovian arrival processes (MAP) and processing times constitute a sequence of independent phase-type distributed random variables. The MAP can accurately capture arrival correlation. Moreover, tools have been developed that characterise MAPs from time-series [4, 5]. Particular subclasses of MAPs include the Poison process, the interrupted Poison process and the renewal arrival process with phase-type distributed renewal times.

The MAP of the order arrival process is described by the generator matrix $\mathbf{Q}_1$ of transitions with arrivals and the generator matrix $\mathbf{Q}_0$ without arrivals. Let Q denote the size of the state space $\mathcal{Q}$ of the order arrival process. The MAP of the product arrival process is described by the generator matrices $\mathbf{P}_0$ and $\mathbf{P}_1$, governing state transitions without and with product arrivals, respectively. Let P denote the size of the state space $\mathcal{P}$ of the product arrival process. Finally, the phase-type distribution — the distribution of the time till absorption of a Markov process with an absorbing state — of the order processing times is characterised by an initial probability vector $\boldsymbol{\tau}$, by the matrix $\mathbf{T}$ which corresponds to non-absorbing transitions and by the column vector $\mathbf{t}'$ of rates to the absorbing state. Let T denote the size of the state space $\mathcal{T}$ of the production process.

We use the environment variable to track the states of both arrival processes as well as of the state of the order processing time. It is notationally convenient to define the environment variable as the triplet (x_o, x_s, x_a) with x_o the state of the order arrival process, x_s the state of the order processing time and x_a the state of the product arrival process. As all state variables are finite, we can easily map the triplets on $\mathcal{K}$. Assuming lexicographical order of the triplets, we get the following matrices.

- Product and order arrivals correspond to marked transitions of the product and order MAPs. Hence we have,

$$\mathbf{\Lambda}_p = \mathbf{I}_Q \otimes \mathbf{I}_T \otimes \mathbf{P}_1,$$
$$\mathbf{\Lambda}_o = \mathbf{Q}_1 \otimes \mathbf{I}_T \otimes \mathbf{I}_P.$$

Here $\mathbf{I}_n$ is the $n \times n$ identity matrix.

- State transitions without arrivals or departures correspond to state transitions of arrival, order and production process. However, there are no transitions of the production process when either order backlog or product inventory are empty. Hence, we get,

$$\begin{aligned}\mathbf{A} &= \mathbf{Q}_0 \otimes \mathbf{I}_T \otimes \mathbf{I}_P + \mathbf{I}_Q \otimes \mathbf{T} \otimes \mathbf{I}_P \\ &\quad + \mathbf{I}_Q \otimes \mathbf{I}_T \otimes \mathbf{P}_0, \\ \hat{\mathbf{A}} &= \mathbf{Q}_0 \otimes \mathbf{I}_T \otimes \mathbf{I}_P + \mathbf{I}_Q \otimes \mathbf{I}_T \otimes \mathbf{P}_0.\end{aligned}$$

- Finally, there is a departure when the phase type reaches its absorbing state. We get,

$$\mathbf{M} = \mathbf{I}_Q \otimes \mathbf{t}'\boldsymbol{\tau} \otimes \mathbf{I}_P.$$

4.2.2 Controlled decoupling inventory

As previously mentioned, the generic inventory model is sufficiently versatile to study decoupling inventory systems following a threshold-based replenishment policy. In such systems, replenishment of semi-finished products starts when the inventory equals the threshold value T_p and stops when the inventory level reaches capacity C_p. In literature, this is referred to as a (s,S)-policy, S being the capacity and s the threshold [20, 3, 11, 7, 17, 18]. We do not use S and s so as to be consistent with the uncontrolled inventory management.

We retain the assumptions on the order arrival process and on the order processing times of the decoupling inventory above. We also retain the notation on the production process. However, now we additionally have to specify the state of the product arrival process when it restarts. Let b_i be the probability that the production process restarts in state $i \in \mathcal{P}$ and let $\mathbf{b} = [b_i]_{i \in \mathcal{P}}$. Remark that the inclusion of a start state allows for introducing setup times. For example, a phase-type distributed setup time prior to production is easily introduced as shown in the numerical examples.

When the inventory level m is equal or below T_p, production is always ongoing. Hence for $m \leq T_p$ the environment variable takes values in,

$$\{(x_o, x_s, x_a) : x_o \in Q, x_s \in \mathcal{T}, x_a \in \mathcal{P}\}.$$

When the inventory level is at least T_p but not full, the environment variable needs to track whether production is on-going or not. For ease of notation, we assume that the state of the production process (as in the preceding section) is augmented with an additional state θ. When the production process is in this state, there is

no production. Hence, for inventory level m and $T_p < m < C_p$, the environment variable takes values in,

$$\{(x_o, x_s, x_a) : x_o \in \mathcal{Q}, x_s \in \mathcal{T}, x_a \in \mathcal{P} \cup \{\theta\}\}.$$

In contrast to the preceding extension of the state description, we are sure that there is no production when the inventory is full. Hence in this case there is no need to track the state of the production process. Hence for inventory level $m = C_p$, the environment variable takes values in,

$$\{(x_o, x_s) : x_o \in \mathcal{Q}, x_s \in \mathcal{T}\}.$$

With the state space defined as above, we now repeat the construction of the matrices $\mathbf{\Lambda}_o$, $\mathbf{\Lambda}_p$, $\mathbf{A}$, $\hat{\mathbf{A}}$ and $\mathbf{M}$. Due to the control of the production process, the generator matrices as well as their sizes are now level-dependent. Introducing the auxiliary matrices and vectors,

$$\widehat{\mathbf{P}}_i = \begin{bmatrix} \mathbf{P}_i & \mathbf{0}_P \\ \mathbf{0}'_P & 0 \end{bmatrix} \quad \widetilde{\mathbf{P}}_i = \begin{bmatrix} \mathbf{P}_i & \mathbf{0}_P \end{bmatrix}$$

$$\mathbf{p}_1 = \begin{bmatrix} \mathbf{P}_1 \mathbf{1}_P \\ 0 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} \mathbf{I}_P \\ \mathbf{b} \end{bmatrix} \quad \mathbf{f} = \begin{bmatrix} \mathbf{0}'_P & 1 \end{bmatrix}$$

The symbols $\mathbf{0}_X$ and $\mathbf{1}_X$ represent column vectors of size X with zeroes and ones, respectively. We get the following generator matrices,

$$\mathbf{\Lambda}_p^{(m)} = \begin{cases} \mathbf{I}_Q \otimes \mathbf{I}_T \otimes \mathbf{P}_1 & \text{for } m < T_p \\ \mathbf{I}_Q \otimes \mathbf{I}_T \otimes \widetilde{\mathbf{P}}_1 & \text{for } m = T_p \\ \mathbf{I}_Q \otimes \mathbf{I}_T \otimes \widehat{\mathbf{P}}_1 & \text{for } T_p < m < C_p - 1 \\ \mathbf{I}_Q \otimes \mathbf{I}_T \otimes \mathbf{p}_1 & \text{for } m = C_p - 1 \end{cases}$$

$$\mathbf{\Lambda}_o^{(m)} = \begin{cases} \mathbf{Q}_1 \otimes \mathbf{I}_T \otimes \mathbf{I}_P & \text{for } m \leq T_p \\ \mathbf{Q}_1 \otimes \mathbf{I}_T \otimes \mathbf{I}_{P+1} & \text{for } T_p < m < C_p \\ \mathbf{Q}_1 \otimes \mathbf{I}_T & \text{for } m = C_p \end{cases}$$

$$\mathbf{A}^{(m)} = \begin{cases} \mathbf{Q}_0 \otimes \mathbf{I}_T \otimes \mathbf{I}_P & \\ \quad + \mathbf{I}_Q \otimes \mathbf{T} \otimes \mathbf{I}_P & \\ \quad + \mathbf{I}_Q \otimes \mathbf{I}_T \otimes \mathbf{P}_0 & \text{for } m \leq T_p \\ \mathbf{Q}_0 \otimes \mathbf{I}_T \otimes \mathbf{I}_{P+1} & \\ \quad + \mathbf{I}_Q \otimes \mathbf{T} \otimes \mathbf{I}_{P+1} & \\ \quad + \mathbf{I}_Q \otimes \mathbf{I}_T \otimes \widehat{\mathbf{P}}_0 & \text{for } T_p < m < C_p \\ \mathbf{Q}_0 \otimes \mathbf{I}_T + \mathbf{I}_Q \otimes \mathbf{T} & \text{for } m = C_p \end{cases}$$

$$\hat{\mathbf{A}}^{(m)} = \begin{cases} \mathbf{Q}_0 \otimes \mathbf{I}_T \otimes \mathbf{I}_P & \\ +\mathbf{I}_Q \otimes \mathbf{I}_T \otimes \mathbf{P}_0 & \text{for } m \leq T_p \\ \mathbf{Q}_0 \otimes \mathbf{I}_T \otimes \mathbf{I}_{P+1} & \\ +\mathbf{I}_Q \otimes \mathbf{I}_T \otimes \widehat{\mathbf{P}}_0 & \text{for } T_p < m < C_p \\ \mathbf{Q}_0 \otimes \mathbf{I}_T & \text{for } m = C_p \end{cases}$$

$$\mathbf{M}^{(m)} = \begin{cases} \mathbf{I}_Q \otimes \mathbf{t}'\boldsymbol{\tau} \otimes \mathbf{I}_P & \text{for } 1 < m \leq T_p \\ \mathbf{I}_Q \otimes \mathbf{t}'\boldsymbol{\tau} \otimes \mathbf{B} & \text{for } m = T_p + 1 \\ \mathbf{I}_Q \otimes \mathbf{t}'\boldsymbol{\tau} \otimes \mathbf{I}_{P+1} & \text{for } T_p + 1 < m < C_p \\ \mathbf{I}_Q \otimes \mathbf{t}'\boldsymbol{\tau} \otimes \mathbf{f} & \text{for } m = C_p \end{cases}$$

4.3 Analysis

4.3.1 Quasi-birth-death process

The studied Markov process is a homogeneous quasi-birth-death process (QBD), see [16]. In the present setting, the so-called level or block-row number, indicates the number of backlogged orders while the phase, i.e. the index within a block element, indicates both the content of the decoupling inventory and the state of the Markovian environment variable. The one-step transitions are restricted to states in the same level (from state $(n,*,*)$ to state $(n,*,*)$) or in two adjacent levels (from state $(n,*,*)$ to state $(n+1,*,*)$ or state $(n-1,*,*)$). Indeed, orders arrive and are processed one by one such that the order backlog increases and decreases in unit steps. We then find that the generator matrix of the Markov process has the following block matrix representation,

$$\mathbf{Q} = \begin{bmatrix} \mathbf{L}'_p & \mathbf{L}_o & \mathbf{0} & 0 & \cdots \\ \mathbf{W} & \mathbf{L}_p & \mathbf{L}_o & 0 & \cdots \\ 0 & \mathbf{W} & \mathbf{L}_p & \mathbf{L}_o & \cdots \\ 0 & 0 & \mathbf{W} & \mathbf{L}_p & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

The blocks are given by,

$$\mathbf{L}_o = \begin{bmatrix} \boldsymbol{\Lambda}_o^{(0)} & 0 & 0 & \cdots & 0 \\ 0 & \boldsymbol{\Lambda}_o^{(1)} & 0 & \cdots & 0 \\ 0 & 0 & \boldsymbol{\Lambda}_o^{(2)} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \boldsymbol{\Lambda}_o^{(C_p)} \end{bmatrix},$$

$$\mathbf{L}_p = \begin{bmatrix} \underline{\mathbf{D}}^{(0)} & \mathbf{\Lambda}_p^{(0)} & 0 & \cdots & 0 \\ 0 & \mathbf{D}^{(1)} & \mathbf{\Lambda}_p^{(1)} & \cdots & 0 \\ 0 & 0 & \mathbf{D}^{(2)} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \mathbf{D}^{(C_p)} \end{bmatrix},$$

$$\mathbf{L}'_p = \begin{bmatrix} \underline{\mathbf{D}}^{(0)} & \mathbf{\Lambda}_p^{(0)} & 0 & \cdots & 0 \\ 0 & \underline{\mathbf{D}}^{(1)} & \mathbf{\Lambda}_p^{(1)} & \cdots & 0 \\ 0 & 0 & \underline{\mathbf{D}}^{(2)} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \underline{\mathbf{D}}^{(C_p)} \end{bmatrix},$$

$$\mathbf{W} = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ \mathbf{M}^{(1)} & 0 & \cdots & 0 & 0 \\ 0 & \mathbf{M}^{(2)} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \mathbf{M}^{(C_p)} & 0 \end{bmatrix}.$$

with $\mathbf{D}^{(m)} = \mathbf{A}^{(m)} - \partial\mathbf{A}^{(m)} - \partial\mathbf{\Lambda}_o^{(m)} - \partial\mathbf{\Lambda}_p^{(m)} - \partial\mathbf{M}^{(m)}$ and $\underline{\mathbf{D}}^{(m)} = \hat{\mathbf{A}}^{(m)} - \partial\hat{\mathbf{A}}^{(m)} - \partial\mathbf{\Lambda}_o^{(m)} - \partial\mathbf{\Lambda}_p^{(m)}$ with $m = (0,1,2,\ldots,C_p)$ being the number of semi-finished products in the inventory. Note that $\partial\mathbf{X}$ represents a diagonal matrix with diagonal elements equal to the row sums of $\mathbf{X}$.

Having defined the different blocks of the QBD process, we now focus on the solution method. Recall that the state of the Markov process is described by the triplet (n,m,i), n is the size of the order backlog, m is the size of the product inventory and i is the state of the environment variable. Let $\pi(n,m,i)$ be the stationary probability of the process to be in state (n,m,i). A well-known method for finding the stationary distribution of QBD processes is the matrix-geometric method. Using the vector notation $\boldsymbol{\pi}_k = (\pi(k,0,0),\pi(k,0,1),\ldots,\pi(k,C_p,K))$, the probability vectors can be expressed as,

$$\boldsymbol{\pi}_k = \boldsymbol{\pi}_0\mathbf{R}^k. \tag{4.1}$$

where the so-called rate matrix $\mathbf{R}$ is the minimal non-negative solution of the non-linear matrix equation

$$\mathbf{R}^2\mathbf{W} + \mathbf{R}\mathbf{L}_p + \mathbf{L}_o = \mathbf{0}. \tag{4.2}$$

Several iterative procedures exist for solving equation (4.2). For example, Gun [8] uses the following simple recursion

$$\mathbf{R} \leftarrow -(\mathbf{L}_o + \mathbf{R}^2\mathbf{W})\mathbf{L}_p^{-1}. \tag{4.3}$$

We compute the rate matrix by implementing the improved iterative algorithm of [16, chapter 8, p.179-187].

4.3.2 Performance measures

Once the steady state probabilities have been determined numerically, we can calculate a number of interesting performance measures for the decoupling inventory system. For ease of notation, we introduce the marginal probability mass functions of the content of the product inventory and the order backlog: $\pi^{(p)}(m) = \sum_{i\in\mathcal{K}}\sum_{n=0}^{\infty}\pi(n,m,i)$ and $\pi^{(o)}(n) = \sum_{i\in\mathcal{K}}\sum_{m=0}^{C_p}\pi(n,m,i)$. We have the following performance measures.

- The mean semi-finished product inventory and the order backlog content: $\mathrm{E}\,Q_p$ and $\mathrm{E}\,Q_o$ respectively,

$$\mathrm{E}\,Q_p = \sum_{m=1}^{C_p}\pi^{(p)}(m)m\,,$$

$$\mathrm{E}\,Q_o = \sum_{n=1}^{\infty}\pi^{(o)}(n)n\,.$$

- The variance of the semi-finished product inventory and the order backlog content: $\mathrm{Var}\,Q_p$ and $\mathrm{Var}\,Q_o$ respectively,

$$\mathrm{Var}\,Q_p = \sum_{m=1}^{C_p}\pi^{(p)}(m)m^2 - (\mathrm{E}\,Q_p)^2\,,$$

$$\mathrm{Var}\,Q_o = \sum_{n=1}^{\infty}\pi^{(o)}(n)n^2 - (\mathrm{E}\,Q_o)^2\,.$$

- As the order backlog queue is infinite, the throughput of the hybrid MT-S/MTO system η equals the order arrival rate:

$$\eta = \boldsymbol{\pi}^{(q)}\mathbf{Q}_1\mathbf{1}_q$$

  The vector $\boldsymbol{\pi}^{(q)}$ is the solution of $\boldsymbol{\pi}^{(q)}(\mathbf{Q}_1+\mathbf{Q}_0) = \mathbf{0}$ and $\boldsymbol{\pi}^{(q)}\mathbf{1}_q = 1$.

- The mean lead time LT (calculated based on Little's theorem) is the average amount of time between the placement of an order and the completion of a finished product:

$$\mathrm{LT} = \frac{\mathrm{E}\,Q_o}{\eta}$$

We now illustrate our approach by means of some numerical examples.

Type of decoupling inventory		
uncontrolled (unctl)	C_p	20
controlled (ctl)	C_p T_p	20 5
Semi-finished product arrivals		
Poisson	λ_p	1
setup times		
constant (= Erlang-10)	μ_p α_p	2 5
exponential (exp)	μ_p	2
Order arrivals		
Poisson	ρ_o	0.6
IPP	σ κ ρ_o	0.8 10 0.6
Order processing		
constant (= Erlang-10)	μ_s α_s	1 10
exponential (exp)	μ_s	1
Cost and profit function		
holding & setup	c_1 c_2	1 5
linear lead time cost	c_3	80
non-linear lead time cost	c_4 ω	500 0.3

Table 4.1: Parameter values of the different studied systems.

4.4 Numerical results

In this section, we analyse the performance of the above defined hybrid MTS/MTO systems. In particular, we study the impact of inventory control, order correlation and the distribution of the setup and order processing time on the mean lead time and inventory level. Moreover, we define a cost function for hybrid MTS/MTO systems with a controlled decoupling inventory and find the optimal inventory capacity and threshold value for specific sets of parameters and cost values.

For further reference, the considered parameter values of all numerical results are listed in Table 4.1. These parameter values are chosen only by way of illustration and are not to be considered as limiting.

4.4.1 Mean inventory level and mean lead time

We here focus on mean inventory level and mean lead time for controlled and uncontrolled inventory management.

First, consider a decoupling inventory with Poisson arrivals of semi-finished products with rate λ_p, no setup times and exponentially distributed order processing times with rate $\mu = 1$. The inventory capacity is $C_p = 20$ and in the case of the

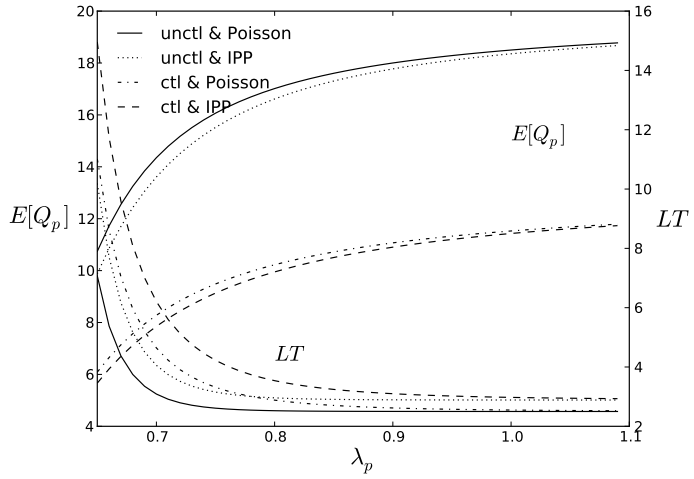


Figure 4.2: Impact of irregular order arrivals on the performance of hybrid MTS/MTO systems with a controlled and uncontrolled decoupling inventory.

controlled inventory the threshold is $T_p = 5$ for all figures below.

To quantify the impact of irregularity in the order process, we further characterise order arrivals by an interrupted Poisson process and benchmark the performance measures against Poisson order arrivals. The interrupted Poisson process is a two-state Markovian process such that new orders arrive in accordance with a Poisson process in its active state while there are no arrivals when it is inactive. This process is fully characterised by the arrival rate in the active state λ_o, by the transition rate α from the active to the inactive state and by the transition rate β from the inactive to the active state. For convenience, we use the more intuitive parametrisation (σ, κ, ρ_o), with

$$\sigma = \frac{\beta}{\alpha+\beta}, \quad \kappa = \frac{1}{\alpha} + \frac{1}{\beta}, \quad \rho_o = \lambda_o \sigma.$$

Here σ is the fraction of time in which the interrupted Poisson process is active, the absolute time parameter κ is the average duration of an active and an inactive period, and ρ_o is the arrival load of the orders.

Figure 4.2 depicts the mean semi-finished product inventory and the mean lead time for both Poisson and interrupted Poisson arrivals of orders versus the product arrival rate λ_p. We set $\sigma = 0.8$ and $\kappa = 10$ which corresponds to moderate correlation. The load of the IPP and the arrival rate of the Poisson process equal $\rho_o = 0.6$. To simplify comparison between controlled and uncontrolled inventory

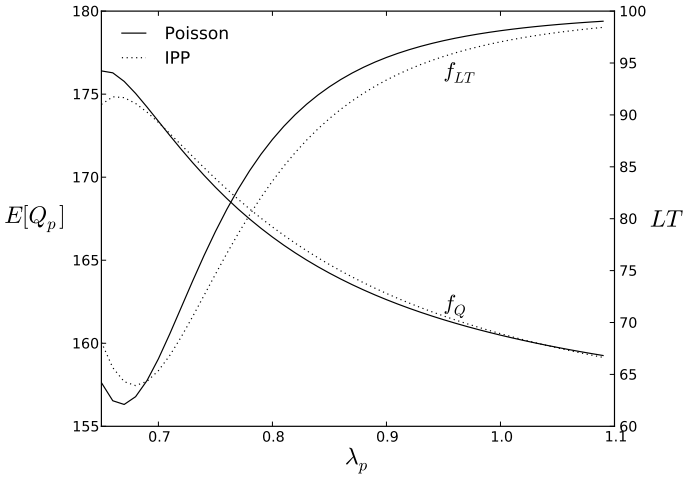
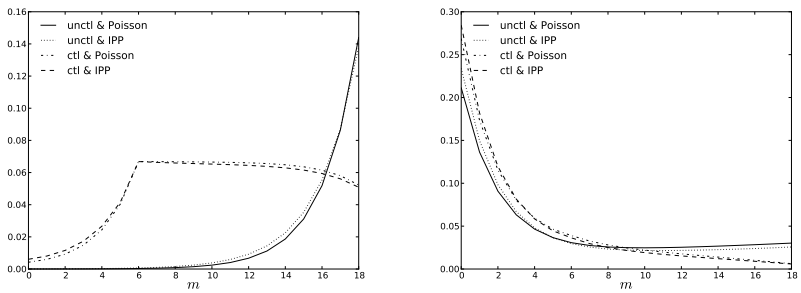


Figure 4.3: Performance difference of hybrid MTS/MTO systems with regular and irregular order arrivals.

management, Figure 4.3 depicts the mean lead time and inventory level of the uncontrolled management, expressed as a fraction of the corresponding performance measure in the controlled case. The fractions for the mean lead time and mean inventory level are respectively defined in % as

$$f_{LT} = \frac{\mathrm{LT}(unctl)}{\mathrm{LT}(ctl)}, \quad f_Q = \frac{\mathrm{E}Q_p(unctl)}{\mathrm{E}Q_p(ctl)}$$

Of course, when the production rate increases, we see an increase of the inventory level and a decrease of the lead time. In absence of setup times, the controlled inventory management differs from the uncontrolled case only by that production restarts at the threshold T_p and not at $C_p - 1$. As such, the threshold mechanism allows to set the trade-off between mean inventory level and mean lead time. Indeed, as Figures 4.2 and 4.3 show, the control considerably decreases the mean inventory level at the cost of increased mean lead times. Correlation in the order process affects these performance measures differently. Figure 4.2 shows that correlation induces a decrease of the mean inventory level and an increase of the mean lead time for both controlled and uncontrolled inventories. The increase of lead times is expected. Correlation means long periods with more arrivals than processable such that order arrivals see more backlog on arrival on average. The decrease of the mean inventory level is not that easily explained. Here, one sees that correlation in the order process induces longer periods where there is backlog, while the conservation of the throughput learns that the periods without backlog are not



(a) (b)

.

Figure 4.4: Probability distribution of the semi-finished product inventory of hybrid MT-S/MTO systems when λ_p *equals* 1.0 *(a) and* 0.62 *(b)*

longer. As the mean inventory level grows to the inventory capacity in the absence of backlog and to some lower value when there is backlog, we observe an overall decrease of the mean inventory level. Finally, note that for increasing λ_p, control no longer matters for the lead time as production is sufficiently fast to ensure the presence of semi-finished products. In contrast, correlation in the order process does not affect the mean inventory level for increasing λ_p as is easily verified on Figures 4.2 and 4.3.

Figures 4.4(a) and 4.4(b) depict the probability mass functions of the inventory level for $\lambda_p = 1.0$ and 0.62, respectively. As in Figure 4.2, we set $\sigma = 0.8$ and $\kappa = 10$ for the IPP and assume that the load of the IPP and the arrival rate of the Poisson process equal $\rho_o = 0.6$. Note that λ_p is only slightly higher than λ_o in Figure 4.4(b) whereas it is considerably larger in Figure 4.4(a). This explains the obvious dissimilarity between the mass functions: for $\lambda_p = 1.0$ (Figure 4.4(a)) the inventory is filled up considerably faster than it is depleted. The inventory is hardly ever empty in this case; the probability mass is concentrated on high inventory levels. Further notice that the threshold is easily discernible. In contrast, the inventory level is most often empty or of limited size for $\lambda_p = 0.62$ (Figure 4.4(b)) as the production rate is only slightly higher than the order arrival load. Reaching the inventory capacity is a rare event in this case such that the probability mass functions for the controlled and uncontrolled systems hardly differ.

We now consider the effect of setup times on performance. As before, we benchmark the controlled inventory management against the uncontrolled case. The inclusion of a phase-type distributed setup time prior to production is easily accomplished as follows. Let $\mathbf{F}$ be the $F \times F$ generator matrix of the non-absorbing transitions of the phase-type distribution, let $\mathbf{f}'$ be the column vector of rates to the absorbing state and let $\boldsymbol{\phi}$ be the row vector of initial state probabilities. Assuming

that production is a Poisson process after the setup, we get the following characterisation of the arrival process:

$$\mathbf{P}_0 = \begin{bmatrix} \mathbf{F} & \mathbf{f}' \\ \mathbf{0}'_F & 0 \end{bmatrix}, \quad \mathbf{P}_1 = \begin{bmatrix} \mathbf{0}_{F\times F} & \mathbf{0}_F \\ \mathbf{0}'_F & \lambda_p \end{bmatrix},$$

and,

$$\mathbf{b} = \begin{bmatrix} \boldsymbol{\phi} & 0 \end{bmatrix}.$$

Figure 4.5 shows the mean inventory level and the mean lead time versus the production rate λ_p for hybrid MTS/MTO systems with controlled and uncontrolled inventory management. The same parameter set is used as for the Poisson order arrivals in Figures 4.2 and 4.3, but we now include a setup time prior to the production in the controlled case. The setup time is either exponentially distributed or Erlang-10 distributed, in either case with mean $\mu_p = 2$. The Erlang-10 under consideration proves to be a good approximation for constant setup times. It is well known that by increasing the number of phases of an Erlang distribution, it is possible to approximate the constant distribution arbitrarily well. Experimentation with Erlang distributions with more than 10 phases revealed that the performance measures at hand are visually indiscernible if we add more phases. As in Figure 4.2, we again see that the control allows to trade in shorter lead times for lower inventory levels. The introduction of setup times further lowers the inventory level but also induces an increase of the mean lead times. Most notable is the limited effect of the distribution of the setup times on mean lead time (the curves visually coincide) and the mean inventory level.

Finally, we study the impact of the order processing time distribution. We again take the same parameters as for the Poison process in Figure 4.2 and now compare exponentially distributed and constant order processing times. Figure 4.6 shows the mean inventory level and mean lead time versus the product arrival rate for hybrid MTS/MTO systems with controlled and uncontrolled inventory management and for exponential and constant order processing times. Again, the constant order processing times are approximated by an Erlang-10 distribution and it was checked that having more phases does not significantly alter the performance measures. Similarly as for the setup-time distributions, we observe that the distribution of the order processing times have only little impact on mean lead time and mean inventory level.

4.4.2 Cost analysis

In view of the limited sensitivity of the mean lead time and mean inventory level with respect to both setup and processing time distributions, we here limit the discussion to exponentially distributed setup and order processing times. Moreover

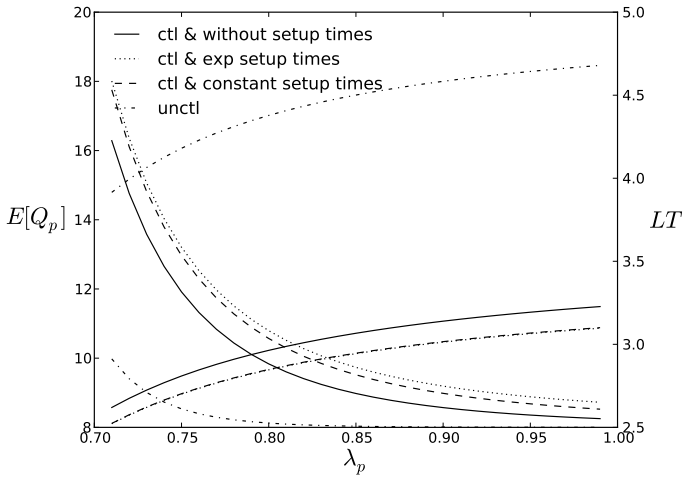


Figure 4.5: Impact of setups on the performance of hybrid MTS/MTO systems with a controlled decoupling inventory and performance difference with the uncontrolled case.

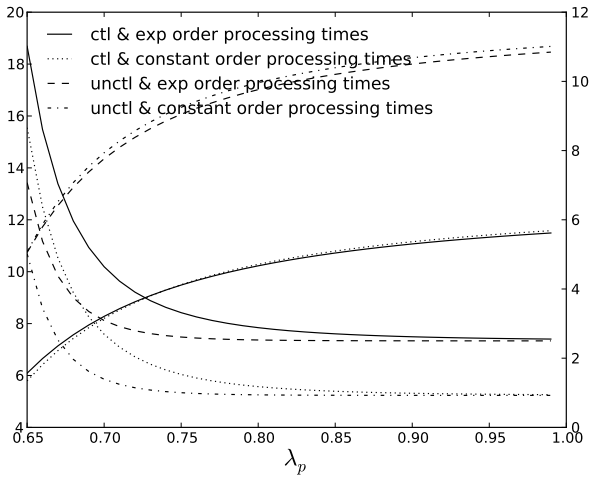


Figure 4.6: Impact of the order processing time distribution on the performance of hybrid MTS/MTO systems with a controlled and uncontrolled decoupling inventory.

we assume Poisson product and order arrivals. Recall that the parameter values of these distributions are listed in Table 4.1.

We now identify the various costs associated to the inventory model with a controlled decoupling inventory. The holding cost relates to the inventory level. Let c_1 be the holding cost per time unit and per item in the inventory, the mean holding cost per item then equals,

$$C_h = \frac{1}{\eta} c_1 \, \mathrm{E} \, Q_p \, .$$

We further associate a fixed cost c_2 every time there is a setup. We have the following setup cost per item,

$$C_s = \frac{1}{\eta} c_2 \gamma,$$

where γ denotes the number of setups per time unit. The latter can be calculated by summing the products of all probabilities of states from which a departure leads to the start of the setup and the rate at which this departure occurs. Hence, we consider the states where the product inventory equals $T_p + 1$, there is no production and the order backlog is nonempty. We have,

$$\gamma = \sum_{n>0}^{\infty} \sum_{x_o \in Q} \sum_{x_s \in \mathcal{T}} \pi(n, T_p + 1, (x_o, x_s, \theta)) t_{x_s}$$

To refer to the state changes in the environment variable, we denote this variable by the triplet $(x_o, x_s, \theta) : x_o \in Q, x_s \in \mathcal{T}$ as defined in Section 4.2.2.

As increasing lead times correspond to diminishing service levels, a cost can be associated with the lead time. We here consider two alternatives. First, we assume that the cost of the lead time is proportional to the lead time. Let c_3 be the cost per time unit of lead time per product, the mean cost related to the lead times then equals,

$$C_\ell = c_3 \, \mathrm{LT} \, .$$

The linear increase of cost in terms of lead times may not very well reflect the real cost associated with lead times. A more realistic function associates zero cost with zero lead times and increasing but bounded costs for increasing lead times. We therefore propose the following cost function,

$$c_4 (1 - \exp(-\omega x)) \, ,$$

with c_4 the maximum cost and where ω describes how fast the cost increases to this maximum. Figure 4.7 compares the linear and non-linear lead-time cost functions for $c_3 = 1$ and $c_4 = 10$. In the case of a non-linear increase of cost, we assume that ω equals 0.2 or 0.5.

