Chapter 6

Queueing systems/networks

This chapter is devoted to the study of a class of models in which customers arrive in some random manner at a service facility and demand service. Upon arrival they are made to wait in queue[1] until it is their turn to be served. Once served they are generally assumed to leave the system (Ross, 2003, p. 475). These models are usually termed *queueing systems*.

If we think for a moment how much time we spend in some form of queue (Kleinrock, 1975, p. 3), we can come with a virtually endless list of queueing systems we deal with on a daily basis. For example,

- supermarkets, post offices, parking lots, call centers of an insurance company and toll booths

(Adan and Resing, 2002, pp. 7–9).

Queueing theory started with research by the Danish mathematician, statistician and engineer Agner Krarup Erlang (1878–1929), when he created models to describe the Copenhagen telephone exchange (https://en.wikipedia.org/wiki/Queueing_theory). Expectedly, the publication of Erlang's first paper in 1909 is often regarded as marking the birth of queueing theory (Cooper, 1981, p. 6).

Queueing models have since been applied to such areas like telecommunications, traffic engineering, computing and in the design of factories, shops, offices and hospitals (https://en.wikipedia.org/wiki/Queueing_theory).

[1]The word queue comes, via French, from the Latin *cauda*, meaning *tail* (https://en.wikipedia.org/wiki/Queueing_theory#Etymology).

6.1 Description of a queueing system

According to Prabhu (1997, p. 1), a queueing system can be essentially described by its:

- *input*;
- *queue discipline*;
- *service mechanism*;
- *cost structure*.

Input

Prabhu (1997, p. 1) also adds that the *input* describes the way customers arrive and join the system.

Customers may:

- come from a finite or infinite source;
- arrive individually or in groups.

The system may also provide waiting space of

- *finite capacity*

and in this case a customer is forced to leave if all the servers and the waiting positions are occupied.

If a waiting line can be formed when customers arrive and demand service, then we are dealing with a

- *waiting* (or *delay*) system.

However, it is possible that customers cannot join the system if it is fully occupied; in this case we are dealing with a

- *loss* system.

Queue discipline

The *queue discipline* refers to the formation of the queue and the way customers behave while waiting in queue (Prabhu, 1997, p. 1).

Various queue disciplines can be adopted, namely:

- *first come, first served* (FCFS) — the simplest (and fairest!) queue discipline, according to which customers are served one at a time and in order of their arrival;
- ***last come, first served* (LCFS)** — customers are also served one at a time, however they are served in reverse order of their arrival.

(See Prabhu, 1997, p. 1; http://en.wikipedia.org/wiki/Queueing_theory#Service_disciplines.)

In a system with priorities, customers with high priority are served first. Two examples of priorities: *rush orders first* and *shortest processing time first* (Adan and Resing, 2002, p. 24).

Priority queues can be of two types (http://en.wikipedia.org/wiki/Queueing_theory#Service_disciplines):

- *non-preemptive*, where a job in service cannot be interrupted;
- *preemptive*, where a job in service can be interrupted by a higher priority job.[2]

Service mechanism

The *service mechanism* describes the arrangements of the m ($1 \leq m \leq \infty$) servers/facilities the customers seek (Prabhu, 1997, p. 2).

If $m = \infty$ then all customers will be served upon arrival and no queue is form.

If m is finite, the servers are arranged in *parallel* and the FCFS discipline has been adopted, then a single waiting line is formed and the first of the m servers to be free serves the customer at the top of the queue (Prabhu, 1997, p. 2).

Prabhu (1997, p. 2) also mentions that, in a system with s servers ($1 \leq m < \infty$) arranged in *series* (or *tandem*), the customers are served by each server in succession and leave the system after being served by the m^{th} server.

Cost structure

Prabhu (1997, p. 2) continues by adding that the *cost structure* refers to:

- the *payments* made by the customers for the service they received;
- the *operating costs* of the system;
- the *holding costs* expressed as a function of the number of customers waiting or as a function of the delay suffered by these customers.

[2]No work is lost in either case.

6.2 Queueing systems and Kendall's notation

In 1953, D.G. Kendall proposed to describe queueing systems using the notation $A/S/s$, where

- A denotes the time between arrivals to the **system**
- S the duration of the services
- m the number of servers

(http://en.wikipedia.org/wiki/Kendall's_notation).

Kendall notation has since been extended to $A/S/s/K/N/D$,[3] where

- K refers to the capacity of the **system**,
- N the size of the population of customers to be served
- D the queueing discipline

(http://en.wikipedia.org/wiki/Kendall's_notation).

When the last three parameters are not specified, we should assume $K = \infty$, $N = \infty$ and $D =$ FCFS.

A: The arrival process

We suppose the successive customers arrive at random time epochs and that the interarrival times are i.i.d. r.v. with common c.d.f. $A(x)$.[4] Moreover, arrivals may refer to individual customers or batch arrivals (Adan and Resing, 2002, p. 23).

Code letters describe the arrival process. A few examples:

Symbol	Distribution of the interarrival times	Obs.
M	Exponential (Memoryless, Markovian)	Poisson arrival process
D	Degenerate (Deterministic or regular)	Fixed interarrival times
E_k	Erlang(k, λ)	
G or GI	General	G also refers to Independent arrivals
M^X	Batch Markov	Batch arrivals of random size X governed by a Poisson process

[3] Also referred to as the Kendall-Lee notation.

[4] The assumption of identically distributed interarrival times is not always valid, however, in most situations it seems reasonable (Prabhu, 1997, p. 3).

When $k = 1$ we are dealing with a Poisson arrival process. For $k \in \{2, 3, \ldots\}$, the interarrival times distribution is a $k-$fold convolution of the exponential distribution; thus, an arrival takes place only after the completion k stages with i.i.d. durations with Exponential(λ) distribution (Prabhu, 1997, p. 5).

S: The service time distribution
The time elapsed while a customer is being served is called her/his service time. We usually assume that the service times are i.i.d. r.v., with common c.d.f. $B(x)$, and that they are independent of the interarrival times.

The previous code letters also apply to the service times.

m: The number of servers
m represents the number of service servers (or channels).

K: The number of places in the system
The capacity of the system, or the maximum number of customers allowed in the system including those in service. When the number is at this maximum, further arrivals are turned away.

N: The calling population
The size of the population from which the customers come.

D: The queue's discipline
As we have previously mentioned, it refers order that jobs in the waiting line are served and three examples are FCFS, LCFS and priority service (preemptive and non-preemptive). Another interesting possibility:

- SIRO (Service In Random Order) — the customers are served in a random order with no regard to arrival order (http://en.wikipedia.org/wiki/Kendall's_notation).

By using the Kendall-Lee notation, we provide all the necessary information to characterize a queueing system, namely the input, the service mechanism and the queue discipline.

Example 6.1 — Kendall's notation
The easiest queueing system to analyze is the $M/M/1$ queue, where interarrival times and service times are exponentially distributed and there is a single-server in the system.

If the single-server admits Markovian arrivals, but the service process is governed by a general distribution, then we are dealing with a $M/G/1$ queue. •

6.3 Performance measures

One of life's most unpleasant (in)activities is waiting in line (Kleinrock, 1975, p. 3). Unsurprisingly, these are important performance measures to be considered while assessing a queueing system:

- $L_s(t)$, the **number of customers in the system at time** t;[5]
- $L_q(t)$, the **number of customers in the queue** (waiting to be served!) **at time** t;
- $W_s(t)$ the time a customer would have to spend in the system if he/she arrived at time t, usually termed **sojourn time**;
- $W_q(t)$, the time a customer would have to wait to be served if he/she arrived at time t, also called **virtual waiting time**.

Note that $W_s(t)$ **is** equal to *virtual waiting time* plus *service time*.

Another important performance measure:

- the **total workload** submitted to the s servers in the interval $[0, t]$, $X(t)$, due to the $N(t)$ arrivals in that period (Prabhu, 1997, p. 5).

In fact, $X(t) = \sum_{i=1}^{N(t)} S_i$, where S_i represents the service time of the i^{th} arrival.

The **ratio**

$$\rho(t) = E\left[\frac{X(t)}{t\,s}\right] \tag{6.1}$$

can be thought as a(n average) measure of congestion in the system (Prabhu, 1997, p. 6). Interestingly enough,

$$\lim_{t \to +\infty} \rho(t) = \rho = \frac{\lambda}{s\,\mu}, \tag{6.2}$$

where λ^{-1} is the expected time between successive arrivals and μ^{-1} the expected service time. (Prove this result using renewal theory!)

Definition 6.2 — **Traffic intensity** (Prabhu, 1997, p. 6; Pacheco, 2002, p. 75)
ρ is called the traffic intensity of the system.[6] It is a (relative) measure of congestion and represents the load offered to each server if the work is divided equally among servers. •

[5] Including the ones being served, if any. Prabhu (1997, p. 7) misleadingly terms this r.v. *queue length*.
[6] Or utilization factor (Kleinrock, 1975, p. 98).

ρ is expressed in *erlang* (the corresponding symbol is E); is a dimensionless quantity used in telephony and named in honor of Agner Krarup Erlang (http://en.wikipedia.org/wiki/Erlang_(unit)).

Remark 6.3 — **Traffic intensity** (Prabhu, 1997, p. 9)
If:

- $\rho > 1$ then the system is over saturated;
- $\rho < 1$ the system is undersaturated;[7]
- $\rho = 1$ then there is a balance between the customers' need for prompt service and the management's desire to avoid an idle system. •

We will be also interested in determining the following **performance measures in the long-run** or equilibrium:

- L_s, the number of customers in the system — an arriving customer sees;
- L_q, the number of customers in the queue (waiting to be served) — an arriving customer sees;
- W_s, the time an arriving customer will spend in the system;
- W_q, the time an arriving customer will spend in the queue waiting to be served.[8]

W_s and W_q influence customer satisfaction, whereas L_s and L_q are particularly important performance measures to resource management (Pacheco, 2002, p. 76).

The distributions of these four r.v. depend on the following parameters:

- λ, the arrival rate;
- μ, the service rate;
- s, the number of (identical) servers in parallel;
- $a = \frac{\lambda}{\mu}$, the (offered) load;[9]

[7] ρ should be strictly less than one for the system to function well (https://en.wikipedia.org/wiki/Queueing_theory#Utilization).

[8] Note that $W_s \overset{st}{=} W_q +$ service time

[9] It corresponds to the expected amount of time a (single) server would take to serve all customers that in the long-run arrive to the system during one unit of time, including blocked customers (Pacheco, 2002, p. 74).

- P_b, the blocking probability;[10]
- $\lambda_e = \lambda \times (1 - P_b)$, the input rate;[11]
- $\rho = \frac{\lambda}{s\,\mu}$, the traffic intensity;
- $\rho_e = \frac{\lambda_e}{s\,\mu} = \rho \times (1 - P_b)$, the carried traffic intensity;
- P_i, the long-run fraction of time in state i.

In addition, relationships between these four performance measures can be obtained by using all these parameters and capitalizing on the following result.

Theorem 6.4 — Little's law (http://en.wikipedia.org/wiki/Little's_law; Ross, 2003, p. 478)
The long-term AVERAGE number of customers in a stable system L is equal to the long-term average effective arrival rate, λ_e, multiplied by the AVERAGE time a customer spends in the system, W — expressed algebraically:

$$L = \lambda_e\, W. \tag{6.3}$$

Consequently:

$$E(L_s) = \lambda_e\, E(W_s); \tag{6.4}$$
$$E(L_q) = \lambda_e\, E(W_q). \tag{6.5}$$

•

Remark 6.5 — Little's law

- Although Little's law looks intuitively reasonable, it is a quite remarkable result (http://en.wikipedia.org/wiki/Little's_law), as it is valid regardless of the
  - arrival process distribution;
  - service distribution;
  - number of servers;
  - service policy (as long as it is not *biased*);
  - etc. •

[10] It is the long-run fraction of customers that are blocked (Pacheco, 2002, p. 75) upon arrival and unable to enter the system.

[11] It is the effective arrival rate which corresponds to the rate at which customers enter the system, thus, we are excluding blocked customers (Pacheco, 2002, p. 75).

6.4 Continuous time Markov chains

A systematic treatment of queues from the point view of stochastic processes is essentially due to Kendall (1951, 1953), according to Prabhu (1997, p. 11). A large class of queueing systems can be indeed studied as continuous time Markov chains (CTMC) (Resnick, 1992, p. 367), namely as birth-death processes. Moreover, since the study of the transient behavior of queueing systems is far from being trivial, we focus on their *equilibrium behavior*, namely derive limiting probabilities of the number of customers an arriving customer sees in the system.

Unsurprisingly, we proceed with

- generalities on CTMC,
- birth-death processes,
- classification of states
- limit behavior of CTMC,

with definitions, propositions, etc., taken from Morais (2003, Chap. 4).[12]

Definition 6.6 — **CTMC** (Ross, 2003, p. 350)
Let $\{X(t) : t \geq 0\}$ be a continuous time stochastic process taking values in the set of non-negative integers (that is, the state space $\mathcal{S} \subseteq \mathbb{N}_0$). If

$$\begin{aligned} P[X(t+s) = j \mid X(s) = i,\, X(u) = x(u),\, 0 \leq u < s] \\ = P[X(t+s) = j \mid X(s) = i], \end{aligned} \tag{6.6}$$

for all $s, t \geq 0$ and non-negative integers i, j and $x(u)$, $0 \leq u < s$, then $\{X(t) : t \geq 0\}$ is said to be a CTMC. If, in addition, the transition probabilities satisfy

$$P[X(t+s) = j \mid X(s) = i] = P[X(t) = j \mid X(0) = i], \tag{6.7}$$

then the CTMC is said to be time-**homogeneous**.[13] •

To motivate another definition of CTMC, let us recall Ross (1983, p. 142), who mentions that, by the Markov property, this stochastic process has the following properties each time it enters state i:

[12] For a more detailed account on CTMC and birth-death processes the reader should refer to textbooks such as Kulkarni (1995) and Ross (1983, 1989, 2003).

[13] Or to have stationary or **homogeneous** transition probabilities. The material in this and the next sections only refers to CTMC with stationary transition probabilities.

- the amount of time spent in state i (sojourn or holding time!) before making a transition into a different state has exponential distribution with parameter ν_i;[14]
- the probability that the process leaves state i and the next state it enters is j equals P_{ij}, where $\sum_{j \neq i} P_{ij} = 1$.[15]

Definition 6.7 — CTMC (bis) (Kulkarni, 1995, pp. 240–241)
Let:

- $\{X(t) : t \geq 0\}$ be a continuous time stochastic process with state space $S \subseteq \mathbb{N}_0$;
- $S_0 = 0$;
- S_n be the time of the n^{th} transition;
- $Y_n = S_n - S_{n-1}$ be the n^{th} sojourn or holding time;
- $X_0 = X(0)$ be the initial state of the process;
- $X_n = X(S_n^+) = X(S_n)$ be the state of the stochastic process immediately after the n^{th} transition;
- $P_{ij} = P[X(S_{n+1}^+) = j \mid X(S_n^+) = i]$.

Then the stochastic process $\{X(t) : t \geq 0\}$ is said to be a CTMC with initial state $X_0 = X(0)$ if it changes states at times $0 < S_1 < S_2 < \ldots$ and the embedded process $\{X_0, (X_n, Y_n) : n \in \mathbb{N}\}$ satisfies

$$P\left[X_{n+1} = j, Y_{n+1} > y \mid (X_n, Y_n) = (i, y_n), (X_{n-1}, Y_{n-1}) = (i_{n-1}, y_{n-1}), \ldots, (X_1, Y_1) = (i_1, y_1), X_0 = i_0\right] = P_{ij}\, e^{-\nu_i \times y}, \tag{6.8}$$

for all non-negative integers $i, j, i_{n-1}, \ldots, i_1, i_0$ and non-negative real numbers $y, y_n, y_{n-1}, \ldots, y_1$.

$\{X_n : n \in \mathbb{N}_0\}$ is usually called the embedded discrete time Markov chain (DTMC) in the CTMC $\{X(t) : t \geq 0\}$. •

Less formally, in a CTMC the succession of states visited still follows a DTMC but now the flow of time is *perturbed* by exponentially distributed sojourn (or holding) times in each state (Resnick, 1992, p. 367).

[14] A state i is called called instantaneous if $\nu_i = +\infty$ (Kulkarni, 1995, p. 241), i.e., the expected sojourn time in state i is equal to 0. From now on, we shall only deal with CTMC with no instantaneous states.

[15] $P_{ii} = 0$ unless state i is an absorbing state — in this case $P_{ii} = 1$ (Kulkarni, 1995, p. 246) and $\nu_i = 0$.

The law of motion of a CTMC is governed by a time-dependent transition probability matrix.

Definition 6.8 — Transition probability matrix (Kulkarni, 1995, p. 243)
Let:

- $\{X(t) : t \geq 0\}$ be a (time-homogeneous) CTMC with state space $\mathcal{S}$;
- $P_{ij}(t) = P[X(t) = j \mid X(0) = i]$, $i, j \in \mathcal{S}$, be the (time-dependent) transition probabilities.

Then

$$\mathbf{P}(t) = [P_{ij}(t)]_{i,j \in \mathcal{S}} \tag{6.9}$$

is called the transition probability matrix (TPM). •

Proposition 6.9 — Characterization of a CTMC (Kulkarni, 1995, Theorem 6.2, p. 244)
A CTMC $\{X(t) : t \geq 0\}$ is fully characterized by its

- TPM, $\mathbf{P}(t)$, and
- initial distribution, that is, the p.f. of $X(0)$ denoted by $\underline{\alpha} = [\alpha_i]_{i \in \mathcal{S}} = [P[X(0) = i]]_{i \in \mathcal{S}}$. •

$\mathbf{P}(t)$ is certainly a stochastic matrix and satisfy the Chapman-Kolmogorov equations (obviously rewritten for a continuous time stochastic process).