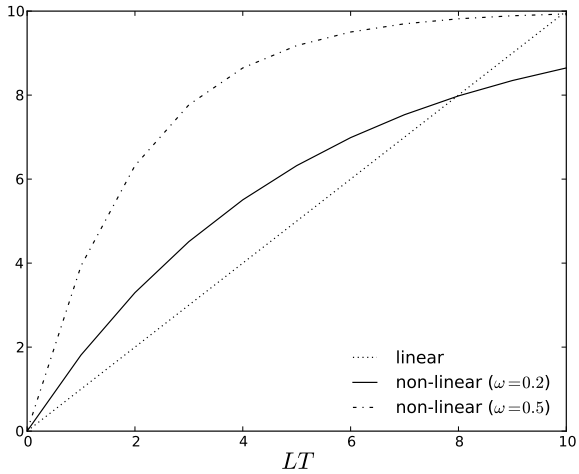


Figure 4.7: Linear versus non-linear lead-time cost function.

Averaging over all lead times yields,

$$\widehat{C}_\ell = c_4(1 - \mathcal{L}^*(\omega))$$

where $\mathcal{L}^*$ is the Laplace transform of the lead time distribution. As orders are processed in order of arrival, one readily observes that the size of the backlog upon completion of a product arrival equals the number of arrivals that occurred during the lead time of this order. Hence, as order arrivals occur according to a Poisson process with rate ρ_o, the probability generating function of the backlog size $U_o(z)$ upon departure can be expressed in terms of the Laplace transform of the lead time as follows,

$$U_o(z) = \mathcal{L}^*(-\rho_o(z-1)) = \sum_{n=0}^{\infty} \pi^{(o)}(n) z^n \,.$$

Here, the last equality follows from the observation that the distribution of the backlog upon departures equals the distribution of the backlog upon arrivals which in turn equals the distribution of the backlog at random times by the PASTA property. Summarizing, we have the following cost,

$$\widehat{C}_\ell = c_4 \left(1 - \sum_{n=0}^{\infty} \pi^{(o)}(n) \left(\frac{\rho_o - \omega}{\rho_o} \right)^n \right) .$$

Finally, in view of the costs defined above, we consider the following two overall cost functions,

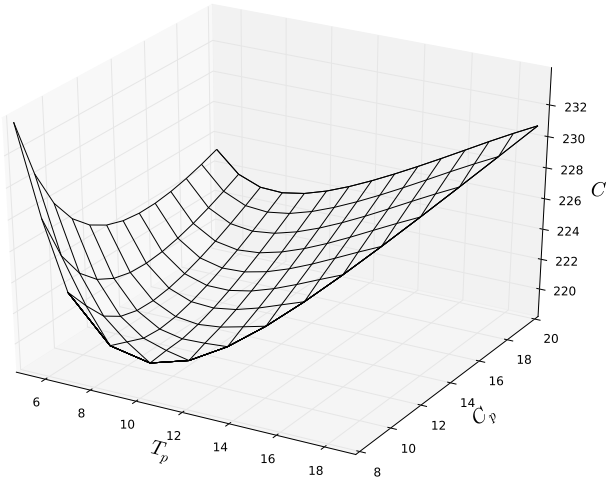


Figure 4.8: Total cost of hybrid MTS/MTO systems with a controlled decoupling inventory and a linear lead-time cost function.

$$C = C_h + C_s + C_\ell \,,$$

and,

$$\widehat{C} = C_h + C_s + \widehat{C}_\ell \,.$$

Figure 4.8 depicts the total cost of hybrid MTS/MTO systems with a linear lead-time cost function. The inventory capacity varies from 8 to 20 and the threshold value T_p from 5 to 19. We assume a holding cost c_1 equal to 1, a setup cost c_2 equal to 5 and a linear lead time cost c_3 equal to 80. As the figure shows, we have a minimum cost when the inventory capacity equals 11 and the threshold value equals 9. Obviously, when the inventory capacity and the threshold value increase, the mean lead time decreases and the mean semi-finished product inventory increases. The number of setups per time unit however increases when the threshold value increases but decreases when the inventory capacity increases.

Figure 4.9 depicts the total cost for the case of a non-linear lead-time cost function. The same holding and setup cost are assumed as in previous figure and we here assume a non-linear lead time cost c_4 equal to 500 and a rate ω equal to 0.3. The inventory capacity also varies from 8 to 20 and the threshold value from 5 to 19. As the figure shows, we have a minimum cost when the inventory capacity equals 11 and the threshold value equals 8.

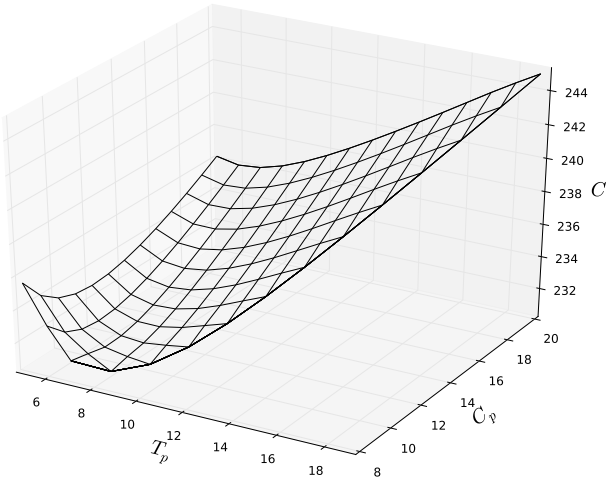


Figure 4.9: Total cost of hybrid MTS/MTO systems with a controlled decoupling inventory and a non-linear lead-time cost function.

4.5 Conclusion

Hybrid make-to-stock (MTS)/ make-to-order (MTO) systems are described as stochastic inventory models with two "queues": the semi-finished product inventory and the order backlog. We rely on matrix-analytic techniques to evaluate the performance of such systems. Our approach allows to account for uncertainty in demand, production and order processing times under non-restrictive stochastic assumptions. Another advantage is its ease of use and computational efficiency in comparison with simulation which makes it an adequate tool for managers. The proposed methodology is sufficiently versatile to account for a continuous review (s,S)-policy in which the replenishment of semi-finished products halts when the inventory is full and restarts when the inventory level drops to a certain threshold value. As the numerical examples show, the distribution of the setup and order processing is shown to have limited impact on the mean lead time and inventory level. However, inventory control and correlation in the order process decreases the mean inventory level at the cost of increased mean lead times. Finally, to capture the trade-off between inventory and service level, we define a cost structure for hybrid MTS/MTO systems with a controlled decoupling inventory.

References

[1] I. Adan and J. Wal. *Combining make to order and make to stock*. OR Spectrum, 20(2), p.73–81, 1998.

[2] A. Arreola-Risa and G.A. DeCroix. *Make-to-order versus make-to-stock in a production inventory system with general production times*, IIE Transactions, 30(8), p.705–716, 1998.

[3] A. Bensoussan, R.H. Liu and S.P. Sethi. *Optimality of an (s, S) policy with compound Poisson and diffusion demands: A quasi-variational inequalities approach*, SIAM Journal on Control and Optimization, 44(5), 1650–1676, 2005.

[4] P. Buchholz. *An EM-Algorithm for MAP Fitting from Real Traffic Data*, Computer Performance Evaluation, Modelling Techniques and Tools, Lecture Notes in Computer Science, 2794, p.218–236, 2003.

[5] G. Casale, E.Z. Zhang and E. Smirni. *KPC-Toolbox: Simple Yet Effective Trace Fitting Using Markovian Arrival Processes*, Fifth International Conference on Quantitative Evaluation of Systems, p.83, 2008.

[6] E. De Cuypere, K. De Turck and D. Fiems. *A queueing theoretic approach to decoupling inventory*, Lecture Notes in Computer Science, 7314, p.150-164, 2012.

[7] Y. Gao, M.L. Wen and S.B. Ding. *(s, S) policy for uncertain single period inventory problem*, International Journal of Uncertainty Fuzziness and knowledge-based systems, 21(6), p.945–953, 2013.

[8] L. Gun. *Experimental results on matrix-analytical solutions techniques: extensions and comparisons*, Stochastic models, 5(4), p.669–682, 1989.

[9] D. Gupta and S. Benjaafar. *Make-to-order, make-to-stock, or delay product differentiation? A common framework for modeling and analysis*, IIE Transactions, 36(6), p.529–546, 2004.

[10] O. Ghrayeb, N. Phojanamongkolkij and B.A. Tan. *A hybrid push/pull system in assemble-to-order manufacturing environment*, Journal of Intelligent Manufacturing, 20, p.379–387, 2009.

[11] U. Gurler and B.Y. Ozkaya. *Analysis of the (s, S) policy for perishables with a random shelf life*, IIE TRANSACTIONS, 40(8), p.759–781, 2008.

[12] A. M. Haghighi and D.P. Mishev. *Queueing Models in Industry and Business*, Nova Science Publishers, Inc., New York, ISBN 978-1-60456-189-0, p.129–148, 2008.

[13] S. Hoekstra, J. Romme and S. Argelo. *Integral logistic structures: developing customer-oriented goods flow*, McGraw-Hill, 1992.

[14] P. Kaminsky and O. Kaya. *Combined make-to-order/make-to-stock supply chains*, IIE Transactions, 41, p.103–119, 2009.

[15] J. Köber and G. Heinecke. *Hybrid Production Strategy between Make-to-order and Make-to-stock - A Case Study at a Manufacturer of Agricultural Machinery with Volative and Seasonal Demand*, 45th CIRP Conference on Manufacturing Systems, 3, p.453–458, 2012.

[16] G. Latouche and Ramaswami. *Introduction to Matrix Analytic Methods in Stochastic Modeling*, 1999.

[17] H.S. Lee and M.M. Srinivasan. *The continuous (s,S) review policy for production/inventory systems with compound Poisson demands*, Technical report, Industrial & OE University of Michigan, Ann Arbor, 1988.

[18] H.S. Lee and M.M. Srinivasan. *The continuous (s,S) review policy for production/inventory systems with Poisson demands and arbitrary processing times*, Technical report, Industrial & OE University of Michigan, Ann Arbor, 1987.

[19] M.M. Srinivasan and H.-S. Lee. *Random review production/inventory systems with compound Poisson demands and arbitrary processing times*. Management Science, 37(7), p.813-833, 1991.

[20] C. Nielsen and C. Larsen. *An analytical study of the Q(s, S) policy applied to the joint replenishment problem*. European Journal of Operational Research, 163(3), p.721–732, 2005.

[21] H. Ohta, T. Hirota and A. Rahim. *Optimal production-inventory policy for make-to-order versus make-to-stock based on the $M/E_r/1$ queueing model*. International Journal of Advanced Manufacturing Technology, 33, p.36–41, 2006.

[22] M. Parlar. *Continuous-review inventory problem with random supply interruptions*, European Journal of Operational Research, 99, p.366–385, 1997.

[23] H. Rafiei and M. Rabbani. *An MADM Framework toward Hierarchical Production Planning in Hybrid MTS/MTO Environments*, World Academy of Science, Engineering and Technology, 3, p.10–25, 2009.

[24] H. Rafiei and M. Rabbani. *Capacity coordination in hybrid make-to-stock/make-to-order production environments*, International Journal of Production Research, 50(3), p.773–789, 2012.

[25] C.A. Soman, D.P. Van Donk and G. Gaalman. *Combined make-to-order and make-to-stock in a food production system*, International Journal of Production Economics, 90, p.223–235, 2004.

[26] C.A. Soman, D.P. Van Donk and G. Gaalman. *Comparison of dynamic scheduling policies for hybrid make-to-order and make-to-stock production systems with stochastic demand*, International Journal of Production Economics, 104, p.441–453, 2006.

5

Stochastic modelling of energy harvesting for low power sensor nodes

Eline De Cuypere, Koen De Turck and Dieter Fiems

presented at QTNA conference, August 2012, Kyoto, Japan.

Abstract. Battery lifetime is a key impediment to long-lasting low power sensor nodes. Energy or power harvesting mitigates the dependency on battery power, by converting ambient energy into electrical energy. This energy can then be used by the device for data collection and transmission. This paper proposes and analyses a queueing model to assess performance of such an energy harvesting sensor node. Accounting for energy harvesting, data collection and data transmission opportunities, the sensor node is modelled as a paired queueing system. The system has two queues, one representing accumulated energy and the other being the data queue. By means of some numerical examples, we investigate the energy-information trade-off.

5.1 Introduction

The problem of battery replacement and disposal is a key impediment to ubiquitous use of wireless sensors networks. Sensor networks are formed by a collection of intercommunicating sensor nodes, collecting spatially distributed data (temperature, humidity, movement, noise, ...). Sensors networks can be used in a large range of

applications, including military, environmental, home and health applications [1]. Despite vast improvements on power consumption and ongoing developments in power management, the lifetime of wireless sensors is largely determined by the energy of on-board batteries [13]. To overcome dependency on batteries, current research effort focusses on wireless devices that extract the necessary energy from their environment [6]. Possible power sources include electromagnetic radiation, thermal energy as well as mechanical energy [8].

The specific dynamics of energy harvesting has also drawn the attention of the modelling community. Sensors being autonomous in deciding which information will be transmitted as well as when to transmit, various authors propose game theoretic models; see e.g. [11] for power control games in wireless networks. Accounting for energy harvesting, Tsuo et al. [15] consider a Bayesian game where each node knows its local energy state. An evolutionary hawk and dove game with harvesting nodes transmitting either at high or low power is studied in [2, 5]. Specifically focussing solar power, optimal energy management for a sensor node that uses a sleep and wakeup strategy for energy conservation is studied by a bargaining game in [12].

Neither of these game-theoretic models assume that acquired data can be temporarily stored at the sensor node. To study data buffering at the sensor node, queueing theoretic modelling applies. Sharma et al. [14] is a recent contribution on such a queueing theoretic approach. These authors analytically study stochastic stability of an energy harvesting node with data buffering and rely on simulation to assess its performance. Also the present contribution investigates a queueing model for a harvesting sensor node. In particular, we assess the performance of an energy harvesting sensor node accounting for uncertainty in data acquisition, in energy harvesting and in transmission opportunities. To this end, we investigate a queueing system with two queues: one queue represents the data buffer and one queue represents the available energy. Maximising versatility of the model at hand while keeping the analysis numerically tractable, we model data acquisition, energy harvesting and transmission by means of Markovian arrival processes; an "arrival" representing some acquired data, some harvested energy and a transmission opportunity (an encounter with another node or a base station) respectively.

Such two-buffer queueing problems are sometimes termed paired queues — pairing refers to the coupling between the queues, service is only possible if both queues are non-empty — and have been studied in various contexts including leaky-bucket access control [16, 17], kitting processes [4] in assembly and decoupling buffers in production systems [3].

Leaky-bucket access control in asynchronous transfer mode, introduces a virtual buffer (a bucket) at sender nodes. The virtual buffer is filled with tokens according to some well behaved process. For every transmission, a token is taken from the bucket and transmission is only allowed if there are tokens present.

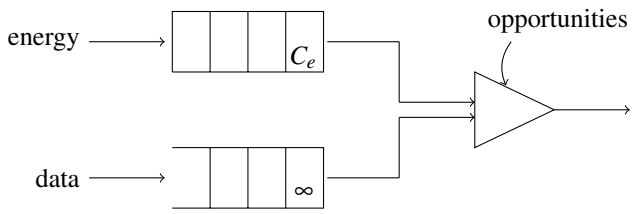


Figure 5.1: Stochastic model of energy harvesting for low power sensor nodes.

Hence, the data buffer and leaky bucket constitute a paired queueing system. Kitting is a particular strategy for supplying materials to an assembly line. Instead of delivering parts in containers of equal parts, kitting collects the necessary parts for a given end product into a specific container, called a kit, prior to arriving at an assembly unit. As kits can only be completed if all parts are present, the part buffers and the kitting operation constitutes a paired queueing system. Finally, decoupling buffers are used to reduce lead times in production systems by buffering semi-finished products at some point in the production process. When there is demand, semi-finished products are taken out of the decoupling buffer and finished according to the demand. Again, paired queueing applies as the second production stage only starts if there are semi-finished products and demand.

Finally, paired queues have also been studied in a more abstract setting. Considering a system with two paired queues, Harrison shows that it is necessary to impose a restriction on the size of the buffer to ensure stability in the operations of a kitting process [7]. Similar observations where made by Latouche [9] who studied the difference of the queue lengths in such a paired queueing system.

The remainder of this paper is organised as follows. The paired queueing model under investigation and the notationally conventions are introduced in the next section. In Section 5.3, the system is analysed as a quasi-birth-death process (QBD). Also, the numerical solution methodology is discussed and relevant performance measures are determined. To illustrate our approach, Section 5.4 considers some numerical examples. Finally, conclusions are drawn in Section 5.5.

5.2 Model description

The energy harvesting sensor node is modelled as a queueing system with two queues, as depicted in Figure 5.1. The energy queue has finite capacity C_e and stores energy extracted from the environment. The data queue keeps track of not yet transmitted data packets and has infinite capacity.

The amount of stored energy is discretised for modelling convenience. We make abstraction of the specifics of energy harvesting apart from the assumption that there is a continuous chance to come by some ‘chunks’ of energy. Therefore,

we assume that energy arrives in accordance with a Markovian arrival process with state space $\mathcal{K}_E$. Let Ω_E^0 and Ω_E^1 denote the generator matrices of this arrival process, governing the state transitions when there are no arrivals and when there is an arrival, respectively. Analogously, the sensor picks up data in accordance with a Markovian arrival process with state space $\mathcal{K}_A$: whenever it picks up data, there is an arrival in the data queue. Let Ω_A^0 and Ω_A^1 denote the generator matrices of this arrival process, governing the state transitions when there are no arrivals and when there is an arrival, respectively.

Data can only be transmitted during transmission opportunities. Moreover, the two queues are paired, meaning that data can only be transmitted if the energy buffer is non-empty. Whenever a transmission occurs, a data packet departs but the level of the energy buffer may or may not decrease (this assumption allows for modelling the dynamics of the battery with fewer states). The arrivals of transmission opportunities being exogenous to the state of the sensor node, the departure process is a marked Markov process with state space $\mathcal{K}_D$. The generator matrices Ω_D^0, Ω_D^1 and Ω_D^2 govern the state transitions of the departure process without transmission opportunities, with a transmission opportunity that leads to a decrease of the energy buffer and with a transmission opportunity that does not lead to such a decrease. Note that for the matrices Ω_E^0, Ω_A^0 and Ω_D^0, diagonal elements are assumed to be zero.

5.3 Analysis

Modulating Markov process For ease of modelling, we first consider the Markov process with state space $\mathcal{K} = \mathcal{K}_E \times \mathcal{K}_A \times \mathcal{K}_D$ that jointly describes the (marked) state changes of energy, arrival and departure processes. In the remainder, let $\mathbf{I}_E$, $\mathbf{I}_A$ and $\mathbf{I}_D$ denote identity matrices with size $|\mathcal{K}_E|$, $|\mathcal{K}_A|$ and $|\mathcal{K}_D|$, respectively. Note that the symbol $\otimes$ denotes the Kronecker's product.

- The matrix $\mathbf{A}$ governs the transitions, when there are neither arrivals nor departures:

$$\mathbf{A} = \Omega_E^0 \otimes \mathbf{I}_A \otimes \mathbf{I}_D + \mathbf{I}_E \otimes \Omega_A^0 \otimes \mathbf{I}_D + \mathbf{I}_E \otimes \mathbf{I}_A \otimes \Omega_D^0 .$$

- The matrix $\mathbf{B}_E$ governs the transitions when there is an arrival in the energy buffer:

$$\mathbf{B}_E = \Omega_E^1 \otimes \mathbf{I}_A \otimes \mathbf{I}_D .$$

- The matrix $\mathbf{B}_A$ governs the transitions when there is an arrival in the data buffer:

$$\mathbf{B}_A = \mathbf{I}_E \otimes \Omega_A^1 \otimes \mathbf{I}_D .$$

- The matrices $\mathbf{C}_1$ and $\mathbf{C}_2$ govern the transitions when there is a departure that drains the energy buffer and that does not drain this buffer, respectively:

$$\mathbf{C}_1 = \mathbf{I}_E \otimes \mathbf{I}_A \otimes \Omega_D^1, \quad \mathbf{C}_2 = \mathbf{I}_E \otimes \mathbf{I}_A \otimes \Omega_D^2.$$

Remark. The matrices $\mathbf{A}$ till $\mathbf{C}_2$ above are defined in terms of the characteristics of the different arrival processes. In the remainder, all results will be expressed in terms of the matrices as defined above. Hence, these results remain valid in the case that the different arrival processes are intercorrelated as well. In that case there is a single marked Markov process, with marks for data arrivals, energy arrivals and transmission opportunities.

Quasi-birth-death process Having defined these transition matrices, we now focus on the queueing model at hand. To be more precise, the energy harvesting sensor node system is a continuous-time Markov process with infinite state space $\mathbb{N} \times \{0,1,2,\ldots,C_e\} \times \mathcal{K}$, $\mathcal{K} = \{0,1,\ldots,K\}$. At any time, the state of the system is described by the triplet $[n,m,i]$, n being the number of data packets available, m being the energy level and i being the state of the modulating process.

The studied Markov process is a homogeneous quasi-birth-death process (QBD), see [10]. In the present setting, the *level* or block-row index, indicates the data packets available while the phase, i.e. the index within a block element, indicates both the energy level and the state of the Markovian environment. The one-step transitions are restricted to states in the same level (from state $(n,*,*)$ to state $(n,*,*)$) or in two adjacent levels (from state $(n,*,*)$ to state $(n+1,*,*)$ or state $(n-1,*,*)$).

We then find that the generator matrix of the Markov process has the following block matrix representation,

$$\mathbf{Q} = \begin{bmatrix} \mathcal{B}_0 & \mathcal{A}_2 & 0 & 0 & \cdots \\ \mathcal{A}_0 & \mathcal{A}_1 & \mathcal{A}_2 & 0 & \cdots \\ 0 & \mathcal{A}_0 & \mathcal{A}_1 & \mathcal{A}_2 & \cdots \\ 0 & 0 & \mathcal{A}_0 & \mathcal{A}_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}. \tag{5.1}$$

The blocks are given by,

$$\mathcal{B}_0 = \begin{bmatrix} \underline{\mathbf{D}} & \mathbf{B}_E & 0 & \cdots & 0 \\ 0 & \underline{\mathbf{D}} & \mathbf{B}_E & \cdots & 0 \\ 0 & 0 & \underline{\mathbf{D}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \underline{\mathbf{D}} \end{bmatrix} \tag{5.2}$$

$$\mathcal{A}_2 = \begin{bmatrix} \mathbf{B}_A & 0 & 0 & \cdots & 0 \\ 0 & \mathbf{B}_A & 0 & \cdots & 0 \\ 0 & 0 & \mathbf{B}_A & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \mathbf{B}_A \end{bmatrix} \tag{5.3}$$

$$\mathcal{A}_0 = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ \mathbf{C}_1 & \mathbf{C}_2 & \cdots & 0 & 0 \\ 0 & \mathbf{C}_1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \mathbf{C}_1 & \mathbf{C}_2 \end{bmatrix} \tag{5.4}$$

$$\mathcal{A}_1 = \begin{bmatrix} \underline{\mathbf{D}} & \mathbf{B}_E & 0 & \cdots & 0 \\ 0 & \mathbf{D} & \mathbf{B}_E & \cdots & 0 \\ 0 & 0 & \mathbf{D} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \mathbf{D} \end{bmatrix}. \tag{5.5}$$

with $\mathbf{D} = \mathbf{A} - \partial\mathbf{A} - \partial\mathbf{C}_1 - \partial\mathbf{C}_2 - \partial\mathbf{B}_A - \partial\mathbf{B}_E$ and $\underline{\mathbf{D}} = \mathbf{D} + \mathbf{C}_1 + \mathbf{C}_2$, where the notation $\partial\mathbf{X}$ represents a diagonal matrix with diagonal elements equal to the row sums of $\mathbf{X}$.

Numerical solution Having defined the different blocks of the QBD process, we now focus on its solution. Recall that the state of the Markov process was described by the triplet $[n,m,i]$; n is the size of the data buffer, m is the size of the energy buffer and i is the state of the modulating process. Let $\pi(n,m,i)$ be the steady state probability to be in state $[n,m,i]$. A well-known method for finding the stationary distribution of QBD processes is the matrix-geometric method. Using the vector notation $\boldsymbol{\pi}_k = (\pi(k,0,0), \pi(k,0,1), \ldots, \pi(k,C_e,K))$, the probability vectors can be expressed as,

$$\boldsymbol{\pi}_k = \boldsymbol{\pi}_0 \mathbf{R}^k. \tag{5.6}$$

where the so-called rate matrix $\mathbf{R}$ is the minimal non-negative solution of the non-linear matrix equation

$$\mathbf{R}^2 \mathcal{A}_0 + \mathbf{R}\mathcal{A}_1 + \mathcal{A}_2 = \mathbf{0}.$$

We compute the rate matrix by implementing the efficient iterative algorithm of [10], chapter 8.

Performance measures Once the steady state probabilities have been determined numerically, we can calculate a number of interesting performance measures for the harvesting energy sensor node. For ease of notation, we introduce the marginal probability mass functions of the energy and the data queue content: $\pi^{(e)}(m) = \sum_{i\in\mathcal{K}}\sum_{n=0}^{\infty}\pi(n,m,i)$ and $\pi^{(d)}(n) = \sum_{i\in\mathcal{K}}\sum_{m=0}^{C_e}\pi(n,m,i)$.

Note that as the data queue is infinite, the throughput of the sensor node system η equals the data arrival rate λ_d. In addition, we have the following performance measures.

- The mean energy queue and the mean data queue: $\mathrm{E}\,Q_e$ and $\mathrm{E}\,Q_d$ respectively,
$$\mathrm{E}\,Q_e = \sum_{m}^{C_e}\pi^{(e)}(m)m\,, \quad \mathrm{E}\,Q_d = \sum_{n}^{\infty}\pi^{(d)}(n)n\,.$$
- The variance of the energy queue and the data queue: $\mathrm{Var}\,Q_e$ and $\mathrm{Var}\,Q_d$ respectively,
$$\mathrm{Var}\,Q_e = \sum_{m}^{C_e}\pi^{(e)}(m)m^2 - (\mathrm{E}\,Q_e)^2\,,$$
$$\mathrm{Var}\,Q_d = \sum_{n}^{\infty}\pi^{(d)}(n)n^2 - (\mathrm{E}\,Q_d)^2\,.$$
- The mean delay L (calculated based on Little's theorem) is the average amount of time between the arrival of a data packet its transmission:
$$L = \frac{\mathrm{E}\,Q_d}{\lambda_d}$$
- As the energy queue has finite capacity, energy harvesting may be blocked. This happens when energy is captured but the queue is full. Hence, blocking corresponds to the loss probability in the energy queue. The loss probability is most easily expressed in terms of the throughput. We have,
$$b_e = \frac{\lambda_e - \eta}{\lambda_e} = \frac{\lambda_e - \lambda_d}{\lambda_e}\,.$$

5.4 Numerical results

We now illustrate our approach by means of some numerical examples.

Poisson arrivals and exponential data transmission opportunities As a first example, the difference between the mean energy queue and the mean data queue versus the capacity C_e is depicted in Figure 5.2(a). We assume that energy units and data units arrive according to a Poisson process with parameter $\lambda_e = 0.6$ and

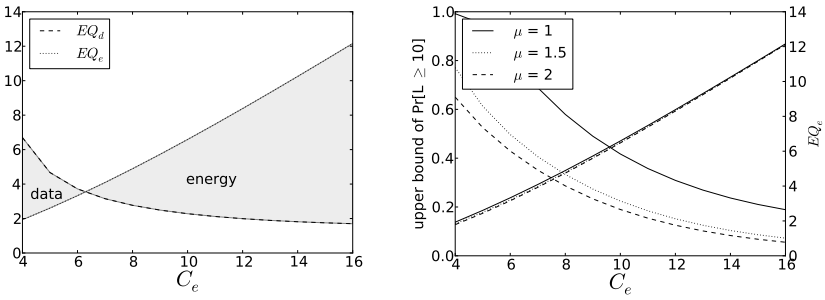
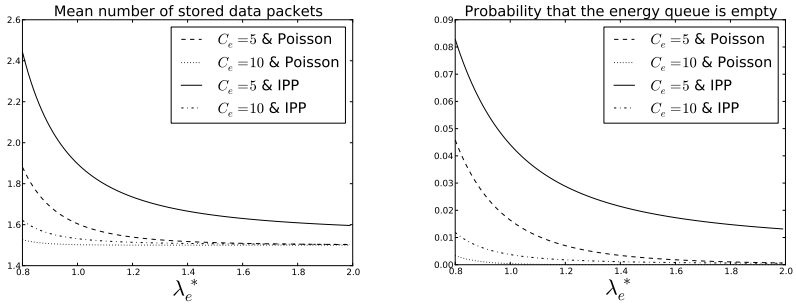


Figure 5.2: There is a trade-off between the mean amount of stored energy and stored data and between the delay.

$\lambda_d = 0.6$, respectively. The probability to use one unit of energy for data transmission p equals 0.8 and the data transmission opportunities are exponentially distributed with service rate μ equal to 1. As the figure shows, the buffer capacity of 6 results more or less on average in the same amount of data and energy in the buffer. Under and above the level, energy and data are on average backlogged, respectively. Obviously, there is on average more amount of energy and less buffer of data as the capacity increases.

Figure 5.2(b) represents the trade-off between the upper bound of the probability to have a delay higher or equal to 10 (left side) and the mean amount of stored energy (right side). Note that we calculated the delay distribution by using the one-sided Chebyshev's inequality. Under the same parameter assumptions of Figure 5.2(a), the upper bound of the probability to have a delay higher or equal to 10 decreases and the mean amount of stored energy increases as the energy capacity increases for each service rate. Indeed, if more buffer capacity is available, it will be used — the energy queue increases such that there is on average less time required to transmit one data unit. Furthermore, we observe a slightly decrease of the amount of energy as the service rate μ increases. Indeed, the more data is transmitted per time unit, the higher the mean amount of energy used to transmit data. Finally, the upper bound probability to have a delay equal or higher than 10 decreases as the service rate increases, as expected.

Markovian arrival process for energy We also quantify the impact of irregular capture of energy. To this end we compare both buffers with Poisson arrivals to corresponding system with interrupted Poisson arrivals for the energy and Poisson arrivals for the data. The arrival interruptions account for inefficiency in the energy harvesting process.



(a) (b)

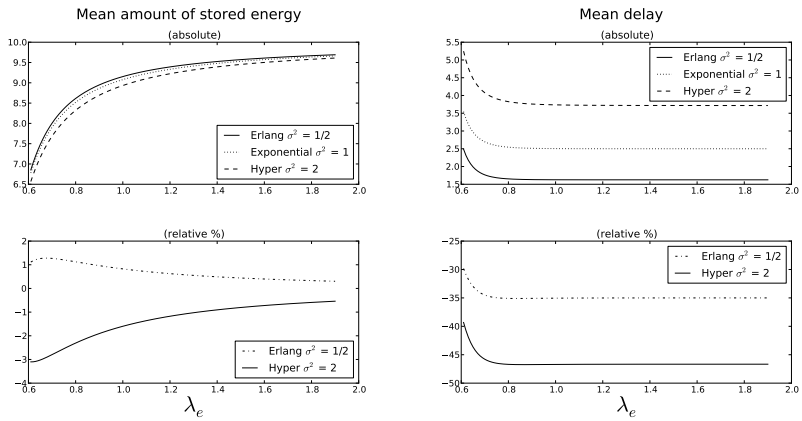
Figure 5.3: Irregular capture of energy results in a higher mean number of stored data packets (a) and a higher probability that the energy queue is full (b).

The interrupted Poisson process considered here is a two-state Markovian process. In the active state, generated energy arrives in accordance with a Poisson process with rate λ_e whereas no new energy arrives in the inactive state. Let α and β denote the rate from the active to the inactive state and vice versa, respectively. We then use the following parameters to characterise the interrupted Poisson process (IPP),

$$\sigma = \frac{\beta}{\alpha+\beta}, \quad \kappa = \frac{1}{\alpha} + \frac{1}{\beta}, \quad \lambda_e^* = \lambda_e \sigma.$$

Note that σ is the fraction of time that the interrupted Poisson process is active, the absolute time parameter κ is the average duration of an active and an inactive period, and λ_e^* is the arrival load of energy.

Figure 5.3(a) shows the mean number of stored data packets versus the arrival load of energy with buffer capacity C_e equal to 5 and 10 for Poisson arrivals as well as for interrupted Poisson arrivals of energy. The probability to use one unit of energy for data transmission equals 0.8 and transmission times are exponentially distributed with service rate μ equal to 1. In addition, we set $\sigma = 0.8$ and $\kappa = 10$ for the interrupted Poisson process (e.g. $\lambda_e = 0.8$ for Poisson arrivals and $\lambda_e = 1.0$ for interrupted Poisson arrivals). The data arrival rate λ_d equals 0.6. As expected, the mean number of stored data packets decreases as the arrival rate of energy increases. Furthermore, the impact of the buffer capacity decreases as the arrival rate of energy λ_e increases. Finally, comparing interrupted Poisson and Poisson processes, burstiness in the energy harvesting process has a negative impact on performance — there is on average more time required to transmit one data unit. Figure 5.3(b) confirms the previous results. Indeed, the probability to have an empty energy queue decreases as the buffer capacity of energy decreases and the probability is higher for interrupted Poisson than for Poisson arrivals.



(a) (b)

Figure 5.4: Given the mean transmission time, the data transmission opportunity distribution has only a limited impact on the mean amount of stored energy (a) and has a significant impact on the mean delay (b) in this case.

Phase-type distributed data transmission opportunities The last numerical example quantifies the impact of the distribution of the data transmission opportunity on the sensor node system. Figure 5.4(a) and 5.4(b) depict the mean amount of energy in the queue and the mean delay of the sensor node system. In both figures, the energy arrival rate λ_e is varied and different values of the variance of the data transmission opportunity distribution are assumed as indicated. The probability to use one unit of energy for data transmission p equals 0.8 and the mean service time equals 1 for all curves. We consider a two-phase hyper-exponential distribution (in which each phase has the same probability to occur) and a two-phase Erlang distribution. Note that two corner cases coincide both with an exponential distribution: a hyper-exponential distribution with unit variance and an Erlang distribution with one phase. Furthermore, the data arrival rate λ_d equals 0.6 and the energy capacity C_e equals 10. Clearly, Figure 5.4(a) and 5.4(b) show respectively that the energy buffer content converges to maximum capacity and the mean delay decreases to a certain value as the energy arrival rate increases. The second plot shows values relative to the exponential distribution. Concerning the mean amount of stored energy, we observe that the data transmission opportunity distribution has no significant effect on this performance measure. However, the difference between σ^2 equal to $1/2$, 1 and 2 for the mean delay remains constant and is significant. Finally, the mean amount of stored energy and the mean delay show respectively a slight decrease and increase as the variance of the data transmission opportunity distribution σ^2 increases.

5.5 Conclusion

In this paper, we analyse the performance of different energy harvesting sensor nodes. In particular, we investigate the impact of irregular capture of energy in the environment as well as the data transmission opportunity distribution on the performance of sensor node systems. In the studied system, both accumulated energy and data needs to be available for transmission. Furthermore, we assume that there is a probability that one unit of energy will whether or not be used to transmit one unit of data. Therefore, the studied sensor node system is modelled as a homogeneous quasi-birth-death process (QBD) and solved with matrix-analytic methods.

As our numerical examples show, there is trade-off to be made between the storage cost of energy and the service level of the sensor node, as expected — e.g. a higher capacity causes on average a higher storage of energy and a smaller time between data availability and data transmission. Furthermore, irregular capture of energy has a negative effect on the performance of the sensor node system. However, system performance is partially insensitive to variation in the data transmission opportunity distribution. Future work will focus on determining the total cost of the studied sensor node system.

References

[1] IF. Akyildiz, W. Su, Y. Sankarasubramaniam and E. Cayirci. *Wireless sensor networks: a survey*, Computer Networks, 38(4), p.393–422, 2002.

[2] E. Altman, D. Fiems, M. Haddad and J. Gaillard. *Semi-Dynamic Hawk and Dove Game Applied to Power Control*, Proceedings of INFOCOM 2012.

[3] E. De Cuypere, K. De Turck and D. Fiems. *Performance analysis of a decoupling stock in a Make-to-Order system*, Proceedings of the 14th IFAC Symposium on Information Control Problems in Manufacturing, 2012.

[4] E. De Cuypere and D. Fiems. *Performance evaluation of a kitting process*, Proceedings of the 17th International Conference on analytical and stochastic modelling techniques and applications, Lecture Notes in Computer Science, 6751, 2011.

[5] M. Haddad, E. Altman, J. Gaillard and D. Fiems. *A Semi-Dynamic Evolutionary Power Control Game*, Proceedings of Networking 2012, Lecture Notes in Computer Science, 2012.

[6] J.M. Gilbert and F. Balouchi. *Comparison of Energy Harvesting Systems for Wireless Sensor Networks*, International Journal of Automation and Computing, 5, p.334, 2008.

[7] J.M. Harrison. *Assembly-Like Queues*, Journal of Applied Probability, 10, p.354–367, 1973.

[8] H.S. Kim, J.-H. Kim and J. Kim. *A Review of Piezoelectric Energy Harvesting Based on Vibration*, International Journal of Precision Engineering and Manufacturing, 12, p.1129–1141, 2012.

[9] G. Latouche. *Queues with paired customers*, Journal of Applied Probability, 18, p.684–696, 1981.

[10] G. Latouche and V. Ramaswami. *Introduction to Matrix Analytic Methods in Stochastic Modeling*, SIAM, 1999.

[11] F. Meshkati, H.V. Poor and S.C. Schwartz. *Energy-Efficient Resource Allocation in Wireless Networks*, Signal Processing Magazine, IEEE, 24, p.58–68, 2007.

[12] D. Niyato and M.M. Rashid and V.K. Bhargave. *Wireless sensor networks with energy harvesting technologies: A game-theoretic approach to optimal energy management*, IEEE Wireless Communications, 14, p.90–96, 2007.

[13] J.A. Paradiso and T. Starner. *Energy scavenging for mobile and wireless electronics*, IEEE Pervasive Computing, 4, p.18–27, 2005.

[14] V. Sharma, U. Mukherji and V. Joseph. *Optimal energy management policies for energy harvesting sensor nodes*, IEEE Transactions on Wireless Communications, 6, p.1326–1336, 2010.

[15] F.Y. Tsuo, H.P. Tan, Y.H. Chew and H.Y. Wei. *Energy-Aware Transmission Control for Wireless Sensor Networks Powered by Ambient Energy Harvesting: A Game-Theoretic Approach*, IEEE International Conference on Communications, 2011.

[16] S. Wittevrongel and H. Bruneel. *A heuristic analytic technique to calculate the cell loss ration in a leaky bucket with bursty input traffic*, AEU - International Journal of Electronics and Communications, 3, p.162–169, 1994.

[17] S. Wittevrongel and H. Bruneel. *Analytic study of the queueing performance and the departure process of a leaky bucket with bursty input traffic*, AEU - International Journal of Electronics and Communications, 1, p.1–10, 1996.

6

Performance evaluation of an energy harvesting sensor node

Eline De Cuypere, Koen De Turck and Dieter Fiems

submitted to Ad Hoc Networks Journal.

Abstract. Battery lifetime is a key impediment to long-lasting low power sensor nodes and networks thereof. Energy harvesting — conversion of ambient energy into electrical energy — has therefore emerged as an alternative to battery power. In this paper, we propose a Markovian model for studying the impact of uncertainty in energy harvesting, energy expenditure, data acquisition and data transmission on the performance of an energy harvesting sensor node. To this end, the energy harvesting sensor node is described as a paired queueing system, one queue corresponding to the energy battery and the other to the data buffer for sensed data. We show that under non-restrictive assumptions on the data acquisition, transmission and energy harvesting processes, performance can be assessed quickly by means of matrix-analytic methods. We illustrate our approach by means of numerical examples and particularly highlight the effects of correlation in energy harvesting.

6.1 Introduction

Sensor networks, formed by collections of intercommunicating sensor nodes (SN), are used to collect and monitor spatially distributed data like temperature, humidity, movement, noise, etc [1, 2, 35]. Sensor networks have a variety of applications including military, environmental, home and health applications, see e.g. Akyildiz et al. [1, 2] for an extensive overview of actual applications and Alemdar and Ersoy [3] for specific applications in healthcare.

A typical SN includes a sensing subsystem, local data processing capability and a data communication subsystem, all drawing power from an on-board battery [26, 30]. As the lifetime of the sensor network mostly depends on the limited energy budget of its SNs, energy conservation has been a major concern in the design of sensor networks since their inception. Indeed, the replacement of batteries is often expensive if not impossible once the SNs are deployed. According to Anastasi et al. [5], controlling the communication subsystem is key to reducing energy consumption. Ideally, the communication subsystem should be switched off when not needed and waken up again when necessary. This basic idea is applied when operating under dynamic power management (DPM). DPM can be integrated into the medium access control (MAC) protocol or may be implemented independently. A detailed explanation of both approaches with a list of low duty cycle MAC protocols and independent sleep/wakeup protocols are respectively given in Section 4.2 and 4.3 of [5].

Despite vast improvements on power consumption and ongoing developments in power management, the limited energy budget of on-board batteries remains an impediment for long-lasting sensor networks. To mitigate or overcome this dependency on batteries, current research effort focusses on the development of sensors that scavenge the necessary energy from their environment [24, 30]. This alternative technique is called energy harvesting. The specific nature of such wireless sensor networks (EH-WSNs) requires a thorough understanding of the energy harvesting dynamics and its impact on performance. This is the subject of the present paper. We first survey related literature.

Within the control community, Sharma et al. [29] study the optimal energy consumption of an EH-WSN that periodically transmits data. The authors mainly tackle existence questions. In particular, they show the existence of an α-discount optimal and average cost optimal control policy assuming finite energy storage capacity. The same control problem is addressed by Yang and Ukulus [40], albeit in a deterministic setting. That is, the amount of energy harvested and the data arrivals are known in advance. Tutuncuoglu and Yener [34] consider optimal transmission policies for short-term throughput maximisation and for transmission completion time minimisation. Based on the concepts of information theory, Ozel and Ulukus [23] derive optimal power allocation for a maximum average throughput and pro-

vide a geometric interpretation for the resulting power allocation. Rajesh et al. [25] find the Shannon capacity of a sensor node with an energy harvesting source and show that the capacity achieving policies are related to throughput optimal policies. They also obtain the capacity when energy conserving sleep-wakeup modes are supported and an achievable rate for a system with inefficiencies in energy storage. Finally, Zhang and Seyedi [42] derive the overall probability of packet loss in the network due to channel errors or lack of energy in the nodes. Based on this result, a near-optimal design for dimensioning storage and harvesting components of sensors is obtained.

Sensors being autonomous in deciding which information to transmit as well as when to transmit, various authors propose game theoretic models; see e.g. [21] for power control games in wireless networks. Tsuo et al. [33] consider a Bayesian game where each node knows its local energy state. An evolutionary hawk and dove game with harvesting nodes transmitting either at high or low power is studied in [4, 10]. With a focus on solar power, Niyato et al. [22] determined the optimal energy management of sensor nodes adopting a sleep-wakeup strategy by means of a bargaining game.

Other authors propose Markovian models to study EH-WSNs. In particular, Jornet and Akyildiz [14] and Seyedi and Sikdar [27, 28] analyse the battery dynamics of a sensor node as a Markovian model with an energy harvesting buffer. Ventura and Chowdhury [36] propose a similar model for an energy harvesting body sensor network and allow for multiple sensor nodes harvesting from the same energy source. Ho et al. [13] and Lee et al. [18] verify statistically that a Markov modulated arrival process is appropriate for describing solar energy harvesting. Sahu et al. [26] study stochastic stability of an energy harvesting node with data buffering and rely on simulation to assess its performance.

The present work most closely relates to the Markovian models above but explicitly accounts for data buffering. This considerably complicates the analysis as the energy harvesting sensor node now consists of two buffers: a finite-capacity buffer modelling the available energy power and an infinite-capacity buffer which tracks the temporarily stored data. Combining versatility and numerical tractability, the energy harvesting sensor node is modelled as a Markovian queueing model with two paired queues. Pairing refers to the coupling between the queues, data transmission is indeed only possible if both queues are nonempty. These systems have been studied in various contexts including leaky-bucket access control [37, 38], kitting processes [6, 9] and decoupling buffers in production systems [7]. While the Markovian setting at hand allows for computationally efficient performance evaluation of the EH-WSN, it is not limiting in terms of versatility. Indeed, the introduction of a Markovian environment variable allows for time-correlation in both energy harvesting and data collection (cfr. infra).

Independently of the availability of sufficient battery power, we assume that the

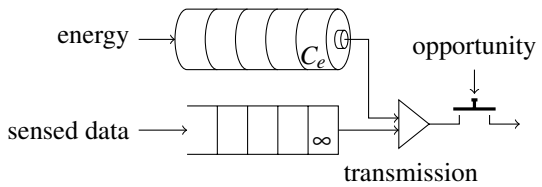


Figure 6.1: Stochastic model of energy harvesting for low power sensor nodes.

EH-WSN cannot send continuously and may require local storage as to await the next transmission opportunity. The introduction of these transmission opportunities is motivated by but not limited to scenarios where a mobile sink is responsible for data collection. A mobile sink moves towards the energy harvesting sensor node and can gather the sensed data only when it is located in the transmission range of the sensor node (and when there is sufficient energy to transmit). As shown by the literature, sink mobility can improve the overall performance of a wireless sensor node network [39, 19, 41, 20, 32]. Turgut and Bölöni [39] worked on the transmission scheduling problem of sensor nodes using mobile sinks. In particular, they develop a graph-theory-based optimal algorithm in order to minimise the energy consumption and the data loss of each node modelled as an autonomous agent. Yun and Xia [41] propose a framework to maximise the network lifetime by using a mobile sink. In this work, data is stored temporarily at the sensor node and is transmitted when the mobile sink is at the most favourable location. Liang et al. [19, 20] incorporate the travel distance of mobile sinks into the problem formulation and proposed heuristics to find a feasible trajectory for each mobile sink so that the network lifetime can be maximised. Ren and Liang [32] formulate an optimisation problem to find an optimal close trajectory for the mobile sink and to schedule the sojourn time at each sojourn location such that the network throughput is maximised.

The remainder of this paper is organised as follows. The EH-WSN model under investigation and the notationally conventions are introduced in the next section. In Section 6.3, the stochastic process at hand is analysed as a quasi-birth-death process (QBD). Also, the numerical solution methodology is discussed and relevant performance measures are determined. To illustrate our approach, Section 6.4 considers various numerical examples. Finally, conclusions are drawn in Section 6.5.

6.2 Model description

Noting that a battery operates very much like a queue — energy chunks being the "customers" in the queue, see [14] — the energy harvesting sensor node is

modelled as a queueing system with two queues as depicted in Figure 6.1. Data transmission is only possible when (i) there is sufficient energy, (ii) there is sensed data available and (iii) there is a transmission opportunity. The stochastic processes that describe data collection and storage, energy harvesting and storage, energy expenditure and transmission opportunities are described below.

Concerning data acquisition, we assume that the sensor picks up data in accordance with a Markovian arrival process with state space $\mathcal{A}$. Let Ω_A^1 and Ω_A^0 denote the generator matrices governing state transitions with and without data packet arrivals respectively. Sensed data is temporarily stored in the data buffer which has infinite capacity.

Remark 1. Here and in the remainder, we assume that generator matrices only collect the transmission rates. Hence, the diagonal elements of the generator matrices of unmarked transmissions like Ω_A^0 are zero. Of course marked transmissions without state change are possible such that the diagonal elements of the generator matrices of the marked transitions like Ω_A^1 may be non-zero.

Analogously, energy harvesting is modelled by a Markovian arrival process with state space $\mathcal{E}$ in accordance with the findings in [13]. Let Ω_E^1 and Ω_E^0 denote the generator matrices governing the state transitions with and without energy arrivals respectively. The energy queue has finite capacity C_e to reflect limitations in energy storage. We however do not subscribe to the energy chunk paradigm where each customer in the energy queue represents a chunk of energy. Instead, we associate queue content with energy levels, the difference being that a packet transmission does not necessarily requires a complete chunk of energy and that an "energy arrival" corresponds to an increase of the energy level. While our modelling assumptions still allow for considering the battery as a storage for energy chunks, dropping the energy chunks in favour for energy levels enables one to describe the dynamics of large batteries with less Markovian states which in turn decreases the numerical complexity of the performance evaluation.

To introduce energy expenditure and transmission, a third marked Markov process is introduced with state space $\mathcal{D}$. This Markov process describes the departures from both energy and data queue and its generator matrices therefore depend on whether or not data and energy is available. When both data buffer content and energy level are non-zero, let Ω_D^e, Ω_D^d, Ω_D^{de} and Ω_D^0 denote the generator matrices governing the transitions when the energy level decreases, when there is a data transmission completed, when the energy level decreases and a data transmission is completed and when there is a state transition with neither an energy drop nor a transmission completion respectively. When there is energy available and no data in the buffer, let $\widehat{\Omega}_D^e$ and $\widehat{\Omega}_D^0$ denote the generator matrices governing the transitions when there is a decrease of energy and when there is no decrease of the energy level. Finally, when there is no energy available, data transmission is also not possible. Hence in these case only non-marked state transitions are possible.

Let $\widetilde{\Omega}_D$ denote the corresponding generator matrix. The following two examples illustrate the versatility of the introduced marked Markov process above.

Example 1. As a first example, consider an exogenous Markov process which neither depends on queue content nor energy level. Let Ω_D be its generator matrix and let its state space D be partitioned into two non-overlapping sets $\mathcal{D}_a$ and $\mathcal{D}_b$. The chain describes the availability of a receiver (like a mobile sink): transmissions occur at a rate μ when the chain is in $\mathcal{D}_a$ (when there is data to send) and there are no transmissions while being in $\mathcal{D}_b$. We further assume that the energy buffer depletes at a rate θ_a during data transmission and at a rate θ_b when there is no transmission. In accordance with [28] and [14], energy is required to communicate with other nodes and to sense, compute and store data. This required amount of energy increases during data transmission such that $\theta_a > \theta_b$. In this case, there are no simultaneous departures from the data and energy queue. Hence, we have,

$$\Omega_D^{de} = 0\,.$$

While there is data and energy, the depletion rates of data and energy queues depend on the state of the exogenous Markov process,

$$\Omega_D^d = \begin{bmatrix} \mu I_a & 0 \\ 0 & 0 \end{bmatrix}, \quad \Omega_D^e = \begin{bmatrix} \theta_a \mathbf{I}_a & 0 \\ 0 & \theta_b \mathbf{I}_b \end{bmatrix},$$

where $\mathbf{I}_a$ and $\mathbf{I}_b$ are identity matrices of size $|\mathcal{D}_a|$ and $|\mathcal{D}_b|$, respectively. In the absence of data there are no data transmissions such that,

$$\widehat{\Omega}_D^e = \theta_b \mathbf{I}_D\,,$$

with $\mathbf{I}_D$ the identity matrix of size $|\mathcal{D}|$. Finally, as the state of the Markov process changes independently of the presence of data and energy, we have,

$$\Omega_D^0 = \widehat{\Omega}_D^0 = \widetilde{\Omega}_D = \Omega_D\,.$$

Example 2. Assuming the energy chunk paradigm, every data transmission requires a single energy chunk from the battery and there is no energy loss when there is no data. Hence, only simultaneous departures from both data and energy queue are possible. This implies $\Omega_D^e = \Omega_D^d = \widehat{\Omega}_D^e = \widehat{\Omega}_D^e = 0$. Adopting the exogenous Markov process with generator matrix Ω_D from the preceding example, state changes of this Markov process do not depend on the presence of data and energy such that,

$$\Omega_D^0 = \widehat{\Omega}_D^0 = \widetilde{\Omega}_D = \Omega_D\,.$$

Again assuming that transmissions occur at a rate μ when the chain is in $\mathcal{D}_a$ and that there are no transmissions while being in $\mathcal{D}_b$, the remaining generator matrix Ω_D^{de} then has the block matrix representation,

$$\Omega_D^{de} = \begin{bmatrix} \mu \mathbf{I}_a & 0 \\ 0 & 0 \end{bmatrix}.$$

6.3 Analysis

We now show that the Markov process at hand is a quasi-birth-death-process (QBD) and derive expressions for a number of performance measures of interest. We first introduce some auxiliary matrices.

6.3.1 Auxiliary matrices

We first describe the transition matrices of the marked Markov process that tracks all state information except the queue content and the energy level. This Markov process has state space $\mathcal{K} = \mathcal{E} \times \mathcal{A} \times \mathcal{D} = \{1, \ldots, K\}$ and its transition matrices depend on the presence of data and energy. Let $\mathbf{I}_E$, $\mathbf{I}_A$ and $\mathbf{I}_D$ denote identity matrices with size $|\mathcal{E}|$, $|\mathcal{A}|$ and $|\mathcal{D}|$, respectively and note that the symbol $\otimes$ denotes the Kronecker's product.

- When both energy level and queue content are non-zero, the unmarked transitions (when there are neither arrivals nor departures) are governed by,

  $$\mathbf{A} = \Omega_E^0 \otimes \mathbf{I}_A \otimes \mathbf{I}_D + \mathbf{I}_E \otimes \Omega_A^0 \otimes \mathbf{I}_D + \mathbf{I}_E \otimes \mathbf{I}_A \otimes \Omega_D^0 \,.$$

  Analogously, when there is energy but no data and when there is neither energy nor data, the unmarked transitions are governed by,

  $$\widehat{\mathbf{A}} = \Omega_E^0 \otimes \mathbf{I}_A \otimes \mathbf{I}_D + \mathbf{I}_E \otimes \Omega_A^0 \otimes \mathbf{I}_D + \mathbf{I}_E \otimes \mathbf{I}_A \otimes \widehat{\Omega}_D^0 \,.$$

  and,

  $$\widetilde{\mathbf{A}} = \Omega_E^0 \otimes \mathbf{I}_A \otimes \mathbf{I}_D + \mathbf{I}_E \otimes \Omega_A^0 \otimes \mathbf{I}_D + \mathbf{I}_E \otimes \mathbf{I}_A \otimes \widetilde{\Omega}_D \,,$$

  respectively.

- The matrix $\mathbf{B}_E$ governs the transitions when there is an arrival in the battery:

  $$\mathbf{B}_E = \Omega_E^1 \otimes \mathbf{I}_A \otimes \mathbf{I}_D \,.$$

- The matrix $\mathbf{B}_A$ governs the transitions when there is an arrival in the data buffer:

  $$\mathbf{B}_A = \mathbf{I}_E \otimes \Omega_A^1 \otimes \mathbf{I}_D \,.$$

- The marked transitions when the energy level drops and/or when there is a transmission again depend on the presence of data and energy. When there is both data and energy, let $\mathbf{C}_D$, $\mathbf{C}_E$ and $\mathbf{C}_{DE}$ denote the generator matrices governing the transitions when there is a transmission, an energy drop or both, respectively:

  $$\mathbf{C}_D = \mathbf{I}_E \otimes \mathbf{I}_A \otimes \Omega_D^d \,,$$

  $$\mathbf{C}_E = \mathbf{I}_E \otimes \mathbf{I}_A \otimes \Omega_D^e \,,$$

$$\mathbf{C}_{DE} = \mathbf{I}_E \otimes \mathbf{I}_A \otimes \Omega_D^{de}.$$

When there is energy but no data, the matrix governing the transitions when the energy level decreases is given by,

$$\widehat{\mathbf{C}}_E = \mathbf{I}_E \otimes \mathbf{I}_A \otimes \widehat{\Omega}_D^{e}.$$

Remark 2. In the remainder, all results will be expressed in terms of the matrices as defined above. These results remain valid when the different arrival processes are intercorrelated as well. In that case there is a single marked Markov process, with marks for data arrivals, energy arrivals and data transmissions.

6.3.2 Quasi-birth-death process

Having defined these transition matrices, we now describe the queueing model at hand as a quasi-birth-death process. Let $Q(t)$ and $C(t)$ be the number of packets and the energy level at time t. Moreover, let $E(t)$, $A(t)$ and $D(t)$ be the state of the energy, data, and transmission process, respectively. The state (in the Markov sense) of the sensor node at time t can then be represented by the vector $[Q(t), C(t), E(t), A(t), D(t)] \in \mathbb{N} \times \mathcal{C} \times \mathcal{K}$. For ease of notation, we further describe the state of the system by the triplet $[n, m, i]$, $n \in \mathbb{N}$ being the number of data packets available, $m \in \mathcal{C}$ being the battery level and $i \in \mathcal{K}$ being the state of the modulating chain. Finally, the energy harvesting sensor node system is assumed to be a continuous-time Markov process with infinite state space $\mathbb{N} \times \mathcal{C} \times \mathcal{K}$.

The studied Markov process is a homogeneous quasi-birth-death process (QBD), see [17]. In the present setting, the *level* or block-row index, indicates the data packets available while the phase, i.e. the index within a block element, indicates both the battery level and the state of the modulating chain. The one-step transitions are restricted to states in the same level (from state $[n,*,*]$ to state $[n,*,*]$) or in two adjacent levels (from state $[n,*,*]$ to state $[n+1,*,*]$ or state $[n-1,*,*]$).

We then find the generator matrix of the Markov process with the following block matrix representation,

$$\mathbf{Q} = \begin{bmatrix} \mathcal{B}_0 & \mathcal{A}_2 & 0 & 0 & \cdots \\ \mathcal{A}_0 & \mathcal{A}_1 & \mathcal{A}_2 & 0 & \cdots \\ 0 & \mathcal{A}_0 & \mathcal{A}_1 & \mathcal{A}_2 & \cdots \\ 0 & 0 & \mathcal{A}_0 & \mathcal{A}_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}. \tag{6.1}$$

The blocks are given by,

$$\mathcal{B}_0 = \begin{bmatrix} \widetilde{\mathbf{A}} & \mathbf{B}_E & 0 & \cdots & 0 \\ \widehat{\mathbf{C}}_E & \widehat{\mathbf{A}} & \mathbf{B}_E & \cdots & 0 \\ 0 & \widehat{\mathbf{C}}_E & \widehat{\mathbf{A}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \widehat{\mathbf{A}} \end{bmatrix}, \tag{6.2}$$

$$\mathcal{A}_0 = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ \mathbf{C}_{DE} & \mathbf{C}_D & \cdots & 0 & 0 \\ 0 & \mathbf{C}_{DE} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \mathbf{C}_{DE} & \mathbf{C}_D \end{bmatrix}, \tag{6.3}$$

$$\mathcal{A}_1 = \begin{bmatrix} \widetilde{\mathbf{A}} & \mathbf{B}_E & 0 & \cdots & 0 \\ \mathbf{C}_E & \mathbf{A} & \mathbf{B}_E & \cdots & 0 \\ 0 & \mathbf{C}_E & \mathbf{A} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \mathbf{A} \end{bmatrix}, \tag{6.4}$$

$$\mathcal{A}_2 = \begin{bmatrix} \mathbf{B}_A & 0 & 0 & \cdots & 0 \\ 0 & \mathbf{B}_A & 0 & \cdots & 0 \\ 0 & 0 & \mathbf{B}_A & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \mathbf{B}_A \end{bmatrix}. \tag{6.5}$$

Having defined the different blocks of the QBD process, we now focus on the solution method. Recall that the state of the Markov process is described by the triplet $[n,m,i]$; n is the size of the data buffer, m is the size of the battery and i is the state of the modulating chain. Let $\pi(n,m,i)$ be the steady state probability to be in state $[n,m,i]$ and let $\boldsymbol{\pi}$ be the vector with elements $\pi(n,m,i)$. The vector $\boldsymbol{\pi}$ satisfies the balance equations,

$$\boldsymbol{\pi}(\mathbf{Q} - \partial\mathbf{Q}) = 0\,.$$

Here the notation $\partial\mathcal{X}$ represents a diagonal matrix with diagonal elements equal to the row sums of $\mathcal{X}$. A well-known method for finding the stationary distribution of QBD processes is the matrix-geometric method. Using the vector notation $\boldsymbol{\pi}_k = (\pi(k,0,1),\ldots,\pi(k,C_e,K))$, the probability vectors can be expressed as,

$$\boldsymbol{\pi}_k = \boldsymbol{\pi}_0\mathbf{R}^k. \tag{6.6}$$

where the so-called rate matrix $\mathbf{R}$ is the minimal non-negative solution of the non-linear matrix equation

$$\mathbf{R}^2\mathcal{A}_0 + \mathbf{R}\bar{\mathcal{A}}_1 + \mathcal{A}_2 = \mathbf{0}\,,$$

with $\bar{\mathcal{A}}_1 = \mathcal{A}_1 - \partial\mathcal{A}_0 - \partial\mathcal{A}_1 - \partial\mathcal{A}_2$. We compute the rate matrix by implementing the improved iterative algorithm of [17, chapter 8, p.179-187]. Once the rate matrix is found, the remaining unknown vector $\boldsymbol{\pi}_o$ is the unique solution of,

$$\boldsymbol{\pi}_0(\mathbf{I}-\mathbf{R})^{-1}\mathbf{1} = 1\,, \quad \boldsymbol{\pi}_0(\bar{\mathcal{B}}_0 + \mathbf{R}\mathcal{A}_2) = 0\,,$$

with $\bar{\mathcal{B}}_0 = \mathcal{B}_0 - \partial\mathcal{A}_2$ and where $\mathbf{I}$ is the identity matrix of size $|\mathcal{K}|(C_e+1)$ and $\mathbf{1}$ is a column vector of ones.

6.3.3 Performance measures

Once the steady state probabilities have been determined numerically, we can calculate a number of interesting performance measures for the harvesting energy sensor node. For ease of notation, we introduce the marginal probability mass functions of the battery level $\pi^{(e)}(m)$ and of the data buffer content $\pi^{(d)}(n)$,

$$\pi^{(e)}(m) = \sum_{i\in\mathcal{K}}\sum_{n=0}^{\infty}\pi(n,m,i)\,,$$

$$\pi^{(d)}(n) = \sum_{i\in\mathcal{K}}\sum_{m=0}^{C_e}\pi(n,m,i)\,.$$

Following performance measures can be computed from the analysis.

- The mean data buffer content $\mathrm{E}\,Q_d$:

$$\mathrm{E}\,Q_d = \sum_{n=1}^{\infty}\pi^{(d)}(n)n\,.$$

- Expressed in percentage, the mean battery level b_e is the average amount of energy in the battery relative to its maximum capacity:

$$b_e = \frac{\mathrm{E}\,Q_e}{C_e}$$

  where

$$\mathrm{E}\,Q_e = \sum_{m=1}^{C_e}\pi^{(e)}(m)m\,.$$

- The variance of the battery level and the data buffer content: $\mathrm{Var}\,Q_e$ and $\mathrm{Var}\,Q_d$ respectively,

$$\mathrm{Var}\,Q_e = \sum_{m=1}^{C_e}\pi^{(e)}(m)m^2 - (\mathrm{E}\,Q_e)^2\,,$$

Battery capacity	C_e	20
Data packet arrivals		
Poisson	λ_d	0.01
Energy reception		
Poisson	λ_e	0.5
IPP	σ_e κ_e λ_e	0.5 10 0.5
Type of transmission		
On-off process	κ_t σ_t	10 0.1
Data transmission		
Exponential (exp)	μ	1.0
Energy expenditure		
Idle Sending	θ_b θ_a	0.5 0.7

Table 6.1: Parameter values of the studied energy harvesting sensor nodes.

$$\operatorname{Var} Q_d = \sum_{n=1}^{\infty} \pi^{(d)}(n) n^2 - (\mathrm{E}\, Q_d)^2 .$$

- The mean data delay LT (calculated based on Little's theorem) is the average amount of time between the arrival of a data packet and its transmission:

$$\mathrm{LT} = \frac{\mathrm{E}\, Q_d}{\lambda} .$$

Here λ is the arrival rate. The latter can be determined from the following expressions,

$$\lambda = \boldsymbol{\tau} \Omega_A^1 \mathbf{1} ,$$

with $\boldsymbol{\tau}$ the unique normalised solution of,

$$\boldsymbol{\tau}(\Omega_A^0 + \Omega_A^1 - \partial\Omega_A^0 - \partial\Omega_A^1) = 0 .$$

6.4 Numerical results

Having established the modelling assumptions and the numerical analysis, we now evaluate the performance of an energy harvesting sensor node that is randomly visited by a mobile sink. Assuming no simultaneous departures from the data and energy queue during data transmission, we adopt the assumptions of Example 1 of Section 6.2. For further reference, the parameter values used throughout this section are displayed in Table 6.1. These parameter values are chosen only by way of illustration and are not to be considered as limiting.

We first assess the impact of irregularity in the energy harvesting process, by comparing Poisson arrivals and interrupted Poisson arrivals of energy. The interrupted Poisson process considered here is a two-state Markov process. In the active state, energy arrives in accordance with a Poisson process with rate λ_e^* whereas no new energy arrives in the inactive state. Let α_e and β_e denote the rate from the active to the inactive state and vice versa, respectively. For convenience, we use the more intuitive parametrisation $(\sigma_e, \kappa_e, \lambda_e)$, with

$$\sigma_e = \frac{\beta_e}{\alpha_e + \beta_e}, \quad \kappa_e = \frac{1}{\alpha_e} + \frac{1}{\beta_e}, \quad \lambda_e = \lambda_e^* \sigma_e .$$

where σ_e is the fraction of time in which the interrupted Poisson process is active, the absolute time parameter κ_e is the average duration of an active and an inactive period, and λ_e is the average arrival rate of harvested energy.

Figure 6.2 depicts the mean battery level b_e (recall that b_e is expressed as a percentage of the total battery capacity) and the mean data delay LT versus the average arrival rate of harvested energy λ_e. Here, for fair comparison, λ_e is the arrival rate for the Poisson process and the average arrival rate for the IPP as defined above. For the IPP, we set $\sigma_e = 0.5$ and $\kappa_e = 10$. Data transmission times are exponentially distributed with service rate $\mu = 1$ and the battery level depletes at a rate $\theta_a = 0.7$ during data transmission and at a rate $\theta_b = 0.5$ when there is no transmission. The availability of the mobile sink is captured by a two-state on-off process. Let α_t and β_t be the rates from on to off and from off to on respectively. Again, the alternative characterisation (σ_t, κ_t) is used, with

$$\sigma_t = \frac{\beta_t}{\alpha_t + \beta_t}, \quad \kappa_t = \frac{1}{\alpha_t} + \frac{1}{\beta_t} .$$

The fraction of time in which the mobile sink is available to receive the data of the sensor node equals $\sigma_t = 10\%$ and the average duration of an available and non-available period equals $\kappa_t = 10$. Finally, we assume that data packets arrive according to a Poisson process with rate $\lambda_d = 0.01$. Figure 6.2 shows that the mean battery level increases and the mean data delay decreases, when the energy harvesting rate increases, as expected. The effect of correlation in the energy harvesting process is less trivial. For low λ_e, we see that the mean battery level is higher when there is correlation. This is easily explained by the fact that we have longer periods where the battery level increases, followed by longer periods where it decreases. When λ_e is high, we notice the opposite effect. Here the finite capacity of the battery comes into play. The battery cannot gain from longer periods of energy harvesting as for higher λ_e the battery fills quickly and excess energy is lost while the battery drains during longer periods without harvesting. For all λ_e, we see that correlation negatively affects the mean waiting time. Correlation in the harvesting process leads to longer periods with and without energy, such that transmissions are more often postponed due to a lack of energy.