Proposition 6.10 — Properties of the TPM; Chapman-Kolmogorov equations (Kulkarni, 1995, Theorem 6.3, p. 253–254)
The TPM, $\mathbf{P}(t)$, of a CTMC $\{X(t) : t \geq 0\}$ has the following properties:

- $P_{ij}(t) \geq 0$, $i, j \in \mathcal{S}$, $\quad t \geq 0$;
- $\sum_{j \in \mathcal{S}} P_{ij}(t) = 1$, $\quad i \in \mathcal{S}$, $t \geq 0$.

Moreover, $\mathbf{P}(0) = \mathbf{I}$ and the Chapman-Kolmogorov equations[16] are written as follows:

$$P_{ij}(t+s) = \sum_{k \in \mathcal{S}} P_{ik}(t) \times P_{kj}(s), \quad i, j \in \mathcal{S}, \quad t, s \geq 0, \tag{6.10}$$

[16]The equations were arrived at independently by both the British mathematician Sydney Chapman (1888–1970) and the Russian mathematician Andrey Kolmogorov (1903–1987) (http://en.wikipedia.org/wiki/Chapman-Kolmogorov_equation).

or, in matrix form,

$$\begin{aligned} \mathbf{P}(t+s) &= \mathbf{P}(t) \times \mathbf{P}(s) && (6.11) \\ &= \mathbf{P}(s) \times \mathbf{P}(t), \quad t, s \geq 0. && (6.12) \end{aligned}$$

•

Exercise 6.11 — Properties of the TPM (Isaacson and Madsen, 1976, p. 231, Exercise 1)
Which of the following matrices have the properties of the TPM for a CTMC?

(a) $\mathbf{P}_1(t) = \begin{bmatrix} e^{-t} & 1-e^{-t} \\ 0 & 1 \end{bmatrix}$

(b) $\mathbf{P}_2(t) = \begin{bmatrix} e^{t} & 1-e^{t} \\ 0 & 1 \end{bmatrix}$

(c) $\mathbf{P}_3(t) = \begin{bmatrix} 1 & 0 \\ 1-te^{-t} & te^{-t} \end{bmatrix}$

(d) $\mathbf{P}_4(t) = \begin{bmatrix} t+e^{-t} & 1-t-e^{-t} \\ 0 & 1 \end{bmatrix}$

(e) $\mathbf{P}_5(t) = \begin{bmatrix} 1-te^{-t} & te^{-t} & 0 & 0 \\ te^{-t} & 1-3te^{-t} & 2te^{-t} & 0 \\ 0 & te^{-t} & 1-2te^{-t} & te^{-t} \\ 0 & 0 & te^{-t} & 1-te^{-t} \end{bmatrix}$

•

Proposition 6.12 — Marginal and joint probabilities
Let:

- $\{X(t) : t \geq 0\}$ be a CTMC with TPM $\mathbf{P}(t) = [P_{ij}(t)]_{i,j \in \mathcal{S}}$;
- $\underline{\alpha} = [\alpha_i]_{i \in \mathcal{S}}$ be the row vector with the initial distribution of the CTMC (i.e., the p.f. of $X(0)$).

Then

$$\begin{aligned} P[X(t) = j] &= \sum_{i \in \mathcal{S}} P[X(0) = i] \times P[X(t) = j \mid X(0) = i] \\ &= \sum_{i \in \mathcal{S}} \alpha_i \times P_{ij}(t), \; j \in \mathcal{S}, && (6.13) \end{aligned}$$

and the row vector with the p.f. of $X(t)$ is given by

$$[P[X(t) = j)]_{j \in \mathcal{S}} = \underline{\alpha} \times \mathbf{P}(t). \tag{6.14}$$

Moreover,

$$P\left[X(t_1) = x(t_1), \ldots, X(t_k) = x(t_k)\right] = \left[\sum_{i \in \mathcal{S}} \alpha_i \times P_{i,x(t_1)}(t_1)\right] \times \prod_{j=2}^{k} P_{x(t_{j-1}),x(t_j)}(t_j - t_{j-1}), \tag{6.15}$$

for $0 \leq t_1 < t_2 < \ldots < t_k$ and $x(t_1), \ldots, x(t_k) \in \mathcal{S}$. •

Definition 6.13 — Instantaneous transition rates (Ross, 2003, p. 362)
For any pair of states i and j $(i \neq j)$, let

$$q_{ij} = \nu_i \times P_{ij}, \tag{6.16}$$

where ν_i is the rate at which the process makes a transition when in state i and P_{ij} is the probability that this transition is into state j in the embedded DTMC. The quantities q_{ij} are called the instantaneous transition rates and represent the rate, when in state i, at which the process makes a transition into state j. •

Remark 6.14 — Instantaneous transition rates and rate diagrams (Kulkarni, 1995, p. 246)
A rate diagram is a directed graph in which each state is represented by a node and there is an arc going from node i to node j (if $q_{ij} > 0$) with q_{ij} written on it.

The rate diagrams helps us visualize the dynamics of the CTMC and are the continuous analogue of the transition diagrams of DTMC. •

Exercise 6.15 — Rate diagram (Kulkarni, 1995, examples 6.1 and pp. 242 and 246–247)
Consider a machine that can be either up (1) or down (0). If the machine is up (resp. down), it fails (resp. is repaired) after an Exp(μ) (resp. Exp(λ)) amount of time. Once this machine is repaired it is good as new.

Let $\{X(t) : t \geq 0\}$ be the state of the machine at time t and draw the corresponding rate diagram. •

Specifying the instantaneous transition rates determines the parameters of the CTMC (Ross, 2003, p. 362). In addition, the instantaneous transition rates are related to the infinitesimal behavior of the transition probabilities, as stated by the next proposition.

Proposition 6.16 — Infinitesimal behavior of the transition probabilities (Ross, 2003, Lemma 6.2, p. 362)
Let $\{X(t) : t \geq 0\}$ be a CTMC with state space $\mathcal{S}$, TPM $\mathbf{P}(t)$ and instantaneous transition rates q_{ij}. Then:

$$\lim_{h \to 0^+} \frac{P_{ij}(h)}{h} = q_{ij}, \quad i \neq j; \tag{6.17}$$

$$\lim_{h \to 0^+} \frac{1 - P_{ii}(h)}{h} = \nu_i. \tag{6.18}$$

•

Proposition 6.17 — Kolmogorov's backward and forward equations (Ross, 2003, Theorem 6.1, pp. 364, 367)
For all states i and j and times $t \geq 0$:

$$\begin{aligned} \frac{d\, P_{ij}(t)}{dt} &= \lim_{h \to 0^+} \frac{P_{ij}(h+t) - P_{ij}(t)}{h} \\ &= \sum_{k \neq i} q_{ik} P_{kj}(t) - \nu_i P_{ij}(t) \quad \text{(backward equations)}; \end{aligned} \tag{6.19}$$

$$\begin{aligned} \frac{d\, P_{ij}(t)}{dt} &= \lim_{h \to 0^+} \frac{P_{ij}(t+h) - P_{ij}(t)}{h} \\ &= \sum_{k \neq j} P_{ik}(t) q_{kj} - P_{ij}(t) \nu_j \quad \text{(forward equations)}. \end{aligned} \tag{6.20}$$

•

Proposition 6.18 — Kolmogorov's backward and forward equations in matrix form (Ross, 2003, p. 388)
Let

$$r_{ij} = \begin{cases} q_{ij}, & i \neq j \\ -\nu_i, & i = j \end{cases} \tag{6.21}$$

and $\mathbf{R} = [r_{ij}]_{i,j \in \mathcal{S}}$.[17] Then the Kolmogorov's backward and forward equations can be written in matrix form:

$$\begin{aligned} \frac{d\, \mathbf{P}(t)}{dt} &= \left[\frac{d\, P_{ij}(t)}{dt} \right]_{i,j \in \mathcal{S}} \\ &= \mathbf{R} \times \mathbf{P}(t) \quad \text{(backward equations)} \qquad (6.22) \\ &= \mathbf{P}(t) \times \mathbf{R} \quad \text{(forward equations)}. \qquad (6.23) \end{aligned}$$

•

[17] $\mathbf{R}$ is usually called the rate matrix (or the infinitesimal generator) of the CTMC.

Proposition 6.19 — Solution of the Kolmogorov's backward and forward equations in matrix form (Ross, 2003, p. 388)
The solution of the matrix differential equations $\frac{d\mathbf{P}(t)}{dt} = \mathbf{R} \times \mathbf{P}(t)$ and $\frac{d\mathbf{P}(t)}{dt} = \mathbf{P}(t) \times \mathbf{R}$ is

$$\begin{aligned} \mathbf{P}(t) &= e^{\mathbf{R}\,t} && (6.24) \\ &= \sum_{n=0}^{+\infty} \frac{\mathbf{R}^n\, t^n}{n!}. && (6.25) \end{aligned}$$

•

Rather than using (6.25) to compute the TPM, we can use the matrix equivalent of the identities

$$\begin{aligned} e^x &= \lim_{n\to+\infty} \left(1 + \frac{x}{n}\right)^n && (6.26) \\ &= \lim_{n\to+\infty} \left[\left(1 - \frac{x}{n}\right)^{-1}\right]^n && (6.27) \end{aligned}$$

to efficiently (derive or) approximate $\mathbf{P}(t)$ when we are dealing with a finite state space.

Proposition 6.20 — Two approximations to $\mathbf{P}(t)$ (Ross, 2003, pp. 389–390)
Since

$$\begin{aligned} \mathbf{P}(t) &= e^{\mathbf{R}\,t} \\ &= \lim_{n\to+\infty} \left(\mathbf{I} + \frac{\mathbf{R}\,t}{n}\right)^n && (6.28) \\ &= \lim_{n\to+\infty} \left[\left(\mathbf{I} - \frac{\mathbf{R}\,t}{n}\right)^{-1}\right]^n, && (6.29) \end{aligned}$$

if we let n be a power of 2, say $n = 2^k$, then we can approximate $\mathbf{P}(t)$ by raising either the matrix $\left(\mathbf{I} + \frac{\mathbf{R}\,t}{n}\right)$ or the matrix $\left(\mathbf{I} - \frac{\mathbf{R}\,t}{n}\right)^{-1}$ to the n^{th} power, which can be accomplished by k matrix multiplications.[18] •

Exercise 6.21 — Two approximations to $\mathbf{P}(t)$
Consider

$$\mathbf{R} = \begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu \end{bmatrix},$$

where $\lambda = 1$ and $\mu = 2$.

[18] For instance, we multiply $\left(\mathbf{I} + \frac{\mathbf{R}\,t}{n}\right)$ by itself to obtain $\left(\mathbf{I} + \frac{\mathbf{R}\,t}{n}\right)^2$ and then multiplying that by itself to obtain $\left(\mathbf{I} + \frac{\mathbf{R}\,t}{n}\right)^4$ and so on.

(a) Use *Mathematica*, in particular the function *MatrixExp*, to obtain $\mathbf{P}(t)$, for $t = 1, 100$.

(b) Compare the exact results in (a) to the approximate ones, obtained by using Proposition 6.20. •

For a more detailed account on other methods to compute the TPM $\mathbf{P}(t)$ of a CTMC with finite state space, the reader is referred to Kulkarni (1995, pp. 261–274).

6.4.1 Birth and death processes

Birth-death processes are special cases of CTMC where the state transitions are of only two types:

- *birth* (or arrival) which increase the state variable by one;
- *death* (or departure) which decrease the state variable by one.

The model's name comes from a common application, the use of such models to represent the current size of a population where the transitions are literally due to births and deaths (http://en.wikipedia.org/wiki/Birth-death_process).

Unsurprisingly, birth and death processes have many applications in demography, queueing theory, performance engineering, epidemiology or in biology — they may be used, for example to study the evolution of bacteria, the number of people with a disease within a population, or the number of customers in line at the supermarket (http://en.wikipedia.org/wiki/Birth-death_process).

Definition 6.22 — Birth and death process (Ross, 2003, p. 352)
Let the state variable $X(t)$ be the number of people in a system at time t. Now, suppose that whenever there are n people in the system

- the time until the next birth/arrival is exponentially distributed, with mean λ_n^{-1} ($n \in \mathbb{N}_0$), and independent of the
- the time until the next death/departure which is exponentially distributed with mean μ_n^{-1} ($n \in \mathbb{N}$).

Then $\{X(t) : t \geq 0\}$ is called a birth and death process, with birth rates $\{\lambda_n : n \in \mathbb{N}_0\}$ and death rates $\{\mu_n : n \in \mathbb{N}\}$. •

Remark 6.23 — Birth and death processes (Ross, 2003, pp. 352–353; Kleinrock, 1975, p. 54)

- A birth and death process is a CTMC with state space $\mathbb{N}_0$ for which transitions only to states $n-1$ and $n+1$ are possible from state n.
- The rates at which the process makes a transition when in state i are:

$$\nu_0 = \lambda_0; \tag{6.30}$$

$$\nu_i = \lambda_i + \mu_i, \quad i \in \mathbb{N}. \tag{6.31}$$

The transition probabilities P_{ij} of the embedded DTMC are equal to:

$$\begin{aligned} P_{01} &= 1; && (6.32)\\ P_{i,i+1} &= P(\text{birth before a death given } i \text{ people in the system}) \\ &= \frac{\lambda_i}{\lambda_i + \mu_i}, \quad i \in \mathbb{N}; && (6.33)\\ P_{i,i-1} &= P(\text{death before a birth given } i \text{ people in the system}) \\ &= \frac{\mu_i}{\lambda_i + \mu_i}, \quad i \in \mathbb{N}. && (6.34) \end{aligned}$$

- In addition, the instantaneous transition rates are given by

$$q_{ij} = \nu_i \times P_{ij} = \begin{cases} \lambda_i, & j = i+1 \\ \mu_i, & j = i-1, \end{cases} \tag{6.35}$$

  for $i \neq j$.

- Given that $X(t) = i$, the probability that:
  - one birth occurs in the interval $(t, t + \Delta t]$ is given by
    $$P_{i,i+1}(\Delta t) = \lambda_i \times \Delta t + o(\Delta t);$$
  - one death occurs in the interval $(t, t + \Delta t]$ is equal to
    $$P_{i,i-1}(\Delta t) = \mu_i \times \Delta t + o(\Delta t);$$
  - no death or birth occur in the interval $(t, t + \Delta t]$ amounts to
    $$P_{i,i}(\Delta t) = 1 - (\lambda_i + \mu_i) \times \Delta t + o(\Delta t).$$

  Consequently,
  - multiple births,
  - multiple deaths,
  - a birth and a death,

  in intervals of infinitesimal range Δt ARE NOT POSSIBLE.

- A birth and death process for which $\mu_n = 0$, $n \in \mathbb{N}$ (resp. $\lambda_n = 0$, $n \in \mathbb{N}_0$), is called a pure birth (resp. pure death) process. •

Exercise 6.24 — Rate diagrams and rate matrices of birth and death processes
Draw the rate diagrams of the following birth and death processes:

(a) Poisson process with arrival rate λ (Kulkarni, 1995, Figure 6.5, p. 249);

(b) pure birth process with birth rates λ_i (Kulkarni, 1995, Figure 6.6, p. 249);

(c) pure death process with death rates μ_i (Kulkarni, 1995, Figure 6.7, p. 250);

(d) general birth and death process with birth and death rates λ_i and μ_i, respectively (Kulkarni, 1995, Figure 6.8, p. 251; http://en.wikipedia.org/wiki/Birth-death_process).

Identify the rate matrices of all these CTMC. •

Proposition 6.25 — Kolmogorov's backward and forward equations for birth and death processes (Ross, 2003, examples 6.10 and 6.12, pp. 364 and 368)
For birth and death processes:

- Kolmogorov's backward $(h+t)$ equations become

$$\frac{d\,P_{0j}(t)}{dt} = \lambda_0\,P_{1j}(t) - \lambda_0\,P_{0j}(t), \quad j \in \mathbb{N}_0 \tag{6.36}$$

$$\frac{d\,P_{ij}(t)}{dt} = \lambda_i\,P_{i+1,j}(t) + \mu_i\,P_{i-1,j}(t) - (\lambda_i+\mu_i)\,P_{ij}(t), \quad i \in \mathbb{N},\, j \in \mathbb{N}_0; \tag{6.37}$$

- Kolmogorov's forward $(t+h)$ equations are given by

$$\frac{d\,P_{i0}(t)}{dt} = P_{i1}(t)\,\mu_1 - P_{i0}(t)\,\lambda_0, \quad i \in \mathbb{N}_0 \tag{6.38}$$

$$\frac{d\,P_{ij}(t)}{dt} = P_{i,j-1}(t)\,\lambda_{j-1} + P_{i,j+1}(t)\,\mu_{j+1} - P_{ij}(t)\,(\lambda_j+\mu_j), \quad i \in \mathbb{N}_0,\, j \in \mathbb{N}. \tag{6.39}$$

•

Exercise 6.26 — Kolmogorov's backward and forward equations for a pure birth process
Write Kolmogorov's backward and forward equations for a pure birth process (Ross, 2003, Example 6.9, p. 364). •

Solving Kolmogorov's backward DIFFERENTIAL equations is feasible, namely for some birth and death processes with finite state space such as the CTMC of the next exercise.

Exercise 6.27 — Solving Kolmogorov's backward differential equations
Suppose that:

- a machine works for an exponential amount of time with mean λ^{-1} before breaking down;
- it takes an exponential amount of time with mean μ^{-1} to repair the machine.

(a) Show that if the machine is in working condition (state 0) at time 0 then the probability that it will be working at time t is equal to

$$P_{00}(t) = \frac{\lambda}{\lambda+\mu} \times e^{-(\lambda+\mu)t} + \frac{\mu}{\lambda+\mu}$$

and

$$P_{10}(t) = \frac{\mu}{\lambda+\mu} - \frac{\mu}{\lambda+\mu} \times e^{-(\lambda+\mu)t}.$$

(Ross, 1989, Example 4c, pp. 263–265; Ross, 2003, Example 6.11, pp. 364–366).