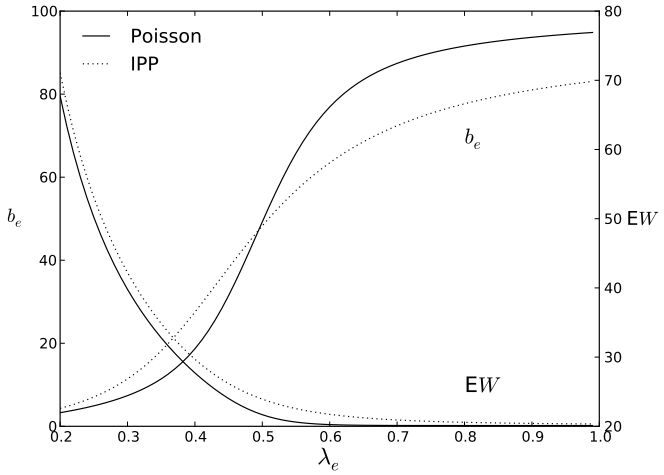


Figure 6.2: The mean battery level and the mean data delay with Poisson and interrupted Poisson energy arrivals.

Figure 6.3 depicts the probability mass functions of the battery level b_e with capacity $C_e = 40$ and for Poisson and interrupted Poisson arrivals with λ_e equal to 0.2, 0.5 and 1.0. The other parameter values are given in Table 6.1. For $\lambda_e = 1.0$, the inventory is filled up considerably faster than it is depleted for both Poisson and IPP harvesting. The battery is hardly ever empty in this case; the probability mass function is concentrated on high battery levels. Note that this situation is obviously the most favourable for the performance of the energy harvesting sensor node. Note also that, as the queue is in overload, the assumption of a limited battery capacity is of main importance as it avoids the model to degrade to an unstable stochastic system [16]. In contrast, when $\lambda_e = 0.2$, the battery level is most often empty or of limited size. Reaching the battery capacity is a rare event in this case. Finally, for $\lambda_e = 0.5$, we have a more or less equal probability to be in one of the different battery levels. Comparing Poisson and IPP harvesting, we see that for $\lambda_e = 1.0$, the probability mass for IPP is less concentrated than for Poisson which confirms the harvesting loss noted in the preceding figure. Although less out-spoken, we see the same for $\lambda_e = 0.2$ which is explained by longer periods of harvesting during which higher energy-levels can be reached.

To further assess the impact of correlation, Figures 6.4 and 6.5 depict the mean battery level and mean data delay for λ_e equal to 0.3, 0.5 and 1.0 versus $\log(\kappa_e)$. We refer to table 6.1 for the remaining parameter values. As Figure 6.4 shows, depending on the energy harvesting rate, the mean battery level decreases, is about

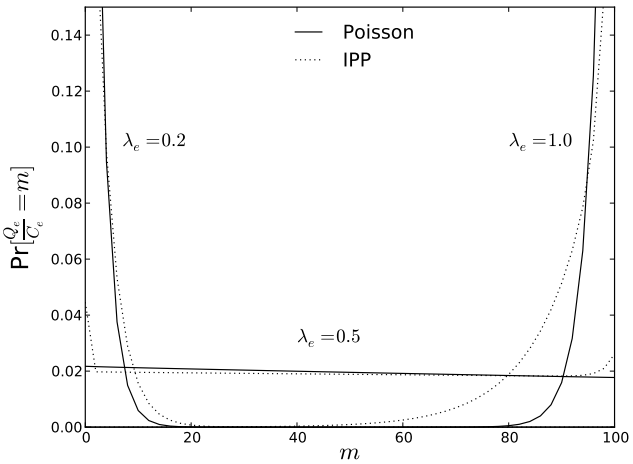


Figure 6.3: Probability mass functions of the battery level for λ_e *equal to* 0.2, 0.5 *and* 1.0.

constant or increases as the absolute time parameter κ_e increases. This confirms and complements the observations of Figure 6.2. When λ_e equals 1.0, the longer the average period in which data is harvested, the higher the probability that energy arrivals cannot be stored in the battery as the maximum capacity is already attained. Hence, an increase in κ_e decreases the mean battery level. In contrast, when λ_e equals 0.3, the battery is hardly ever full such that the longer the average period in which harvested energy does and does not arrive, the larger the mean battery level. Lastly, when λ_e equals 0.5, κ_e has no significant influence on the mean battery level in the considered set of parameter values. Notice that this value is equal to the mean depletion rate in absence of data. In Figure 6.5, the mean data delay increases as the absolute time parameter κ_e increases. Indeed, a larger κ_e induces longer periods without energy, such that the average number of data packets waiting in the buffer and their waiting times increases.

In Figure 6.6, we depict the mean data delay and the mean battery level with interrupted Poisson energy arrivals versus the data arrival rate λ_d. We assume transmission rates μ equal to 0.8, 1.0 and 1.2. The other parameter values are given in Table 6.1. As expected, the mean data delay increases and the mean battery level decreases as the data arrival rate increases. Also, both effects are enhanced when the data transmission rate decreases and the data arrival rate increases.

Finally, we study the impact of the absolute time parameter of the mobile sink availability process κ_t on the performance of an energy harvesting sensor node. Figure 6.7 and 6.8 depict the mean battery level and the mean data delay for σ_t

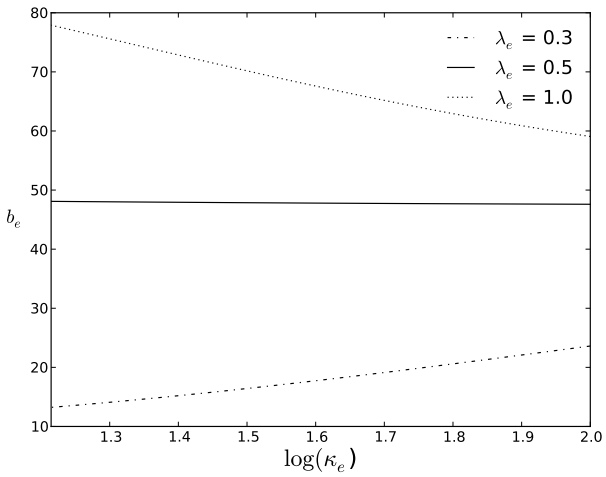


Figure 6.4: The mean battery level for different energy harvesting rates.

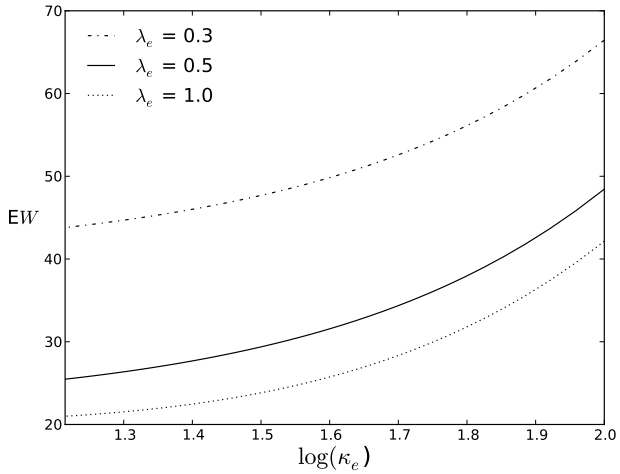


Figure 6.5: The mean data delay for different energy harvesting rates.

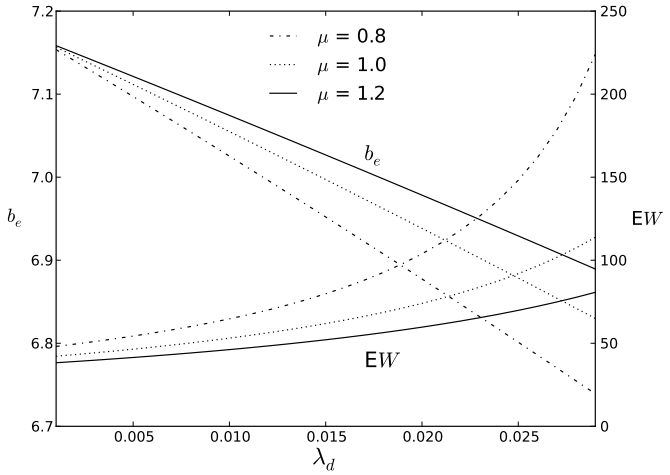


Figure 6.6: The mean data delay and the mean battery level for different transmission rates.

equal to 0.025, 0.05 and 0.1 versus κ_t. The remaining parameter values are given in Table 6.1. As the figures show, the larger the fraction of time in which the mobile sink is available, the more data packets can be transmitted, the more energy is depleted. Hence, the mean battery level increases and the mean data delay decreases as σ_t increases. In Figure 6.7, we further observe that the values for κ_t and σ_t have but a small impact on the mean battery level. We observe a decrease followed by an increase of mean battery level as κ_e increases. This contrasts with Figure 6.8, where a significant increase of the mean waiting time is observed for increasing κ_t. Indeed, the longer the periods in which the mobile sink is not available to transmit, the more data packets wait in the buffer on average.

6.5 Conclusion

In this paper, we analysed the performance of an energy harvesting sensor node under uncertainty in energy harvesting, energy depletion, data acquisition and data transmission. To this end, energy harvesting sensor nodes are described as stochastic models with two queues: the battery and the data packet backlog. Independently of the battery level, the sensor node can transmit sensed data to the receiver only when it is located in the transmission range of the sensor node. The introduction of these limited transmission opportunities is motivated by but not limited to scenarios where a mobile sink is responsible for data collection. Hence, data transmission is only possible when there is sufficient energy, a data packet avail-

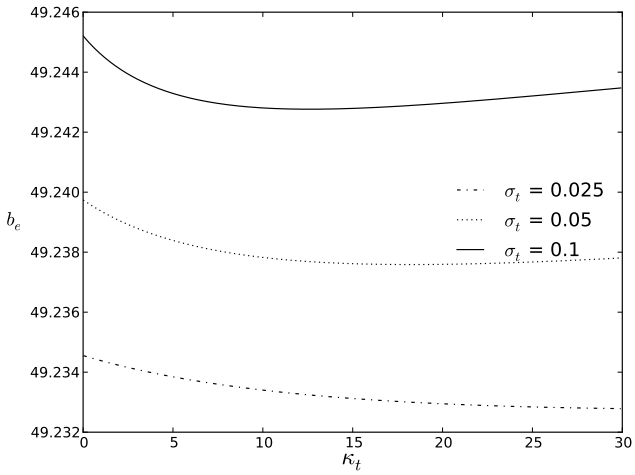


Figure 6.7: The mean battery level for different fractions of time that the mobile sink is available.

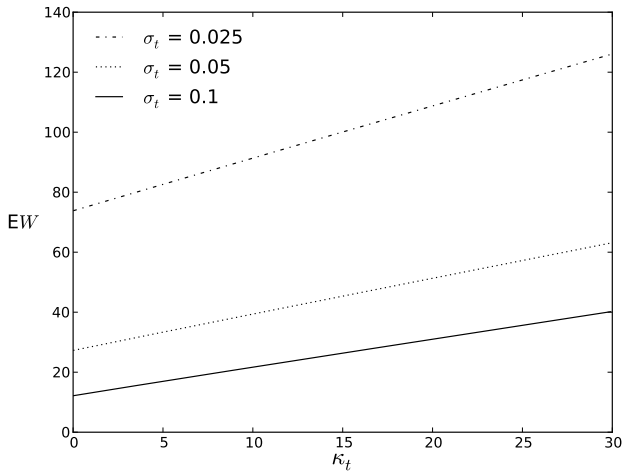


Figure 6.8: The mean data delay for different fractions of time that the mobile sink is available.

able and a transmission opportunity. Methodologically, the developed queueing system is analysed as a homogeneous quasi-birth-death process (QBD) and solved with matrix-analytic methods. By means of numerical examples, we evaluated the impact of different parameters on the performance of an energy harvesting sensor node. Correlation in the energy harvesting process decreases the performance of the energy harvesting sensor node: data packets wait longer on average. Also, if the energy harvesting rate is high, correlation induces long periods with more energy arrivals than can be stored in the battery with finite capacity. Hence, the mean battery level decreases.

References

[1] I.F. Akyildiz, W. Su, Y. Sankarasubramaniam and E. Cayirci. *Wireless sensor networks: a survey*, Computer Networks, 38(4), p.393–422, 2002.

[2] I.F. Akyildiz and M.C. Vuran. *Wireless sensor networks*, Wiley, 2010.

[3] H. Alemdar and C. Ersoy. *Wireless sensor networks for healthcare: A survey*, Computer Networks, 54(15), p.2688–2710, 2010.

[4] E. Altman, D. Fiems, M. Haddad and J. Gaillard. *Semi-Dynamic Hawk and Dove Game Applied to Power Control*, Proceedings of INFOCOM 2012.

[5] G. Anastasi, M. Conti, M. Di Francesco and A. Passarella. *Energy conservation in wireless sensor networks: a survey*, Ad Hoc Networks, 7, p.537–568, 2009.

[6] E. De Cuypere, K. De Turck and D. Fiems. *Performance analysis of a kitting process as a paired queue*, Hindawi Publishing Corporation Mathematical Problems in Engineering, 2013, vol. 2013, Article ID 843184, 10 pages, http://dx.doi.org/10.1155/2013/843184.

[7] E. De Cuypere, K. De Turck and D. Fiems. *Performance analysis of a decoupling stock in a Make-to-Order system*, Proceedings of the 14th IFAC Symposium on Information Control Problems in Manufacturing, 2012.

[8] E. De Cuypere, K. De Turck and D. Fiems. *Stochastic Modelling of Energy Harvesting for Low Power Sensor Nodes*, Proceedings of QTNA, Kyoto, Japan, 2012.

[9] E. De Cuypere and D. Fiems. *Performance evaluation of a kitting process*, Proceedings of the 17th International Conference on analytical and stochastic modelling techniques and applications, Lecture Notes in Computer Science, 6751, 2011.

[10] M. Haddad, E. Altman, J. Gaillard and D. Fiems. *A Semi-Dynamic Evolutionary Power Control Game*, Proceedings of Networking 2012, Lecture Notes in Computer Science, 2012.

[11] J.M. Gilbert and F. Balouchi. *Comparison of Energy Harvesting Systems for Wireless Sensor Networks*, International Journal of Automation and Computing, 5, p.334, 2008.

[12] J.M. Harrison. *Assembly-Like Queues*, Journal of Applied Probability, 10, p.354–367, 1973.

[13] C.K. Ho, P.D. Khoa and P.C. Ming. *Markovian models for harvested energy in wireless communications*, Proceedings of ICSS, Amsterdam, 2010.

[14] J.S. Jornet and I.F. Akyildiz. *Joint Energy Harvesting and Communication Analysis for Perpetual Wireless Nanosensor Networks in the Terahertz Band*, IEEE Transactions on nanotechnology, 11, p.570–580, 2012.

[15] H.S. Kim, J.-H. Kim and J. Kim. *A Review of Piezoelectric Energy Harvesting Based on Vibration*, International Journal of Precision Engineering and Manufacturing, 12, p.1129–1141, 2012.

[16] G. Latouche. *Queues with paired customers*, Journal of Applied Probability, 18, p.684–696, 1981.

[17] G. Latouche and V. Ramaswami. *Introduction to Matrix Analytic Methods in Stochastic Modeling*, SIAM, 1999.

[18] P. Lee, Z.A. Eu, M. Han and H. Tan. *Empirical modeling of a solar powered energy harvesting wireless sensor node for time-slotted operation*, Proceedings of WCNC, p.179–184, Cancun, 2011.

[19] W. Liang, J. Luo and X. Xu. *Prolonging network lifetime via a controlled mobile sink in wireless sensor networks*, Proceedings of GLOBECOM, IEEE, 2010.

[20] W. Liang and J. Luo. *Network lifetime maximization in sensor networks with multiple mobile sinks*, Proceedings of LCN, IEEE, 2011.

[21] F. Meshkati, H.V. Poor and S.C. Schwartz. *Energy-Efficient Resource Allocation in Wireless Networks*, Signal Processing Magazine, IEEE, 24, p.58–68, 2007.

[22] D. Niyato, M.M. Rashid and V.K. Bhargave. *Wireless sensor networks with energy harvesting technologies: A game-theoretic approach to optimal energy management*, IEEE Wireless Communications, 14, p.90–96, 2007.

[23] O. Ozel and S. Ulukus. *Information-theoretic analysis of an energy harvesting communication system*, IEEE 21st International Symposium on Personal, Indoor and Mobile Radio Communications Workshops, p.330–335, Istanbul, Turkey, 26-30 Sept. 2010.

[24] J.A. Paradiso and T. Starner. *Energy scavenging for mobile and wireless electronics*, IEEE Pervasive Computing, 4, p.18–27, 2005.

[25] R. Rajesh, V. Sharma and P. Viswanath. *Information capacity of Energy harvesting*, Proceedings of the 2011 IEEE International Symposium on Information Theory, p.2363–2367, St. Petersburg, Russia, 31 July-5 August 2011.

[26] A. Sahu, E.B. Fernandez, M. Cardei and M. Vanhilst. *A pattern for a sensor node*, Proceedings of the 17th conference on Pattern Languages of Programs, 2010.

[27] A. Seyedi and B. Sikdar. *Performance modelling of transmission schedulers capable of energy harvesting*, Proceedings of ICC, Cape Town, 2010.

[28] A. Seyedi and B. Sikdar. *Energy efficient transmission strategies for body sensor networks with energy harvesting*, IEEE Transactions on Communications, 58(7), p.2116–2126, 2010.

[29] V. Sharma, U. Mukherji and V. Joseph. *Optimal energy management policies for energy harvesting sensor nodes*, IEEE Transactions on Wireless Communications, 6, p.1326–1336, 2010.

[30] S. Sudevalayam and P. Kulkarni. *Energy Harvesting Sensor Nodes: Survey and Implications*, Communications Surveys Tutorials, IEEE, 13, p.443–461, 2011.

[31] V. Raghunathan, S. Ganeriwal and M. Srivastava, *Emerging techniques for long lived wireless sensor networks*, IEEE Communication Magazine, p.108–114, April 2006.

[32] X. Ren and W. Liang. *Delay-Tolerant Data Gathering in Energy Harvesting Sensor Networks With a Mobile Sink*, Proceedings of GLOBECOM, IEEE, 2012.

[33] F.Y. Tsuo, H.P. Tan, Y.H. Chew and H.Y. Wei. *Energy-Aware Transmission Control for Wireless Sensor Networks Powered by Ambient Energy Harvesting: A Game-Theoretic Approach*, IEEE International Conference on Communications, 2011.

[34] K. Tutuncuoglu and A. Yener, *Optimum Transmission Policies for Battery Limited Energy Harvesting Nodes*, IEEE Transactions on Wireless Communications, 11, p.1180–1189, 2012.

[35] Y.K. Tan and S.K. Panda. *Review of Energy Harvesting Technologies for Sustainable Wireless Sensor Network*, Sustainable Wireless Sensor Networks, INTECH Publisher, p.15–43, 2010.

[36] J. Ventura and K. Chowdhury. *Markov modelling of energy harvesting body sensor networks*, Proceedings of PIMRC, p.2168–2172, Toronto, 2011.

[37] S. Wittevrongel and H. Bruneel. *A heuristic analytic technique to calculate the cell loss ration in a leaky bucket with bursty input traffic*, AEU - International Journal of Electronics and Communications, 3, p.162–169, 1994.

[38] S. Wittevrongel and H. Bruneel. *Analytic study of the queueing performance and the departure process of a leaky bucket with bursty input traffic*, AEU - International Journal of Electronics and Communications, 1, p.1–10, 1996.

[39] D. Turgut and L. Bölöni. *Heuristic approaches for transmission scheduling in sensor networks with multiple mobile sinks*, The Computer Journal, 54(3), p.332–344, 2009.

[40] J. Yang and S. Ulukus, *Optimal Packet scheduling in an Energy Harvesting Communication System*, IEEE Transactions on Communications, 60(1), p.220–230, 2012.

[41] Y. Yun and Y. Xia, *Maximizing the lifetime of wireless sensor networks with mobile sink in delay-tolerant applications*, IEEE Trans. Mobile Computing, 9, p.1308–1318, 2010.

[42] S. Zhang and A. Seyedi. *Analysis and Design of Energy Harvesting Wireless Sensor Networks with Linear Topology*, Proceedings of ICC 2011, Kyoto, Japan, 5-9 June 2011.

7

A Maclaurin-series expansion approach to multiple paired queues

Eline De Cuypere, Koen De Turck and Dieter Fiems

published in Operations Research Letters, 2014, vol.42, no.3, p.203-207.

Abstract. Motivated by kitting processes in assembly systems, we consider a Markovian queueing system with K paired finite-capacity buffers. Pairing means that departures from the buffers are synchronised and that service is interrupted if any of the buffers is empty. To cope with the inherent state-space explosion problem, we propose an approximate numerical algorithm which calculates the first N coefficients of the Maclaurin-series expansion of the steady-state probability vector in $O(KNM)$ operations, M being the size of the state space.

7.1 Introduction

We consider a system of K queues, each queue having finite capacity. Let C_i denote the capacity of the ith queue. Moreover, for each of the queues, customers arrive in accordance with an independent Poisson process, let $\lambda_i > 0$ denote the arrival rate in queue i. Departures from the different queues are paired which means that there are simultaneous departures from all queues with rate μ as long as all queues are non-empty. If one of the queues is empty, there are no departures.

The queueing system at hand is motivated by kitting processes in assembly

systems. A kitting process collects the necessary parts for a given end product in a container prior to assembly. While conceptually simple, kitting comes with many advantages. Kitting clearly mitigates storage space requirements at the assembly station since no part inventories need to be kept there. Moreover, parts are placed in proper positions in the container such that assembly time reductions can be realised [3, 9]. A kitting process is obviously related to a paired queueing system: the inventories of the different parts that go into the kit correspond to the different buffers, the kitting time corresponds to the service time and kitting is blocked if one or more parts are missing.

Paired queueing systems have been studied by various authors. Harrison [4] studies stability of paired queueing under very general assumptions: $K \geq 2$ infinite-capacity buffers, generally distributed interarrival times at the different buffers and generally distributed service times. He shows that it is necessary to impose a restriction on the size of the buffer to ensure stability of the queueing system. In particular, the distribution of the vector of waiting times (in the different queues) of the components of a paired customer is shown to be defective. The inherent instability was also demonstrated in [8] where the excess — the difference between the queue sizes — is studied in the two-queue case. Assuming finite capacity buffers, Hopp and Simon developed a model for a two-buffer kitting process with exponentially distributed processing times for kits and Poisson arrivals [5]. The exponential service times and Poisson arrival assumptions were later relaxed in [12] and [2], respectively. For paired queueing systems with more than two finite buffers, the size of the state-space of the associated Markov process grows quickly, even for the case of Poisson arrivals and exponential service times. Hence, most authors focus on approximations; a recent account on approximations of multi-buffer paired queueing systems can be found in [10]. Also the present letter investigates approximations for multi-buffer paired queueing systems. In particular, we propose a numerical evaluation method for Markovian paired queueing systems which relies on a Maclaurin-series expansion of the steady-state probability vector. For an overview on the technique of series expansions in stochastic systems, which is known under the names light traffic analysis or stochastic perturbation, we refer the reader to the surveys in [1, 7]. Finally, we note that the paired queueing system somewhat resembles a fork-join queueing system; see e.g. [6] and the references therein. However, in fork-join queueing systems both arrivals and departures in the different buffers are synchronised, which leads to entirely different dynamics.

7.2 Maclaurin-series expansion

As arrivals in the different queues are modelled by Poisson processes and the service time distribution is exponential, the state of the system is described by a vector $\mathbf{i} \in \mathcal{C}$ whose kth element corresponds to the queue size of the kth buffer. Here $\mathcal{C} = \mathcal{C}_1 \times \ldots \times \mathcal{C}_K$ denotes the state space of this continuous-time Markov process (CTMC), with $\mathcal{C}_k = \{0, 1, \ldots, C_k\}$ being the set of possible levels of queue k. Let $\pi(\mathbf{i})$ be the steady-state probability of state $\mathbf{i}$, $\mathbf{i} \in \mathcal{C}$. These steady-state probabilities satisfy the following set of balance equations,

$$\begin{aligned}\pi(i_1, i_2, \ldots, i_K) \left(\mu \prod_{\ell=1}^{K} \mathbf{1}_{\{i_\ell > 0\}} + \sum_{\ell=1}^{K} \mathbf{1}_{\{i_\ell < C_\ell\}} \lambda_\ell \right) = \\ \pi(i_1+1, i_2+1, \ldots, i_K+1) \mu \prod_{\ell=1}^{K} \mathbf{1}_{\{i_\ell < C_\ell\}} \\ + \sum_{\ell=1}^{K} \pi(i_1, \ldots, i_{\ell-1}, i_\ell - 1, i_{\ell+1}, \ldots, i_K) \lambda_\ell \mathbf{1}_{\{i_\ell > 0\}}, \quad (7.1)\end{aligned}$$

for all $\mathbf{i} = (i_1, i_2, \ldots, i_K) \in \mathcal{C}$ and where $\mathbf{1}_{\{x\}}$ is the indicator function which equals one if x is true and equals zero otherwise. While the former system of equations is easily solved if there are only a few queues with low capacity, the size of the state space explodes for even a moderate number of queues and reasonable queue capacities and a direct solution is computationally infeasible.

To mitigate this state space explosion problem, we rely on a Maclaurin-series expansion in μ. It is shown in the appendix that $\pi(\mathbf{i})$ is analytic in $\mu = 0$ and therefore admits the representation,

$$\pi(\mathbf{i}) = \sum_{n=0}^{\infty} \pi_n(\mathbf{i}) \mu^n,$$

for $0 \leq \mu < \mu_0$ and for $\mathbf{i} \in \mathcal{C}$. Here μ_0 is a non-negative value for which a lower bound is provided in the appendix.

Substituting the former expression in the balance equations yields,

$$\begin{aligned}\sum_{n=0}^{\infty} \pi_n(i_1, i_2, \ldots, i_K) \mu^n \left(\mu \prod_{\ell=1}^{K} \mathbf{1}_{\{i_\ell > 0\}} + \sum_{\ell=1}^{K} \mathbf{1}_{\{i_\ell < C_\ell\}} \lambda_\ell \right) = \\ \sum_{n=0}^{\infty} \pi_n(i_1+1, i_2+1, \ldots, i_K+1) \mu^{n+1} \prod_{\ell=1}^{K} \mathbf{1}_{\{i_\ell < C_\ell\}} \\ + \sum_{n=0}^{\infty} \sum_{\ell=1}^{K} \pi_n(i_1, \ldots, i_{\ell-1}, i_\ell - 1, i_{\ell+1}, \ldots, i_K) \lambda_\ell \mu^n \mathbf{1}_{\{i_\ell > 0\}}. \quad (7.2)\end{aligned}$$

For $\mathbf{i} \in \mathcal{C}^* = \mathcal{C} \setminus \{[C_1, C_2, \ldots, C_K]\}$, comparing the terms in μ^0 on both sides of the former equation yields,

$$\pi_0(i_1, i_2, \ldots, i_K) = 0\,, \tag{7.3}$$

whereas comparing the terms in μ^n for $n > 0$ gives,

$$\begin{aligned}\pi_n(i_1, i_2, \ldots, i_K) = \frac{1}{\sum_{\ell=1}^{K} \mathbf{1}_{\{i_\ell < C_\ell\}} \lambda_\ell} \times \\ \Bigg(\mathbf{1}_{\{n>0\}} \pi_{n-1}(i_1+1, i_2+1, \ldots, i_K+1) \prod_{\ell=1}^{K} \mathbf{1}_{\{i_\ell < C_\ell\}} \\ + \sum_{\ell=1}^{K} \pi_n(i_1, \ldots, i_{\ell-1}, i_\ell - 1, i_{\ell+1}, \ldots, i_K) \lambda_\ell \mathbf{1}_{\{i_\ell > 0\}} \\ - \mathbf{1}_{\{n>0\}} \pi_{n-1}(i_1, i_2, \ldots, i_K) \prod_{\ell=1}^{K} \mathbf{1}_{\{i_\ell > 0\}} \Bigg)\,. \end{aligned} \tag{7.4}$$

For $\mathbf{i} = \mathbf{c} \doteq [C_1, C_2, \ldots, C_K]$, such a comparison does not yield an expression for $\pi_n(\mathbf{i})$. To determine the remaining unknown, we invoke the normalisation condition:

$$\sum_{\mathbf{i} \in \mathcal{C}} \pi_0(\mathbf{i}) = 1\,, \quad \sum_{\mathbf{i} \in \mathcal{C}} \pi_n(\mathbf{i}) = 0\,.$$

Solving for $\pi_n(\mathbf{c})$ then yields,

$$\pi_0(\mathbf{c}) = 1\,, \quad \pi_n(\mathbf{c}) = -\sum_{\mathbf{i} \in \mathcal{C}^*} \pi_n(\mathbf{i})$$

Once the series expansions of the steady state distribution has been obtained, the expansion of various performance measures directly follows. Let $\mathbf{X} \sim \pi$, then for a performance measure $J = \mathrm{E}[f(\mathbf{X})]$ we have,

$$J = \sum_{\mathbf{i} \in \mathcal{C}} f(\mathbf{i}) \pi(\mathbf{i}) = \sum_{\mathbf{i} \in \mathcal{C}} f(\mathbf{i}) \sum_{n=0}^{\infty} \pi_n(\mathbf{i}) \mu^n = \sum_{n=0}^{\infty} \sum_{\mathbf{i} \in \mathcal{C}} f(\mathbf{i}) \pi_n(\mathbf{i}) \mu^n = \sum_{n=0}^{\infty} J_n \mu^n\,, \tag{7.5}$$

for $0 \leq \mu < \mu_0$ with,

$$J_n = \sum_{\mathbf{i} \in \mathcal{C}} f(\mathbf{i}) \pi_n(\mathbf{i})\,.$$

The interchange of the summations is justified by the finiteness of $\mathcal{C}$ and the convergence of $\sum_n \pi_n(\mathbf{i}) \mu^n$ for all $\mathbf{i} \in \mathcal{C}$. As such, any term J_n in the expansion of a performance measure J can be calculated from the corresponding vector π_n of the expansion of the steady-state vector. Performance measures of interest include amongst others the ℓth order moment of the queue content of the kth queue ($f(\mathbf{i}) = i_k^\ell$), the blocking probability ($f(\mathbf{i}) = 1 - \prod_{j=1}^{K} \mathbf{1}_{\{i_j > 0\}}$) and the throughput ($f(\mathbf{i}) = \mu \prod_{j=1}^{K} \mathbf{1}_{\{i_j > 0\}}$).

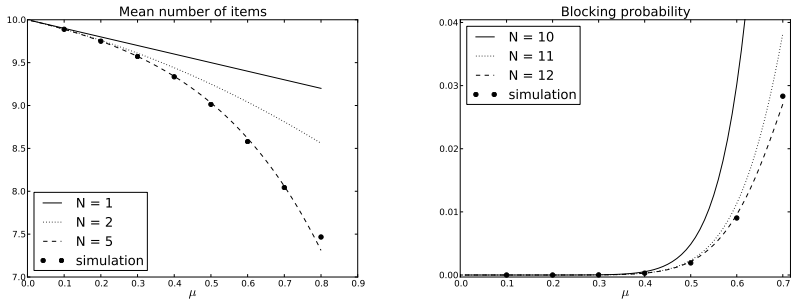
Computational complexity From (7.4), calculation of $\pi_n(\mathbf{i})$ takes at most $K+2$ additions and one division (assuming the rate sums are known). Hence, the computational complexity of calculating π_n is $O(KM)$, with $M = |\mathcal{C}|$ the size of the state space. Having the same complexity for every additional term in the expansion, calculating the first N coefficients then has complexity $O(KMN)$.

As the size of the state space is very large, limited memory consumption is equally important. To limit memory consumption to the size of storing only one steady-state vector one can proceed as follows. Assuming one is mainly interested in the expansion of a number of performance measures, note that once the nth term of the expansion of the steady state vector is determined, the corresponding terms in the expansions of various performance measures can be determined as well; see (7.5). Hence, there is no need to keep track of previous terms of the expansion of steady-state probabilities unless they are required for further calculations of coefficients of steady state probabilities. From (7.4) one sees that $\pi_n(\mathbf{i})$ is expressed in terms of $\pi_{n-1}(\mathbf{j})$, with $\mathbf{j}$ larger then $\mathbf{i}$ (lexicographically). This means that the coefficients of the vector π_{n-1} can be overwritten progressively during the calculation of π_n and memory for only one vector of size M is needed.

7.3 Numerical results

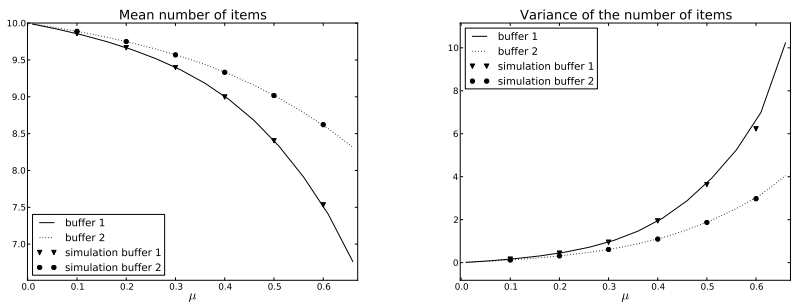
To illustrate our series expansion approach, we now assess its accuracy by means of some numerical examples. First, consider a system with $K = 5$ paired queues, each queue having capacity $C = 10$. Moreover, the arrival intensity at each queue is equal to $\lambda = 1$. Hence, the paired queueing system is symmetric and performance measures are equal for all queues. Figures 7.1(a) and 7.1(b) depict the mean queue content and the blocking probability in a queue versus the service rate μ, respectively. For both figures, series expansions of various orders are depicted as indicated ($N = 1,2,5$ for Figure 7.1(a) and $N = 10,11,12$ for Figure 7.1(b)), as well as simulation results which allow for assessing the accuracy of the series expansions. As expected, the mean queue content decreases and the blocking probability increases as the service rate μ increases. Moreover, for $\mu = 0$, the queues are completely filled as there is no service. From Figure 7.1(a), it is observed that the approximation method at hand is accurate for low orders of the expansion ($N = 5$) whereas more terms are needed to accurately determine the blocking probability ($N = 12$); see Figure 7.1(b). As the computation time of the series expansion is linear in the number of terms in the expansion, accurately assessing the blocking probability takes more than twice the computation time of assessing the mean queue content.

Figure 7.2(a) depicts the mean of the queue content of the first and second queue out of 5 paired queues, whereas Figure 7.2(b) depicts the corresponding variances. For both figures, the expansion of order $N = 20$ is compared with sim-



(a) (b)

Figure 7.1: Mean queue content (a) and blocking probability (b) for a symmetric paired queueing system.



(a) (b)

Figure 7.2: Mean (a) and variance (b) of the queue content of an asymmetric paired queueing system.

ulation results. The capacity equals 10 for all queues, and the arrival intensity in all but the first queue equals $\lambda_i = 1$, $i = 2, \ldots, 5$. The arrival rate in the first queue is lowered to $\lambda_1 = 0.8$. In comparison with the symmetric paired queueing system of Figure 7.1(a), the mean queue content increases for the second queue. This does not come as a surprise. Decreasing the arrival rate in the first queue implies that this queue is empty more often, thereby blocking service in the other queues. Finally, note that the variance increases for increasing μ, $\mu = 0$ corresponds to the case that the queue content deterministically equals the queue capacity for all queues, hence the variance is zero.

Appendix: Convergence of the power series

We now justify the series expansion. The basic ideas in this section date back to the seminal work of Schweitzer [11]. The series expansion is validated by explicitly constructing such an expansion. We first introduce some additional notation and the basic notion of the deviation matrix of a CTMC.

Let $\boldsymbol{\pi}^{(\mu)}$ denote the steady state solution $[\pi(\mathbf{i})]_{\mathbf{i}\in\mathcal{C}}$ of the balance equations. We have made the dependence of $\boldsymbol{\pi}^{(\mu)}$ on μ explicit for ease of notation. The balance equations can then be written in matrix notation as follows,

$$\boldsymbol{\pi}^{(\mu)} Q^{(\mu)} = \boldsymbol{\pi}^{(\mu)} (Q_0 + \mu Q_1) = 0\,, \tag{7.6}$$

where $Q^{(\mu)}$ is the $|\mathcal{C}| \times |\mathcal{C}|$ generator matrix of the CTMC and where Q_0 and Q_1 are known matrices that do not depend on μ. In view of the system assumptions it is readily seen that $Q^{(0)} = Q_0$ only has one recurrent state, i.e. $\mathbf{c}$ (the *full state*) is recurrent and all the others are transient. Therefore, the stationary vector $\boldsymbol{\pi}^{(0)}$ exists, with state $\pi^{(0)}(\mathbf{c}) = 1$ and $\pi^{(0)}(\mathbf{i}) = 0$ for $\mathbf{i} \in \mathcal{C}^*$.

Let D_0 be the deviation matrix of the CTMC with generator matrix Q_0,

$$D_0 = \int_0^\infty (P_0(t) - \Pi_0)dt\,. \tag{7.7}$$

Here the family $\{P_0(t) = \exp(Q_0 t), t \geq 0\}$ is the Markov semigroup of the CTMC, and $\Pi_0 = \lim_{t\to\infty} P_0(t) = \mathbf{1}'\boldsymbol{\pi}^{(0)}$, $\mathbf{1}'$ being a column vector of ones. As the state-space $\mathcal{C}$ is finite, the deviation matrix is well defined. Moreover, the deviation matrix satisfies $D_0\mathbf{1}' = 0$ — the row sums are zero — and,

$$D_0 Q_0 = Q_0 D_0 = \Pi_0 - I\,. \tag{7.8}$$

Theorem 1. *The solution $\boldsymbol{\pi}^{(\mu)}$ of the CTMC adheres to the following power series expansion,*

$$\boldsymbol{\pi}^{(\mu)} = \sum_{k=0}^{\infty} \left(\boldsymbol{\pi}^{(0)} (Q_1 D_0)^k\right) \mu^k\,, \tag{7.9}$$

for $0 \leq \mu < \mu_0$, μ_0^{-1} being the spectral radius of $Q_1 D_0$. Moreover, μ_0 is bounded from below by μ_0^* and μ_1^*,

$$\mu_0^* \;=\; \left(2\int_0^\infty \left(1 - \prod_{k=1}^{K} F(t; C_k, \lambda_k)\right) dt\right)^{-1} \;\geq\; \left(2\sum_{k=1}^{K} \frac{C_k}{\lambda_k}\right)^{-1} \;=\; \mu_1^*\,,$$

with F being the Erlang distribution,

$$F(t; C_k, \lambda_k) = 1 - \sum_{n=0}^{C_k - 1} \frac{1}{n!} e^{-\lambda_k t} (\lambda_k t)^n\,.$$

Proof Multiplying (7.6) by D_0 and invoking (7.8) yields,

$$\boldsymbol{\pi}^{(\mu)}(Q_0+\mu Q_1)D_0 = \boldsymbol{\pi}^{(\mu)}(\Pi_0 - I) + \boldsymbol{\pi}^{(\mu)}\mu Q_1 D_0 = 0\,.$$

Moreover, we have $\boldsymbol{\pi}^{(\mu)}\Pi_0 = \boldsymbol{\pi}^{(\mu)}\mathbf{1}'\boldsymbol{\pi}^{(0)} = \boldsymbol{\pi}^{(0)}$, such that,

$$\boldsymbol{\pi}^{(\mu)}(I-\mu Q_1 D_0) = \boldsymbol{\pi}^{(0)}\,.$$

The spectral radius of $\mu Q_1 D_0$ is μ/μ_0. Hence for $\mu < \mu_0$, $(I-\mu Q_1 D_0)$ is invertible and the Neumann series converges to the inverse,

$$\sum_{k=0}^{\infty}(\mu Q_1 D_0)^k = (I-\mu Q_1 D_0)^{-1}\,.$$

Combining the previous expressions immediately yields the series expansion (7.9).

As all elements but the last column of Π_0 are zero, only the last column of D_0 may contain negative values; see (7.7). Moreover, the row sums of D_0 are zero, hence the last column is equal in absolute value to the sum of the other columns. The entries in the last column of D_0 have the following interpretation,

$$[D_0]_{\mathbf{ic}} = -\int_0^{\infty}(1-[P_0(t)]_{\mathbf{ic}})dt = -\mathbb{E}[T_{\mathbf{i}}]\,,$$

where $T_{\mathbf{i}}$ is a random variable denoting the time it takes to reach the full state $\mathbf{c}$ from state $\mathbf{i}$ (assuming no departures). This interpretation shows that $\gamma \doteq E[T_{\mathbf{0}}] \geq E[T_{\mathbf{i}}]$ for all $\mathbf{i} \in \mathcal{C}$ where $\mathbf{0}$ denotes the empty state.

The time to fill up the ith queue is Erlang distributed with C_i stages and rate λ_i and the time to fill up all queues is the maximum of K Erlang distributed random variables. Therefore, the cumulative distribution of $T_{\mathbf{0}}$ is the product of K Erlang distributions and γ is calculated by integrating this distribution,

$$\gamma = \int_0^{\infty}\left(1-\prod_{k=1}^{K}F(t;C_k,\lambda_k)\right)dt\,.$$

Moreover, the maximum of K non-negative random variables is bounded from above by the sum of these random variables, which yields the following crude upper bound for γ,

$$\gamma \leq \sum_{k=1}^{K}\frac{C_k}{\lambda_k}\,, \tag{7.10}$$

the kth term in the sum on the right-hand side corresponding to the mean time to fill up the kth queue.

As the row sums of Q_1 are zero ($Q^{(\mu)}$ is a generator matrix for every μ), we have $Q_1\Pi_0 = 0$. Moreover, for any induced matrix norm, we have $\|Q_1 D_0\| \geq \mu_0^{-1}$. Therefore, we find,

$$\mu_0^{-1} \leq \|Q_1 D_0\| = \|Q_1(D_0+\gamma\Pi_0)\| \leq \|Q_1\|\|D_0+\gamma\Pi_0\|\,.$$

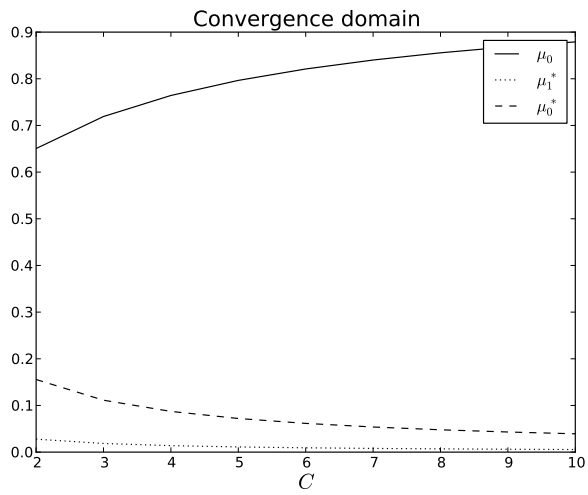


Figure 7.3: spectral radius μ_0 and lower bounds μ_0^* and μ_1^*.

Particularly using the maximum absolute row sum norm, we have $\|Q_1\| = 2$; $[Q_1]_{\mathbf{ii}} = -1$ if all queues are non-empty in state $\mathbf{i}$ and 0 if this not the case such that the corresponding row sums equal 2 and 0 respectively. In view of the definition of γ, one easily verifies that the matrix $D_0 + \gamma\Pi_0$ has no negative entries. Recalling that D_0 has zero row sums, this shows that all row sums of $D_0 + \gamma\Pi_0$ equal γ: $\|D_0 + \gamma\Pi_0\| = \gamma$ and,

$$\frac{1}{\mu_0} \leq 2\gamma \doteq \frac{1}{\mu_0^*},$$

which proves the lower bound μ_0^* for μ_0. The lower bound μ_1^* follows from $\mu_0^{-1} \leq 2\gamma$ and the crude bound (7.10) for γ.

To illustrate Theorem 1, Figure 7.3 depicts μ_0, the spectral radius of $Q_1 D_0$ and the lower bounds μ_0^* and μ_1^* for a system with $K = 3$ paired queues, each queue having a varying capacity from 2 to 10. As the figure shows, the bounds are much smaller than the convergence radius. It should be noted that both bounds are easy-to-derive but also rather loose bounds on the convergence radius. The bounds above can be made tighter by (1) not relying on the submultiplicative property of the matrix norm; (2) a matrix norm which is more adapted to this model. Both these approaches quickly lead to lengthy calculations and we consider them to be outside of the scope of the paper.

Acknowledgements

The second author is a post-doctoral fellow with the Research Foundation, Flanders. This research has been funded by the Interuniversity Attraction Poles Programme initiated by the Belgian Science Policy Office.

References

[1] B. Błaszczyszyn, T. Rolski and V. Schmidt. *Advances in Queueing: Theory, Methods and Open Problems*, chapter Light-traffic approximations in queues and related stochastic models, CRC Press, Boca Raton, Florida, 1995.

[2] E. De Cuypere and D. Fiems. *Performance evaluation of a kitting process*, Proceedings of the 18th International Conference on Analytical and Stochastic Modelling Techniques and Applications (ASMTA 2011), p. 175–188, Venice, June 2011.

[3] B. Johansson and M. Johansson. *High automated kitting system for small parts: a case study from the Volvo Uddevalla plant*, Proceedings of the 23rd International Symposium on Automotive Technology and Automation, p. 75–82, Vienna, Austria, 1990.

[4] J. Harrison. *Assembly-like queues*, Journal of Applied Probability, 10, p.354–367, 1973.

[5] W.J. Hopp and J.T. Simon. *Bounds and heuristics for assembly-like queues*. Queueing Systems, 4, p.137–156, 1989.

[6] S.S. Ko and R.F. Serfozo. *Response times in M/M/s fork-join networks*, Advances in Applied Probability, 36(3), p.854–871, 2004.

[7] I. Kovalenko. *Rare events in queueing theory. A survey*. Queueing systems, 16(1), p.1–49, 1994.

[8] G. Latouche. *Queues with paired customers*, Journal of Applied Probability, 18(3), p.684–696, 1981.

[9] R. Ramakrishnan and A. Krishnamurthy. *Analytical approximations for kitting systems with multiple inputs*, Asia-Pacific Journal of Operations Research, 25(2), p.187–216, 2008.

[10] R. Ramakrishnan and A. Krishnamurthy. *Performance evaluation of a synchronization station with multiple inputs and population constraints*. Computers & Operations Research, 39, p.560–570, 2012.

[11] P.J. Schweitzer. *Perturbation Theory and Finite Markov Chains*, Journal of Applied Probability, 5(2), p.401–413, 1968.

[12] M. Takahashi, H. Osawa and T. Fujisawa. *On a synchronization queue with two finite buffers*, Queueing Systems, 36, p.107–23, 2000.

8

A Maclaurin-series expansion approach to coupled queues with phase-type distributed service times

Eline De Cuypere, Koen De Turck, Sabine Wittevrongel and Dieter Fiems

submitted to the European Journal of Operations Research.

Abstract. We propose an efficient numerical scheme for the evaluation of large-scale Markov processes, under the condition that their generator matrix reduces to a triangular matrix when a certain rate is sent to zero. The methodology at hand is motivated by coupled queueing systems. Such systems are a natural abstraction for kitting processes in assembly systems and consist of multiple parallel buffers which are coupled in the sense that departures from the different buffers are synchronised and that there cannot be a service if any of the buffers is empty. As multiple customer buffers are involved, the Markovian description of the system obviously suffers from the state-space explosion problem. To cope with this problem, a numerical algorithm is presented which calculates the coefficients of the Maclaurin-series expansion of the steady-state probability vector. While the series expansion is a regular perturbation problem for the coupled queueing system with exponential service times, it is a singular perturbation problem if the service time are phase-type distributed. By means of numerical examples, we show that

the series expansion technique combined with a simple heuristic provides a high numerical accuracy.

8.1 Introduction

Coupled queueing systems arise as a convenient abstraction for kitting processes. A kitting process collects the necessary parts for a given end product in a container prior to assembly. While conceptually simple, kitting comes with many advantages. It clearly mitigates storage space requirements at the assembly station since no part inventories need to be kept there. Moreover, parts are placed in proper positions in the container such that assembly time reductions can be realised [17, 23]. A kitting process is obviously related to a coupled queueing system: the inventories of the different parts that go into the kit correspond to the different buffers, the kitting time corresponds to the service time and kitting is blocked if one or more parts are missing [9, 11].

There is considerable literature on the performance analysis of kitting systems with two buffers. Hopp and Simon [16] developed a model for a two-part kitting process with Poisson arrivals and exponentially distributed kit processing times. They found accurate bounds for the buffer capacities of both parts. Explicitly accounting for finite buffer capacities, Som et al. [26] further refined the results of Hopp and Simon. The exponential service times and Poisson arrival assumptions were later relaxed in [29] and [9]. Although results from the analysis of kitting systems with two buffers are useful, practical kitting systems as well as other applications with coupled queues typically involve more than two buffers. Such systems become however easily cumbersome and mathematically intractable even for a moderate number of buffers and reasonable buffer capacities. Indeed, the state-space explosion problem prohibits an exact analysis of such systems. Hence, approximation techniques have been proposed. Bonomi [7], Liu and Perros [13] and Baynat and Dallery [4] used a decomposition approach to analyse several independent two-buffer kitting systems. Ramakrishnan and Krishnamurthy [23, 24] studied kitting systems as a fork/join synchronisation station. In both works, they constructed and analysed a queueing system with two buffers and applied an aggregation-based approach to approximate the system with more than two buffers. A closed form approximation for the throughput and the mean queue length is derived in terms of the input parameters.

In this paper, we approximate large-scale finite kitting systems which relies on a Maclaurin-series expansion of the steady-state probability vector. This means that the Markov process of interest is transformed in a set of Markov processes parametrised by a certain variable known as the perturbation parameter. Such approximations go by different names including the power series method and the perturbation technique. One has to distinguish between regular and singular per-

turbation. In regular perturbation problems, the Markov process is irreducible when the perturbation parameter is set to zero. Hence, a unique solution of the stationary distribution of the Markov process can be found. This is not the case for singular perturbation problems. Indeed, if the Markov process is decomposable when the parameter is set to zero, the unperturbed part of the operator has no inverse and an approximation cannot be obtained [1, 20]. To cope with this inversion problem, several authors provided methods which calculate the coefficients of the Laurent series expansion of the deviation matrix of the Markov process. Schweitzer and Stewart [28] derived a recurrent formula for the calculation of the terms of the series for the case of linear perturbation. These results were generalised to the case of analytic perturbation by Korolyuk and Turbin [19] and by Avrachenkov [3]. In Avrachenkov's work, three related methods to determine the coefficients of the Laurent series are suggested. These three methods, based on the recursive solution of the infinite set of fundamental equations, depend to some extent on prior knowledge of the order of the pole at the singularity. This order of the pole can be determined by using for instance the combinatorial method of Hassin and Haviv [14].

This paper particularly focusses on the singular perturbation problem that arises in Markov processes for kitting processes with phase-type distributed service times when the service times are scaled up. The remainder of the paper is organised as follows. In the next section, the coupled queueing model at hand is described and the series expansion technique is introduced. For completeness we not only focus on the singular perturbation but also discuss the case of regular perturbation (exponential service times). In Sections 8.3 and 8.4, we prove a decoupling result for the regular perturbation and evaluate the regular and singular perturbation approach numerically, respectively. Finally, conclusions are drawn in Section 8.5.

8.2 Analysis

In this paper, we study the kitting process with K buffers, depicted in Figure 8.1. Each buffer has a finite capacity — let C_ℓ denote the capacity of buffer ℓ, $\ell = \{1,\ldots,K\}$ — and models the inventory of parts of a single type. New parts arrive at the buffers and, if both buffers are nonempty, a kit is assembled by collecting a part from each buffer. Arrivals at the buffers are modelled according to independent Poisson processes — let λ_ℓ denote the arrival rate in queue ℓ — and the consecutive kit assembly times (or service times) constitute a sequence of independent and identically phase-type distributed random variables.

A random variable has a phase-type distribution with M phases if its distribution has the representation,

$$F(x) = 1 - \mathbf{a}\exp(xA)\mathbf{1}',$$

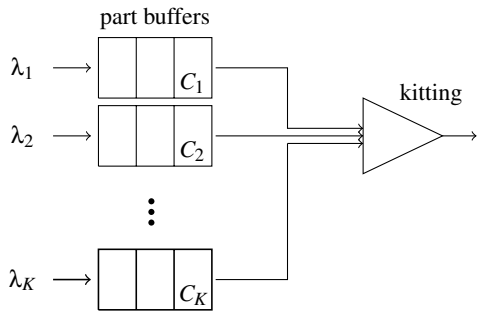


Figure 8.1: Kitting process with K queues

where **a** is a (row) probability vector of size M, where **1** is a row vector of ones and where A is an $M \times M$ matrix with negative entries on the diagonal, non-negative entries elsewhere and negative row-sums. A random variable has a phase-type distribution if it is the time until absorption of a finite Markov process with state-space $\mathcal{M} = \{1, 2, \ldots, M\}$. The vector **a** collects the probabilities of the initial state of this Markov process, the non-diagonal entries of the matrix A are the transition rates between non-absorbing states, and the absolute value of the row sums denote the rates to the absorbing state. For further use, let a_i be the ith element of **a** and let α_{ij} $(i \neq j)$ be the ijth element of the matrix A. Moreover, let α_{i0} denote the rate from state i to absorption,

$$\alpha_{i0} = -\sum_{j=1}^{M} \alpha_{ij}\,.$$

8.2.1 Regular perturbation

We first consider the case of regular perturbation, noting that a phase-type distribution with one phase corresponds to an exponential distribution. Let μ be the rate of this exponential distribution.

When the kit assembly time distribution is exponential, the state of the system is described by a vector $\mathbf{i} \in \mathcal{C}$ whose ℓth element corresponds to the queue size of the ℓth buffer. Here, $\mathcal{C} = \mathcal{C}_1 \times \ldots \times \mathcal{C}_K$ denotes the state space of this Markov process, with $\mathcal{C}_\ell = \{0, 1, \ldots, C_\ell\}$ being the set of possible levels of buffer ℓ. Let $\pi(\mathbf{i})$ be the steady-state probability of being in state $\mathbf{i}$ for this chain, $\mathbf{i} \in \mathcal{C}$. These steady-state probabilities satisfy the following set of balance equations,

$$\pi(\mathbf{i})\left(\mu\prod_{\ell=1}^{K}\mathbf{1}_{\{i_\ell>0\}} + \sum_{\ell=1}^{K}\mathbf{1}_{\{i_\ell<C_\ell\}}\lambda_\ell\right) = \pi(\mathbf{i}+\mathbf{1})\,\mu\prod_{\ell=1}^{K}\mathbf{1}_{\{i_\ell<C_\ell\}} + \sum_{\ell=1}^{K}\pi(\mathbf{i}-\mathbf{e}_\ell)\lambda_\ell\mathbf{1}_{\{i_\ell>0\}}\,, \quad (8.1)$$

for all $\mathbf{i} = (i_1, i_2, \ldots, i_K) \in \mathcal{C}$ and where $\mathbf{1}_{\{x\}}$ is the indicator function which equals one if x is true and equals zero otherwise. The symbol $\mathbf{e}_\ell$ represents a row vector with zero-elements except the ℓth element which is equal to one. Further, recall that $\mathbf{1}$ represents a row vector of ones.

The former system of equations has $C = (C_1 + 1) \times \ldots \times (C_K + 1)$ unknowns. Hence, even for a moderate number of buffers and reasonable buffer capacities the size of the state space is very large. As direct computation of the steady-state probability vector has an asymptotic complexity of $O(C^3)$, we focus on approximating the performance measures of interest by means of a series expansion approach.

To this end, we make the dependence of the steady state probabilities on μ explicit and introduce the Maclaurin-series expansion of the steady-state probabilities around $\mu = 0$,

$$\pi(\mathbf{i}) = \sum_{n=0}^{\infty} \pi_n(\mathbf{i})\mu^n \,, \tag{8.2}$$

for $\mathbf{i} \in \mathcal{C}$. Substitution of the former expression in the balance equation (8.1), comparing terms in μ^n for $n = 01, 2, \ldots$ and solving for $\pi_n(\mathbf{i})$ yields,

$$\pi_0(\mathbf{i}) = \frac{\sum_{\ell=1}^{K} \pi(\mathbf{i} - \mathbf{e}_\ell)\lambda_\ell \mathbf{1}_{\{i_\ell > 0\}}}{\sum_{\ell=1}^{K} \mathbf{1}_{\{i_\ell < C_\ell\}}\lambda_\ell} \,, \tag{8.3}$$

and,

$$\begin{aligned}\pi_n(\mathbf{i}) = \frac{1}{\sum_{\ell=1}^{K} \mathbf{1}_{\{i_\ell < C_\ell\}}\lambda_\ell} \times \Bigg(& \mathbf{1}_{\{n>0\}}\pi_{n-1}(\mathbf{i}+\mathbf{1}) \prod_{\ell=1}^{K} \mathbf{1}_{\{i_\ell < C_\ell\}} \\ & + \sum_{\ell=1}^{K} \pi_n(\mathbf{i} - \mathbf{e}_\ell)\lambda_\ell \mathbf{1}_{\{i_\ell > 0\}} - \mathbf{1}_{\{n>0\}}\pi_{n-1}(\mathbf{i}) \prod_{\ell=1}^{K} \mathbf{1}_{\{i_\ell > 0\}} \Bigg) \,,\end{aligned} \tag{8.4}$$

for $\mathbf{i} \in \mathcal{C}^{\triangleright} = \mathcal{C} \setminus \{\mathbf{c}\}$ with $\mathbf{c} = [C_1, C_2, \ldots, C_K]$. Evaluating (8.3) in lexicographical order shows,

$$\pi_0(\mathbf{i}) = 0 \,, \tag{8.5}$$

for $\mathbf{i} \in \mathcal{C}^{\triangleright}$ while (8.4) allows for calculating all $\pi_n(\mathbf{i})$ for $\mathbf{i} \in \mathcal{C}^{\triangleright}$ in lexicographical order once the $n-1$st terms are known. Finally, for the terms of the stationary probabilities of state $\mathbf{c}$ we invoke the normalisation condition, yielding,

$$\pi_0(\mathbf{c}) = 0 \,, \quad \pi_n(\mathbf{c}) = -\sum_{\mathbf{i} \in \mathcal{C}^{\triangleright}} \pi_n(\mathbf{i}) \,.$$

Remark 1. In order for a series expansion to make sense, the stationary vector is required to be analytic in a neighbourhood of $\mu = 0$. For finite state spaces (in contrast to infinite ones, see e.g. [1, 15]), this is fairly easy to establish. Finding the steady state distribution is in this case essentially a finite-dimensional eigenproblem. If a matrix depends analytically on a parameter, then the corresponding

eigenvalues and eigenvectors are also analytic in case of null-space perturbation [2]. Another possible path towards proving analyticity is via V-uniform ergodicity of the unperturbed Markov process with generator $Q^{(0)}$ (see a.o [1]), which is equivalent to the existence of a spectral gap (the distance between eigenvalue 0 of the generator matrix $Q^{(0)}$ and the eigenvalue that is its nearest neighbour). For finite Markov processes, there is a spectral gap as long as there is only one recurrent class. This means that the Markov process is irreducible when the perturbation parameter is set to zero.

Remark 2. The numerical complexity of the algorithm is $O(CKN)$ where N is the number of terms in the series expansion, C is the size of the state space and K is the number of buffers. This immediately follows from the observation that we have to calculate N terms of the C stationary probabilities. For the calculation of each term in the expansion, we sum $O(K)$ terms.

8.2.2 Singular perturbation

As for the coupled queueing system with exponential service times, we now propose an efficient numerical scheme for the evaluation of kitting processes with phase-type service times. We assume that the service times are scaled with factor μ^{-1} and again consider the series expansion around $\mu = 0$. Note that the rescaled service times are phase-type distributed with generator matrix μA and initial probability vector $\mathbf{a}$.

Let $\mathcal{C}^*$ be the subset of $\mathcal{C}$ such that all buffers are nonempty. For all $\mathbf{i} \in \mathcal{C} \setminus \mathcal{C}^*$, at least one buffer is empty meaning that there is no ongoing service. Hence $\mathbf{i}$ captures the state of the Markov process. In contrast, for $\mathbf{i} \in \mathcal{C}^*$, service is ongoing meaning that the service process is in some state $j \in \mathcal{M}$. Therefore, the state space of the Markov process with phase-type service times is $(\mathcal{C} \setminus \mathcal{C}^*) \cup (\mathcal{C}^* \times \mathcal{M})$.

With a slight abuse of notation, let $\pi(\mathbf{i})$ be the steady state probability of state $\mathbf{i} \in \mathcal{C} \setminus \mathcal{C}^*$ and let $\pi(\mathbf{i}, j)$ be the steady state probability of state $(\mathbf{i}, j) \in \mathcal{C}^* \times \mathcal{M}$. Finally, let $\mathbf{c} = [C_1, \ldots, C_K]$ as in the case of exponential service times and — for ease of exposition — assume $C_k > 1$ for $k = 1, \ldots, K$.

We can now write down the balance equations:

$$\pi(\mathbf{i}) \sum_{\ell=1}^{K} \mathbf{1}_{\{i_\ell < C_\ell\}} \lambda_\ell = \sum_{\ell=1}^{K} \pi(\mathbf{i} - \mathbf{e}_\ell) \lambda_\ell \mathbf{1}_{\{i_\ell > 0\}} + \prod_{\ell=1}^{K} \mathbf{1}_{\{i_\ell < C_\ell\}} \mu \sum_{k=1}^{M} \pi(\mathbf{i} + \mathbf{1}, k) \alpha_{k0} \,,$$

for $\mathbf{i} \in C \setminus \mathcal{C}^*$ and

$$\pi(\mathbf{i}, j) \left(\sum_{\ell=1}^{K} \mathbf{1}_{\{i_\ell < C_\ell\}} \lambda_\ell + \mu \sum_{k=0, k \neq j}^{M} \alpha_{jk} \right) = \prod_{\ell=1}^{K} \mathbf{1}_{\{i_\ell < C_\ell\}} \mu \sum_{k=1}^{M} \pi(\mathbf{i} + \mathbf{1}, k) \alpha_{k0} a_j$$
$$+ \sum_{\ell=1}^{K} \pi(\mathbf{i} - \mathbf{e}_\ell, j) \lambda_\ell \mathbf{1}_{\{i_\ell > 1\}} + \sum_{\ell=1}^{K} \pi(\mathbf{i} - \mathbf{e}_\ell) \lambda_\ell \mathbf{1}_{\{i_\ell = 1\}} a_j + \mu \sum_{k=1, k \neq j}^{M} \pi(\mathbf{i}, k) \alpha_{kj} \,,$$

for $\mathbf{i} \in \mathcal{C}^*$ and $j \in \mathcal{M}$.

Proceeding as for the kitting system with exponential service times, we introduce the following Maclaurin series expansions,

$$\pi(\mathbf{i}) = \sum_{n=0}^{\infty} \pi_n(\mathbf{i})\mu^n\,, \quad \pi(\mathbf{i}, j) = \sum_{n=0}^{\infty} \pi_n(\mathbf{i}, j)\mu^n\,,$$

for $\mathbf{i} \in \mathcal{C} \setminus \mathcal{C}^*$ and $\mathbf{i} \in \mathcal{C}^*$, respectively. Plugging the above expansions in the balance equations and comparing terms in μ^n yields,

$$\pi_n(\mathbf{i}) = \left(\sum_{\ell=1}^{K} \mathbf{1}_{\{i_\ell < C_\ell\}}\lambda_\ell\right)^{-1} \left(\prod_{\ell=1}^{K} \mathbf{1}_{\{i_\ell < C_\ell\}} \sum_{k=1}^{M} \pi_{n-1}(\mathbf{i}+\mathbf{1}, k)\alpha_{k0} + \sum_{\ell=1}^{K} \pi_n(\mathbf{i}-\mathbf{e}_\ell)\lambda_\ell \mathbf{1}_{\{i_\ell > 0\}}\right), \tag{8.6}$$

for $\mathbf{i} \in \mathcal{C} \setminus \mathcal{C}^*$ and $n = 0, 1, \ldots$,

$$\begin{aligned}\pi_n(\mathbf{i}, j) = \left(\sum_{\ell=1}^{K} \mathbf{1}_{\{i_\ell < C_\ell\}}\lambda_\ell\right)^{-1} \Bigg(&-\pi_{n-1}(\mathbf{i}, j) \sum_{k=0, k\neq j}^{M} \alpha_{jk} \\ &+ \prod_{\ell=1}^{K} \mathbf{1}_{\{i_\ell < C_\ell\}} \sum_{k=1}^{M} \pi_{n-1}(\mathbf{i}+\mathbf{1}, k)\alpha_{k0} a_j + \sum_{\ell=1}^{K} \pi_n(\mathbf{i}-\mathbf{e}_\ell, j)\lambda_\ell \mathbf{1}_{\{i_\ell > 1\}} \\ &+ \sum_{\ell=1}^{K} \pi_n(\mathbf{i}-\mathbf{e}_\ell)\lambda_\ell \mathbf{1}_{\{i_\ell = 1\}} a_j + \sum_{k=1, k\neq j}^{M} \pi_{n-1}(\mathbf{i}, k)\alpha_{kj}\Bigg),\end{aligned} \tag{8.7}$$

for $\mathbf{i} \in \mathcal{C}^* \setminus \{\mathbf{c}\}$. Here, we assumed $\pi_{-1}(\mathbf{i}) = \pi_{-1}(\mathbf{i}, j) = 0$ for all $\mathbf{i} \in \mathcal{C}$ and $j \in \mathcal{M}$. As for the regular perturbation, the former set of equations allow for recursive calculation of the nth term in the expansion of all stationary probabilities in lexicographical, once the $n-1$st terms are known.