(b) Consider $\lambda = 1$, $\mu = 2$ and $t = 10$ and compare $\mathbf{P}(t)$ to its approximations $\left(\mathbf{I} + \frac{\mathbf{R}t}{n}\right)^n$ and $\left[\left(\mathbf{I} - \frac{\mathbf{R}t}{n}\right)^{-1}\right]^n$, where $n = 2^{10}$. •

Solving Kolmogorov's forward DIFFERENTIAL equations is also possible in certain cases, namely for pure birth processes, as shown by Proposition 6.28 and Exercise 6.30.

Moreover, Kolmogorov's forward differential equations are in fact DIFFERENTIAL-DIFFERENCE equations; they can always be solved, at least in principle, by recurrence, that is, successive substitution (Cooper, 1981, p. 16).

Proposition 6.28 — Solving Kolmogorov's forward equations for pure birth processes (Ross, 1989, Proposition 4.1, p. 266)
Let $\{X(t) : t \geq 0\}$ be a pure birth process with rates λ_i, $i \in \mathbb{N}_0$. Then the entries of the TPM can be obtained recursively:

$$P_{ii}(t) = e^{-\lambda_i t}, \quad i \in \mathbb{N}_0; \tag{6.40}$$

$$P_{ij}(t) = \lambda_{j-1} \times e^{-\lambda_j t} \times \int_0^t e^{\lambda_j s} P_{i,j-1}(s)\, ds, \quad i \in \mathbb{N}_0, \quad j = i+1, i+2, \ldots; \tag{6.41}$$

and $P_{ij}(t) = 0$, for $j = 0, 1, \ldots, i$. •

Exercise 6.29 — Solving Kolmogorov's forward equations for pure birth processes

Prove Proposition 6.28 (Ross, 1989, p. 266). •

Exercise 6.30 — Solving Kolmogorov's forward equations for a Yule process

The Yule process is a pure birth process having rates $\lambda_j = j\lambda$, $j \in \mathbb{N}_0$.

(a) Use Proposition 6.28 to prove that, for fixed $i \in \mathbb{N}$,

$$P_{ij}(t) = \binom{j-1}{i-1} \left(e^{-\lambda t}\right)^i \left(1 - e^{-\lambda t}\right)^{j-i}, \quad j = i, i+1, \ldots$$

(Ross, 1989, pp. 266–267).

(b) Give a probabilistic interpretation to the result (Ross, 1983, pp. 144–145). •

Kolmogorov's forward DIFFERENTIAL equations are also easy to derive and handle when we are dealing with pure death processes, such as the ones in exercises 6.31 and 6.32.

Exercise 6.31 — Verifying Kolmogorov's forward differential equations

Admit the size of a population at time t, $X(t)$, can be be described by a pure death process with rates $\mu_k = k\mu$, $k = 0, 1, \ldots, n$, where n ($n \in \mathbb{N}$) represents the initial number of individuals.

(a) Write Kolmogorov's forward differential equations in terms of $P_k(t) \equiv P_{nk}(t) = P[X(t) = k \mid X(0) = n]$.

(b) Show that

$$P_k(t) = \binom{n}{k} \left(e^{-\mu t}\right)^k \left(1 - e^{-\mu t}\right)^{n-k}, \quad k = 0, 1, \ldots, n,$$

verifies the Kolmogorov's forward equations written in (a). •

Exercise 6.32 — Kolmogorov's forward differential equations for a pure death process

There are n_0 ($n_0 \in \mathbb{N}$) seals in an isolated cove; there are all sick and have to be captured and taken from the cove to be treated.

Let $X(t)$ be the number of (uncaptured) seals in the isolated cove at time t and admit that $\{X(t) : t \geq 0\}$ is a pure death process with rates $\mu_k = k\mu$, $k \in \{0, 1, \ldots, n_0\}$.

(a) Derive Kolmogorov's forward equations in terms of $P_k(t) \equiv P_{n_0,k}(t) = P[X(t) = k \mid X(0) = n_0]$.

(b) Argue that the solution to these equations is

$$P_k(t) = P[X(t) = k \mid X(0) = n_0] = \binom{n_0}{k} (p_t)^k (1 - p_t)^{n_0 - k},$$

and identify p_t.

(c) Compute $E[X(t)| X(0) = n_0]$.

(d) Let T_c be the time needed to capture all the seals. Derive the p.d.f. of T_c. •

The p.g.f. method, also called $z-$transform method, is frequently used to reduce the Kolmogorov's forward differential equations to a single PARTIAL DIFFERENTIAL equation, whose solution can be derived for some birth and death processes.

Let:

- $\{X(t) : t \geq 0\}$ be a birth and death process such that $X(0) = i$;
- $P_j(t) \equiv P[X(t) = j \mid X(0) = i]$ be the p.f. of the r.v. $(X(t) \mid X(0) = i)$;
- $P(z,t) = E\left[z^{X(t)} \mid X(0) = i\right], |z| \leq 1$, be the p.g.f. of $(X(t) \mid X(0) = i)$.

Then multiplying the j^{th} Kolmogorov's forward differential equation in (6.39) by z^j and summing up in j (Kulkarni, 1995, p. 279), we get a single equation:

$$\sum_{j \in \mathcal{S}} z^j \times \frac{d\, P_j(t)}{dt} = \sum_{j \in \mathcal{S}} z^j \times \left[P_{j-1}(t)\, \lambda_{j-1} + P_{j+1}(t)\, \mu_{j+1} - P_j(t)\, (\lambda_j + \mu_j)\right]. \tag{6.42}$$

By noting that

$$\sum_{j \in \mathcal{S}} z^j \times \frac{d\, P_j(t)}{dt} = \frac{\partial P(z,t)}{\partial t} \tag{6.43}$$

and that, depending of the birth and death rates, the right term of (6.42) can be written in terms of $P(z,t)$ and

$$\frac{\partial P(z,t)}{\partial z} = \sum_{j \in \mathcal{S}} j z^{j-1} \times P_j(t) = \sum_{j \in \mathcal{S}} (j+1) z^j \times P_{j+1}(t), \tag{6.44}$$

(6.42) is nothing but a (first order) PARTIAL differential equation whose solution is the p.g.f. of the r.v. $(X(t) \mid X(0) = i)$.

Exercise 6.33 — **Solving Kolmogorov's forward equations via the p.g.f. method** (Kleinrock, 1975, Exercise 2.10(a)–(d), p. 81)
Admit $\{X(t) : t \geq 0\}$ is a Yule process — i.e., a pure birth process with birth rates $\lambda_k = k\lambda$, for $k \in \mathbb{N}_0$ — with $X(0) = 1$.

(a) Derive Kolmogorov's forward equations in terms of $P_k(t) \equiv P_{1k}(t) = P[X(t) = k \mid X(0) = 1]$.

(b) After having rewritten the Kolmogorov's forward equations derived in (a) as a partial differential equation in terms of the p.g.f. of the r.v. $(X(t) \mid X(0) = 1)$, $P(z,t) = E\left[z^{X(t)} \mid X(0) = 1\right]$, verify that

$$P(z,t) = \frac{z\, e^{-\lambda t}}{1 - (1 - e^{-\lambda t}) \times z},\ |z| \leq 1,$$

satisfies that partial differential equation (Cooper, 1981, Exercise 6a)b), p. 34).

(c) Identify the distribution of $(X(t)|\ X(0) = 1)$ and compute $E[X(t)|\ X(0) = 1]$. •

Exercise 6.34 — **Solving Kolmogorov's forward equations via the p.g.f. method (bis)** (Kleinrock, 1975, Exercise 2.12, p. 82)
Let:

- $\{X(t) : t \geq 0\}$ be a birth and death process with $X(0) = 0$ and rates $\lambda_k = \lambda$, $k \in \mathbb{N}_0$ and $\mu_k = k\mu$, $k \in \mathbb{N}$;
- $P_j(t) \equiv P_{0j}(t)$ be the p.f. of the r.v. $(X(t) \mid X(0) = 0)$.

(a) Derive Kolmogorov's forward equations in terms of $P_j(t)$.

(b) After having rewritten the Kolmogorov's forward equations derived in (a) as a partial differential equation in terms of the the p.g.f. of $(X(t) \mid X(0) = 0)$, verify that

$$P(z,t) = \exp\left[-\frac{\lambda \times (1 - e^{-\mu t}) \times (1 - z)}{\mu}\right],\ |z| \leq 1,$$

satisfies that partial differential equation (Cooper, 1981, pp. 32–33).

(c) Rewrite $P(z,t)$ as a power series to identify $P_j(t)$ and calculate $\lim_{t \to +\infty} P_j(t)$ (Cooper, 1981, p. 33). •

Exercise 6.35 — Solving Kolmogorov's forward equations via the p.g.f. method (bis, bis) (Kleinrock, 1975, Exercise 2.14, pp. 82–83)
Let:

- $\{X(t) : t \geq 0\}$ a birth and death process with $X(0) = 1$ and rates $\lambda_k = k\lambda$, $k \in \mathbb{N}_0$ and $\mu_k = k\mu$, $k \in \mathbb{N}$;
- $P_j(t) \equiv P_{1j}(t)$ be the p.f. of the r.v. $(X(t) \mid X(0) = 1)$.

(a) Derive Kolmogorov's forward equations in terms of $P_j(t)$ and a partial differential equation satisfied by the p.g.f. of $(X(t) \mid X(0) = 1)$ (Kulkarni, 1995, Example 6.24, pp. 278–279).

(b) Verify that $P(z,t) = \frac{\mu\left[1-e^{(\lambda-\mu)t}\right]-\left[\lambda-\mu\, e^{(\lambda-\mu)t}\right]z}{\mu-\lambda\, e^{(\lambda-\mu)t}-\lambda\left[1-e^{(\lambda-\mu)t}\right]z}$ satisfies the partial differential equation derived in (a).

(c) Calculate the expected value and the variance of $(X(t) \mid X(0) = 1)$.

(d) After having rewritten $P(z,t)$ as a power series, show that

$$P_j(t) = \begin{cases} \alpha(t), & j = 0 \\ [1-\alpha(t)] \times [1-\beta(t)] \times [\beta(t)]^{j-1}, & j \in \mathbb{N}, \end{cases}$$

and obtain expressions for $\alpha(t)$ and $\beta(t)$ (Kulkarni, 1995, Example 6.24, p. 281).

(e) Find the extinction probability, $\lim_{t\to+\infty} P_0(t)$. •

6.4.2 Classification of states

The concepts of accessibility, communication, irreducibility, transience and recurrence for CTMC can be defined in the same lines as for DTMC.[19] Consequently, these concepts are going to be briefly discussed.

Definition 6.36 — CTMC and accessibility, communication, irreducibility, transience and recurrence (Kulkarni, 1995, definitions 6.2–6.8, pp. 283–285)
Let:

- $\{X(t) : t \geq 0\}$ be a CTMC with state space $\mathcal{S}$, TPM $\mathbf{P}(t)$ and initial state i;
- S_1 be the time of the first jump of this stochastic process;
- $T_j = \inf\{t \geq S_1 : X(t) = j\}$ be the first time the CTMC enters state $j \in \mathcal{S}$;
- $T_i = \inf\{t \geq S_1 : X(t) = i\}$ be the first time the CTMC returns to state $i \in \mathcal{S}$;
- $f_{ij} = P[T_j < +\infty \mid X(0) = i]$ be the probability that the first visit to state j (resp. the first return to the initial state i if $j = i$) occurs in finite time;
- $\mu_{ij} = E[T_j \mid X(0) = i]$ be the expected time until the first visit to state j (resp. the first return to the initial state i if $j = i$).

Then, for $i, j \in \mathcal{S}$:

- state j is said to be accessible from state i, i.e., $i \to j$, if $P_{ij}(t) > 0$ for some $t \geq 0$;
- states i and j are said to communicate, i.e., $i \leftrightarrow j$, if $i \to j$ and $j \to i$;[20]
- a set of states $C \subset \mathcal{S}$ is said to be a communicating class if

  (i) $i, j \in C \;\Rightarrow\; i \leftrightarrow j$

  (ii) $i \in C, i \leftrightarrow j \;\Rightarrow\; j \in C$;
- a communicating class $C \subset \mathcal{S}$ is said to be closed if $i \in C, j \notin C \;\Rightarrow\; i \not\to j$;
- the CTMC is said to be irreducible if its state space $\mathcal{S}$ is a single closed communicating class, i.e., if all states communicate with each other; otherwise, the CTMC is called reducible;

[19] **Please refer to Morais (2013, Chap. 3, Section 3.3).**
[20] Two states that communicate are obviously said to be in the same class.

- state i is said to be recurrent if $f_{ii} = 1$;
- state i is called transient if $f_{ii} < 1$;
- a recurrent state i is said to be
  (i) positive recurrent if $\mu_{ii} < +\infty$
  (ii) null recurrent if $\mu_{ii} = +\infty$. •

Remark 6.37 — Periodicity (Kulkarni, 1995, p. 287)
T_j is a continuous r.v. and, thus, if state j is accessible from state i ($i \to j$) then it is possible to visit j at any time $t > 0$ starting from i.[21] Consequently, the notion of PERIOD of a state of a CTMC DOES NOT EXIST. •

Since CTMC can be alternatively described in terms of holding times and an embedded DTMC, can accessibility, communication, irreducibility, transience and recurrence be defined in terms of such DTMC?

YES!

This is indeed possible if we are dealing with what is called a regular CTMC, i.e., with no instantaneous states.

Definition 6.38 — Regular CTMC (Ross, 1983, p. 142)
A CTMC is said to be regular if with probability one, the number of transitions in any finite length time is also finite, that is, if $\sup_{i \in \mathcal{S}} \nu_i < +\infty$. •

Proposition 6.39 — Accessibility, communication, irreducibility, transience and recurrence redefined for CTMC (Kulkarni, 1995, theorems 6.8 and 6.9, pp. 284–285)
Let:

- $\{X(t) : t \geq 0\}$ be a regular CTMC with state space $\mathcal{S}$ and TPM $\mathbf{P}(t) = [P_{ij}(t)]_{i,j \in \mathcal{S}}$;
- $\mathbf{R} = [r_{ij}]_{i,j \in \mathcal{S}}$ be the associated rate matrix (or infinitesimal generator), where $r_{ij} = q_{ij} = \nu_i \times P_{ij}$ $(i \neq j)$ and $r_{ij} = -\nu_i$ $(i = j)$;
- $\{X_n : n \in \mathbb{N}_0\}$ be the embedded DTMC with TPM $\mathbf{P} = [P_{ij}]_{i,j \in \mathcal{S}}$, where[22]

[21] That is, if $\exists s > 0 : P_{ij}(s) > 0$ then $P_{ij}(t) > 0, \forall t > 0$.
[22] This is because the quantities are undefined when $\nu_i = 0$ (Kulkarni, 1995, p. 284).

$$P_{ij} = \begin{cases} \frac{q_{ij}}{\nu_i}, & \nu_i \neq 0,\ i \neq j \\ 0, & \nu_i \neq 0,\ i = j \\ 0, & \nu_i = 0,\ i \neq j \\ 1, & \nu_i = 0,\ i = j. \end{cases} \tag{6.45}$$

Then

$\{X(t) : t \geq 0\}$		$\{X_n : n \in \mathbb{N}_0\}$
$i \to j$	$\Leftrightarrow$	$i \to j$
$i \leftrightarrow j$	$\Leftrightarrow$	$i \leftrightarrow j$
C is a communicating class	$\Leftrightarrow$	C is a communicating class
MC is irreducible	$\Leftrightarrow$	MC is irreducible
i is recurrent	$\Leftrightarrow$	i is recurrent
i is transient	$\Leftrightarrow$	i is transient

•

Remark 6.40 — Transience and recurrence redefined for CTMC (Kulkarni, 1995, p. 285)
Immediate consequences of Proposition 6.39:

- recurrence and transience are class properties;
- the criteria (necessary and sufficient conditions for...) to test recurrence and transience of a DTMC[23] can be used to establish the recurrence and transience of the embedded DTMC and therefore of the CTMC. •

Needless to say that positive and null recurrence cannot be defined in terms of that embedded DTMC because those two concepts rely on the holding times. However, the next proposition establishes a criterion for positive (resp. null) recurrence somewhat related to a result related to the positive recurrence of the DTMC.

Proposition 6.41 — Criterion for positive (resp. null) recurrence (Kulkarni, 1995, Theorem 6.10, p. 285)
Let:

- $\{X(t) : t \geq 0\}$ be an irreducible and recurrent CTMC with state space $\mathcal{S}$;

[23] These are necessary and sufficient conditions for recurrence and transience: state i is recurrent iff $\sum_{n=1}^{+\infty} P_{ii}^n = +\infty$; state i is transient iff $\sum_{n=1}^{+\infty} P_{ii}^n < +\infty$.

- $\{X_n : n \in \mathbb{N}_0\}$ be the recurrent embedded DTMC with TPM $\mathbf{P} = [P_{ij}]_{i,j \in \mathcal{S}}$;
- $\underline{\pi}$ be a positive solution to $\underline{\pi} = \underline{\pi} \times \mathbf{P}$.

Then the CTMC is positive (resp. null) recurrent IFF $\sum_{i \in \mathcal{S}} \frac{\pi_i}{\nu_i} < +\infty$ (resp. $\sum_{i \in \mathcal{S}} \frac{\pi_i}{\nu_i} = +\infty$). •

Proposition 6.41 also proves that positive and null recurrence are class properties in the CTMC setting (Kulkarni, 1995, p. 286).[24]

[24]Please refer to Kulkarni (1995, Example 6.28, pp. 286–287) for a positive recurrent CTMC with null recurrent embedded DTMC and vice-versa.

6.4.3 Limit behavior of CTMC

Computing the TPM $\mathbf{P}(t)$ for a fixed finite t is not a trivial problem to handle, algebraically or numerically (Kulkarni, 1995, p. 282). Expectedly, we shift our focus to the study of the behavior of $\mathbf{P}(t)$ as $t \to +\infty$. But can we determine $\lim_{t\to+\infty} \mathbf{P}(t)$?

YES!

What follows provides answers to questions, such as:

- when does $P_{ij}(t)$ have a limit as $t \to +\infty$?
- how to compute $\lim_{t\to+\infty} P_{ij}(t)$?

(Kulkarni, 1995, p. 282).