For $\mathbf{i} = \mathbf{c}$, we fell back on the normalisation condition in the regular case. This was possible as there was only one remaining unknown — $\pi_n(\mathbf{c})$ — for every term in the series expansion. In this case however, there remain M unknown terms: $\pi_n(\mathbf{c}; 1), \ldots, \pi_n(\mathbf{c}; M)$. Plugging the expansions in the balance equation for $\mathbf{i} = \mathbf{c}$ and comparing terms in μ^n yields,

$$\pi_{n-1}(\mathbf{c}, j) \sum_{k=0, k\neq j}^{M} \alpha_{jk} = \sum_{\ell=1}^{K} \pi_n(\mathbf{c}-\mathbf{e}_\ell, j)\lambda_\ell + \sum_{k=1, k\neq j}^{M} \pi_{n-1}(\mathbf{c}, k)\alpha_{kj}\,, \tag{8.8}$$

for $n = 0, 1, \ldots$. These expressions however do not allow to calculate the remaining unknowns. Therefore, we proceed as follows.

Let $\mathcal{C}^\diamondsuit$ be the set of states $(\mathbf{i}, j)$, with $\mathbf{i}$ lexicographically larger than $\mathbf{c}-\mathbf{1}$ and with $j \in \mathcal{M}$. Assuming that the probabilities $\pi_{n-1}(\mathbf{c}; 1), \ldots, \pi_{n-1}(\mathbf{c}; M)$ are not known, equation (8.7) still allows to calculate all $\pi_n(\mathbf{i}, j)$ for $\mathbf{i} \in \mathcal{C}^* \setminus \mathcal{C}^\diamondsuit$ and $j \in \mathcal{M}$ but no longer allows to determine $\pi_n(\mathbf{i}, j)$ for $\mathbf{i} \in \mathcal{C}^\diamondsuit$, and $j \in \mathcal{M}$.

For $\mathbf{i} \in \mathcal{C}^{\diamondsuit}$, we therefore express $\pi_n(\mathbf{i}, j)$ in terms of the probabilities $\pi_{n-1}(\mathbf{c}, \ell)$ as follows,

$$\pi_n(\mathbf{i}, j) = \beta_n(\mathbf{i}, j; 0) + \sum_{\ell=1}^{M} \beta_n(\mathbf{i}, j; \ell)\pi_{n-1}(\mathbf{c}, \ell)\,. \tag{8.9}$$

In view of equation (8.7), the terms $\beta_n(\mathbf{i}, j; \ell)$ in expression (8.9) adhere,

$$\begin{aligned}
\beta_n(\mathbf{c}-\mathbf{1}, j; 0) = &\left(\sum_{\ell=1}^{K} \lambda_\ell\right)^{-1} \Bigg(-\pi_{n-1}(\mathbf{c}-\mathbf{1}, j) \sum_{k=0, k\neq j}^{M} \alpha_{jk} \\
&+ \sum_{\ell=1}^{K} \pi_n(\mathbf{c}-\mathbf{1}-\mathbf{e}_\ell, j)\lambda_\ell \mathbf{1}_{\{C_\ell > 2\}} + \sum_{\ell=1}^{K} \pi_n(\mathbf{c}-\mathbf{1}-\mathbf{e}_\ell)\lambda_\ell \mathbf{1}_{\{C_\ell = 2\}} a_j \\
&+ \sum_{k=1, k\neq j}^{M} \pi_{n-1}(\mathbf{c}-\mathbf{1}, k)\alpha_{kj} \Bigg)\,,
\end{aligned} \tag{8.10}$$

$$\beta_n(\mathbf{c}-\mathbf{1}, j; k) = \left(\sum_{\ell=1}^{K} \lambda_\ell\right)^{-1} \left(\alpha_{k0} a_j\right)\,, \tag{8.11}$$

$$\begin{aligned}
\beta_n(\mathbf{i}, j; 0) = &\left(\sum_{\ell=1}^{K} \mathbf{1}_{\{i_\ell < C_\ell\}} \lambda_\ell\right)^{-1} \Bigg(\sum_{k=1, k\neq j}^{M} \pi_{n-1}(\mathbf{i}, k)\alpha_{kj} - \pi_{n-1}(\mathbf{i}, j) \sum_{k=0}^{M} \alpha_{jk} \\
&+ \sum_{\ell=1}^{K} \pi_n(\mathbf{i}-\mathbf{e}_\ell, j)\lambda_\ell \mathbf{1}_{\{i_\ell > 1, \mathbf{i}-\mathbf{e}_\ell < \mathbf{c}-\mathbf{1}\}} \\
&+ \sum_{\ell=1}^{K} \beta_n(\mathbf{i}-\mathbf{e}_\ell, j; 0)\lambda_\ell \mathbf{1}_{\{i_\ell > 1, \mathbf{i}-\mathbf{e}_\ell \geq \mathbf{c}-\mathbf{1}\}} + \sum_{\ell=1}^{K} \pi_n(\mathbf{i}-\mathbf{e}_\ell)\lambda_\ell \mathbf{1}_{\{i_\ell = 1\}} a_j \Bigg)\,,
\end{aligned} \tag{8.12}$$

$$\beta_n(\mathbf{i}, j; k) = \left(\sum_{\ell=1}^{K} \mathbf{1}_{\{i_\ell < C_\ell\}} \lambda_\ell\right)^{-1} \left(\sum_{\ell=1}^{K} \beta_n(\mathbf{i}-\mathbf{e}_\ell, j; k)\lambda_\ell \mathbf{1}_{\{i_\ell > 1, \mathbf{i}-\mathbf{e}_\ell \geq \mathbf{c}-\mathbf{1}\}}\right)\,, \tag{8.13}$$

for $\mathbf{i} \in \mathcal{C}^{\diamondsuit}$ and $j \in \mathcal{M}$. Clearly, we can now calculate all $\beta_n(\mathbf{i}, j; k)$ in lexicographical order.

Finally, plugging equation (8.9) in (8.8) yields a set off equations for the remaining unknowns $\pi_{n-1}(\mathbf{c}, j)$, $j \in \mathcal{M}$:

$$\begin{aligned}
\pi_{n-1}(\mathbf{c}, j) \sum_{k=0, k\neq j}^{M} \alpha_{jk} = &\sum_{k=1, k\neq j}^{M} \pi_{n-1}(\mathbf{c}, k)\alpha_{kj} + \sum_{\ell=1}^{K} \Bigg(\beta_n(\mathbf{c}-\mathbf{e}_\ell, j; 0) \\
&+ \sum_{k=1}^{M} \beta_n(\mathbf{c}-\mathbf{e}_\ell, j; k)\pi_{n-1}(\mathbf{c}, k) \Bigg)\lambda_\ell\,.
\end{aligned} \tag{8.14}$$

Using arguments from Hassin and Haviv [14], one can show that the former set of equations has rank $M-1$. Complementing this set with the normalisation condition,

$$\sum_{j \in \mathcal{M}} \pi_0(\mathbf{c}, j) = 1\,,$$

$$\sum_{j\in\mathcal{M}} \pi_n(\mathbf{c},j) = -\sum_{\mathbf{i}\in\mathcal{C}\setminus\mathcal{C}^*} \pi_n(\mathbf{i}) - \sum_{\mathbf{i}\in\mathcal{C}^*\setminus\{\mathbf{c}\}} \sum_{j\in\mathcal{M}} \pi_n(\mathbf{i},j)\,. \tag{8.15}$$

allows for determining $\pi_{n-1}(\mathbf{x},j)$, for $j\in\mathcal{M}$. Note that the right-hand side in the second expression of equation (8.15) only contains known terms.

Summarising, assuming that the $n-1$st term is calculated apart from the elements $\pi_{n-1}(\mathbf{c},j)$, $j\in\mathcal{M}$, we obtain the nth order terms (apart from the elements $\pi_n(\mathbf{c},j)$, $j\in\mathcal{M}$) and the elements $\pi_{n-1}(\mathbf{c},j)$, $j\in\mathcal{M}$ as follows,

1. Calculate the nth terms in lexicographical order by equations (8.6) and (8.7) up to but excluding state $(\mathbf{c}-\mathbf{1},1)$.
2. Calculate the terms $\beta_n(\mathbf{i},j;k)$ by equations (8.10) to (8.13) in lexicographical order for all $\mathbf{i}\in\mathcal{C}^{\diamondsuit}\setminus\{\mathbf{c}\}$, $j\in\mathcal{M}$ and $k\in\mathcal{M}\cup\{0\}$.
3. Solve the system of equations (8.14) together with the normalisation condition given in (8.15).
4. Use equation (8.9) to calculate $\pi_n(\mathbf{i},j)$ for $\mathbf{i}\in\mathcal{C}^{\diamondsuit}\setminus\{\mathbf{c}\}$ and $j\in\mathcal{M}$.

Remark 3. In contrast to regular perturbation, the Markov process in this section has multiple ergodic classes for $\mu=0$, implying that there exists no unique stationary distribution. In fact, for $\mu=0$ there are M absorbing states (all queues full and the service process in any of its M states). Nevertheless, the stationary distribution is analytic in a deleted neighbourhood of $\mu=0$ and there exists a unique analytic continuation for $\mu=0$. Practically, the singular perturbation reflects in not having enough equations to solve term by term in the expansion, by consecutively equating terms in μ^n. It is however possible to find the terms of the expansion by combining the equations one gets for μ^n until μ^{n+k} for some integer k. k is the order of the Laurent series expansion of the deviation matrix of the Markov process and can be determined by solving a combinatorial problem [14]. In this particular case, we have $k=1$.

Remark 4. The numerical complexity of the algorithm is $O((C+2^K)(K+M)N+M^3N)$. The first step has complexity $O(C(K+M))$, similar as for regular perturbations. The second step has numerical complexity $O(2^KM(K+M))$ as we need to calculate $O(2^KM)$ different β's for each term in the expansion. The third step corresponds to the solution of system of M equations, which has complexity $O(M^3)$. Finally, the last step has complexity $O(2^KM^2)$.

8.3 Decoupling result

While scrutinising numerical results of the algorithm, we noticed a peculiar pattern in the case of exponential service times, which we will explain and establish in the following. To this end, we first derive the series expansion of the mean queue

content of a $M/M/1/C$ queue with arrival rate λ and departure rate μ, for small μ. As almost anything about this queueing system can be derived in closed-form, the mean queue content not being an exception, this derivation is rather straightforward. Indeed, recall that the mean buffer content Q is equal to [8]:

$$Q = \frac{\rho}{1-\rho} - \frac{(C+1)\rho^{C+1}}{1-\rho^{C+1}},$$

where $\rho = \lambda/\mu$. As we are interested in small μ, we introduce $r = \rho^{-1} = \mu/\lambda$ and write in powers of r:

$$\begin{aligned} Q &= -\frac{1}{1-r} + \frac{C+1}{1-r^{C+1}} \\ &= -\sum_{k=0}^{\infty} r^k + (C+1) + \sum_{n=1} (C+1) r^{(C+1)n}. \end{aligned} \tag{8.16}$$

This leads to repeating coefficients in the series expansion in r: C, -1, -1, $\cdots$, -1, C, -1, $\cdots$.

We noticed this exact series expansion for the first few terms of the mean queue content of any queue in a coupled queueing system. This can be explained as follows. Assume without loss of generality that $C_1 \leq C_2 \leq \cdots \leq C_K$ and suppose we are interested in the mean queue content of the ith queue. For series expansions up to μ^n, with $n < C_1$, we find the same series expansion as for the single M/M/1/C_i queue with arrival rate λ_i and service rate μ. This is because of the n events rule: the nth order coefficient is determined by sample paths in which n or fewer departures occur. This means that the smallest queue never gets empty (hence no queue gets empty) and thus the ith queue considered in isolation is indistinguishable from said M/M/1/C_i queue. It is possible to take this argument a bit further: for a series expansion of the mean content of the ith queue up to order n, we can consider an adapted coupled queueing system that has a size that is certainly not larger than the original system and includes: all queues j for which $C_j \leq n$ plus the ith queue itself, and compute the series expansion for this adapted system. Hence, for the smallest queue, the expansion up to the order C_2 follow the pattern of Equation (8.16).

This result is not limited to just the mean queue content, but holds for any performance measure that can be derived from the marginal distribution of a single queue.

8.4 Numerical results

We now assess the accuracy of the perturbation approach by means of several numerical examples.

To establish the regions in which the results of the numerical scheme are accurate enough, we propose a simple heuristic which compares the Nth and the $2N$th order expansions. Let $f_N(\mu)$ be the Nth order expansion in μ, we then accept our Nth order approximation provided if

$$\left| \frac{f_{2N}(\mu) - f_N(\mu)}{f_{2N}(\mu)} \right| < \varepsilon, \tag{8.17}$$

or equivalently,

$$1 - \varepsilon < \left| \frac{f_N(\mu)}{f_{2N}(\mu)} \right| < 1 + \varepsilon. \tag{8.18}$$

We can thus establish for each expansion order the region in which the inequality of the heuristic holds, and denote it as the *heuristic convergence region*. In the plots, we render these regions with a short vertical line. We take an error term ε equal to 10^{-4}. Consider a system with $K = 5$ coupled queues, each queue having capacity $C = 10$ and exponential service times. Moreover, the arrival streams at each queue are Poisson with rate $\lambda = 1$. Figures 8.2 and 8.3 depict the mean queue content and the blocking probability in log scale versus the exponentially distributed service rate μ, respectively. The blocking probability is the probability that service is blocked because at least one of the queues is empty. For both figures, series expansions of various orders are depicted as indicated ($N = 1, 2, 5$ for Figure 8.2 and $N = 12, 15, 18$ for Figure 8.3), as well as simulation results which allow for assessing the accuracy of the series expansions. As expected, the mean queue content decreases and the blocking probability increases as the service rate μ increases. Moreover, for $\mu = 0$, the queues are completely filled as there is no service. From Figure 8.2, it is observed that low orders of the expansion of the mean queue content suffice for even quite large μ, whereas more terms are needed to accurately determine the blocking probability; see Figure 8.3. This is because the blocking probability is a rare event for low values of μ, and hence more terms are required to increase the accuracy. The regions for which the inequality of the heuristic holds in Figure 8.2 go up to $\mu = 0.03$ for $N = 1$, up to $\mu = 0.09$ for $N = 2$ and up to $\mu = 0.29$ for $N = 5$ while the regions go up to $\mu = 0.17$ for $N = 12$, up to $\mu = 0.35$ for $N = 15$ and up to $\mu = 0.45$ for $N = 18$. As the computation time of the series expansion is linear in the number of terms in the expansion, accurately assessing the blocking probability takes more than twice the computation time of assessing the mean queue content.

We also show what can be obtained by merely using the decoupling result of Section 8.3 (hence without any computational cost at all). In Figure 8.4, the mean

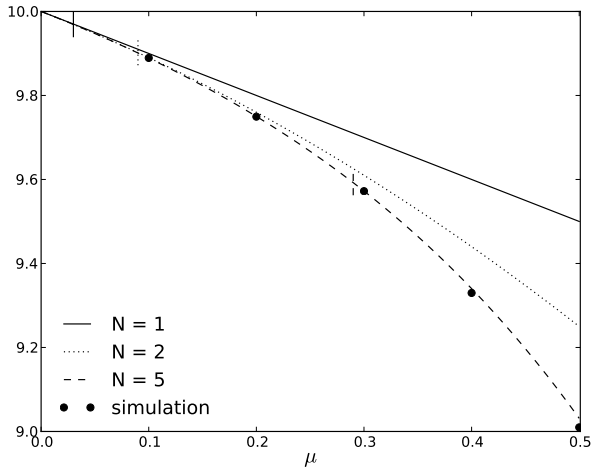


Figure 8.2: Mean queue content of a coupled queueing system with exponential service times.

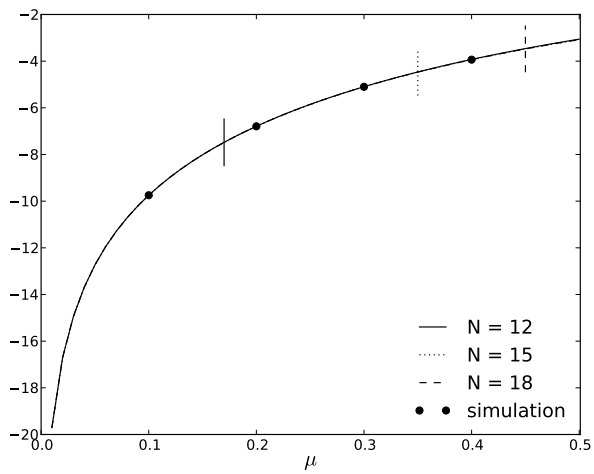


Figure 8.3: Blocking probability (in log scale) of a coupled queueing system with exponential service times.

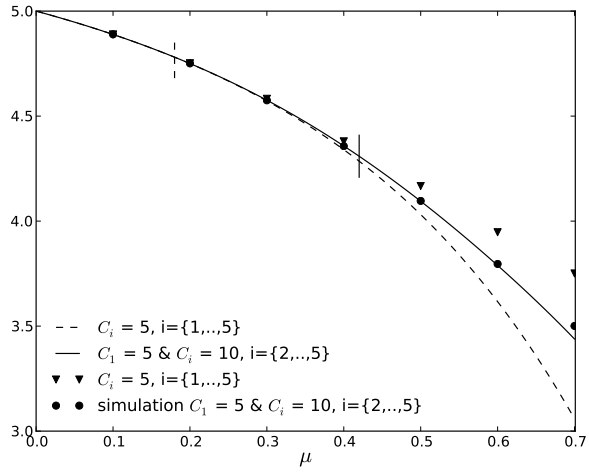


Figure 8.4: Mean queue content of a coupled queueing system with exponential service times, using only the decoupling result.

number of items of the queue with capacity $C_1 = 5$ of a 5 coupled queueing system versus an exponential service rate is depicted. We notice an excellent correspondence with the simulation results up to $\mu = 0.18$ for 5 coupled queues with capacity $C_i = 5$, $i = 1, \ldots, 5$ and up to $\mu = 0.42$ for 5 coupled queues with capacity $C_1 = 5$ and $C_i = 10$, $i = 2, \ldots, 5$. This is partially due to the fact that we can use the expansion up to order 10 in the asymmetric case instead of up to 5 in the symmetric case such that a more accurate expansion is found to approximate the $M/M/1/C_1$ queue.

Instead of exponential service times, we now assume coupled queueing systems with phase-type service times. Figure 8.5 depicts the mean queue content of a coupled queueing system with a three-phase hyperexponential service time distribution versus the service rate μ. As in previous figures, we assume 5 queues of capacity 10 and a Poisson arrival rate of 1 for all queues. The phases have the same probability to occur and we assume a mean service rate equal to 2μ. As Figure 8.5 shows, the regions for which the inequality of the heuristic holds in Figure 8.5 go up to $\mu = 0.06$ for $N = 2$, up to $\mu = 0.09$ for $N = 3$ and up to $\mu = 0.29$ for $N = 4$. Comparing the results of the approximation method with those of the simulation, we can derive that the performance assessment is highly accurate in the heuristically determined region.

In Figure 8.6, different Poisson arrival rates for all queues (resp. equal to 1.0, 1.5 and 2.0) are considered. We assume the same parameter values as in Figure 8.5 and show the mean queue content. The expansion is of order $N = 3$. As expected,

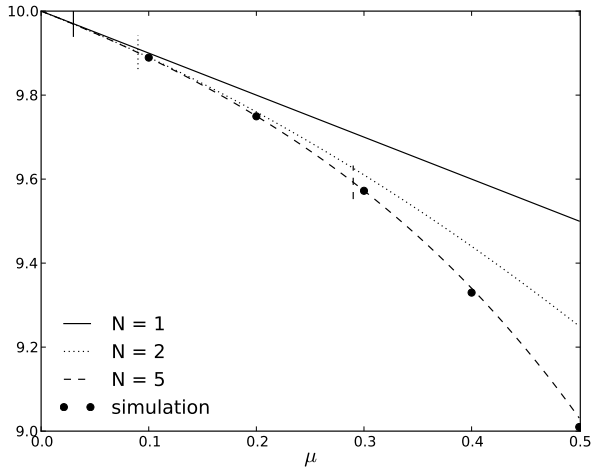


Figure 8.5: Mean queue content of a coupled queueing system with three-phase hyperexponential service times.

the higher the arrival rate, the larger the mean queue content. Also, the regions for which the inequality of the heuristic holds increases as the arrival rate increases.

Finally, Figure 8.7 depicts the mean queue content of a coupled queueing system with a two-phase hyperexponential service time distribution versus the mean service rate. The phases have probability $\frac{1}{40}$ and $1 - \frac{1}{40}$ to occur and the mean service rate is equal to μ. The expansion is of order $N = 20$. The other parameter values are the same as in Figure 8.5. For sake of clarity, we here only show performance results with a value between 8 and 10. As the figure shows, a higher value of the variance decreases the mean queue content.

8.5 Conclusion

To evaluate the performance of large-scale coupled queueing systems, we propose a numerical algorithm which calculates the coefficients of the Maclaurin-series expansion of the steady-state probability vector. Coupling means that service is only possible when none of the queues are empty. In this paper, we consider both regular and singular perturbation problems when the coupled queueing system has respectively exponential and phase-type service times. As shown by the numerical results, the Maclaurin-series expansion combined with the proposed heuristic give a quite good approximation of the studied coupled queueing system in the regular as well as in the singular case.

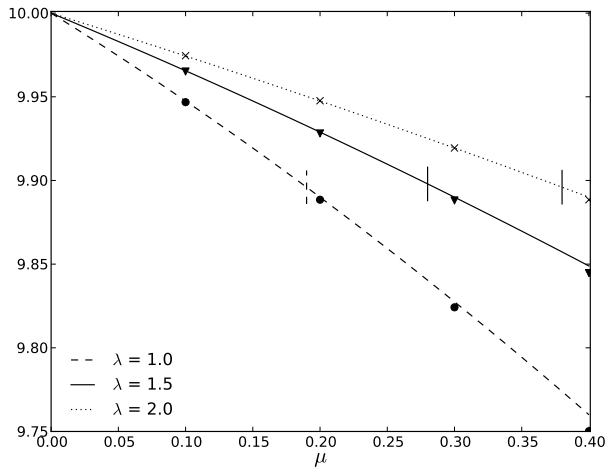


Figure 8.6: Mean queue content of a coupled queueing system with hyperexponential service times and different arrival rates.

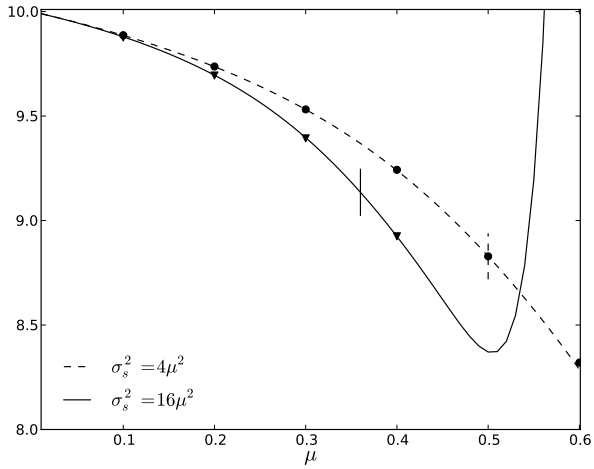


Figure 8.7: Mean queue content of a coupled queueing system with hyperexponential service times and different values of the variance.

References

[1] E. Altman, K.E. Avrachenkov and R. Nunez-Queija. *Perturbation analysis for denumerable Markov chains with application to queueing models*, Advances in Applied Probability, 36(3), p.839–853, 2004.

[2] K.E. Avrachenkov and M. Haviv. *Perturbation of null spaces with application to the eigenvalue problem and generalized inverses*, Linear Algebra and its Applications, 369, p.1–25, 2003.

[3] K.E. Avrachenkov. *Analytic Perturbation Theory and its Applications. Chapter 2: Analytic Perturbation of Singular Linear Systems*, PhD thesis. School of Mathematics, Faculty of Information Technology, University of South Australia, 1999.

[4] B. Baynat and Y. Dallery. *Approximate analysis of multi-class synchronized closed queueing networks*, In: IEEE international workshop on modelling, analysis and simulation of computer and telecommunication systems, 1995.

[5] B. Błaszczyszyn. *Factorial-moment expansion for stochastic systems*, Stochastic Processes and their Applications, 56, p.321–335, 1995.

[6] B. Błaszczyszyn, T. Rolski and V. Schmidt. *Advances in Queueing: Theory, Methods and Open Problems*, chapter Light-traffic approximations in queues and related stochastic models, CRC Press, Boca Raton, Florida, 1995.

[7] F. Bonomi. *An approximate analysis for a class of assembly-like queues*, Queueing Systems, 1, p.289–309, 1987.

[8] J. Cohen. *The single server queue*, North-Holland Pub. Co., Amsterdam, 1969.

[9] E. De Cuypere and D. Fiems. *Performance evaluation of a kitting process*, Proceedings of the 18th International Conference on Analytical and Stochastic Modelling Techniques and Applications (ASMTA 2011), p.175–188, Venice, June 2011.

[10] E. De Cuypere, K. De Turck and D. Fiems. *Algorithmic approach to series expansions around transient Markov chains with applications to paired queuing systems*, Proceedings of the 6th International Conference on Performance Evaluation Methodologies and Tools, Valuetools 2012, Cargèse, p.38–44, October 2012.

[11] E. De Cuypere, K. De Turck and D. Fiems. *Performance analysis of a kitting process as a paired queue*, Hindawi Publishing Corporation Mathematical Problems in Engineering, vol. 2013, Article ID 843184, 2013.

[12] E. De Cuypere, K. De Turck and D. Fiems. *A Maclaurin-series expansion approach to multiple paired queues*, Operations Research Letters, 42(3), p.203–207, 2014.

[13] Y.C. Liu and H.G. Perros. *Approximate analysis of a closed fork/join model*, European Journal of Operational Research, 53(3), p.382–392, 1991.

[14] R. Hassin and M. Haviv. *Mean passage times and nearly uncoupled Markov chains*, SIAM Journal on Discrete Mathematics, 5(3), p.386–397, 1992.

[15] B. Heidergott and A. Hordijk. *Taylor Series Expansions for Stationary Markov Chains*, Advances in Applied Probability, 35(4), p.1046–1070, 2003.

[16] W.J. Hopp and J.T. Simon. *Bounds and heuristics for assembly-like queues*, Queueing Systems, 4, p.137–156, 1989.

[17] B. Johansson and M. Johansson. *High automated kitting system for small parts: a case study from the Volvo Uddevalla plant*, Proceedings of the 23rd International Symposium on Automotive Technology and Automation, p.75–82, Vienna, Austria, 1990.

[18] I. Kovalenko. *Rare events in queueing theory. A survey*, Queueing systems, 16(1), p.1–49, 1994.

[19] V.S. Korolyuk and A.F. Turbin, *Mathematical foundations of the state lumping of large systems*, Naukova Dumka, Kiev, 1978, (in Russian), translated by Kluwer Aademic Publishers, Dordrecht, Boston, 1993.

[20] J.B. Lasserre. *A Formula for Singular Perturbations of Markov Chains Journal of Applied Probability*, 31(3), p.829–833, 1994.

[21] G. Latouche. *Queues with paired customers*, Journal of Applied Probability, 18(3), p.684–696, 1981.

[22] S. Meyn and R.L. Tweedie. *Markov Chains and Stochastic Stability, 2nd edition*, Cambridge University Press, 2009.

[23] R. Ramakrishnan and A. Krishnamurthy. *Analytical approximations for kitting systems with multiple inputs*, Asia-Pacific Journal of Operations Research, 25(2), p.187–216, 2008.

[24] R. Ramakrishnan and A. Krishnamurthy. *Performance evaluation of a synchronization station with multiple inputs and population constraints*, Computers & Operations Research, 39, p.560–570, 2012.

[25] M. Reiman and B. Simon. *Open queueing systems in light traffic*, Mathematics of operations research, 14(1), p.26–59, 1989.

[26] P. Som, W. Wilhelm and R. Disney. *Kitting process in a stochastic assembly system*, Queueing Systems, 17, p.471–490, 1994.

[27] P.J. Schweitzer. *Perturbation theory and finite Markov chains*, Journal of Applied Probability, 5(2), p.401–413, 1968.

[28] P.J. Schweitzer and G.W. Stewart. *The Laurent expansion of pencils that are singular at the origin*, Linear Algebra Appl., v.183, p. 237-254, 1993.

[29] M. Takahashi, H. Osawa and T. Fujisawa. *On a synchronization queue with two finite buffers*, Queueing Systems, 36, p.107–23, 2000.

9

Opinion propagation in medium-sized populations

Eline De Cuypere, Sabine Wittevrongel, Koen De Turck and Dieter Fiems

submitted to Performance Evaluation.

Abstract. We study the dynamics of opinion propagation in a medium-sized population with low population turnover. Opinion spreading is modelled by a Markovian non-standard Susceptible-Infected-Recovered (SIR) epidemic model with stochastic arrivals, departures, infections and recoveries. The system performance is evaluated by two complementary approaches: a numerical but approximate solution approach which relies on Maclaurin-series expansions of the stationary solution of the Markov process and a fluid limit approach. Both methods are evaluated numerically. Moreover, convergence to the fluid limit is proved, and explicit expressions for the fixed points of the differential equations are obtained for the case of linearly increasing infection and arrival rates.

9.1 Introduction

Given the rapid growth of companies in the internet sector that base their revenue model on advertisement (such as Google, Facebook etc.) [12] and the ascent of social networks in particular, the study of opinion spreading is a trending topic,

and there is a strong interest in understanding how new opinions spread through a community. Apart from these economic considerations, the analysis of opinion spreading can improve our comprehension of social relations among individuals, both online and offline.

This paper studies opinion propagation by drawing parallels with the spreading of diseases [3, 15]. Indeed, opinion propagation bears some similarity to the spread of an infectious disease, and particularly to Kermack and McKendrick's classical compartmental SIR model for such propagation [21]. The acronym SIR stands for susceptible (S), infectious (I) and recovered (R), and refers to the possible states that an individual can be in, the possible transitions between these states following the order $S \to I \to R$. In particular, the SIR model assumes that if a healthy individual encounters an infected individual, there is a chance that the healthy individual gets infected. An infected individual then recovers from the disease after some time, making him/her immune for the infection. This process can be directly reformulated in terms of the propagation of opinions on a particular topic: a susceptible or non-opinioned individual has yet to form an opinion on the topic, whereas infected or opinioned individuals do have such an opinion. Susceptible individuals may form an opinion when they encounter infected individuals. Finally, individuals loose their interest in the topic after some time and stop spreading their opinion. These individuals have recovered.

The SIR epidemic model has been predominantly applied to study disease, see e.g. [15, 16, 17, 19], but there is some prior work in the areas of rumour and opinion propagation as well. Concerning the latter, Zhao extends the SIR model for opinion spreading in social networks by including a hibernator state in which individuals temporary interrupt infecting others [33]. Moreno et al. study SIR like spreading of rumours explicitly accounting for the network structure [26]. Bettencourt et al. [4] draw parallels between epidemics and idea diffusion by applying several epidemiological models to empirical data. Woo et al. [31] show the plausibility to describe the mechanism of violent topic diffusion in web forums by using a SIR model while Fan et al. [13] propose an extended SIR model for opinion dynamics in which individuals can have a positive or negative opinion about a topic. Finally, some preliminary results on series expansion techniques (cfr. infra) for stochastic SIR models for opinion propagation were presented in [9].

SIR models are not the only models for studying opinion propagation in literature. For example, the threshold model starts from a random directed graph where each node selects a random threshold [23]. Opinion propagation then evolves deterministically: a node becomes active (gets an opinion) if the fraction of its active neighbours exceeds its threshold. In contrast to the threshold model, the dynamics of the voter model are stochastic: each node changes its state to the state of a random neighbour [24]. The Sznajd model [30] assumes more complex interactions

between nodes: a node and a neighbouring node are selected at random. If the neighbouring node is undecided, it adopts the opinion with some probability. If both have the same opinion, they try to convince the other neighbours. Finally, if they have different opinions, nothing happens. Specifically focussing on tweet propagation, Yang et al. [32] propose a factor graph model based on the analysis of the factors influencing the user's retweet behaviour. Kawamoto et al. [20] model the information diffusion as a random multiplicative process, with a particular focus on retweet behaviour. In [18], the traditional Susceptible-Infected-Susceptible (SIS) epidemic model is studied in order to predict retweeting trends.

In this paper, we focus on a compartmental Markovian SIR model for opinion spreading in medium-sized populations. While the total population of internet users easily qualifies as large, the size of an online community — say, of people contributing to an online forum or of people tweeting and retweeting some hash tag — is often not that large. Moreover, these communities hardly remain constant over time, with individuals joining and leaving all the time. Epidemics on medium-sized populations are also interesting from a mathematical point of view. Indeed, when the population is large, the fluid limit is accurate and the dynamics of the epidemic are described by a set of differential equations. In contrast, when the population size is small, the size of the state space of the epidemic Markov process is small such that the steady-state probability vector is easily calculated. For medium-sized populations, the fluid-limit is not yet accurate, while direct calculation of the stationary vector is computationally infeasible.