Example/Exercise 6.42 — Limit behavior of $\mathbf{P}(t)$

(a) The CTMC described in Exercise 6.27 has TPM equal to

$$\mathbf{P}(t) = \begin{bmatrix} \frac{\lambda}{\lambda+\mu} & -\frac{\lambda}{\lambda+\mu} \\ -\frac{\mu}{\lambda+\mu} & -\frac{\mu}{\lambda+\mu} \end{bmatrix} \times e^{-(\lambda+\mu)t} + \begin{bmatrix} \frac{\mu}{\lambda+\mu} & \frac{\lambda}{\lambda+\mu} \\ \frac{\mu}{\lambda+\mu} & \frac{\lambda}{\lambda+\mu} \end{bmatrix},$$

thus,

$$\lim_{t\to+\infty} \mathbf{P}(t) = \begin{bmatrix} \frac{\mu}{\lambda+\mu} & \frac{\lambda}{\lambda+\mu} \\ \frac{\mu}{\lambda+\mu} & \frac{\lambda}{\lambda+\mu} \end{bmatrix},$$

obviously independent of the initial state of the CTMC (Kulkarni, 1995, Example 6.25, p. 282).

(b) Consider the CTMC described in Kulkarni (1995, Example 6.13, pp. 261–262), with five states and the following rate matrix

$$\begin{bmatrix} -\lambda_1 & 0 & \lambda_1 & 0 & 0 \\ 0 & -\lambda_2 & 0 & \lambda_2 & 0 \\ 0 & \mu_1 & -(\mu_1+\lambda_2) & 0 & \lambda_2 \\ \mu_2 & 0 & 0 & -(\mu_2+\lambda_1) & \lambda_1 \\ 0 & 0 & \mu_2 & \mu_1 & -(\mu_1+\mu_2) \end{bmatrix},$$

where $\lambda_1 = 1$, $\lambda_2 = 2$, $\mu_1 = 0.1$ and $\mu_2 = 0.15$ (Kulkarni, 1995, Example 6.26, p. 283).

After having drawn the rate diagram of this CTMC, use the *Mathematica* function *MatrixExp* to obtain $\mathbf{P}(t)$ and investigate the limit behavior of this TPM.

(c) Consider the CTMC from Exercise 6.35 — now with $X(0) = i$ and $\lambda > \mu$. It can be shown that

$$\lim_{t\to+\infty} P_{ij}(t) = \begin{cases} \left(\frac{\mu}{\lambda}\right)^i, & j = 0 \\ 0, & j \in \mathbb{N}, \end{cases}$$

thus, the limiting probabilities are dependent of the initial state (Kulkarni, 1995, Example 6.27, p. 283). •

After this example/exercise, we proceed with results concerning the limit behavior of the TPM $\mathbf{P}(t)$ of a general CTMC.

Proposition 6.43 — Limit behavior of $\mathbf{P}(t)$ (Kulkarni, 1995, theorems 6.11–6.12 and Corollary 6.3, pp. 287–288)
Let $\{X(t) : t \geq 0\}$ be a CTMC. Then:

- $\lim_{t\to+\infty} P_{jj}(t) = \frac{1}{\nu_j \times \mu_{jj}}$, where $1/\mu_{jj}$ is taken to be 0 if $\mu_{jj} = +\infty$;
- $\lim_{t\to+\infty} P_{ij}(t) = \begin{cases} \frac{f_{ij}}{\nu_j \times \mu_{jj}}, & \nu_j > 0 \\ f_{ij}, & \nu_j = 0, \end{cases}$
  where, once again, $1/\mu_{jj}$ is taken to be 0 if $\mu_{jj} = +\infty$;
- if j is a transient or null recurrent state of the CTMC then $\lim_{t\to+\infty} P_{ij}(t) = 0$, for all $i \in \mathcal{S}$. •

Now, we turn our attention to the limit behavior of positive recurrent (i.e., ergodic), irreducible CTMC. Expectedly, it depends on the stationary distribution of the embedded DTMC.

Theorem 6.44 — Limiting behavior of irreducible, positive recurrent CTMC (Kulkarni, 1995, Theorem 6.13, p. 288; Ross, 1983, p. 152)
Let:

- $\{X(t) : t \geq 0\}$ be an irreducible, positive recurrent CTMC;
- $\{X_n : n \in \mathbb{N}_0\}$ be the embedded DTMC;
- $\underline{\pi} = [\pi_j]_{j\in\mathcal{S}}$ be the unique stationary distribution of the embedded DTMC.[25]

[25] I.e., $\pi_j = \sum_{i\in\mathcal{S}} \pi_i P_{ij}$, $j \in \mathcal{S}$, and $\sum_{j\in\mathcal{S}} \pi_j = 1$; in other words $\underline{\pi} = \underline{\pi}\mathbf{P}$.

Then the limiting probabilities

$$P_j = \lim_{t \to +\infty} P_{ij}(t) \tag{6.46}$$

are given by

$$P_j = \frac{\frac{\pi_j}{\nu_j}}{\sum_{k \in \mathcal{S}} \frac{\pi_k}{\nu_k}}, \quad j \in \mathcal{S}. \tag{6.47}$$

•

Remark 6.45 — Limiting behavior of irreducible, positive recurrent CTMC (Ross, 1983, p. 152)

- P_j also equals the long-run proportion of time the CTMC is in state j.
- If the initial state is chosen according to the limiting probabilities $\{P_j : j \in \mathcal{S}\}$, then $P[X(t) = j] = \sum_{i \in \mathcal{S}} P_i \times P_{ij}(t) = P_j$, for all t, i.e., the resultant CTMC is stationary.[26] •

The next theorem gives one method of computing the limiting distribution of $X(t)$ in terms of the rate matrix.

Theorem 6.46 — Limiting distribution of an irreducible, positive recurrent CTMC in terms of its rate matrix (Kulkarni, 1995, Theorem 6.11, p. 289; Ross, 1983, p. 152)
Let $\{X(t) : t \geq 0\}$ be an irreducible, positive recurrent CTMC with rate matrix $\mathbf{R}$. Then the limiting distribution, represented by the row vector $\underline{P} = [P_j]_{j \in \mathcal{S}}$, is given by the unique non negative solution to

$$\begin{cases} \underline{P} \times \mathbf{R} = \underline{0} \\ \sum_{j \in \mathcal{S}} P_j = 1. \end{cases} \tag{6.48}$$

•

Remark 6.47 — Limiting distribution of an irreducible, positive recurrent CTMC in terms of its rate matrix

- $\underline{P} \times \mathbf{R} = \underline{0}$ can be written as

$$P_j \times \nu_j = \sum_{i \in \mathcal{S}} P_i \times q_{ij}, \quad j \in \mathcal{S} \tag{6.49}$$

(Ross, 1983, p. 152), where $q_{ii} = 0$.

[26] In fact, $P[X(t) = j] = \sum_{i \in \mathcal{S}} P_i \times P_{ij}(t) = \sum_{i \in \mathcal{S}} [\lim_{s \to +\infty} P_{ki}(s)] \times P_{ij}(t) = \lim_{s \to +\infty} \sum_{i \in \mathcal{S}} P_{ki}(s) \times P_{ij}(t) = \lim_{s \to +\infty} P_{kj}(s+t) = P_j$.

- $P_j \times \nu_j =$ rate at which the process leaves state j,

  because P_j is the proportion of time the process is in state j and when it is in state j it leaves at rate ν_j (Ross, 1983, p. 153).

- $\sum_{i \in \mathcal{S}} P_i \times q_{ij} =$ rate at which the process enters state j,

  because P_i is the proportion of time the process is in state i and when it is in state i it departs to state j at rate q_{ij} (Ross, 1983, p. 153).

- Since equations (6.49) can be thought as a statement of the equality of the rate at which the process leaves and enters state j, they are sometimes referred to as *balance equations* (Ross, 1983, p. 153).

- An irreducible CTMC is positive recurrent IFF there is a solution to the system of equations (6.48).[27] Hence, like in the DTMC setting, by solving these equations, we are automatically guaranteed positive recurrence of the CTMC. •

Exercise 6.48 — Limiting distribution of an irreducible, positive recurrent CTMC in terms of its rate matrix

Derive and solve the balance equations of the CTMC with the following rate matrices:

(a) $\begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu \end{bmatrix}$ (Kulkarni, 1995, Example 6.29, p. 290);

(b) $\begin{bmatrix} -\lambda_1 & 0 & \lambda_1 & 0 & 0 \\ 0 & -\lambda_2 & 0 & \lambda_2 & 0 \\ 0 & \mu_1 & -(\mu_1 + \lambda_2) & 0 & \lambda_2 \\ \mu_2 & 0 & 0 & -(\mu_2 + \lambda_1) & \lambda_1 \\ 0 & 0 & \mu_2 & \mu_1 & -(\mu_1 + \mu_2) \end{bmatrix}$,

where $\lambda_1 = 1$, $\lambda_2 = 2$, $\mu_1 = 0.1$ and $\mu_2 = 0.15$ (Kulkarni, 1995, Example 6.30, p. 291).[28] •

Let us now determine the limiting probabilities for a birth and death process (Ross, 1983, p. 153), with rates λ_n, $n \in \mathbb{N}_0$, and μ_n, $n \in \mathbb{N}$. These are obtained by equating the rate at which the process leaves a state with the rate at which it enters that state,[29] as follows:

[27] See Kulkarni (1995, Theorem 6.15, p. 290).

[28] Try not to solve (b) by hand...

[29] This is the result of taking limits as $t \to +\infty$ throughout Kolmogorov's forward equations (6.38) – (6.39), setting $\lim_{t \to +\infty} \frac{d P_{ij}(t)}{dt} = \lim_{t \to +\infty} \frac{d P_j(t)}{dt} = 0$ (because if $\frac{d P_{ij}(t)}{dt}$ converges then it must converge to 0) and $\lim_{t \to +\infty} P_{ij}(t) = P_j$, and normalizing so that $\sum_{j \in \mathcal{S}} P_j = 1$ (Cooper, 1981, p. 21).

State	Rate at which process leaves state	=	Rate at which process enters state
0	$P_0\lambda_0$	=	$P_1\mu_1$
$n \in \mathbb{N}$	$P_n(\lambda_n + \mu_n)$	=	$P_{n-1}\lambda_{n-1} + P_{n+1}\mu_{n+1}$

and then rewriting and solving these equations in terms of P_0 we get the limiting probabilities in the following proposition.

Proposition 6.49 — Limiting probabilities for a birth and death process (Ross, 1983, p. 154)
Let $\{X(t) : t \geq 0\}$ be a birth and death process with rates λ_n, $n \in \mathbb{N}_0$, and μ_n, $n \in \mathbb{N}$. Then

$$P_0 = \frac{1}{1 + \sum_{n=1}^{+\infty} \frac{\lambda_0 \lambda_1 \dots \lambda_{n-1}}{\mu_1 \mu_2 \dots \mu_n}} \tag{6.50}$$

$$\begin{aligned} P_j &= \frac{\lambda_{j-1}}{\mu_j} P_{j-1} \\ &= P_0 \times \frac{\lambda_0 \lambda_1 \dots \lambda_{j-1}}{\mu_1 \mu_2 \dots \mu_j}, \quad j \in \mathbb{N}. \end{aligned} \tag{6.51}$$

•

For an account on the limiting behavior of reducible CTMC, the reader should refer to Kulkarni (1995, pp. 296–299).

Exercise 6.50 — Limiting probabilities for a birth and death process
Prove Proposition 6.49 (Ross, 1983, p. 154). •

Exercise 6.51 — Limiting probabilities for a birth and death process
A taxi company has one mechanic who replaces fuel pumps when they fail. Assume:

- the waiting time in days until a fuel pump fails is exponentially distributed with parameter $\frac{1}{300}$;
- the company has 1000 cars;
- the repair time for each car is exponentially distributed with expected repair time of $\frac{1}{4}$ days.

Find the long-run distribution for $X(t)$, the number of cars with a broken fuel pump at time t, by considering $\{X(t) : t \geq 0\}$ a process where a birth corresponds to a broken fuel pump and a death corresponds to a repaired fuel pump (Isaacson and Madsen, 1976, Example VII.3.5, p. 246).[30] •

[30] Note that the rates are given by $\lambda_n = \frac{1000-n}{300}$, **for** $n = 0, 1, \dots, 999$, and $\mu_{n+1} = 4$, for $n = 1, \dots, 1000$.

Remark 6.52 — Existence of limiting probabilities for a birth and death process

- Equation (6.50) shows us what condition is needed for the limiting probabilities for a birth and death process to exist:

$$\sum_{n=1}^{+\infty} \frac{\lambda_0 \, \lambda_1 \, \ldots \, \lambda_{n-1}}{\mu_1 \, \mu_2 \, \ldots \, \mu_n} < +\infty \tag{6.52}$$

  (Ross, 1983, p. 154); we are simply requiring that $P_0 > 0$ (Kleinrock, 1975, p. 93).

- We should also note that the condition for the existence of limiting probabilities for a birth and death process is met whenever the sequence $\{\frac{\lambda_k}{\mu_k} : k \in \mathbb{N}\}$ remains below the unit from some k onwards, i.e., if

$$\exists k_0 : \frac{\lambda_k}{\mu_k} < 1, \forall k \geq k_0 \tag{6.53}$$

  (Kleinrock, 1975, p. 94).

  Simply stated, in order for those expressions to represent a probability distribution we have to place a condition on the birth and death rates that essentially says that the system occasionally empties (Kleinrock, 1975, p. 93). •

Remark 6.53 — Classification of states of a birth and death process (Kleinrock, 1975, pp. 93–94)
Let

$$S_1 = \sum_{n=1}^{+\infty} \frac{\lambda_0 \, \lambda_1 \, \ldots \, \lambda_{n-1}}{\mu_1 \, \mu_2 \, \ldots \, \mu_n} \tag{6.54}$$

$$S_2 = \sum_{n=1}^{+\infty} \frac{\mu_1 \, \mu_2 \, \ldots \, \mu_n}{\lambda_0 \, \lambda_1 \, \ldots \, \lambda_n}. \tag{6.55}$$

Then, all states will be:

- positive recurrent (i.e., ergodic) iff $S_1 < +\infty$ and $S_2 = +\infty$;
- null recurrent iff $S_1 = +\infty$ and $S_2 = +\infty$;
- transient iff $S_1 = +\infty$ and $S_2 < +\infty$;

It is the ergodic case that gives rise to the equilibrium/limiting probabilities and that is of most interest to our studies. •

Exercise 6.54 — (Existence of) limiting probabilities for a birth and death process (Ross, 1993, Exercise 5.13, p. 179)
The size of a biological population is assumed modeled as a birth and death process — for which immigration is not allowed when the population size is N or larger — with rates

$$\lambda_i = \begin{cases} k\lambda + \theta, & i = 0, 1, \ldots, N-1 \\ k\lambda, & i = N, N+1, \ldots \end{cases}$$

and $\mu_k = k\mu$, $k \in \mathbb{N}$.

Determine the proportion of time that immigration is restricted, in case $N = 3$, $\lambda = \theta = 1$ and $\mu = 2$. •

Exercise 6.55 — (Existence of) limiting probabilities for a birth and death process (bis)
After having established conditions that guarantee the existence of limiting probabilities, obtain (in case it is possible) those probabilities for the birth and death processes with the following birth and death rates λ_i, $i \in \mathbb{N}_0$, and μ_k, $k \in \mathbb{N}$:

(a) $\lambda_i \equiv \lambda$ and $\mu_k \equiv \mu$;

(b) $\lambda_i \equiv \lambda$ and $\mu_k = k\mu$;

(c) $\lambda_i \equiv \lambda$ and $\mu_k = \begin{cases} k\mu, & k = 1, 2, \ldots, c \\ c\mu, & k = c+1, c+2, \ldots \end{cases}$, where $c \in \mathbb{N}$;

(d) $\lambda_i = i\lambda$ and $\mu_k \equiv \mu$;

(e) $\lambda_i = \alpha^i \lambda$ and $\mu_k \equiv \mu$, with $0 < \alpha < 1$
(Kleinrock and Gail, 1996, p. 71);

(f) $\lambda_i = \begin{cases} (M-n)\lambda, & i = 1, 2, \ldots, M \\ 0, & i = M+1, M+2, \ldots \end{cases}$ and $\mu_k \equiv \mu$
(Ross, 1983, Example 5.5(b), p. 155). •

Interestingly enough, some of the birth and death processes described in Exercise 6.55 are in fact related to queueing systems we shall study in Section 6.5.

6.5 Birth and death queueing systems in equilibrium

In this section, we narrowed the class of queueing systems to the ones that can be modelled as birth and death processes — also called *birth and death queues* — the $M/M/s/s + c/K/FCFS$ queueing systems.

Let us remind the reader that these systems have the following characteristics:

- the customers arrive according to a Poisson process at rate λ $(0 < \lambda < +\infty)$;
- the service times are i.i.d. r.v., with exponential distribution with mean μ^{-1} and independent of the interarrival times;
- s $(1 \leq s \leq +\infty)$ servers are set in parallel;
- the waiting room has capacity c, where $0 \leq c \leq +\infty$;
- there is a source of K $(1 \leq K \leq +\infty)$ customers to be served;
- the queue discipline is FCFS.

Since the times between consecutive arrivals are all exponentially distributed r.v., these systems enjoy two very crucial properties. To state them we have to consider the following r.v. and limits (assuming they exist), as defined by Kulkarni (2011, p. 191):

- $X(t) = L_s(t)$ is the number of customers in the system at time t;
- X_n is the number of customers left behind by the n^{th} departure;
- X_n^* is the number of customers in the system as seen by the n^{th} entry (excluding the entering customer) into the system;
- $\hat{X}_n$ is the number of customers in the system as seen by the n^{th} arrival (excluding the arriving customer) to the system;
- $p_j = \lim_{t \to +\infty} P[X(t) = j]$ is the long-run fraction of the time that the system has j customers;
- $\pi_j = \lim_{n \to +\infty} P[X_n = j]$ is the long-run fraction of departures that leave behind j customers in the system;
- $\pi_j^* = \lim_{n \to +\infty} P[X_n^* = j]$ is the long-run fraction of entering customers that see j customers ahead of them in the system;
- $\hat{\pi}_j = \lim_{n \to +\infty} P[\hat{X}_n = j]$ is the long-run fraction of arrivals that see j customers ahead of them in the system.

It is natural to enquire how are these quantities related (Kulkarni, 2011, p. 191).

Proposition 6.56 — Relating π_j and π_j^* (Kulkarni, 2011, p. 192)
Admit the arrivals and departures take place one at a time and that π_j and π_j^* exist. Then

$$\pi_j = \pi_j^*, \, j \in \mathbb{N}_0, \tag{6.56}$$

that is, the long-run fraction of departures that leave behind j customers in the system coincides with the long-run fraction of entries that see j customers ahead of them in the system. •

Exercise 6.57 — Relating π_j and π_j^*
Argue that Proposition 6.56 is valid (Kulkarni, 2011, p. 192). •

Proposition 6.58 — Relating p_j and $\hat{\pi}_j$ (PASTA, Poisson Arrivals See Time Averages) (Kulkarni, 2011, p. 192; Pacheco, 2002, p. 76)
Admit that the customers arrive according to a Poisson process and that $\{N(t+s) : s \geq 0\}$ is independent of $\{X(u) : 0 \leq u \leq t\}$.[31] Then

$$p_j = \hat{\pi}_j, \, j \in \mathbb{N}_0, \tag{6.57}$$

i.e., the fraction of time the system spends in state j coincides with the long-run fraction of customers that find at arrival j customers in the system. •

Since queueing systems with Poisson arrivals possess the PASTA property, arriving customers find on average the same situation in the queueing system as an outside observer looking at the system at an arbitrary point in time (Adan and Resing, 2002, p. 27).

Never forget that the PASTA property is only true when we deal with Poisson arrivals. For instance, in a $D/D/1$ system which is empty at time 0, and with arrivals at $1, 3, 5, \ldots$ and unitary service times, every arriving customer finds an empty system, whereas the fraction of time the system is empty is $\frac{1}{2}$ (Adan and Resing, 2002, p. 27).

Let us proceed and characterize L_s, L_q, W_s and W_q, the four performance measures of a birth and death queue in the long-run or equilibrium.

[31] That is, the arrivals after t (or at t) do not interfere with the number of customers in the system up to time t.

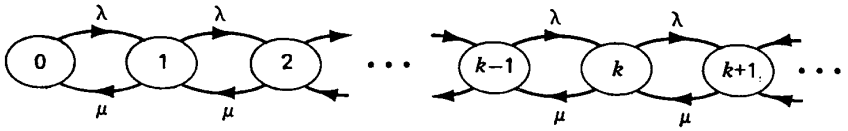
6.5.1 M/M/1, the classical queueing system

The celebrated $M/M/1$ queue is the simplest non trivial interesting queueing system (Kleinrock, 1975, p. 94).

An $M/M/1$ queue may be described by a birth and death process with rates:

$$\lambda_k = \lambda,\ k \in \mathbb{N}_0 \tag{6.58}$$

$$\mu_k = \mu,\ k \in \mathbb{N} \tag{6.59}$$



(Kleinrock, 1975, pp. 94–95).

Furthermore, the necessary and sufficient condition for the ergodicity in the $M/M/1$ system is simply written in terms of the traffic intensity:[32]

$$\rho = \frac{\lambda}{\mu} < 1 \tag{6.60}$$

(Kleinrock, 1975, p. 95).[33] Needless to say that the next results are stated assuming that $\rho < 1$ and refer to $L_s(t)$, $L_q(t)$, $W_s(t)$ and $W_q(t)$ in equilibrium, L_s, L_q, W_s and W_q (respectively).

Proposition 6.59 — M/M/1: distribution of L_s (Kleinrock, 1975, p. 96)
The steady state probability of finding k customers in the $M/M/1$ system only depends on λ and μ through their ratio ρ and is given by:

$$P(L_s = k) = \rho^k\,(1-\rho),\ k \in \mathbb{N}_0, \tag{6.61}$$

i.e., $L_s \sim \text{Geometric}^*(1-\rho)$. •

[32]Note that in this case $\rho = \rho_e$ because the $M/M/1$ system has a waiting area with infinite capacity.

[33]Expectedly, when $\rho \geq 1$, the performance measures $L_s(t)$, $L_q(t)$, $W_s(t)$ and $W_q(t)$ tend to $+\infty$, as $t \to +\infty$.

Exercise 6.60 — M/M/1: distribution of L_s
Prove Proposition 6.59 (Kleinrock, 1975, pp. 95–96). •

Exercise 6.61 — M/M/1: characteristics of L_s
Consider an $M/M/1$ queueing system.

(a) Plot the p.f. of L_s for $\rho = \frac{1}{2}$ (Kleinrock, 1975, Figure 3.2, p. 97).

(b) Obtain the expected value and the variance of L_s as a function of ρ.

(c) Plot $E(L_s)$ (Kleinrock, 1975, Figure 3.3, p. 97) to show that this performance measure grows in an unbounded fashion with ρ (Kleinrock, 1975, p. 98).

(d) Show that L_s stochastically increases with the arrival rate, λ, and with the expected service time, μ^{-1}.[34] •

Proposition 6.62 — M/M/1: distribution of L_q
The equilibrium probability of finding k customers waiting to be served in the $M/M/1$ system equals:

$$P(L_q = k) = \begin{cases} P(L_s \leq 1) = 1 - \rho^2, & k = 0 \\ P(L_s = k+1) = \rho^{k+1}\,(1-\rho), & k \in \mathbb{N}. \end{cases} \tag{6.62}$$

•

Exercise 6.63 — M/M/1: distribution of L_q
Prove Proposition 6.62. •

Proposition 6.64 — M/M/1: distribution of W_s
Since the service times are memoryless in the $M/M/1$ queueing system, we get:[35]

$$(W_s \mid L_s = k) \sim \text{Gamma}(k+1, \mu),\ k \in \mathbb{N}_0; \tag{6.63}$$

$$W_s \sim \text{Exponential}(\mu(1-\rho)). \tag{6.64}$$

•

[34] A r.v. X, whose distribution depends on the parameter θ, is said to stochastically increase with θ if $P_\theta(X > x)$ is an increasing function of θ, for all $-\infty < x < +\infty$.

[35] Given that upon arrival a customer finds k customers in the $M/M/1$ system, he/she will leave this system after the completion of $1 + (k-1) + 1$ services: the service that have already started when the customer arrived; the ones of the $k-1$ customers waiting to be served when the customer arrived; and her/his own service.

Exercise 6.65 — M/M/1: distribution of W_s
Prove Proposition 6.64.[36] •

Proposition 6.66 — M/M/1: distribution of W_q
For the $M/M/1$ queueing system, W_q is a mixed r.v. with the following characteristics:

$$(W_q \mid L_s = 0) \stackrel{st}{=} 0; \tag{6.65}$$

$$(W_q \mid L_s = k) \sim \text{Gamma}(k, \mu),\ k \in \mathbb{N}; \tag{6.66}$$

$$(W_q \mid W_q > 0) \sim \text{Exponential}(\mu(1-\rho)); \tag{6.67}$$

$$F_{W_q}(t) = \begin{cases} 0, & t < 0 \\ 1-\rho, & t = 0 \\ (1-\rho) + \rho \times F_{Exp(\mu(1-\rho))}(t), & t > 0. \end{cases} \tag{6.68}$$

•

Exercise 6.67 — M/M/1: distribution of W_q
Prove Proposition 6.66. •

Exercise 6.68 — M/M/1 queueing system
Consider an $M/M/1$ queueing system and draw the graphs of the following parameters in terms of ρ:

(a) the limiting probability that the system is empty;

(b) $E(W_q)$;

(c) $E(W_s)$ (Kleinrock, 1975, Figure 3.4, p. 97). •

Exercise 6.69 — M/M/1 queueing system (bis)
Prove that:

(a) $V(L_s) = \frac{\rho}{(1-\rho)^2}$ (Prabhu, 1997, p. 15);

(b) $V(L_q) = \frac{\rho^2(1+\rho-\rho^2)}{(1-\rho)^2}$ (Pacheco, 1990, p. 36);

(c) $V(W_s) = \frac{1}{[\mu(1-\rho)]^2}$ (Prabhu, 1997, p. 16);

(d) $V(W_q) = \frac{\rho(2-\rho)}{[\mu(1-\rho)]^2}$ (Prabhu, 1997, p. 16). •

[36] Apply the total probability law to prove (6.163).

The following table condenses the distributions and expected values of the four performance measures of an (ergodic) $M/M/1$.

M/M/1	
Rates	$\lambda_k = \lambda, \quad k \in \mathbb{N}_0$ $\mu_k = \mu, \quad k \in \mathbb{N}$
L_s	$P(L_s = k) = \rho^k (1-\rho), \quad k \in \mathbb{N}_0$ $E(L_s) = \frac{\rho}{1-\rho}$
L_q	$P(L_q = k) = \begin{cases} 1-\rho^2, & k = 0 \\ \rho^{k+1}(1-\rho), & k \in \mathbb{N} \end{cases}$ $E(L_q) = \frac{\rho^2}{1-\rho}$
W_s	$(W_s \mid L_s = k) \sim \text{Gamma}(k+1, \mu),\ k \in \mathbb{N}_0$ $W_s \sim \text{Exponential}(\mu(1-\rho))$ $E(W_s) = \frac{1}{\mu(1-\rho)}$
W_q	$(W_q \mid L_s = k) \sim \text{Gamma}(k, \mu),\ k \in \mathbb{N}$ $F_{W_q}(t) = \begin{cases} 0, & t < 0 \\ 1-\rho, & t = 0 \\ (1-\rho) + \rho \times F_{Exp(\mu(1-\rho))}(t), & t > 0 \end{cases}$ $(W_q \mid W_q > 0) \sim \text{Exponential}(\mu(1-\rho))$ $E(W_q) = \frac{\rho}{\mu(1-\rho)}$

Exercise 6.70 — M/M/1 queueing system (bis, bis)
Admit that defective items from a production line arrive to the repair shop of the same factory according to a Poisson process with constant rate λ. The repair shop has a single-server who completes repairs after independent and exponentially distributed times with expected value equal to 3 minutes.

(a) Determine the distribution of L_s.

(b) The manager of the production line wishes that the probability of having more than 5 defective items waiting for repair does not exceed 10% and that the probability of having an idle server in the repair shop does not exceed 30%. Identify the arrival rates that satisfy both conditions. •

Exercise 6.71 — More on the M/M/1 queueing model
Passengers arrive to a passport control area in a very small airport according to a Poisson process having rate equal to 30 passengers per hour. The passport control has a sole officer who completes checks after independent and exponentially distributed times with expected value equal to 1.5 minutes.

(a) Obtain the probability that the server is idle.

(b) Calculate the expected number of passengers in the passport control area.

(c) What is the probability that passengers form a queue and its expected size?

(d) Determine not only the expected time an arriving passenger spends in the passport control area, but also the expected value this passenger waits to be served.

(e) What is the probability that a passenger waits at least 10 minutes until her/his passport starts to be checked by the officer? •

Exercise 6.72 — More on the M/M/1 queueing model (bis)
People arrive to a phone booth according to a Poisson process with rate 1/10 persons per minute and the durations of the phone calls are independent and exponentially distributed r.v. with common expected value equal to 3 minutes.

(a) What is the probability that someone has to wait to make a phone call?

(b) Determine the expected size of the queue.

(c) The phone company will install another phone booth in the same area if the expected waiting time is of at least 3 minutes. Calculate the increase in the arrival rate that justifies the installation of the second phone booth.

(d) Obtain the probability that a customer has to wait more than 10 minutes to start her/his phone call.

(e) What is the probability that a person does not spend more than 10 minutes from arrival to the system until the end of the phone call.

(f) Calculate the percentage of time the phone booth is being used. •

Exercise 6.73 — More on the M/M/1 queueing model (bis)
Vehicles arrive to a car wash according to a Poisson process with rate 5 vehicles per hour and the durations of the car washes are independent and exponentially distributed r.v. with expected value equal to 10 minutes. Admit that the car wash has a waiting area with infinite capacity.

(a) What is the probability that a vehicle has to wait to be washed?

(b) Determine the expected number of vehicles that have to wait to be washed.

(c) Compute the standard deviation of the time spent in queue waiting for the vehicle to be washed.

(d) What is the percentage of time the car wash machines are nor working? •

If not all arriving customers decide to join the system we have the phenomenon of BALKING (Prabhu, 1997, p. 30) or a queueing model with discouragement (Reynolds, 1968) or DISCOURAGED ARRIVALS. We are particularly interested in the case where the decision to join (or not!) is made on the basis on the number of customers found in the system upon arrival (Prabhu, 1997, p. 30). The next exercise illustrates an $M/M/1$ with balking (or discouraged arrivals); the $M/M/s$ with balking is addressed later on.

Exercise 6.74 — M/M/1 queueing system with discouraged arrivals
Consider an $M/M/1$ queueing system where arrivals tend to get discouraged when more and more people are present in the system. One possible way to model this effect is to consider an HARMONIC DISCOURAGEMENT OF ARRIVALS with respect to the number present in the system, i.e., having birth rates equal to

$$\lambda_k = \frac{\lambda}{k+1},\ k \in \mathbb{N}_0,$$

and keep the death rates equal to $\mu_k = \mu$, $k \in \mathbb{N}$ (Kleinrock, 1975, p. 99).

(a) Draw the rate diagram of this birth and death process (Kleinrock, 1975, Figure 3.5, p. 100).

(b) Verify that the process is ergodic if $\frac{\lambda}{\mu} < +\infty$.

(c) Show that the limiting probabilities are given by

$$P_k = e^{-\lambda/\mu}\frac{(\lambda/\mu)^k}{k!},\ k \in \mathbb{N}_0$$

(Kleinrock, 1975, pp. 99–100), i.e., $L_s \sim \text{Poisson}(\lambda/\mu)$. •

6.5.2 M/M/∞, the queueing system with responsive servers

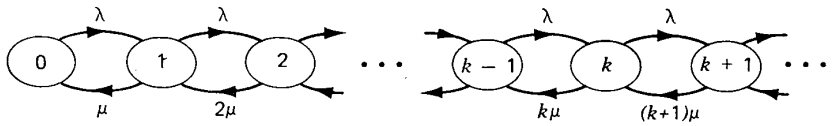
The $M/M/\infty$ can be thought as a system where there is always a new server for each arriving customer (Kleinrock, 1975, p. 101).[37]

This queueing system can be obviously described by a birth and death process with rates

$$\lambda_k = \lambda, \quad k \in \mathbb{N}_0, \tag{6.69}$$

$$\mu_k = k\mu, \quad k \in \mathbb{N}, \tag{6.70}$$

and the ergodic condition is simply $\frac{\lambda}{\mu} < +\infty$



(Kleinrock, 1975, p. 101).

Proposition 6.75 — M/M/∞: distribution of L_s (Kleinrock, 1975, p. 101)
$L_s \sim \text{Poisson}(\lambda/\mu)$. •

Exercise 6.76 — M/M/∞: distribution of L_s
Prove Proposition 6.75. •

Suffice to say that we have not to wait in this system and the time spent in the system coincides with the duration of the service, thus,

$$L_q \overset{st}{=} 0 \tag{6.71}$$

$$W_s \sim \text{Exponential}(\mu) \tag{6.72}$$

$$W_q \overset{st}{=} 0. \tag{6.73}$$

Since this system is quite simple to describe in the equilibrium state we are tempted to state the transient behavior of the number of customers in the system at time t.

Proposition 6.77 — M/M/∞: transient behavior of number of customers
Let $X(t)$ be number of customers in the $M/M/\infty$ system at time t. Then

$$(X(t) \mid X(0) = 0) \sim \text{Poisson}(\lambda\,(1 - e^{-\mu t})/\mu). \tag{6.74}$$

•

[37] It may also be interpreted as a system with a responsive server who accelerates her/his service rate linearly (Kleinrock, 1975, p. 101), to avoid any customers waiting.

Exercise 6.78 — M/M/∞: transient behavior of number of customers
Prove Proposition 6.77, by deriving the Kolmogorov's forward equations and the associated partial differential equation. •

M/M/∞	
Rates	$\lambda_k = \lambda, \quad k \in \mathbb{N}_0$ $\mu_k = k\mu, \quad k \in \mathbb{N}$
L_s	$L_s \sim \text{Poisson}(\lambda/\mu)$
L_q	$L_q \overset{st}{=} 0$
W_s	$W_s \sim \text{Exp}(\mu)$
W_q	$W_q \overset{st}{=} 0$
$X(t) =$ number of customers in the system at time t $(X(t) \mid X(0) = 0) \sim \text{Poisson}(\lambda\,(1 - e^{-\mu t})/\mu)$	

Even though the $M/G/\infty$ queueing system[38] cannot be modeled as a birth and death process, we digress and state the transient and limit behavior of its number of customers.

Proposition 6.79 — M/G/∞: transient and limit behavior of number of customers (Pacheco, 2002, p. 91)
Let $X(t)$ be number of customers in the $M/G/\infty$ system at time t. Then

$$(X(t) \mid X(0) = 0) \quad \sim \quad \text{Poisson}\left(\lambda \int_0^t [1 - G(t-s)]\, ds\right) \tag{6.75}$$

$$\lim_{t \to +\infty} (X(t) \mid X(0) = 0) \quad \sim \quad \text{Poisson}(\lambda/\mu). \tag{6.76}$$

•

The limit distribution of the number of customers in the $M/G/\infty$ system is insensitive to the distribution of the service time — it only depends on its mean (Adan and Resing, 2002, p. 111).

Exercise 6.80 — M/G/∞: transient and limit behavior of the number of customers
Prove Proposition 6.79 (Pacheco, 2002, p. 91). •

[38] This system is associate with a Poisson arrival process with rate λ and service time distribution function G with finite expected value μ^{-1}.