The contributions of this paper are twofold. First, we investigate the performance of opinion propagation in a Markovian framework by an approximate solution technique for Markov processes which relies on Maclaurin-series expansions of the steady-state probability vector. This technique was recently applied to study kitting processes [10]; a kitting process is a type of multi-buffer queueing system in which service is synchronised between the different buffers and temporarily blocked if one of the buffers is empty [8]. The epidemic process under consideration generalises the Markovian SIR process in various ways. We assume that the population size is bounded, but individuals join and leave the population over time to account for the dynamic formation of online communities. Moreover, the assumptions on infection and recovery rates are relaxed and individuals are allowed to move from susceptible to recovered directly as community members not necessarily want to spread the opinion to others. Secondly, in addition to the Maclaurin-series approach, we consider a fluid limit of the Markov process at hand and formally prove convergence. Fluid limits are a popular mathematical technique (see e.g. [11], [28]) which (when a good scaling is found) allow for focussing on the salient features of the stochastic process while discarding 'second-order fluctuations' around this main trend. In the present paper, it helps to make the link with more standard deterministic SIR models. We like to mention that the fluid scaling

under study (arrival rates and location capacity are sent to infinity), differs significantly and therefore complements the Maclaurin-series expansion limit (which holds for low departure rates). We thus aim to view this computationally cumbersome Markov model from different limiting cases, and gain new insights by combining them. We also note that the derivation of the fluid limit as performed in this paper also lends itself naturally to refinements in the form of diffusion results, but this is considered to be outside of the scope of the current paper.

The remainder of this paper is organised as follows. Section 9.2 introduces the opinion spreading model at hand as well as some particular examples, discussed further on. In Section 9.3, the balance equations are derived and the numerical series expansion approach is explained. Next, we find a fluid limit for the epidemic Markov process in Section 9.4. To illustrate both approaches, Section 9.5 considers various numerical examples. Finally, conclusions are drawn in Section 9.6.

9.2 Model description

We consider an opinion propagation system as depicted in Figure 9.1. There are at most L individuals in the community, each individual either being recovered (r), infected (i) or susceptible (s) (this particular ordering instead of the traditional s, i, r will prove useful for the Maclaurin analysis of Section 9.3). Let $X_k(t)$ be the number of individuals of type $k \in \mathcal{K} = \{r,i,s\}$ at time t, and let $\mathbf{X}(t) = (X_r(t), X_i(t), X_s(t)) \in \mathcal{L} = \{(x_r, x_i, x_s) \in \mathbb{N}^3 | x_i + x_r + x_s \leq L\}$. For any $\mathbf{x} \in \mathcal{L}$, x_k is the number of individuals of type $k \in \mathcal{K}$ and $||\mathbf{x}|| = |x_r| + |x_i| + |x_s| = x_r + x_i + x_s$ is the L_1 norm which corresponds to the total number of individuals. We consider a Markovian opinion propagation system, the number of individuals of the different types being the state of the Markov process. We make the following assumptions on the arrival, infection and recovery rates of the Markov process.

- For, $||\mathbf{X}(t)|| < L$, there is a new arrival of type $k \in \mathcal{K}$ in the interval $[t, t+dt)$ with probability $\lambda_k(\mathbf{X}(t))dt + o(dt)$. The total arrival rate in state $\mathbf{x} \in \mathcal{L}$ is denoted by $\lambda(\mathbf{x}) = \lambda_r(\mathbf{x}) + \lambda_i(\mathbf{x}) + \lambda_s(\mathbf{x})$. To simplify notation, assume $\lambda(\mathbf{x}) = \lambda_k(\mathbf{x}) = 0$ for $||\mathbf{x}|| \geq L$ and $k \in \mathcal{K}$.

- There is a departure of an individual of type $k \in \mathcal{K}$ in the interval $[t, t+dt)$ with probability $\mu X_k(t)dt + o(dt)$. Hence, the residence time of any individual is exponentially distributed with mean $1/\mu$.

- A single susceptible (infected, susceptible) individual gets infected (recovers, recovers) in the interval $[t, t+dt)$ with probability $\alpha_{si}(\mathbf{X}(t))dt + o(dt)$ ($\alpha_{ir}(\mathbf{X}(t))dt + o(dt)$, $\alpha_{sr}(\mathbf{X}(t))dt + o(dt)$). To simplify further notation, we assume $\alpha_{si}(\mathbf{x}) = \alpha_{sr}(\mathbf{x}) = 0$ for $x_s = 0$ and $\alpha_{ir}(\mathbf{x}) = 0$ for $x_i = 0$. There can be no infection or recovery if there are no individuals that can get infected or that can recover.

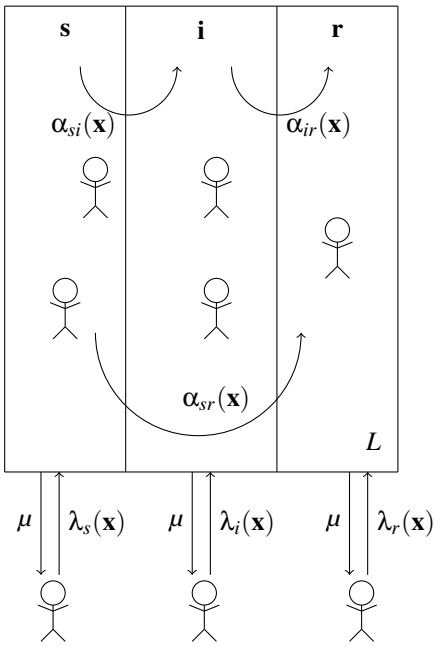


Figure 9.1: Opinion propagation model.

The Maclaurin-series expansion in Section 9.3 further requires that for every $\mathbf{x} \in \mathcal{L}$, (i) the total arrival rate $\lambda(\mathbf{x})$ is non-zero, (ii) the infection rate $\alpha_{si}(\mathbf{x})$ and the refusing rate $\alpha_{sr}(\mathbf{x})$ are non-zero for $x_s > 0$, and (iii) the recovery rate $\alpha_{ir}(\mathbf{x})$ is non-zero for $x_i > 0$.

We study this Markov model via its generator Q,

$$Qf(\mathbf{x}) := \lim_{t \to 0} \frac{1}{t} \left(\mathrm{E}[f(\mathbf{X}(t)) | X(0) = \mathbf{x}] - f(\mathbf{x}) \right),$$

for any bounded and measurable function $f : \mathcal{L} \to \mathbb{R}$. It is straightforward to deduce from the informal description above that

$$\begin{aligned} Qf(\mathbf{x}) = &\sum_{k \in \mathcal{K}} \lambda_k(\mathbf{x})[f(\mathbf{x}+\mathbf{e}_k) - f(\mathbf{x})] + \sum_{k \in \mathcal{K}} \mu\, x_k [f(\mathbf{x}-\mathbf{e}_k) - f(\mathbf{x})] \\ &+ \sum_{(j,k) \in \mathcal{K}^*} \alpha_{jk}(\mathbf{x})[f(\mathbf{x}-\mathbf{e}_j+\mathbf{e}_k) - f(\mathbf{x})], \end{aligned} \tag{9.1}$$

where $\mathbf{e}_r := [1,0,0]$, $\mathbf{e}_i := [0,1,0]$, $\mathbf{e}_s := [0,0,1]$ and $\mathcal{K}^* := \{(s,i),(i,r),(s,r)\}$.

Let $\pi(\cdot)$ denote the stationary distribution of the Markov process (which is guaranteed to exist as the state space is finite and uni-chain), and — for further use — let $\mathbf{X}$ denote a generic random variable distributed according to π. From the

above generator representation we derive the following set of balance equations:

$$\begin{aligned}
&\pi(\mathbf{x})\Big(||\mathbf{x}||\mu+\lambda(\mathbf{x})+\alpha_{si}(\mathbf{x})+\alpha_{ir}(\mathbf{x})+\alpha_{sr}(\mathbf{x})\Big)\\
&=\pi(x_r+1,x_i,x_s)\mu(x_r+1)+\pi(x_r,x_i+1,x_s)\mu(x_i+1)+\pi(x_r,x_i,x_s+1)\mu(x_s+1)\\
&+\pi(x_r-1,x_i,x_s)\lambda_r(x_r-1,x_i,x_s)+\pi(x_r,x_i-1,x_s)\lambda_i(x_r,x_i-1,x_s)\\
&+\pi(x_r,x_i,x_s-1)\lambda_s(x_r,x_i,x_s-1)+\pi(x_r,x_i-1,x_s+1)\alpha_{si}(x_r,x_i-1,x_s+1)\\
&+\pi(x_r-1,x_i+1,x_s)\alpha_{ir}(x_r-1,x_i+1,x_s)\\
&+\pi(x_r-1,x_i,x_s+1)\alpha_{sr}(x_r-1,x_i,x_s+1)\,,
\end{aligned}\tag{9.2}$$

for $\mathbf{x}\in\mathcal{L}$. Here and in the remainder, we follow the convention that $\pi(\mathbf{x})=0$ if $\mathbf{x}\notin\mathcal{L}$.

Prior to introducing the series expansions and fluid approximations, we first introduce two particular examples of the Markov process. In Section 9.5 where we present various numerical results, we return to these examples.

9.2.1 Constant infection and arrival rates

In the most basic setting, individuals arrive according to a Poisson process with the (state-independent) parameter λ_k, $k\in\mathcal{K}$. Furthermore, if we assume that each susceptible individual has a constant probability to get infected and to recover, i.e. $\alpha_{si}(\mathbf{x})=x_s\alpha_{si}$ and $\alpha_{sr}(\mathbf{x})=x_s\alpha_{sr}$, and that each infected individual has a constant probability to recover, i.e. $\alpha_{ir}(\mathbf{x})=x_i\alpha_{ir}$, then the mean number of each type in steady state can be calculated explicitly.

Indeed, as the departure rate of each individual is equal to μ, the total number of individuals $||\mathbf{X}||$ is distributed as the queue content of a classic $M/M/L/L$ queue, with arrival rate $\lambda=\lambda_r+\lambda_i+\lambda_s$ and departure rate μ, for which the steady-state distribution can be found in every queueing-theory textbook,

$$\mathbb{P}[||\mathbf{X}||=n]=\frac{\frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n}{\sum_{m=0}^{L}\frac{1}{m!}\left(\frac{\lambda}{\mu}\right)^m}\,.\tag{9.3}$$

Let p_k denote the probability that a random departing individual is of type $k\in\mathcal{K}$. An individual leaves the system as a susceptible individual, provided it arrived in the system as a susceptible individual and it leaves the system prior to infection or recovery. Hence, we have,

$$p_s=\frac{\lambda_s}{\lambda}\frac{\mu}{\alpha_{si}+\mu+\alpha_{sr}}\,,$$

the first factor being the probability that an arriving individual is susceptible and the second factor being the probability that a susceptible individual leaves prior to

infection or recovery. Analogously, an individual departs as an infected individual if (i) it arrives as susceptible, gets infected and does not recover or (ii) it arrives as infected and does not recover. Hence, we find,

$$p_i = \frac{\lambda_i}{\lambda}\frac{\mu}{\alpha_{ir}+\mu} + \frac{\lambda_s}{\lambda}\frac{\alpha_{si}}{(\alpha_{si}+\alpha_{sr}+\mu)}\frac{\mu}{(\alpha_{ir}+\mu)}\,.$$

Finally, an individual leaves as recovered if it does not leave the system as a susceptible or infected individual, hence we have,

$$\begin{aligned} p_r = 1 - p_i - p_s =& 1 - \frac{\lambda_s}{\lambda}\frac{\mu}{\alpha_{si}+\mu+\alpha_{sr}} - \frac{\lambda_i}{\lambda}\frac{\mu}{\alpha_{ir}+\mu} \\ & - \frac{\lambda_s}{\lambda}\frac{\alpha_{si}}{(\alpha_{si}+\alpha_{sr}+\mu)}\frac{\mu}{(\alpha_{ir}+\mu)}\,. \end{aligned}$$

Note that the rate at which an individual leaves the system does not depend on the type of the individual. Therefore p_k is the probability that a random individual in the system is of type k. Moreover, individuals change type, independent of other individuals. Hence, the distribution of the number of individuals of the different types, conditioned on the total number of individuals in the system, is a multinomial distribution with parameters p_r, p_i and p_s. Combining this observation with equation (9.3), yields,

$$\mathbb{P}[\mathbf{X}=\mathbf{x}] = \mathbb{P}[||\mathbf{X}|| = ||\mathbf{x}||]\,\frac{||\mathbf{x}||!}{x_r!\,x_i!\,x_s!}\,p_r^{x_r}\,p_i^{x_i}\,p_s^{x_s} = \frac{\left(\frac{\lambda}{\mu}\right)^{||\mathbf{x}||}\frac{p_r^{x_r}\,p_i^{x_i}\,p_s^{x_s}}{x_r!\,x_i!\,x_s!}}{\sum_{m=0}^{L}\frac{1}{m!}\left(\frac{\lambda}{\mu}\right)^m}\,,$$

and,

$$\mathbb{P}[X_k=n] = \sum_{\ell=n}^{L}\binom{\ell}{n}\mathbb{P}[||\mathbf{X}||=\ell]\,p_k^n\,(1-p_k)^{\ell-n} = \frac{\sum_{\ell=n}^{L}\left(\frac{\lambda}{\mu}\right)^{\ell}\frac{p_k^n}{n!}\frac{(1-p_k)^{\ell-n}}{(\ell-n)!}}{\sum_{m=0}^{L}\frac{1}{m!}\left(\frac{\lambda}{\mu}\right)^m}\,.$$

9.2.2 A model for opinion spreading

The following example is more complex and does not have a simple solution. We will rely on either a Maclaurin-series approach or on a fluid limit to estimate performance.

We propose the following infection and recovery rates. It is reasonable to assume that non-opinioned individuals are more likely to form an opinion if there are more opinioned individuals. Therefore, we assume that the infection rate of susceptible individuals is an affine function of the number of infected individuals,

$$\alpha_{si}(\mathbf{x}) = \left(\alpha_{si}^0 + \alpha_{si}^1 x_i\right)x_s\,.$$

Moreover, the effects of other individuals on shifting to neutral (recovered) are neglected. Therefore the rate at which non-opinioned and opinioned individuals shift to neutral are constant, which implies,

$$\alpha_{ir}(\mathbf{x}) = \alpha_{ir}^0 x_i\,, \quad \alpha_{sr}(\mathbf{x}) = \alpha_{sr}^0 x_s\,.$$

Finally, having many individuals in the community, is likely to attract new opinioned and non-opinioned individuals. Hence, we assume that the arrival rates of the susceptible and infected individuals are affine functions of the total number of individuals,

$$\lambda_i(\mathbf{x}) = \lambda_i^0 + \lambda_i^1 ||\mathbf{x}||\,, \quad \lambda_s(\mathbf{x}) = \lambda_s^0 + \lambda_s^1 ||\mathbf{x}||\,,$$

whereas neutral individuals arrive at a constant rate,

$$\lambda_r(\mathbf{x}) = \lambda_r^0\,.$$

Note that the assumptions above are a generalisation of Kermack and McKendrick's SIR model. In Kermack and McKendrick's setting, the population is fixed ($\lambda_i^0 = \lambda_i^1 = \lambda_s^0 = \lambda_s^1 = \lambda_r^0 = \mu = 0$), the infection rate is proportional to the number of infected individuals ($\alpha_{si}^0 = 0$), and individuals do not recover without being infected ($\alpha_{sr}^0 = 0$).

9.3 Maclaurin-series expansions

While the system of equations (9.2) is easily solved when the maximum number of individuals is limited, the state space size already explodes for relatively small L and a direct solution is computationally infeasible. Indeed, numerical computation of the steady-state vector has an asymptotic time complexity of $O(M^3)$, where $M = \binom{L+3}{3} \sim L^3/6$ is the size of the state space. We introduce the numerical Maclaurin-series approach in generic terms in the subsection below and survey related work in Subsection 9.3.2. We then tailor the method to the Markov process at hand in Subsection 9.3.3 and derive performance measures in Subsection 9.3.4.

9.3.1 Methodology

Notice that the generator matrix of the Markov process at hand can be decomposed as follows,

$$Q_\mu = \mathbf{Q}^{(\mathbf{0})} + \mu \mathbf{Q}^{(\mathbf{1})}\,,$$

with neither $\mathbf{Q}^{(\mathbf{0})}$ nor $\mathbf{Q}^{(\mathbf{1})}$ depending on μ. Moreover, $\mathbf{Q}^{(\mathbf{0})}$ is still a proper generator matrix, it is the generator matrix of the Markov process when there are no departures. As the size of the state space does not allow for a direct calculation of $\boldsymbol{\pi}$ as the normalised solution of,

$$\boldsymbol{\pi} Q_\mu = 0\,,$$

we introduce the Maclaurin-series expansion of the steady state vector in μ,

$$\boldsymbol{\pi} = \sum_{n=0}^{\infty} \mu^n \boldsymbol{\pi}^{(n)}.$$

In order for such an expansion to make sense, the vector $\boldsymbol{\pi}$ is required to be analytic in a neighbourhood of $\mu = 0$. For finite state spaces (in contrast to infinite ones, see e.g. [2, 14]), this is fairly easy to establish. Finding the steady-state distribution is in this case essentially a finite-dimensional eigenproblem. If a matrix depends analytically on a parameter, then the corresponding eigenvalues and eigenvectors are also analytic in case of null-space perturbation [1]. Another possible path towards proving analyticity is via V-uniform ergodicity of the unperturbed Markov process with generator $\mathbf{Q}^{(\mathbf{0})}$ (see a.o [2]), which is equivalent to the existence of a spectral gap (the distance between eigenvalue 0 of the generator matrix $\mathbf{Q}^{(\mathbf{0})}$ and the eigenvalue that is its nearest neighbour). For finite Markov processes, there is a spectral gap as long as there is only one recurrent class for $\mu = 0$. Hence, provided that $\mathbf{Q}^{(\mathbf{0})}$ is a generator matrix with one recurrent class (this condition is also denoted as 'regular perturbation', as opposed to 'singular perturbation'), the expansion makes sense and $\boldsymbol{\pi}^{(0)}$ is the normalised solution of the equation,

$$\boldsymbol{\pi}^{(0)} \mathbf{Q}^{(\mathbf{0})} = 0\,.$$

Every subsequent term $\boldsymbol{\pi}^{(n)}$ can be found by identifying equal powers of μ in the equation,

$$\sum_{n=0}^{\infty} \mu^n \boldsymbol{\pi}^{(n)} (\mathbf{Q}^{(\mathbf{0})} + \mu \mathbf{Q}^{(\mathbf{1})}) = 0\,,$$

which leads to the following equation for $\boldsymbol{\pi}^{(n+1)}$,

$$\boldsymbol{\pi}^{(n+1)} \mathbf{Q}^{(\mathbf{0})} = -\boldsymbol{\pi}^{(n)} \mathbf{Q}^{(\mathbf{1})}\,.$$

The normalisation condition of $\boldsymbol{\pi}$ implies that the elements of $\boldsymbol{\pi}^{(n)}$ sum to 0 for each $n > 0$. This fact and the former equation allows for recursively solving the terms in the expansion.

As we now have to solve a linear system of equations for each term $\boldsymbol{\pi}^{(n)}$ in the expansion, plus an additional vector-matrix product, it appears that we have not gained very much. However, if we impose the extra condition that $\mathbf{Q}^{(\mathbf{0})}$ is triangular for some ordering of the state space (either upper or lower triangular), then the resulting linear systems of equations can be solved by backward substitutions, with considerably reduced computational complexity. As a worst case, its computation time is $O(M^2)$. However, the number of transitions from a state is typically much smaller than the state space, such that the computation time typically is $O(M)$.

9.3.2 Related work

Prior to applying the series expansion method to the Markov model at hand, we survey work on series expansions of stochastic models. For an overview, we refer the reader to [22] and [5]. The first work seems to be by Schweitzer in 1968 [29]. Ever since, it has been applied in many forms and flavours, and is known under various names such as perturbations, light-traffic expansions, Taylor-series expansions and so on. This technique is in principle not confined to the Markov framework (see e.g. [6], which utilises Palm theory), although many interesting examples indeed fall within this framework.

There are roughly three methods to establish series expansions of stochastic models. The first makes use of the direct derivation sketched above and forms the basis of the computational method that we propose in this paper and will evaluate in the next subsection. The second makes use of sample-path arguments. Consider the case that μ denotes a particular event rate. For example, for light-traffic approximations, μ denotes the arrival rate; in the worked-out example of Section 9.3, the parameter denotes the service rate, and hence constitutes a 'low-service rate' approximation. An important result for this strand of research is what we can call the *n events rule*, which states that for an nth order expansion, only sample paths with n or fewer of such events must be considered. This can be intuited from the non-rigorous reasoning that a sample path containing n such events has a probability of order μ^n. However due to the fact that the number of sample paths is uncountable and thus the probability of every individual path is zero, making this rigorous is non-trivial. For series expansions revolving around a Poisson process with a small rate, to which the examples in this work essentially belong, this was made rigorous by Reiman and Simon [27]. Important work extending this to a Palm calculus context was presented in [6].

The third approach to series expansions is via the following updating formula, which has been established in general Markov settings, see eg. [14]:

$$\boldsymbol{\pi} = \boldsymbol{\pi}_0 \sum_{k=0}^{\infty} [\mu \mathbf{Q}^{(\mathbf{1})} D]^k.$$

where D denotes the deviation matrix of $\mathbf{Q}^{(\mathbf{0})}$. In this case, a successful application revolves around finding this deviation matrix D, defined as follows:

$$D = \int_0^{\infty} ([P_0(t)]_{ij} - \Pi_0) dt\,, \tag{9.4}$$

with $P_0(t)$ the Markov semigroup of the continuous-time Markov process and $\Pi_0 = \lim_{t\to\infty} P_0(t)$.

As the matrix D is closely related to Poisson's equation for Markov processes, this technique is sometimes also denoted as such [25]. Note that the matrix D pertains to the unperturbed Markov process, so that in this updating formula we

see another justification for the n events rule. Indeed, as the events are in fact nothing else than the transitions recorded in $\mathbf{Q}^{(\mathbf{1})}$, transitions which do not occur in $\mathbf{Q}^{(\mathbf{0})}$ and hence nor in D, it follows that in the vector $\boldsymbol{\pi}^{(n)} = \boldsymbol{\pi}^{(0)}(\mathbf{Q}^{(\mathbf{1})}D)^n$, only those states that can be reached with n events (or less) can be non-zero. To the best of our knowledge, a rigourous identification of the sample-path method and the updating formula has not yet been attempted.

9.3.3 Application

In view of the method described above, let $\pi_n(\mathbf{x})$ be the nth component of the expansion,

$$\pi(\mathbf{x}) = \sum_{n=0}^{\infty} \pi_n(\mathbf{x})\mu^n\,,$$

for $\mathbf{x} \in \mathcal{L}$. Substituting the former expression in the balance equations (9.2), we get

$$\begin{aligned}
&\sum_{n=0}^{\infty} \pi_n(\mathbf{x})\mu^n \Big(||\mathbf{x}||\mu + \lambda(\mathbf{x}) + \alpha_{si}(\mathbf{x}) + \alpha_{ir}(\mathbf{x}) + \alpha_{sr}(\mathbf{x}) \Big) \\
&= \mathbf{1}_{\{||\mathbf{x}||<L\}} \Bigg(\sum_{n=0}^{\infty} \pi_n(x_r+1,x_i,x_s)\mu^{n+1}(x_r+1) + \sum_{n=0}^{\infty} \pi_n(x_r,x_i+1,x_s)\mu^{n+1}(x_i+1) \\
&+ \sum_{n=0}^{\infty} \pi_n(x_r,x_i,x_s+1)\mu^{n+1}(x_s+1) \Bigg) + \sum_{n=0}^{\infty} \pi_n(x_r-1,x_i,x_s)\lambda_r(x_r-1,x_i,x_s)\mu^n \\
&+ \sum_{n=0}^{\infty} \pi_n(x_r,x_i-1,x_s)\lambda_i(x_r,x_i-1,x_s)\mu^n + \sum_{n=0}^{\infty} \pi_n(x_r,x_i,x_s-1)\lambda_s(x_r,x_i,x_s-1)\mu^n \\
&+ \sum_{n=0}^{\infty} \pi_n(x_r,x_i-1,x_s+1)\alpha_{si}(x_r,x_i-1,x_s+1)\mu^n \\
&+ \sum_{n=0}^{\infty} \pi_n(x_r-1,x_i+1,x_s)\alpha_{ir}(x_r-1,x_i+1,x_s)\mu^n \\
&+ \sum_{n=0}^{\infty} \pi_n(x_r-1,x_i,x_s+1)\alpha_{sr}(x_r-1,x_i,x_s+1)\mu^n\,. \qquad (9.5)
\end{aligned}$$

For $\mathbf{x} \in \mathcal{L}^* = \mathcal{L} \setminus \{(L,0,0)\}$, comparison of the terms in μ^0 on both sides of the former equation yields,

$$\begin{aligned}
&\pi_0(\mathbf{x}) \Big(\lambda(\mathbf{x}) + \alpha_{si}(\mathbf{x}) + \alpha_{ir}(\mathbf{x}) + \alpha_{sr}(\mathbf{x}) \Big) \\
&= \sum_{n=0}^{\infty} \pi_0(x_r-1,x_i,x_s)\lambda_r(x_r-1,x_i,x_s) + \sum_{n=0}^{\infty} \pi_0(x_r,x_i-1,x_s)\lambda_i(x_r,x_i-1,x_s) \\
&+ \sum_{n=0}^{\infty} \pi_0(x_r,x_i,x_s-1)\lambda_s(x_r,x_i,x_s-1)
\end{aligned}$$

$$
\begin{aligned}
&+\sum_{n=0}^{\infty}\pi_0(x_r,x_i-1,x_s+1)\alpha_{si}(x_r,x_i-1,x_s+1)\\
&+\sum_{n=0}^{\infty}\pi_0(x_r-1,x_i+1,x_s)\alpha_{ir}(x_r-1,x_i+1,x_s)\\
&+\sum_{n=0}^{\infty}\pi_0(x_r-1,x_i,x_s+1)\alpha_{sr}(x_r-1,x_i,x_s+1)\,.
\end{aligned}
$$

Evaluating this expression in lexicographical order shows,

$$\pi_0(\mathbf{x}) = 0\,, \tag{9.6}$$

for $\mathbf{x} \in \mathcal{L} \setminus \{(L,0,0)\}$. By the normalisation condition of $\boldsymbol{\pi}_0$ we further get π_0 $(L,0,0) = 1$. This result is not unexpected. In the absence of departures, the population size reaches its boundary and no new arrivals are possible. Moreover, every individual in the population will recover after some time, such that there are but recovered individuals.

Comparison of the terms in μ^n for $n > 0$ in equation (9.5) gives,

$$
\begin{aligned}
\pi_n(\mathbf{x}) = \frac{1}{\Delta(\mathbf{x})}\Bigg(&\mathbf{1}_{\{||\mathbf{x}||<L\}}\Big(\pi_{n-1}(x_r+1,x_i,x_s)(x_r+1)+\pi_{n-1}(x_r,x_i+1,x_s)(x_i+1)\\
&+\pi_{n-1}(x_r,x_i,x_s+1)(x_s+1)\Big)+\pi_n(x_r-1,x_i,x_s)\lambda_r(x_r-1,x_i,x_s)\\
&+\pi_n(x_r,x_i-1,x_s)\lambda_i(x_r,x_i-1,x_s)+\pi_n(x_r,x_i,x_s-1)\lambda_s(x_r,x_i,x_s-1)\\
&+\pi_n(x_r,x_i-1,x_s+1)\alpha_{si}(x_r,x_i-1,x_s+1)\\
&+\pi_n(x_r-1,x_i+1,x_s)\alpha_{ir}(x_r-1,x_i+1,x_s)\\
&+\pi_n(x_r-1,x_i,x_s+1)\alpha_{sr}(x_r-1,x_i,x_s+1)-\pi_{n-1}(\mathbf{x})||\mathbf{x}||\Bigg)\,,
\end{aligned}
$$

for $\mathbf{x} \in \mathcal{L} \setminus \{(L,0,0)\}$ with,

$$\Delta(\mathbf{x}) = \Big(\lambda(\mathbf{x})+\alpha_{si}(\mathbf{x})+\alpha_{ir}(\mathbf{x})+\alpha_{sr}(\mathbf{x})\Big)\,.$$

As detailed in [10], we can use the above equation to compute new terms very efficiently, by iterating over the state space in lexicographic order, as on the RHS only entries of either order $n-1$ or lexicographically smaller entries of order n are present. Moreover, the assumptions on the arrival process assure that $\Delta(\mathbf{x}) > 0$ for all $\mathbf{x} \in \mathcal{L} \setminus \{(L,0,0)\}$. As for the 0th order term, the normalisation condition is used to find the nth order expansion of $\pi(L,0,0)$,

$$\pi_n(L,0,0) = -\sum_{\mathbf{x}}\pi_n(\mathbf{x})\,.$$

9.3.4 Performance measures

Once the series expansion of the steady-state distribution has been obtained, the expansion of various performance measures directly follows. Let $\mathbf{X} \sim \pi$, then for a performance measure $J = \mathrm{E}[f(\mathbf{X})]$, we have

$$J = \sum_{\mathbf{x}\in\mathcal{L}} f(\mathbf{x})\pi(\mathbf{x}) = \sum_{\mathbf{x}\in\mathcal{L}} f(\mathbf{x}) \sum_{n=0}^{\infty} \pi_n(\mathbf{x})\mu^n = \sum_{n=0}^{\infty}\sum_{\mathbf{x}\in\mathcal{L}} f(\mathbf{x})\pi_n(\mathbf{x})\mu^n = \sum_{n=0}^{\infty} J_n\mu^n\,, \quad (9.7)$$

with

$$J_n = \sum_{\mathbf{x}\in\mathcal{L}} f(\mathbf{x})\pi_n(\mathbf{x})\,.$$

The interchange of the summations is justified by the finiteness of $\mathcal{L}$ and the convergence of $\sum_n \pi_n(\mathbf{x})\mu^n$ for all $\mathbf{x} \in \mathcal{L}$. As such, any term J_n in the expansion of a performance measure J can be calculated from the corresponding vector $\boldsymbol{\pi}_n$ of the expansion of the steady-state vector $\boldsymbol{\pi}$. Performance measures of interest include amongst others the jth order moment of the number of individuals of type $k \in \mathcal{K}$ ($f(\mathbf{x}) = {x_k}^j$).

9.4 Fluid limit

In this section, we develop a fluid limit for the model described in this contribution, relying on the monograph of Ethier and Kurtz [11], and on the survey article by Darling and Norris [7].

9.4.1 Convergence for the generic epidemic process

Let $\{X_r^\varepsilon(t), X_i^\varepsilon(t), X_s^\varepsilon(t)\}$ denote the continuous-time Markov process indexed by the scaling parameter ε, which affects the system in the following way: step sizes are scaled by ε, whereas the transition rates are scaled by ε^{-1}. We assume that the Markov process takes values in a compact set $U \subset \mathbb{R}^3$, where $U := \{\mathbf{x} \in \mathbb{R}^3 : \mathbf{x} \geq \mathbf{0}, ||\mathbf{x}|| \leq L\}$.

In particular, the generator of the scaled process is as follows,

$$Q_\varepsilon f(\mathbf{x}) = \sum_{k\in\mathcal{K}} \varepsilon^{-1}\bar{\lambda}_k(\mathbf{x})[f(\mathbf{x}+\varepsilon\mathbf{e}_k) - f(\mathbf{x})] + \sum_{k\in\mathcal{K}} \varepsilon^{-1}\mu x_k[f(\mathbf{x}-\varepsilon\mathbf{e}_k) - f(\mathbf{x})] \quad (9.8)$$

$$+ \sum_{(j,k)\in\mathcal{K}^*} \varepsilon^{-1}\bar{\alpha}_{jk}(\mathbf{x})[f(\mathbf{x}-\varepsilon\mathbf{e}_j+\varepsilon\mathbf{e}_k) - f(\mathbf{x})], \quad (9.9)$$

for suitable functions $\bar{\lambda}_k(\cdot)$ and $k \in \mathcal{K}$, $\bar{\alpha}_{jk}(\cdot)$, $(j,k) \in \mathcal{K}^*$, which we require to be Lipschitz continuous on U. Note that this means in particular that the arrival rates $\bar{\lambda}_k(\cdot)$ must not go discontinuously to zero at the border ∂U of U, and therefore

we smoothen these functions in a certain manner. In terms of numerical results, the exact manner with which these are smoothened is not important, as we have found that the fluid limit gives the best results when the system steers clear from the boundary as $\varepsilon \to 0$, that is

$$\mathbb{P}[d(\mathbf{X}^{\varepsilon}(t), \partial U) < \delta] \to 0,$$

where $d(\cdot,\cdot)$ denotes the Euclidean distance.

Let us introduce the transition rates $q_{\varepsilon}(\mathbf{x},\mathbf{x}')$ corresponding to Q_{ε} in the obvious manner. By 'Tayloring', we find that

$$Q_{\varepsilon} f(\mathbf{x}) \to \sum_{k\in\mathcal{K}} \bar{\lambda}_k(\mathbf{x})\partial_{x_k} f(\mathbf{x}) - \sum_{k\in\mathcal{K}} \mu x_k \partial_{x_k} f(\mathbf{x}) \tag{9.10}$$

$$+ \sum_{(j,k)\in\mathcal{K}^*} \bar{\alpha}_{jk}(\mathbf{x})[-\partial_{x_j} + \partial_{x_k}] f(\mathbf{x}) + O(\varepsilon^2). \tag{9.11}$$

Note that the limit is exact when f is a linear function, as the second order derivatives of f, which feature in the $O(\varepsilon^2)$ term, are evidently all zero. It is well-known that a generator of this type has a deterministic solution that can be formulated in terms of the following system of (non-linear) differential equations:

$$\begin{aligned}
\dot{x}_s(t) &= \lambda_s(\mathbf{x}(t)) - \bar{\alpha}_{si}(\mathbf{x}(t)) - \bar{\alpha}_{sr}(\mathbf{x}(t)) - \mu x_s(t);\\
\dot{x}_i(t) &= \lambda_i(\mathbf{x}(t)) + \bar{\alpha}_{si}(\mathbf{x}(t)) - \bar{\alpha}_{ir}(\mathbf{x}(t)) - \mu x_i(\tau);\\
\dot{x}_r(t) &= \lambda_r(\mathbf{x}(t)) + \bar{\alpha}_{ir}(\mathbf{x}(t)) + \bar{\alpha}_{sr}(\mathbf{x}(t)) - \mu x_r(t)\,,
\end{aligned}$$

which has in general no closed-form solution but can be solved efficiently with a suitable numerical procedure for differential equations. Let us denote this system of equations in shorthand as $\dot{\mathbf{x}} = \mathbf{b}(\mathbf{x})$, for a suitably defined vector field $\mathbf{b}(\cdot)$ on U, with Lipschitz constant K.