$\mathbf{M/G/\infty}$
$(X(t) \mid X(0) = 0) \sim \text{Poisson}\left(\lambda \int_0^t [1 - G(t-s)]\, ds\right)$
$\lim_{t \to +\infty} (X(t) \mid X(0) = 0) \sim \text{Poisson}(\lambda/\mu)$

Exercise 6.81 — **M/G/∞: transient behavior of the number of customers** Users arrive at a library according to a Poisson process with rate equal to 3 users per minute and spend in the library an amount of time with Uniform$(10, 210)$ distribution.

What is the expected number of users in the library two hours after it opened (Pacheco, 2002, p. 91)? •

Exercise 6.82 — **M/G/∞: transient behavior of the number of customers (bis)** According to historical data regarding the National Gallery, the interarrival and visit times (all in minutes) have Exponential(λ) and Gama(α, β) distributions, respectively.

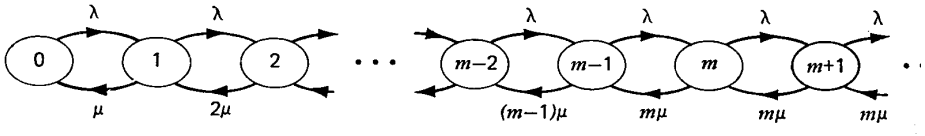
Considering $\lambda = 1$, $\alpha = 3$, $\beta = 0.1$ and that the National Gallery opens at 10AM, obtain an approximate value for the probability that the number of visitors exceeds 200 at 3PM. •

6.5.3 M/M/m, the m–server case

Once again we consider a queueing system with an unlimited waiting area and a constant arrival rate; this system provides a maximum of m servers, is within the reach of a birth and death formulation and leads to

$$\lambda_k = \lambda, \quad k \in \mathbb{N}_0 \tag{6.77}$$

$$\mu_k = \begin{cases} k\mu, & k = 1, \ldots, m \\ m\mu, & k = m+1, m+2, \ldots \end{cases} \tag{6.78}$$



(Kleinrock, 1975, p. 102).

From these birth and death rates, it is easily seen that the condition for ergodicity is written, expectedly, in terms of the traffic intensity:

$$\rho = \frac{\lambda}{m\mu} < 1. \tag{6.79}$$

Proposition 6.83 — M/M/m: distribution of L_s (Kleinrock, 1975, pp. 102–103)
The limit probability of finding k customers in the $M/M/m$ system depends, once again, on λ and μ through the traffic intensity $\rho = \frac{\lambda}{m\mu}$:

$$P(L_s = k) = \begin{cases} P_0 \frac{(m\rho)^k}{k!}, & k = 0, 1, \ldots, m-1 \\ P_0 \frac{m^m \rho^k}{m!}, & k = m, m+1, \ldots, \end{cases} \tag{6.80}$$

where $P_0 = P(L_s = 0) = \left[\sum_{k=0}^{m-1} \frac{(m\rho)^k}{k!} + \frac{(m\rho)^m}{m!(1-\rho)}\right]^{-1}$. •

Proposition 6.84 — M/M/m: Erlang's C formula or delay probabilities (Kleinrock, 1975, p. 103)
In a $M/M/m$ system, the long-run fraction of customers that are delayed is equal to

$$C(m, m\rho) = P(\text{delay}) = P\left(L_s \geq m\right) = \frac{\frac{(m\rho)^m}{m!(1-\rho)}}{\sum_{k=0}^{m-1} \frac{(m\rho)^k}{k!} + \frac{(m\rho)^m}{m!(1-\rho)}}. \tag{6.81}$$

This probability is usually referred to as *Erlang's C formula* (or Erlang's second formula).[39] Expectedly, the p.f. of L_s can be written in terms of $C(m, m\rho)$:

[39] Some authors, such as Kleinrock (1975, p. 103), represent this probability by $C(m, \rho)$ instead of $C(m, m\rho)$. We prefer the notation of Pacheco (2002, p. 80).

$$P(L_s = k) = \begin{cases} \frac{m!}{k!}\,(1-\rho)(m\rho)^{k-m} C(m, m\rho), & k = 0, 1, \ldots, m-1 \\ (1-\rho)\,\rho^{k-m}\,C(m, m\rho), & k = m, m+1, \ldots. \end{cases} \tag{6.82}$$

•

Proposition 6.85 — M/M/m: distribution of L_q
The equilibrium probability of finding k customers waiting in line in the $M/M/m$ queueing system is simply given by:

$$P(L_q = k) = \begin{cases} P(L_s \leq m) = 1 - \rho\, C(m, m\rho), & k = 0 \\ P(L_s = m + k) = (1-\rho)\,\rho^k\, C(m, m\rho), & k \in \mathbb{N}. \end{cases} \tag{6.83}$$

•

Proposition 6.86 — M/M/m: distribution of W_s
The distribution of W_s conditional $L_s = k$, depends on the fact that the arriving customer is immediately served or not:

$$(W_s \mid L_s = k) \quad \sim \quad \begin{cases} \text{Exp}(\mu),\ k = 0, \ldots, m-1, \\ \text{Exp}(\mu) \star \text{Gamma}(k - m + 1, m\mu),\ k = m, m+1, \ldots, \end{cases} \tag{6.84}$$

where $\star$ represents the convolution (or sum of two independent r.v.).

The SURVIVAL function of W_s has two expressions, depending on whether ρ is equal to $\frac{m-1}{m}$ or not:

$$1 - F_{W_s}(t) = \begin{cases} [1 + \mu t\, C(m, m\rho)]\, e^{-\mu t}, & t \geq 0, \quad \rho = \frac{m-1}{m} \\ \left[1 + \frac{e^{\mu[1-m(1-\rho)]t}}{1-m(1-\rho)} \times C(m, m\rho)\right] e^{-\mu t}, & t \geq 0, \quad \rho \neq \frac{m-1}{m}. \end{cases} \tag{6.85}$$

•

Proposition 6.87 — M/M/m: distribution of W_q
Once more W_q is a mixed r.v. In this case:

$$(W_q \mid L_q = k) \quad \sim \quad \text{Gamma}(k - m + 1, m\mu),\ k = m, m+1, \ldots; \tag{6.86}$$

$$(W_q \mid W_q > 0) \quad \sim \quad \text{Exponential}(m\mu(1-\rho)); \tag{6.87}$$

$$1 - F_{W_q}(t) \quad = \quad \begin{cases} 1, & t < 0 \\ C(m, m\rho), & t = 0 \\ C(m, m\rho) \times [1 - F_{Exp(m\mu(1-\rho))}(t)], & t > 0. \end{cases} \tag{6.88}$$

•

M/M/m	
Rates	$\lambda_k = \lambda, \quad k \in \mathbb{N}_0$ $\mu_k = \begin{cases} k\mu, & k = 1, \ldots, m \\ m\mu, & k = m+1, m+2, \ldots \end{cases}$
L_s	$P(L_s = k) = \begin{cases} \frac{m!}{k!}(1-\rho)(m\rho)^{k-m} C(m, m\rho), & k = 0, 1, \ldots, m-1 \\ (1-\rho)\,\rho^{k-m}\, C(m, m\rho), & k = m, m+1, \ldots \end{cases}$ $C(m, m\rho) = P(L_s \geq m) = \frac{\frac{(m\rho)^m}{m!(1-\rho)}}{\sum_{k=0}^{m-1} \frac{(m\rho)^k}{k!} + \frac{(m\rho)^m}{m!(1-\rho)}}$ $C(1, \rho) = \rho$ $C(2, 2\rho) = \frac{2\rho^2}{1+\rho}$ $E(L_s) = m\rho + \frac{\rho}{1-\rho} C(m, m\rho)$
L_q	$P(L_q = k) = \begin{cases} 1 - \rho\, C(m, m\rho), & k = 0 \\ (1-\rho)\,\rho^k\, C(m, m\rho), & k \in \mathbb{N} \end{cases}$ $E(L_q) = \frac{\rho}{1-\rho} C(m, m\rho)$
W_s	$(W_s \mid L_s = k) \sim \begin{cases} \text{Exp}(\mu),\ k = 0, \ldots, m-1, \\ \text{Exp}(\mu) \star \text{Gamma}(k-m+1, m\mu),\ k = m, m+1, \ldots \end{cases}$ $1 - F_{W_s}(t) = \begin{cases} [1 + \mu t\, C(m, m\rho)]\, e^{-\mu t}, & t \geq 0, \quad \rho = \frac{m-1}{m} \\ \left[1 + \frac{e^{\mu[1-m(1-\rho)]t}}{1-m(1-\rho)} \times C(m, m\rho)\right] e^{-\mu t}, & t \geq 0, \quad \rho \neq \frac{m-1}{m} \end{cases}$ $E(W_s) = \frac{1}{\mu} + \frac{C(m, m\rho)}{m\mu(1-\rho)}$
W_q	$(W_q \mid L_q = k) \sim \text{Gamma}(k-m+1, m\mu),\ k = m, m+1, \ldots$ $(W_q \mid W_q > 0) \sim \text{Exponential}(m\mu(1-\rho))$ $1 - F_{W_q}(t) = \begin{cases} 1, & t < 0 \\ C(m, m\rho), & t = 0 \\ C(m, m\rho) \times F_{Exp(\mu(1-\rho))}(t), & t > 0 \end{cases}$ $E(W_q) = \frac{C(m, m\rho)}{m\mu(1-\rho)}$

Exercise 6.88 — **M/M/m: characteristics of L_s, L_q, W_s and W_q**
Prove propositions 6.83–6.87 and derive the following results referring to the performance measures L_s, L_q, W_s and W_q, and to $C(m, m\rho)$:

(a) $E(L_s) = m\rho + \frac{\rho}{1-\rho}\, C(m, m\rho)$;

(b) $E(L_q) = \frac{\rho}{1-\rho}\, C(m, m\rho)$;

(c) $E(W_s) = \frac{1}{\mu} + \frac{C(m,m\rho)}{m\mu(1-\rho)}$;

(d) $E(W_q) = \frac{C(m,m\rho)}{m\mu(1-\rho)}$;

(e) $C(1,\rho) = \rho$; $C(2,2\rho) = \frac{2\rho^2}{1+\rho}$. •

Exercise 6.89 — M/M/m: properties of $C(m, m\rho)$
Draw graphs to illustrate the following properties of Erlang's C formula:

(a) $C(m, m\rho) \uparrow \rho$;

(b) $\lim_{\rho\to 0^+} C(m, m\rho) = 0$;

(c) $\lim_{\rho\to 1^-} C(m, m\rho) = 1$;

(d) $C(m, m\rho) \downarrow m$;

(e) $0 < C(m, m\rho) < C(1, \rho) = \rho$, $m = 2, 3, \ldots$. •

Exercise 6.90 — M/M/m queueing system
A system has two servers, who attend to customers in a FCFS basis and whose service times are independent and exponentially distributed r.v. with mean value 1.8 minutes. Considering that customers arrive to the system according to a Poisson process with rate equal to 1 customer per minute and that the system, compute:

(a) the probability that there are more than 10 customers in the system;

(b) the expected time a customer spends in line waiting to be served;

(c) the expected number of customers in the system;

(d) the probability that exactly one server is idle. •

Exercise 6.91 — M/M/m queueing system (bis)
A small public office has two officers, who service times are independent and exponentially distributed r.v. with rate equal to 60 visitors per hour. Admit that the times between consecutive arrivals of visitors are i.i.d. r.v. exponentially distributed with parameter equal to 100 visitors per hour and calculate:

(a) the probability that there are more than 4 visitors in the system;

(b) the expected number of visitors in the system;

(c) the expected time a visitor spends in the system. •

Exercise 6.92 — M/M/m queueing system (bis, bis) (Hsu, 2011, p. 361, Exercise 9.13)
A corporate computing center has two computers of the same capacity. The jobs arriving at the center are of two types, internal and external jobs. These jobs arrive according to two independent Poisson processes with rates 18 internal jobs per hour and 15 external jobs per hour. The service times are i.i.d. r.v. exponentially distributed with mean 3 minutes.

(a) Find the average waiting time per arriving job, when the two computers handle both types of jobs.

(b) Obtain the average waiting time per arriving internal job, when one computer is used exclusively for internal jobs and the other for external jobs. Comment the result in light of (a). •

Exercise 6.93 — M/M/m queueing system (bis, bis, bis)
A department has three secretaries, who process requests that arrive according to a Poisson process with rate equal to 20 requests per 8 hours. Assume that the processing times are independent and exponentially distributed r.v. with expected value equal to 40 minutes.

(a) What is the percentage of time all (resp. at least one of) the secretaries are busy?

(b) Obtain the expected time one waits for a request to be completely processed

(c) Admit that due to financial problems one of the secretaries had to be fired. Recompute the quantities in (a) and (b). •

Exercise 6.94 — Designing a M/M/m queueing system
Airplanes arrive to an airport according to a Poisson process having rate equal to 18 airplanes per hour, and their times in a runway during landing are independent and exponentially distributed r.v. with expected value equal to 2 minutes.

Derive the number of runways the airport should have so that the probability that an arriving airplane waits to land (*delay probability*) does not exceed 0.20. •

Exercise 6.95 — A M/M/m queueing system with impatient customers (Isaacson and Madsen, 1976, Example VII.3.4, pp. 245–246)
Assume:

- customers arrive at a ticket counter with m windows according to a Poisson process with parameter 6 per minute;

- customers are served on a first-come-first-served basis;
- service times are independent and exponentially distributed with mean $\frac{1}{3}$ of a minute.

(a) What is the minimum number of windows needed to guarantee that the line does not get infinitely long?

(b) Assume $m = 4$ and that we are dealing with impatient customers who:

- wait for service if $L_s \leq 4$;
- wait for service with probability $\frac{1}{2}$ if $L_s = 5$;
- leave if $L_s \geq 6$.

What is the distribution of L_s? •

The system described in Exercise 6.95 in another example of a queueing system with BALKING — consequently, we close this subsection readdressing the issue of BALKING (see Prabhu, 1997, pp. 30–32). We shall also introduce the reader to $M/M/m$ queueing systems with RENEGING (see Prabhu, 1997, pp. 32–35).

Remark 6.96 — The M/M/m system with balking
Let us remind the reader that:

- the term *balk* (or *baulk*)[40] means to hesitate or to be unwilling to go on;
- in a queueing system with balking, the probability that an arriving customer joins the system depends on the number of customers, say k, found in the system upon arrival; let us represent such probability by b_k.[41]

In practice $\{b_k,\ k \in \mathbb{N}_0\}$ is an non-increasing sequence — the more customers found upon arrival, the *less* prone is the arriving customer to join the queueing system. A few examples taken from Prabhu (1997, p. 30):

(i) $b_k = \frac{1}{k-m+2}$, for $k \geq m$ ($b_k = 1,\ k < m$);

(ii) $b_k = \alpha^{k-m+1}$, for $k \geq m$ ($b_k = 1,\ k < m$), where $\alpha \in (0, 1)$;

(iii) $b_k = 1$, for $k < C$, and $b_k = 0$, for $k \geq C$ (where C is some critical value). •

[40] Variant spelling of *balk*.
[41] $1 - b_k$ is the balking probability (Prabhu, 1997, p. 30).

Exercise 6.97 — The M/M/m system with balking
Capitalizing on the fact that, for a $M/M/m$ system with balking, $\{L_s(t),\, t \geq 0\}$ is a birth and death process with rates

$$\lambda_k = \lambda\, b_k, \quad k \in \mathbb{N}_0, \tag{6.89}$$

$$\mu_k = \begin{cases} k\mu, & k = 1, \ldots, m \\ m\mu, & k = m+1, m+2, \ldots \end{cases} \tag{6.90}$$

(Prabhu, 1997, p. 31), derive:

(a) conditions that ensure the existence of limiting probabilities for this birth and death process;[42]

(b) the limiting probabilities referring to L_s.[43] •

Remark 6.98 — The M/M/m system with balking (bis)
In general, balking behavior can be characterized by a r.v. D defined as the maximum number of customers an arriving customer would be prepared to wait for (Prabhu, 1997, p. 31). Thus, according to Prabhu (1997, p. 32),

$$b_k = P(D \geq k),\, k \in \mathbb{N}_0, \tag{6.91}$$

and in the previous examples we have:

(i) $P(D = k) = \frac{1}{(k-m+2)(k-m+3)}$, $k \geq m-1$;

(ii) $P(D = k) = (1-\alpha)\alpha^{k-m+1}$, $k \geq m-1$;

(iii) $D = C - 1$.

D can be thought as the psychological effect that the observed number of customers in the system has on the arriving customer (Prabhu, 1997, p. 32). •

Reneging occurs when a customer decides to leave the system after having waited for sometime to be served (Prabhu, 1997, p. 32).

[42] The series $\sum_{k=m}^{+\infty} \beta_k\, \rho^{k-m}$ should converge, and note that $\beta_m = 1$, $\beta_k = \prod_{j=m}^{k-1} b_j$ $(k \geq m+1)$ and $\rho = \frac{\lambda}{m\mu}$.

[43] $P_0 = \left[\sum_{k=0}^{m-1} \frac{(m\rho)^k}{m!} + \frac{(m\rho)^m}{m!} \sum_{k=m}^{+\infty} \beta_k\, \rho^{k-m}\right]^{-1}$; $P_k = P_0 \times \frac{(m\rho)^k}{k!}$, for $k = 1, \ldots, m-1$; and $P_k = P_m\, \beta_k \rho^{k-m}$, for $k = m, m+1, \ldots$.

Remark 6.99 — The M/M/m system with reneging
If we admit that the times until reneging by the successive customers are i.i.d. r.v. with Exponential(ν) distribution, then $\{L_s(t),\, t \geq 0\}$ is a birth and death process with rates

$$\lambda_k = \lambda, \quad k \in \mathbb{N}_0 \tag{6.92}$$

$$\mu_k = \begin{cases} k\mu, & k = 1, \ldots, m \\ m\mu + (k-m)\nu, & k = m+1, m+2, \ldots \end{cases} \tag{6.93}$$

(Prabhu, 1997, pp. 32–33).

The resulting system is called a $M/M/m$ system with reneging at rate ν (Prabhu, 1997, p. 33). •

Exercise 6.100 — The M/M/1 system with reneging (Kleinrock and Gail, 1996, p. 93)
Consider a $M/M/1$ system, with parameters λ and μ, and reneging at rate ν.

(a) Draw the rate diagram and express P_{k+1} in terms of P_k (Kleinrock and Gail, 1996, p. 94).

(b) Derive P_k, when $\nu = \mu$ (Kleinrock and Gail, 1996, p. 94). •

Exercise 6.101 — The M/M/m system with reneging
Consider a $M/M/m$ system with reneging at rate ν and let

$$\rho = \frac{\lambda}{m\mu}, \quad \delta = \frac{\lambda}{\nu}, \quad \beta_m = 1 \quad \beta_k = \prod_{j=1}^{k-m} \left(j + \frac{\delta}{\rho} \right). \tag{6.94}$$

Prove that:

(a) the existence of limiting probabilities for this birth and death process is ensured if the series $\sum_{k=m}^{+\infty} \beta_k \, \rho^{k-m}$ converges;

(b) the p.f. of L_s is given by

$$P_0 = \left[\sum_{k=0}^{m-1} \frac{(m\rho)^k}{m!} + \frac{(m\rho)^m}{m!} \sum_{k=m}^{+\infty} \beta_k \, \delta^{k-m} \right]^{-1}$$

$$P_k = \begin{cases} P_0 \times \frac{(m\rho)^k}{k!}, & k = 1, \ldots, m-1 \\ P_m \, \beta_k \rho^{k-m}, & k = m, m+1, \ldots \end{cases}$$

(Prabhu, 1997, p. 33).

Derive:

(c) the distribution of the virtual waiting time of a customer, who arrived at time t and decided to stay until he/she is served (Prabhu, 1997, p. 34);

(d) the transient and the equilibrium distributions of the waiting time of a customer, who arrived at time t (Prabhu, 1997, pp. 34–35). •

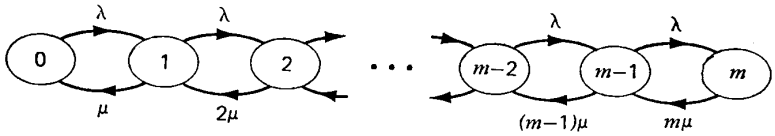
6.5.4 M/M/m/m, the m–server pure loss system

The $M/M/m/m$ queueing system is a $m-$server system with no waiting area: each newly arriving customer is given her/his private server; however, if a customer arrives when all servers are occupied, that customer is lost (Kleinrock, 1975, p. 105). Unsurprisingly, this queueing system also called a $m-$server pure loss system.

This queueing system can be modeled as birth and death process with

$$\lambda_k = \begin{cases} \lambda, & k = 0, 1, \ldots, m-1 \\ 0, & k = m, m+1, \ldots \end{cases} \tag{6.95}$$

$$\mu_k = \begin{cases} k\mu, & k = 1, \ldots, m \\ 0, & k = m+1, m+2, \ldots \end{cases} \tag{6.96}$$



(Kleinrock, 1975, p. 105).

Since we are dealing with a finite state space ($\mathcal{S} = \{0, 1, \ldots, m\}$), ergodicity is obviously assured as long as the traffic intensity $\rho = \frac{\lambda}{m\mu}$ is finite, and this condition can be written in terms of the offered load:

$$m\rho = \frac{\lambda}{\mu} < +\infty. \tag{6.97}$$

Proposition 6.102 — M/M/m/m: distribution of L_s (Kleinrock, 1975, p. 105)
The limit probability of finding k customers in the $M/M/m/m$ queueing system depends on the offered load $m\,\rho = \frac{\lambda}{\mu}$:

$$P(L_s = k) = \begin{cases} P_0 \frac{(m\rho)^k}{k!}, & k = 0, 1, \ldots, m \\ 0, & k = m+1, m+2, \ldots, \end{cases} \tag{6.98}$$

where $P_0 = P(L_s = 0) = \left[\sum_{k=0}^{m} \frac{(m\rho)^k}{k!}\right]^{-1}$. •

Proposition 6.103 — M/M/m/m: Erlang's B formula or blocking probabilities
In a $M/M/m/m$ system, the long-run fraction of lost≡blocked customers is equal to

$$P(L_s = m) = \frac{\frac{(m\rho)^m}{m!}}{\sum_{k=0}^{m} \frac{(m\rho)^k}{k!}} \tag{6.99}$$

$$= B(m, m\rho). \tag{6.100}$$

It is usually referred to as *Erlang's B formula* (or Erlang's first formula) and it was first derived by Erlang in 1917 (Kleinrock, 1975, p. 106).

The (equilibrium) distribution of L_s is sometimes written in terms of $B(m, m\rho)$:

$$P(L_s = k) = \begin{cases} \frac{m!}{k!\,(m\rho)^{m-k}} \times B(m, m\rho), & k = 0, 1, \ldots, m \\ 0, & k = m+1, m+2, \ldots \end{cases} \tag{6.101}$$

•

Exercise 6.104 — **M/M/m/m: distributions of L_s, L_q, W_s and W_q**
Prove propositions 6.102–6.103 and show that $E(L_s) = m\rho[1 - B(m, m\rho)]$. •

We are dealing, once more, with a system where there is no wait — in this case because there is no waiting area and, thus, arriving customers who find all the m servers busy are lost. As a consequence:

$$L_q \overset{st}{=} 0 \tag{6.102}$$

$$W_s \sim \text{Exp}(\mu) \tag{6.103}$$

$$W_q \overset{st}{=} 0. \tag{6.104}$$

M/M/m/m	
Rates	$\lambda_k = \begin{cases} \lambda, & k = 0, 1, \ldots, m-1 \\ 0, & k = m, m+1, \ldots \end{cases}$ $\mu_k = \begin{cases} k\mu, & k = 1, \ldots, m \\ 0, & k = m+1, m+2, \ldots \end{cases}$
L_s	$P(L_s = k) = \begin{cases} \frac{\frac{(m\rho)^k}{k!}}{\sum_{k=0}^{m} \frac{(m\rho)^k}{k!}} = \frac{m!}{k!\,(m\rho)^{m-k}} \times B(m, m\rho), & k = 0, 1, \ldots, m \\ 0, & k = m+1, m+2, \ldots \end{cases}$ $B(m, m\rho) = \frac{\frac{(m\rho)^m}{m!}}{\sum_{k=0}^{m} \frac{(m\rho)^k}{k!}}$ $B(1, \rho) = \frac{\rho}{1+\rho}$ $B(2, 2\rho) = \frac{2\rho^2}{1+2\rho+2\rho^2}$ $E(L_s) = m\rho[1 - B(m, m\rho)]$
L_q	$L_q \overset{st}{=} 0$
W_s	$W_s \sim \text{Exp}(\mu)$
W_q	$W_q \overset{st}{=} 0$

Exercise 6.105 — M/M/m/m system
Answer the questions in Exercise 6.91, considering that the small public office has no waiting area. •

Exercise 6.106 — M/M/m/m: properties of $B(m, m\rho)$
Draw graphs to illustrate the following properties of Erlang's B formula:

(a) $B(m, m\rho) \uparrow \rho$;

(b) $\lim_{\rho \to 0^+} B(m, m\rho) = 0$;

(c) $\lim_{\rho \to +\infty} B(m, m\rho) = 1$;

(d) $B(m, m\rho) \downarrow m$;

(e) $0 < B(m, m\rho) < B(1, \rho) = \frac{\rho}{1+\rho}$, $m = 2, 3, \ldots$. •

Exercise 6.107 — Erlang's B and C formulae
Prove the following results.

(a) Erlang's B formula can be obtained in a recursive way:

$$\begin{aligned} B(m, m\rho) &= B(m, \lambda/\mu) \\ &= \begin{cases} \frac{\rho}{1+\rho}, & m = 1 \\ \frac{m\rho \times B(m-1,\lambda/\mu)}{m+m\rho \times B(m-1,\lambda/\mu)}, & m = 2, 3, \ldots \end{cases} \end{aligned} \tag{6.105}$$

(b) Erlang's C formula is related with Erlang's B formula as follows:

$$C(m, m\rho) = \frac{m \times B(m, m\rho)}{m - m\rho \times [1 - B(m, m\rho)]} \tag{6.106}$$

$$C(m, \lambda/\mu) = \frac{1}{1 + (m - m\rho) \times [m\rho \times B(m-1, \lambda/\mu)]^{-1}}, \tag{6.107}$$

where $B(0, \lambda/\mu) = 1$.

(c) $C(m, m\rho) > B(m, m\rho)$.

(d) $C(m, m\rho) = \frac{1}{1+\frac{1-\rho}{\rho} \times \frac{m-1-m\rho \times C(m-1,m\rho)}{(m-1-m\rho) \times C(m-1,m\rho)}}$, for $m > m\rho + 1$. •

The computation of values of $B(m, m\rho)$ maybe problematic for large values of m, such as $m = 150$. However, let us remind the reader that there is a simple and numerically stable recursion for the blocking probabilities in a $M/M/m/m$, formula (6.105), derived in Exercise 6.107.

Exercise 6.108 — Recursion for Erlang's B formula
Admit that:

- customers arrive according to a Poisson process at a parking lot near a small shopping center with a rate of 60 cars per hour,
- the mean parking time is 2.5 hours and the parking lot offers place to 150 cars,
- when the parking lot is full, an arriving customer has to park his car somewhere else

(Adan and Resing, 2002, p. 114).

(a) Write an expression for the fraction of customers finding all places occupied on arrival.

(b) Use *Mathematica* to obtain such value.

(c) Use *Mathematica* to implement the recursion in formula (6.105). •

Exercise 6.109 — M/M/m/m system (bis) (Hsu, 2011, p. 364, Exercise 9.20)
An air freight terminal has 2 loading docks. An aircraft which arrives when all docks are full is diverted to another terminal. The aircrafts arrive to the air freight terminal according to a Poisson process with rate 3 aircrafts per hour. The service times are i.i.d. r.v. exponentially distributed with mean equal to 2 hours.

(a) Find the proportion of arriving aircrafts that are diverted to another terminal.

(b) Obtain the average number of aircrafts in the air freight terminal. •

Exercise 6.110 — M/M/m/m system (bis, bis) (Kulkarni, 2011, p. 203, Example 6.8)
Consider a telephone switch modelled by a $M/M/m/m$ queueing system, with an arrival rate of $\lambda = 4$ calls per minute and a service rate of $\mu = 5$ calls per minute per server.

(a) What is the expected number of calls in the switch in the steady state?

(b) How large should the switch capacity be if we desire to lose no more than 10% of the incoming calls? •

Exercise 6.111 — M/M/m/m vs. M/M/m system

A small town bank has m employees working at the counter. The stream of customers is Poisson with rate λ and the service times are i.i.d. r.v. with exponential distribution and mean μ^{-1}.

(a) Admit the customers leave the bank if they are not immediately served.

(i) For which values of λ and μ this system reaches equilibrium?

(ii) Derive the steady state p.f. of the number of customers in the system an arriving customer sees.

(iii) Determine the probability of losing customers and find the mean number of customers found in this system, when $m = 2$, $\lambda = 4$ and $\mu = 5$.

(b) Now, consider that after some marketing "manoeuvres" the customers are willing to wait for their service.

Obtain the expected value of the number of customers in the system and compare it with the previous value, admitting once again that $\lambda = 4$ and $\mu = 5$. Comment the results. •

6.5.5 M/M/m/m+d, the m server with finite storage

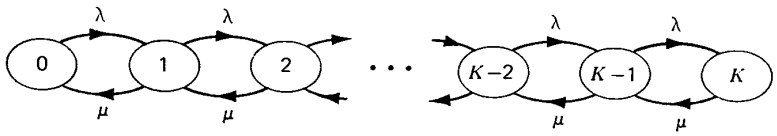
The $M/M/m/m+d$ queueing system is a $m-$server system which can accommodate only a finite number $m+d$ $(m,d \in \mathbb{N})$[44] of customers at a time (Prabhu, 1997, p. 24), so that if a customer arrives — when all the m servers and all the d positions in the waiting area are occupied —, he/she will in fact be refused entry to the system and depart immediately without service (Kleinrock, 1975, p. 103). Newly arriving customers will continue to be generated according to a Poisson process; however, only those who arrive and find the queueing system with strictly less than $m+d$ customers will be allowed entry (Kleinrock, 1975, p. 103).

This queueing system is also aptly termed the $m-$server *loss-delay system*. It can be modelled once again as a birth and death process by effectively *turning off* the Poisson input as soon as the system fills up (Kleinrock, 1975, p. 104):

$$\lambda_k = \begin{cases} \lambda, & k = 0, 1, \ldots, m+d-1 \\ 0, & k = m+d, m+d+1, \ldots \end{cases} \tag{6.108}$$

$$\mu_k = \begin{cases} k\mu, & k = 1, \ldots, m-1 \\ m\mu, & k = m, m+1, \ldots, m+d \\ 0, & k = m+d+1, \ldots \end{cases} \tag{6.109}$$

Remark 6.112 — **M/M/m/m+d: rate diagram**
The corresponding rate diagram when $m = 1$ is



$(K \equiv m+d)$

(Kleinrock, 1975, p. 104). •

We are dealing once again with a finite state space ($\mathcal{S} = \{0, 1, \ldots, m+d\}$), therefore ergodicity is immediately assured as long as $\rho = \frac{\lambda}{m\mu} < \infty$.

Proposition 6.113 — **M/M/m/m+d: distribution of L_s** (Prabhu, 1997, p. 25)
The limit probabilities depend on the traffic intensity $\rho = \frac{\lambda}{m\mu}$ and read as follows:

$$P(L_s = k) = \begin{cases} P_0 \frac{(m\rho)^k}{k!}, & k = 0, 1, \ldots, m-1 \\ P_0 \frac{m^m \rho^k}{m!}, & k = m, m+1, \ldots, m+d, \end{cases} \tag{6.110}$$

where $P_0 = P(L_s = 0) = \left[\sum_{k=0}^{m-1} \frac{(m\rho)^k}{k!} + \frac{(m\rho)^m}{m!} \sum_{k=0}^{d} \rho^k\right]^{-1}$. •

[44] When $d = 0$ we are dealing with a $M/M/m/m$ pure loss system.

Proposition 6.114 — M/M/m/m+d: loss-delay probabilities (Pacheco, 2002, p. 80)

In this system of finite storage, the loss-delay probability is the long-run fraction of customers that are either blocked or delayed:[45]

$$C_d(m, m\rho) = P(L_s \geq m) \tag{6.111}$$

$$= \frac{\frac{(m\rho)^m}{m!} \sum_{k=0}^{d} \rho^k}{\sum_{k=0}^{m-1} \frac{(m\rho)^k}{k!} + \frac{(m\rho)^m}{m!} \sum_{k=0}^{d} \rho^k}. \tag{6.112}$$

The p.f. of L_s can written in terms of $C_d(m, m\rho)$:

$$P(L_s = k) = \begin{cases} \frac{m!}{k!\,(m\rho)^{m-k} \sum_{i=0}^{d} \rho^i} C_d(m, m\rho), & k = 0, 1, \ldots, m-1 \\ \frac{\rho^{k-m}}{\sum_{i=0}^{d} \rho^i} C_d(m, m\rho), & k = m, m+1, \ldots, m+d. \end{cases} \tag{6.113}$$

•

Proposition 6.115 — M/M/m/m+d: distribution of L_q

The equilibrium probability of finding k customers waiting to be served in a $M/M/m/m+d$ queueing system equals:

$$P(L_q = k) = \begin{cases} P(L_s \leq m) = 1 - \left(1 - \frac{1}{\sum_{i=0}^{d} \rho^i}\right) C_d(m, m\rho), & k = 0 \\ P(L_s = m + k) = \frac{\rho^k}{\sum_{i=0}^{d} \rho^i} C_d(m, m\rho), & k = 1, \ldots, d. \end{cases} \tag{6.114}$$

•

Exercise 6.116 — M/M/m/m+d: loss-delay probabilities

Draw graphs to illustrate and comment the following properties of the loss-delay probabilities:

(a) $C_d(m, m\rho) \uparrow d$;

(b) $C_d(m, m\rho) \uparrow \rho$;

(c) $C_d(m, m\rho) \downarrow m$;

(d) $0 < C_d(m, m\rho) < C_d(1, \rho) = \left[1 + \frac{1}{1+\rho \sum_{k=0}^{d} \rho^k}\right]^{-1}$, $m = 2, 3, \ldots$, $d \in \mathbb{N}_0$. •

Exercise 6.117 — M/M/m/m+d: loss-delay probabilities

Prove that for a $M/M/1/1+d$ system the loss-delay probabilities are given by:

[45] We adopted a slightly different notation than Pacheco (2002, p. 80).

$$C_d(1,\rho) = \begin{cases} \frac{\rho-\rho^{d+2}}{1-\rho^{d+2}}, & \rho \neq 1 \\ \frac{d+1}{d+2}, & \rho = 1 \end{cases} \tag{6.115}$$

(Pacheco, 2002, p. 80). •

Proposition 6.118 — M/M/m/m+d: blocking probabilities (Pacheco, 2002, p. 78)
In a $M/M/m/m+d$ system the blocking probability corresponds to the long-run fraction of customers who find $m+d$ clients in the system upon arrival. The blocking probability is usually denoted by P_b, $B(m, m\rho, d)$ or $B_d(m, m\rho)$; it equals P_{m+d}, i.e.,

$$B_d(m, m\rho) = \frac{\frac{(m\rho)^m}{m!}\rho^d}{\sum_{j=0}^{m-1}\frac{(m\rho)^j}{j!} + \frac{(m\rho)^m}{m!}\sum_{k=0}^{d}\rho^k}. \tag{6.116}$$

Furthermore:

- $B_d(m, m\rho) \downarrow d$ (decreases with the size of the waiting room d);
- $B_d(m, m\rho) \downarrow m$ (decreases with the number of servers m);
- $B_d(m, m\rho) \uparrow \rho$ (increases with the offered traffic intensity or the offered load). •

Exercise 6.119 — M/M/1/1+d: blocking probabilities (Pacheco, 2002, p. 78)
Prove that the blocking probability in a $M/M/1/1+d$ system is given by

$$B_d(1,\rho) = \begin{cases} \frac{1}{d+2} & \rho = 1 \\ \frac{\rho^{d+1}(1-\rho)}{1-\rho^{d+2}} & \rho \neq 1. \end{cases} \tag{6.117}$$

•

Remark 6.120 — M/M/m/m+d: input rate and the carried traffic intensity
The input rate and the carried traffic intensity are equal to

$$\begin{aligned} \lambda_e &= \lambda \times (1 - P_b) \\ &= \lambda \times [1 - P(L_s = m+d)] \\ &= \lambda \times [1 - B_d(m, m\rho)] \\ &= \lambda \times \left[1 - \frac{\rho^d}{\sum_{i=0}^{d}\rho^i} C_d(m, m\rho)\right] \qquad (6.118) \\ \rho_e &= \rho \times (1 - P_b) \\ &= \frac{\lambda_e}{m\mu}, \qquad (6.119) \end{aligned}$$

respectively. •

We have to be a bit more careful in the derivation of the distribution of W_q and $W_s \stackrel{d}{=} (W_q + \text{service})$ and take advantage of the following result.