We show an error bound for the fluid limit using Proposition 4.2 from [7]. Consider the following events:

$$\begin{aligned}
\Omega_0 &:= \left\{|\mathbf{X}^{\varepsilon}(0) - \mathbf{x}(0)| \le \delta\right\},\\
\Omega_1 &:= \left\{\int_0^{t_0} |\beta(\mathbf{X}^{\varepsilon}(t)) - \mathbf{b}(\mathbf{x}(0))| \le \delta\right\},\\
\Omega_2 &:= \left\{\int_0^{t_0} \gamma(\mathbf{X}^{\varepsilon}(t)) \le A(\varepsilon)t_0\right\}.
\end{aligned}$$

where

$$\beta(\mathbf{x}) = \sum_{\mathbf{x}'} q(\mathbf{x},\mathbf{x}')(\mathbf{x}' - \mathbf{x}),$$

and

$$\gamma(\mathbf{x}) = \sum_{\mathbf{x}'} q(\mathbf{x},\mathbf{x}')|\mathbf{x}' - \mathbf{x}|^2.$$

We have per Proposition 4.2 that

$$\mathbb{P}\left(\sup_{t\le t_0}|\mathbf{X}^\varepsilon(t)-\mathbf{x}|>\kappa(\varepsilon)\right)\le 4A(\varepsilon)t_0/\delta^2+\mathbb{P}(\Omega_0^c\cup\Omega_1^c\cup\Omega_2^c).$$

where $\delta=\kappa(\varepsilon)e^{-Kt_0}/3$.

A simple calculation shows that $\mathbb{P}(\Omega_1^c)=0$ and also $\mathbb{P}(\Omega_2^c)=0$ if we choose

$$\varepsilon^{-1}A(\varepsilon)=\sum_{k\in\mathcal{K}}[\max_{\mathbf{x}}\bar{\lambda}_k(\mathbf{x})+\max_{\mathbf{x}}\mu x_k]+2\max_{(j,k)\in\mathcal{K}^*}\bar{\alpha}_{jk}(\mathbf{x}).$$

If we choose $\kappa(\varepsilon)$ such that $\kappa(\varepsilon)^2/\varepsilon\to 0$, and if $\mathbf{X}^\varepsilon(0)$ to $\mathbf{x}(0)$ then we have indeed convergence to the fluid limit.

9.4.2 Equilibrium points for the opinion spreading model

In this section, we calculate the equilibrium points for the opinion spreading model introduced in Subsection 9.2.2 by solving $\mathbf{b}(\mathbf{x})=\mathbf{0}$. We have

$$0=\left(\lambda_s^0+\lambda_s^1||\mathbf{x}||\right)-\left(\alpha_{si}^0+\alpha_{si}^1x_i\right)x_s-\alpha_{sr}^0x_s-\mu x_s\,; \tag{9.12}$$

$$0=\left(\lambda_i^0+\lambda_i^1||\mathbf{x}||\right)+\left(\alpha_{si}^0+\alpha_{si}^1x_i\right)x_s-\alpha_{ir}^0x_i-\mu x_i\,; \tag{9.13}$$

$$0=\lambda_r^0+\alpha_{ir}^0x_i+\alpha_{sr}^0x_s-\mu x_r\,. \tag{9.14}$$

where $||\mathbf{x}||=x_r+x_i+x_s$. By subtracting (9.12) from (9.13), we can eliminate the quadratic part in $\bar{\alpha}_{si}(\mathbf{x})$. Then, we find solutions for x_i and x_r by solving this new linear equation together with equation (9.14). These solutions are substituted in equation (9.12) such that we get the following quadratic equation in x_s:

$$\begin{aligned}
f(x_s)=\;&x_s^2\frac{\alpha_{si}^1}{a}\left(-\mu^2+\lambda_s^1\alpha_{sr}^0+\lambda_i^1\alpha_{sr}^0-\mu\alpha_{sr}^0+\lambda_s^1\mu+\lambda_i^1\mu\right)\\
&+x_s\left(\frac{1}{a}\left(-(\lambda_s^1)^2\mu+\lambda_s^1\mu^2-(\lambda_s^1)^2\alpha_{ir}^0-\lambda_s^1\lambda_i^1\mu\right.\right.\\
&+\lambda_s^1\alpha_{ir}^0\mu-\lambda_s^1\alpha_{ir}^0\lambda_i^1+\alpha_{si}^1\lambda_s^1\lambda_r^0+\alpha_{si}^1\lambda_i^1\lambda_r^0+\alpha_{si}^1\mu\lambda_s^0\\
&\left.\left.+\alpha_{si}^1\mu\lambda_i^0\right)+\lambda_s^1-\alpha_{sr}^0-\mu-\alpha_{si}^0\right)\\
&+\left(\frac{\lambda_s^1}{a}\left(-\mu\lambda_s^0-\mu\lambda_i^0-\alpha_{ir}^0\lambda_s^0-\mu\lambda_r^0-\alpha_{ir}^0\lambda_i^0-\alpha_{ir}^0\lambda_r^0\right)+\lambda_s^0\right)
\end{aligned}$$

with $a=\alpha_{ir}^0\lambda_s^1+\alpha_{ir}^0\lambda_i^1+\lambda_s^1\mu+\lambda_i^1\mu-\alpha_{ir}^0\mu-\mu^2$.

To examine the stability of the system, we define the Jacobian matrix at each of the equilibrium points from the above quadratic equation. We have

$$J = \begin{bmatrix} \frac{\mathrm{d}F_s}{\mathrm{d}x_s} & \frac{\mathrm{d}F_i}{\mathrm{d}x_s} & \frac{\mathrm{d}F_r}{\mathrm{d}x_s} \\ \frac{\mathrm{d}F_s}{\mathrm{d}x_i} & \frac{\mathrm{d}F_i}{\mathrm{d}x_i} & \frac{\mathrm{d}F_r}{\mathrm{d}x_i} \\ \frac{\mathrm{d}F_s}{\mathrm{d}x_r} & \frac{\mathrm{d}F_i}{\mathrm{d}x_r} & \frac{\mathrm{d}F_r}{\mathrm{d}x_r} \end{bmatrix}$$

where F_s, F_i and F_r are respectively equal to equations (9.12), (9.13) and (9.14). We get

$$J = \begin{bmatrix} \lambda_s^1 - \alpha_{si}^0 - \alpha_{si}^1 x_i - \mu & \lambda_i^1 + \alpha_{si}^0 + \alpha_{si}^1 x_i & \alpha_{sr} \\ \lambda_s^1 - \alpha_{si}^1 x_s & \lambda_i + \alpha_{si}^1 x_s - \alpha_{ir} - \mu & \alpha_{ir} \\ \lambda_s^1 & \frac{\lambda_i}{L} & -\mu \end{bmatrix}. \tag{9.15}$$

Note that if all eigenvalues have a negative real part, the equilibrium point is stable, otherwise the equilibrium point is unstable. Numerical results will be given in the next section.

9.5 Numerical results

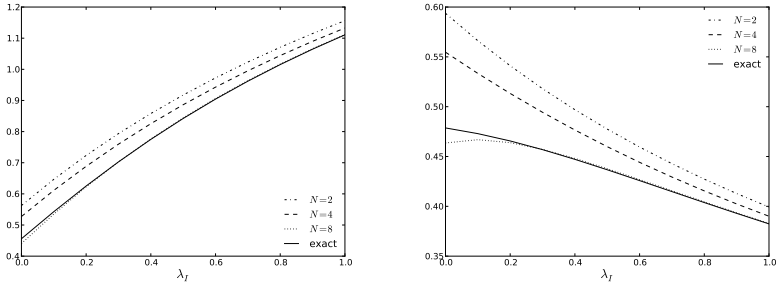
To illustrate our numerical approach, we now assess the accuracy of the series expansion technique and the fluid limit by means of several numerical examples.

9.5.1 Constant infection and arrival rates

First, consider the first example as described in Subsection 9.2.1. Recall, that for this particular example, the solution can be calculated explicitly. We here compare the accuracy of the series expansion with the exact result. To this end, Figures 9.2(a) and 9.2(b) depict the mean number of infected and susceptible individuals, respectively, versus the arrival rate of infected individuals λ_i. The maximum population size is $L = 50$ and we further assume $\lambda_r = \lambda_s = 1$ and $\alpha_{ir} = \alpha_{si} = \alpha_{sr} = 1$. Moreover, the departure rate μ is set to 0.05. The exact result is compared with the approximation by an Nth order expansion in μ for $N = 2$, $N = 4$ and $N = 8$. For $N = 8$, we observe that the approximation is accurate, apart from a slight deviation for small λ_i. In contrast, for $N = 2$ and $N = 4$, the results of the series expansions are clearly not accurate for the considered parameter settings.

To establish the regions in which the results of the series expansion are accurate enough, we propose a simple heuristic which compares the Nth and the $2N$th order expansions. Let $f_N(\mu)$ be the Nth order expansion in μ, we then accept our Nth order approximation provided if

$$\left| \frac{f_{2N}(\mu) - f_N(\mu)}{f_{2N}(\mu)} \right| < \varepsilon, \tag{9.16}$$



(a) (b)

Figure 9.2: Mean number of infected (a) and susceptible (b) individuals.

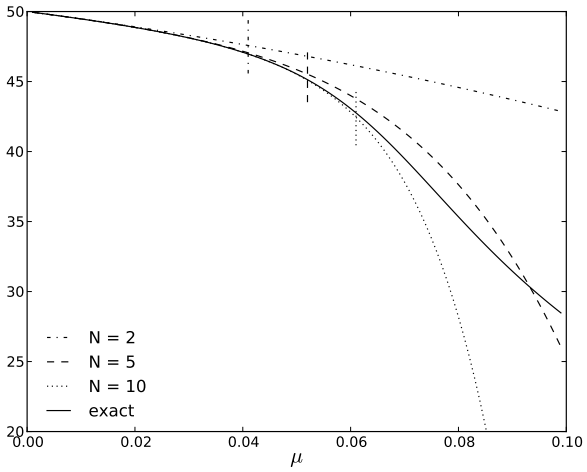


Figure 9.3: Mean number of recovered individuals.

or equivalently,

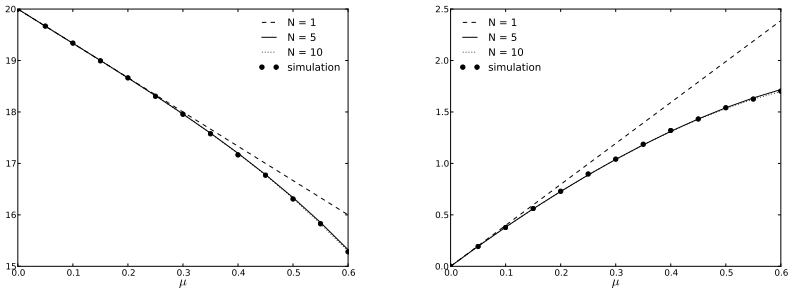
$$1-\varepsilon < \left| \frac{f_N(\mu)}{f_{2N}(\mu)} \right| < 1+\varepsilon. \tag{9.17}$$

Let Ω_N denote the region where these inequalities hold. In Figure 9.3, we divide the region where these inequalities hold and do not hold for $N = \{2,5,10\}$ with $\varepsilon = 0.01$ by means of a line. The same parameter settings as in Figures 9.2(a) and 9.2(b) are considered and the arrival rate of infected individuals is assumed to be equal to 3. As can be observed, the regions for which the inequality of the heuristic hold go up to $\mu = 0.041$ for $N = 2$, up to $\mu = 0.052$ for $N = 5$ and up to $\mu = 0.061$ for $N = 10$. Comparing the results of the approximation method with the exact result, we observe that the performance assessment is accurate in the heuristically determined region.

9.5.2 Opinion spreading model

We now consider the opinion spreading model as described in Section 9.2.2. In this case, the exact solution cannot be calculated and we rely on simulation to assess the accuracy of our results. Figures 9.4(a), 9.4(b) and 9.5 depict respectively the mean number of recovered individuals, the mean number of infected individuals and the mean number of susceptible individuals versus the lifetime rate μ varying from 0 to 0.6. Moreover, the maximum population size equals $L = 20$, the arrival rates λ_k^0 where $k = \{r,i,s\}$ and λ_i^1 where $i = \{i,s\}$ are respectively equal to 3 and 0.1 and the rates at which an individual changes of type, α_{si}^0, α_{si}^1,α_{sr}^0 and α_{ir}^0, are equal to 3. Series expansions of various orders N are depicted as indicated ($N = 1$, 5, 10), as well as simulation results. As expected, the mean number of recovered individuals decreases and the mean number of infected and susceptible individuals increase as the departure rate increases. Moreover, for $\mu = 0$, the population consists only of recovered individuals as their lifetime is infinite such that all individuals recover eventually. As the figures show, the approximation for $N = 5$ is already accurate for the mean number of recovered, infected and susceptible individuals.

We also consider the fluid approximation of the opinion spreading model. In Figure 9.6(a), we depict the stable equilibrium point of the quadratic equations given in Subsection 9.4.2. The parameter settings are the same as in Figure 9.4 and 9.5 but for a varied λ_i^1 from 0 to 1 and a constant $\mu = 0.7$. By evaluating the eigenvalues of the Jacobian matrix (9.15) at the equilibrium points, we can determine the stability of these solutions. In Figure 9.6(b), we show the unstable equilibrium as well. As indicated on the figure, the equilibrium points of the dotted line give stable solutions while the equilibrium points of the solid line are unstable nodes. Indeed, the equilibrium points depicted by the solid line have at least one eigenvalue that has a positive real part while the eigenvalues derived from the dotted line all have a negative real part.



(a) (b)

Figure 9.4: Mean number of recovered (a) and infected (b) individuals.

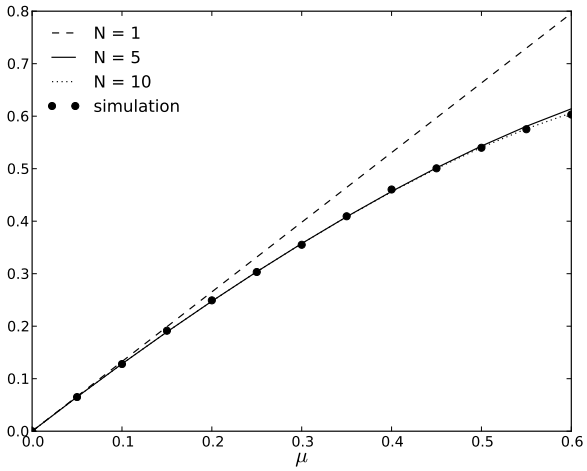
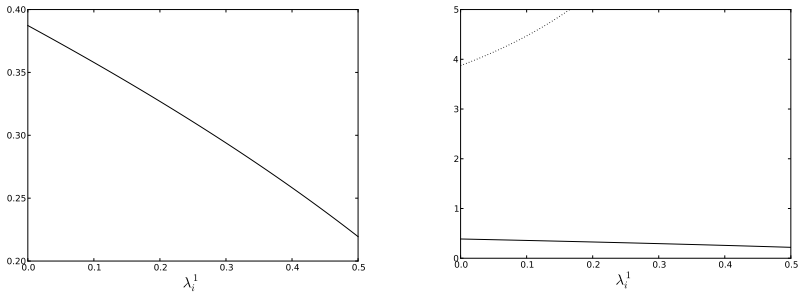
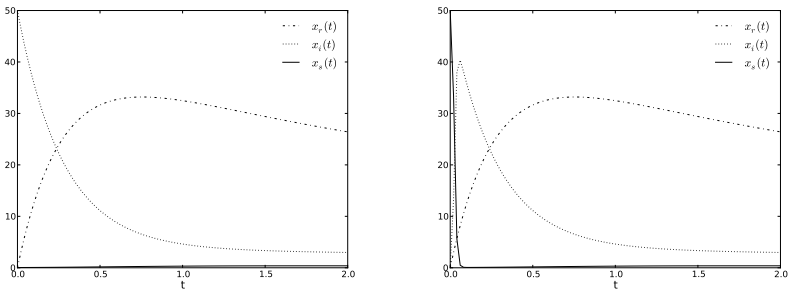


Figure 9.5: Mean number of susceptible individuals.



(a) (b)

Figure 9.6: Stable equilibrium point of x_s (a) and stable and unstable equilibrium points of x_s (b) for $\lambda_k^0 = 3$, $\mu = 0.7$ $\alpha_{si}^0 = \alpha_{si}^1 = \alpha_{sr}^0 = \alpha_{ir}^0 = 3$.



(a) (b)

Figure 9.7: Fluid model for $M = 100$, $\lambda_k^0 = 3$, $\lambda_i^1 = 0.1$ $\mu = 0.7$, $\alpha_{si}^0 = \alpha_{si}^1 = \alpha_{sr}^0 = \alpha_{ir}^0 = 3$ and $L = 50$.

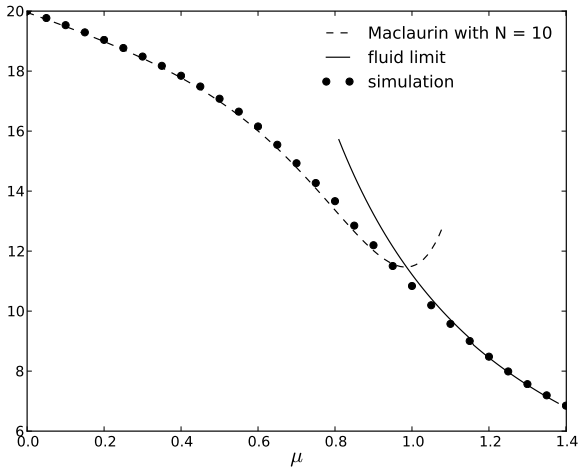


Figure 9.8: Mean number of recovered individuals calculated by the Maclaurin-series expansion (left) and the fluid limit (right).

The fluid limit also allows for evaluating the evolution to equilibrium. The transient behaviour predicted by the fluid approximation is shown in Figures 9.7(a) and 9.7(b). We assume the same parameter settings except for the maximum number of individuals L which is now equal to 50 instead of 20. The start values are respectively $x_r(0) = 0, x_i(0) = 50$ and $x_s(t) = 0$ and $x_r(0) = 0, x_i(0) = 0$ and $x_s(t) = 50$ in Figures 9.7(a) and 9.7(b). As the figures show, a large number of infected or susceptible individuals quickly lead to a relative large number of recovered individuals.

Finally, we combine series- and fluid approximations. In Figure 9.8, we depict the mean number of recovered individuals as calculated by the Maclaurin-series expansion with $N = 10$ for μ varying from 0.0 to 0.95 as well as the mean number of recovered individuals as calculated by the fluid limit of the system for μ varying from 0.75 to 1.45. Moreover, both approximations are compared with simulation results. The other parameters are the same as in Figure 9.4 and 9.5. This figure clearly demonstrates that both approximations are complementary: the series expansion and the fluid limit approximate well the mean number of recovered individuals for low and high values of μ, respectively. It can be seen that the accuracy of the series expansion deteriorates as μ increases while the accuracy of the fluid limit deteriorates for decreasing μ. In this case, combining both approximations yields a good approximation for the whole range of μ going from 0 to 1.45 (from 0 to 0.83 with the series expansions and from 0.83 to 1.45 with the fluid limit).

9.6 Conclusion

In this paper, we evaluate the propagation of an opinion in a size-limited population that has a low population turnover. Furthermore, we assume that individuals can have either no opinion (S), an opinion that they want to spread (I) or an opinion that they don't want to transmit or no opinion as they become neutral or lose their interest in the topic (R). Moreover, the evaluation method at hand allows for arrival rates of the three types as well as for rates at which an individual changes type that are state-dependent.

To cope with the inherent state space explosion, we propose an approximative numerical algorithm for the Markovian epidemic process. In particular, a numerical algorithm is applied which calculates the first N coefficients of the Maclaurin-series expansion of the steady-state probability vector. From the numerical results, we show that the series expansion approach gives us a good approximation for the opinion model in a heuristically determined region. Complementary to the series expansion approach, we derive a fluid limit of the Markov process where the arrival rates of the three types and the population size are sent to infinity.

References

[1] K.E. Avrachenkov and M. Haviv. *Perturbation of null spaces with application to the eigenvalue problem and generalized inverses*, Linear Algebra and its Applications, 369, p.1-25, 2003.

[2] E. Altman, K.E. Avrachenkov and R. Nunez-Queija. *Perturbation analysis for denumerable Markov chains with application to queueing models*, Advances in Applied Probability, 36(3), p.839–853, 2004.

[3] H. Andersson and R.M. May. *Infectious Diseases of Humans*, Oxford University Press, Oxford, UK, 1992.

[4] L.M.A. Bettencourt, A. Cintron-Arias, D.I. Kaiser and C. Castillo-Chavez. *The power of a good idea: Quantitative modelling of the spread of ideas from epidemiological models*, Physica A: Statistical Mechanics and its Applications, 364, p.513–536, 2006.

[5] B. Błaszczyszyn, T. Rolski and V. Schmidt. *Advances in Queueing: Theory, Methods and Open Problems*, Chapter Light-traffic approximations in queues and related stochastic models, CRC Press, Boca Raton, Florida, 1995.

[6] B. Błaszczyszyn. *Factorial-moment expansion for stochastic systems*, Stochastic Processes and their Applications, 56, p.321–335, 1995.

[7] R.W.R. Darling and J.R. Norris. *Differential equation approximations for Markov chains*, Probability Surveys, 5, p.37–79, 2008.

[8] E. De Cuypere and D. Fiems. *Performance evaluation of a kitting process*. Proceedings of the 18th International Conference on Analytical and Stochastic Modelling Techniques and Applications (ASMTA 2011), p. 175–188, Venice, June 2011.

[9] E. De Cuypere, K. De Turck, S. Wittevrongel and D. Fiems. *Markovian SIR model for opinion propagation*, Proceedings of the 25th International Teletraffic Congress (ITC 2013), Shanghai, September 2013.

[10] K. De Turck, E. De Cuypere, S. Wittevrongel and D. Fiems. *Algorithmic approach to series expansions around transient Markov chains with applications to paired queuing systems*, International Conference on Performance Evaluation Methodologies and Tools (VALUETOOLS 2012), p.38–44, France, October 2012.

[11] S.N. Ethier and T.G. Kurtz. *Markov Processes: Characterization and Convergence*, Wiley and Sons, Second Edition, 2005.

[12] D.S. Evans. *The online advertising industry: economics, evolution and privacy*, Journal of Economic Perspectives, 23(3), p.37–60, 2009.

[13] P. Fan, H. Wang, P. Li, W. Li and Z. Jiang. *Analysis of opinion spreading in homogeneous networks with signed relationships*, Journal of statistical mechanics-theory and experiment, DOI = 10.1088/1742-5468/2012/08/P08003, August 2012.

[14] B. Heidergott, A. Hordijk, and N. Leder. *Series Expansions for Continuous-Time Markov Processes*, Operations Research, 58.3, 756–767, DOI=10.1287/opre.1090.0738, http://dx.doi.org/10.1287/opre.1090.0738, May 2010.

[15] H.W. Hethcote. *The mathematical of infectiours diseases*, SIAM Review, 42, p.599–653, 2000.

[16] A.L. Hill, D.G. Rand, M.A. Nowak and N.A. Christakis. *Infectious Disease Modeling of Social Contagion in Networks*, Plos computational Biology, 6, 11, November 2010.

[17] V. Isham, J. Kaczmarska and M. Nekovee. *Spread of information and infection on finite random networks*, Physical review E, 83(4), 2, April 2011.

[18] Y. Li, Z. Feng, H. Wang, S. Kong, and L. Feng. *ReTweetp: Modeling and Predicting Tweets Spread Using an Extended Susceptible-Infected-Susceptible*

Epidemic Model, DASFAA 2013, W. Meng et al. (Eds.), Part II, LNCS 7826, Springer-Verlag Berlin Heidelberg 2013, p.454–457, 2013.

[19] G. Katriel and L. Stone. *Pandemic Dynamics and the Breakdown of Herd Immunity*, Plos one, 5-3, 2010.

[20] T. Kawamoto. *A stochastic model of tweet diffusion on the Twitter network*, Physica A-statistical mechanics and its applications, 392, p.3470–3475, 2013.

[21] W.O. Kermack and A.G. McKendrick. *A Contribution to the Mathematical Theory of Epidemics*, Proceedings of the Royal Society of London A, 115, p.700–721, 1927.

[22] I. Kovalenko. *Rare events in queueing theory, a survey*, Queueing systems. 16(1), p.1–49, 1994.

[23] D. Kempe, J. Kleinberg and E. Tardos. *Maximizing the spread of influence through a social network*, Proceedings of KDD, p.137–146, 2003.

[24] C.J. Kuhlmana,V.S.A. Kumara and S.S. Ravib. *Controlling opinion propagation in online networks*, Computer Networks, 57, 10, p.2121–2132, July 2013.

[25] S. Meyn and R.L. Tweedie. *Markov Chains and Stochastic Stability*, 2nd edition, Cambridge University Press, 2009.

[26] Y. Moreno, M. Nekovee and A.F. Pacheco. *Dynamics of rumor spreading in complex networks*, Physical review E, 69, 6, 2, June 2004.

[27] M. Reiman and B. Simon. *Open queueing systems in light traffic. Mathematics of operations research*, 14(1), p.26–59, 1989.

[28] P. Robert. *Stochastic Networks and Queues*, Springer, Berlin, 2003.

[29] P.J. Schweitzer. *Perturbation theory and finite Markov chains*, Journal of Applied Probability, 5(2), p.401–413, 1968.

[30] A.M. Timpanaro and C.P.C. Pradot. *Generalized Sznajd model for opinion propagation*, Physical review E, 80, 021119, DOI: 10.1103/PhysRevE.80.021119, 2009.

[31] J. Woo, J. Son and H. Chen. *An SIR model for violent topic diffusion in social media*, IEEE International Conference on Intelligence and Security Informatics (ISI 2011), p.15–19, 9–12 July, Beijing, China, 2011.

[32] Z. Yang, J. Guo, J. Tang, L. Zhang and Z. Su. *Understanding retweeting behaviors in social networks*, In: CIKM, 2010.

[33] L. Zhao, J. Wang, Y. Chen, Q. Wang, J. Cheng and H. Cui. *SIHR rumor spreading model in social networks*, Physica A-statistical mechanics and its applications, 391, 7, p.2444–2453, April 2012.

10
Conclusions

In this dissertation, we investigated a particular type of Markovian queueing systems, namely queueing systems with shared service. Shared service means that there is a departure in every queue upon service completion and that service is only possible when each queue is nonempty. To gain insights into the dynamics of such systems under uncertainty, we used and developed state-of-the-art modelling and numerical solution techniques. More specifically, we exploited structural properties of the Markov processes that describe queueing systems with shared service, so as to speed up computation of the steady-state solution.

In Chapter 2 we investigated a two-part kitting system in a Markovian setting. As kits can only be compiled when both parts are available, the kitting process is modelled as a two-queue system with shared service. The introduction of a Markovian environment allows the study of kitting performance under non-restrictive stochastic assumptions like bursty part arrivals and phase-type distributed kit assembly times. As most of the entries in the generator matrix equal zero, we exploited the sparse property of the generator matrix of the Markov process that describe such systems by using the GMRES method. Although this method performs better than the LU decomposition with respect to speed, increasing the number of queues (or increasing the state space) leads to inefficient calculations of the steady-state performance measures.

Another structural property of some Markov processes we considered is the repetitive block-triangular structure of their generator matrices. Such processes are called quasi-birth-death (QBD) and can be solved efficiently by the matrix-geometric methods. Restricting ourselves to a system with two queues, the queue

content of one queue can be identified as the level of the QBD and the queue content of the other queue as the phase of the QBD. In this dissertation, we assumed the number of levels and phases to be respectively infinite and finite. Accounting for the repetitive finite block structure of the QBD process when calculating the steady-state probability vector reduces significantly the computational complexity. If we move beyond two queues, the phase can still describe the queue content of all but one queue and we still obtain a QBD. Note however that the state-space explosion translates into increasing block sizes and matrix-geometric methods are no longer computationally efficient.

In Chapter 3 and 4, hybrid MTS/MTO systems with order backlog are analysed as homogeneous QBD processes and solved by matrix-geometric methods. As order processing can only start when there is an order and a semi-finished product available, service is shared. Moreover, a threshold-based control policy was implemented: production of semi-finished products starts when the inventory level drops below a certain value, referred to as the threshold value, and stops when the inventory level reaches maximum capacity. Under uncertainty of demand, production and service times, the performance analysis of hybrid MTS/MTO systems with and without a threshold-based control policy was conducted. As shown in the numerical results, the setup time and order processing time distribution have a limited impact on the mean lead time and inventory level. However, inventory control and correlation in the order process decreases the mean inventory level at the cost of increased mean lead times.

Similar to hybrid MTS/MTO systems, energy harvesting sensor nodes are analysed as homogeneous QBD processes and solved by matrix-geometric methods in Chapter 5 and 6. A rechargeable battery operates very much like a queue, customers being discretised as chunks of energy. As a sensor node requires both sensed data and energy for transmission, shared service can again be identified. The performance of such sensor nodes is evaluated under uncertainty in energy capture, energy consumption, data acquisition and data transmission. We also accounted for the transmission range of the sensor node by assuming limited time periods in which data can be transmitted. As shown in the numerical results, correlation in the energy harvesting process decreases the performance of the energy harvesting sensor node: data packets wait longer on average. Also, if the energy harvesting rate is high, correlation induces long periods with more energy arrivals than can be stored in the battery with finite capacity. Hence, the mean battery level decreases.

As stated earlier, neither of the former solution methods allows for investigating systems with many queues. Therefore, we developed an approximation technique based on a Maclaurin-series expansion of the steady-state probability vector. When the departure rate is sent to zero, the resulting generator matrix has an upper-triangular structure. In this case, the solution at zero is trivial as there

is only one final state. The calculation of the higher order terms in the series expansion results in a computational complexity proportional to the size of the state space.

Chapter 7 studied kitting systems with exponential service times as regular perturbation problems. Indeed, the Markov process is irreducible when the perturbation parameter is sent to zero which means that we can find one unique steady-state solution. In this Chapter, a proof of convergence of the series expansion and a lower bound on the convergence radius are provided. The convergence domain is illustrated by a numerical example.

Chapter 8 builds on the results of Chapter 7 by proposing an efficient numerical scheme for the evaluation of singular perturbation problems. In this case, kitting systems with phase-type service times are considered. When there is no service, all queues will eventually be full but the system will remain in one of the phases of the service process. Hence, the perturbation is singular. As shown in the numerical results of both Chapters 7 and 8, the Maclaurin-series expansion combined with a proposed heuristic give a quite good approximation of the studied queueing systems with shared service in the regular as well as in the singular case.

Chapter 9 studied single opinion propagation systems as a Markovian SIR epidemic model. Accounting for limited population size, opinion spreading can be captured by a multidimensional Markov process which is very similar to the Markov process of the queueing systems with shared service. Although individuals leave the system one by one, i.e. there is no shared service, we showed that the developed numerical algorithm can be utilised to approximate the steady-state probability vector of the epidemic model. Assuming exponential departure rates, the perturbation problem is regular. The fluid limit of the Markov process is also derived. As shown by the numerical examples, the Maclaurin-series expansion and the fluid limit complement each other: both methods approximate well the studied performance measures for low and high values of the perturbation parameter, respectively.

10.1 Future work

In this dissertation, applications in inventory management and telecommunications motivate the study of queueing systems with shared service. The developed queueing models and considered analysis techniques can however be extended to other applications. An example of a sector using queueing systems with shared service is that of healthcare operations management. Indeed, different types of customers (i.e. nurses, patients, beds, medicine etc.) need to be in place for a service (i.e. surgery) to proceed. Other possible future work concerns the extension of the already developed queueing models. For example, the studied kitting processes can be extended to and compared with the threshold-based case. This

analysis would determine the optimal replenishment strategy given a specific set of parameter values. The studied kitting processes can also be complemented with a queue of outstanding orders in order to further comprehend the complexity of the dynamics of hybrid push-pull systems. Finally, we could expand the single opinion propagation model to a model with multiple opinions. As in the case of single opinion, system performance can then be evaluated by two complementary approaches: the Maclaurin-series expansion and the fluid limit of the system at hand. Concerning the developed Maclaurin-series expansion, we foresee that the conditions can be relaxed to a block triangular structure for the unperturbed part of the generator matrix. Indeed, these additional complexities are expected to be manageable if the size of the blocks remains small.