Proposition 6.121 — **Relating the $M/M/m/m+d$ and $M/M/m/m+d-1$ systems** (Pacheco, 2002, p. 84)
Let $L_s^{M/M/m/m+d}$ (resp. $L_s^{M/M/m/m+d-1}$) be the number of customers found in the $M/M/m/m+d$ (resp. $M/M/m/m+d-1$) system by an arriving customer. Then

$$P(L_s^{M/M/m/m+d} = k | L_s^{M/M/m/m+d} < m+d) = P(L_s^{M/M/m/m+d-1} = k). \tag{6.120}$$

•

For the expressions of the c.d.f. of W_s (resp. W_q) written as mixtures of r.v., the reader is referred to Pires (1990, equations (4.86–4.88)) (resp. Pacheco, 2002, pp. 85–86). However, we should refer in passing that:

$$\begin{aligned}
F_{W_q}(0) &= P(W_q = 0) \\
&= P(L_s < m | L_s < m+d) \\
&= 1 - C_{d-1}(m, m\rho); \\
P(0 < W_q \leq t) &= \sum_{k=m}^{m+d-1} F_{Gamma(k-m+1,m\mu)}(t) \times P(L_s = k | L_s < m+d) \\
&= \sum_{k=m}^{m+d-1} F_{Gamma(k-m+1,m\mu)}(t) \times \frac{\rho^{k-m}}{\sum_{i=0}^{d-1} \rho^i} C_{d-1}(m, m\rho).
\end{aligned}$$

Exercise 6.122 — **M/M/m/m+d: expectations of L_q, L_s, W_q and W_s**
Derive the following results referring to L_q, L_s, W_q and W_s:

(a) $E(L_q) = \begin{cases} \frac{d}{2} C_d(m, m\rho), & \rho = 1 \\ \frac{\rho}{1-\rho} \times \frac{1-(d+1)\rho^d + d\rho^{d+1}}{1-\rho^{d+1}} \times C_d(m, m\rho), & \rho \neq 1; \end{cases}$

(b) $E(L_s) = m\rho_e + E(L_q) = m\rho_e + \begin{cases} \frac{d}{2} C_d(m, m\rho), & \rho = 1 \\ \frac{\rho}{1-\rho} \times \frac{1-(d+1)\rho^d + d\rho^{d+1}}{1-\rho^{d+1}} \times C_d(m, m\rho), & \rho \neq 1; \end{cases}$

(c) $E(W_q) = \frac{1}{\lambda_e} E(L_q) = \begin{cases} \frac{d+1}{2m\mu} C_{d-1}(m, m\rho), & \rho = 1 \\ \frac{1}{1-\rho} \times \frac{1-(d+1)\rho^d + d\rho^{d+1}}{1-\rho^d} \times \frac{C_{d-1}(m,m\rho)}{m\mu}, & \rho \neq 1. \end{cases}$

(d) $E(W_s) = \frac{1}{\lambda_e} E(L_s) = \frac{1}{\mu} + E(W_q)$
$= \frac{1}{\mu} + \begin{cases} \frac{d+1}{2m\mu} C_{d-1}(m, m\rho), & \rho = 1 \\ \frac{1}{1-\rho} \times \frac{1-(d+1)\rho^d + d\rho^{d+1}}{1-\rho^d} \times \frac{C_{d-1}(m,m\rho)}{m\mu}, & \rho \neq 1. \end{cases}$

•

M/M/m/m+d	
Rates	$\lambda_k = \begin{cases} \lambda, & k = 0, 1, \ldots, m+d-1 \\ 0, & k = m+d, m+d+1, \ldots \end{cases}$
	$\mu_k = \begin{cases} k\mu, & k = 1, \ldots, m-1 \\ m\mu, & k = m, m+1, \ldots, m+d \\ 0, & k = m+d+1, \ldots \end{cases}$
Input rate	$\rho_e = \frac{\lambda}{m\mu} \times \left[1 - \frac{\rho^d}{\sum_{i=0}^{d} \rho^i} C_d(m, m\rho)\right]$
L_s	$P(L_s = k) = \begin{cases} \frac{m!}{k!\,(m\rho)^{m-k} \sum_{i=0}^{d} \rho^i} C_d(m, m\rho), & k = 0, 1, \ldots, m-1 \\ \frac{\rho^{k-m}}{\sum_{i=0}^{d} \rho^i} C_d(m, m\rho), & k = m, \ldots, m+d \end{cases}$
	$C_d(m, m\rho) = P(L_s \geq m) = \frac{\frac{(m\rho)^m}{m!} \sum_{k=0}^{d} \rho^k}{\sum_{k=0}^{m-1} \frac{(m\rho)^k}{k!} + \frac{(m\rho)^m}{m!} \sum_{k=0}^{d} \rho^k}.$
	$C_d(1, \rho) = \begin{cases} \frac{\rho - \rho^{d+2}}{1-\rho^{d+2}}, & \rho \neq 1 \\ \frac{d+1}{d+2}, & \rho = 1 \end{cases}$
	$E(L_s) = m\rho_e + \begin{cases} \frac{d}{2} C_d(m, m\rho), & \rho = 1 \\ \frac{\rho}{1-\rho} \times \frac{1-(d+1)\,\rho^d + d\,\rho^{d+1}}{1-\rho^{d+1}} \times C_d(m, m\rho), & \rho \neq 1 \end{cases}$
L_q	$P(L_q = k) = \begin{cases} P(L_s \leq m) = 1 - \left(1 - \frac{1}{\sum_{i=0}^{d} \rho^i}\right) C_d(m, m\rho), & k = 0 \\ P(L_s = m+k) = \frac{\rho^k}{\sum_{i=0}^{d} \rho^i} C_d(m, m\rho), & k \in \mathbb{N} \end{cases}$
	$E(L_q) = \begin{cases} \frac{d}{2} C_d(m, m\rho), & \rho = 1 \\ \frac{\rho}{1-\rho} \times \frac{1-(d+1)\,\rho^d + d\,\rho^{d+1}}{1-\rho^{d+1}} \times C_d(m, m\rho), & \rho \neq 1; \end{cases}$
W_s	$E(W_s) = \frac{1}{\mu} + \begin{cases} \frac{d+1}{2m\mu} C_{d-1}(m, m\rho), & \rho = 1 \\ \frac{1}{1-\rho} \times \frac{1-(d+1)\,\rho^d + d\,\rho^{d+1}}{1-\rho^d} \times \frac{C_{d-1}(m, m\rho)}{m\mu}, & \rho \neq 1. \end{cases}$
W_q	$E(W_q) = \begin{cases} \frac{d+1}{2m\mu} C_{d-1}(m, m\rho), & \rho = 1 \\ \frac{1}{1-\rho} \times \frac{1-(d+1)\,\rho^d + d\,\rho^{d+1}}{1-\rho^d} \times \frac{C_{d-1}(m, m\rho)}{m\mu}, & \rho \neq 1. \end{cases}$

Exercise 6.123 — **M/M/m/m+d system** (Kulkarni, 2011, p. 202, Example 6.7) Consider a call center modelled by a $M/M/6/6+4$ with an arrival rate of $\lambda = 60$ calls per hour and a service rate of $\mu = 10$ calls per hour per server.

(a) Compute the limiting distribution of the number of calls in the system in steady state.

(b) Obtain the expected number of calls on hold.

(c) What fraction of the calls are lost? •

Exercise 6.124 — M/M/m/m+d system (bis)

For safety reasons the number of visitors to an exhibition room cannot exceed 2 and not more than 2 visitors are allowed to be the waiting room.

Admit visitors arrive to the exhibition room according to a Poisson process having rate equal to 6 visitors per 15 minutes and that the time spent by each visitor in the exhibition room is an exponentially distributed r.v. with mean equal to 4 minutes.

(a) Determine the probability of finding the exhibition room empty.

(b) What happens to the number of visitors found in the waiting room if its capacity is reduced to 1? Justify and comment your result. •

6.5.6 Birth and death queues in equilibrium, with finite customer population

In this subsection, we no longer consider a Poisson input process with an infinite user population, but rather a FINITE population of possible users that act independently of each other (Kleinrock, 1975, p. 106).

A customer is either in the system[46] or not, and in some sense are expected to *arrive* after a exponentially distributed time with rate λ (Kleinrock, 1975, p. 106).

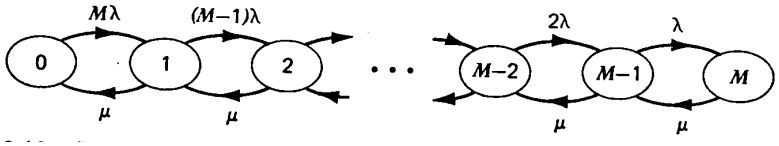
These self regulating systems can be modelled quite appropriately with birth and death processes (Kleinrock, 1975, p. 106), as illustrated in the following exercises.

Exercise 6.125 — M/M/1//m, the single-server system with finite customer population

We model the single-server system with finite customer population with a birth and death process with rates

$$\lambda_k = \lambda(m-k), \quad k = 0, 1, \ldots, m \tag{6.121}$$

$$\mu_k = \mu, \quad k = 1, \ldots, m \tag{6.122}$$



$(M \equiv m)$

(Kleinrock, 1975, pp. 106–107).

(a) If $\rho = \frac{\lambda}{\mu} < +\infty$ then this system reaches equilibrium.

Derive and write in terms of ρ:

(i) the p.f. of L_s (Kleinrock, 1975, p. 107) and its expectation;

(ii) the p.f. of L_q and prove that $E(L_q) = E(L_s) - (1 - P_0)$;

(iii) the expressions for the c.d.f. of W_s and W_q;

(b) Prove that the input rate is equal to $\lambda_e = \sum_{k=0}^{m} \lambda_k P_k = m\,[\lambda - E(L_s)]$.

(c) Use Little's law to write expressions for the expected values of W_s and W_q. •

[46] Consisting of a queue and m servers who provide independent and exponentially distributed services with mean μ^{-1}.

Exercise 6.126 — M/M/m//m, the m–server system with finite customer population[47]

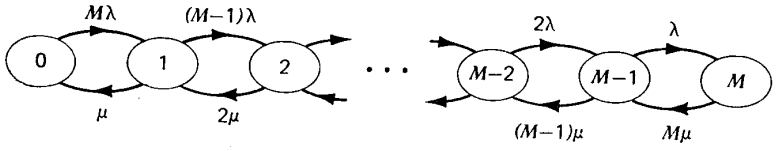
In the $m-$server system with finite customer population, a separate server is provided for each customer in the system (Kleinrock, 1975, p. 107), thus,

$$\begin{aligned} L_q &\overset{st}{=} 0 \\ W_s &\sim \text{Exponential}(\mu) \\ W_q &\overset{st}{=} 0. \end{aligned}$$

The associated birth and death process has rates

$$\lambda_k = \lambda(m-k), \quad k = 0, 1, \ldots, m \tag{6.123}$$

$$\mu_k = k\mu, \quad k = 1, \ldots, m \tag{6.124}$$



$(M \equiv m)$

(Kleinrock, 1975, p. 107).

(a) Once again this system reaches equilibrium if $\rho = \frac{\lambda}{m\mu} < +\infty$.

Derive and write the p.f. of L_s and its expectation (Kleinrock, 1975, p. 108).

(b) Confirm that $L_s \sim \text{Binomial}\left(n, 1 - \frac{1}{1+m\rho}\right)$.[48]

(c) Prove that the input rate is given by $\lambda_e = \sum_{k=0}^{m} \lambda_k P_k = \frac{m\lambda}{1+m\rho}$.

(d) Use Little's law to confirm that $E(W_s) = \frac{1}{\lambda_e} E(L_s) = \frac{1}{\mu}$. •

Exercise 6.127 — M/M/m/m+d/m+d+e, the m–server system with finite customer population and finite storage

Depending on the values of d and e ($d, e \in \mathbb{N}_0 \cup \infty$), this rather general model will reduce to all the previous birth and death queueing systems (Kleinrock, 1975, p. 108), with the exception of the ones with balking or reneging.

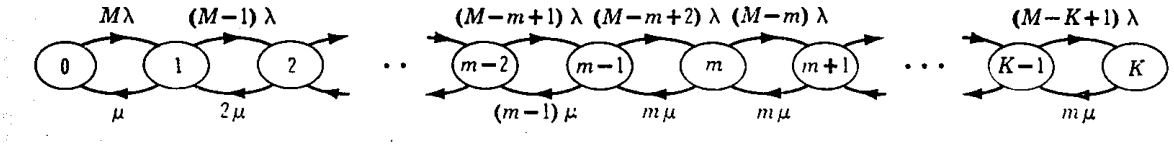
[47] Kleinrock (1975, p. 107) represents this system as $M/M/\infty//m$ and terms it a *finite customer population — "infinite" number of servers.*

[48] Curiously enough, Kleinrock (1975, p. 108) fails to note this important fact.

The rates of the associated birth and death process are equal to

$$\lambda_k = \lambda(m+d+e-k), \quad k=0,1,\ldots,m+d-1 \tag{6.125}$$

$$\mu_k = \begin{cases} k\mu, & k=1,\ldots,m \\ m\mu, & k=m+1,\ldots,m+d \end{cases} \tag{6.126}$$



$(K \equiv m+d,\ M \equiv m+d+e)$

(Kleinrock, 1975, pp. 108–109).

(a) If $\rho = \frac{\lambda}{m\mu} < +\infty$ and $m+d < +\infty$ then this system reaches equilibrium.

Derive the p.f. of L_s (Kleinrock, 1975, p. 109) and write in terms of ρ.

(b) Prove that, for $d=0$,

$$P_k = \frac{\binom{m+e}{k}(m\rho)^k}{\sum_{i=0}^{m}\binom{m+e}{i}(m\rho)^i},\ k=0,1,\ldots,m,$$

that is, L_s has a distribution known as the Engset distribution[49] (Kleinrock, 1975, p. 109). •

[49]For more details, the reader is referred to http://en.wikipedia.org/wiki/Erlang_(unit)#Engset_formula and http://en.wikipedia.org/wiki/T._O._Engset

6.6 Markovian queueing systems in equilibrium

The purpose of this section is to go (slightly) beyond the birth and death queueing systems and deal with systems such as the single-server queues:

- $M/E_r/1$, with Poisson arrivals and Erlang$(r, r\mu)$ service density;
- $E_r/M/1$, with Erlang$(r, r\lambda)$ interarrival times and exponential service time density.

In the systems $M/E_r/1$ and $E_r/M/1$, the stochastic process $\{L_s(t),\, t \geq 0\}$[50] is not Markovian because we are dealing with the Erlang distribution which has not the memoryless property, as mentioned by Prabhu (1997, p. 59), unless $r = 1$.

The reader is probably wondering why the Erlang distribution was chosen to model the service or the interarrival times.

Firstly, the exponential distribution is not the appropriate candidate to model service or interarrival times.

Secondly, the Erlang model is a two-parameter family of absolutely continuous distributions which accommodates the exponential distribution and several other unimodal distributions due to its shape parameter.

Thirdly, the coefficient of variation of the Erlang$(r, \bullet)$ distribution equals $\frac{1}{\sqrt{r}}$, $r \in \mathbb{N}$. Therefore it is a suitable model if the sample data suggests a coefficient of variation not larger than the unit. Moreover, if we hold the mean constant to λ^{-1}, as $r \to +\infty$ then the distribution converges to the mean λ^{-1}, as illustrated by Exercise 6.128.

Exercise 6.128 — Erlangian input/service (limit distribution)
Prove that the sequence of r.v. $\{X_r,\, r \in \mathbb{N}\}$, where $X_r \sim$ Erlang$(r, r\lambda)$, converges in distribution to λ^{-1}. •

Finally, the Erlang distribution E_r can be interpreted as a $r-$fold convolution of the Exponential distribution, thus, an E_r distributed service (resp. interarrival) time can be viewed as being offered in r successive phases (resp. stages), according to Prabhu (1997, p. 59). This author also adds that, by making use of this interpretation of the Erlang distribution E_r, we deal with a bivariate Markov process in which $L_s(t)$ is one of the two state variables.

[50] Recall that $L_s(t)$ represents the number of customers in the system at time t.

Curiously enough, the method of stages/phases leads to what can be considered a more general birth and death processes, where steps beyond nearest neighbours are allowed and whose structure still permits explicit solutions (Kleinrock, 1975, p. 115) of the balance equations. This ingenious method is another proof of the brilliance of A.K. Erlang and dates from the beginning of the XX century (Kleinrock, 1975, p. 119).

The $M/E_r/1$ and $E_r/M/1$ systems are particular cases of the $M/G/1$ and $G/M/1$ systems and therefore they are going to be discussed in much more detail in future sections using the embedded Markov chain approach.

6.6.1 M/E_r/1, the single-server system with Erlangian service times

To fully describe the dynamics of a $M/E_r/1$ system in equilibrium, we need to specify

- L_s, the number of customers found in this system by an arriving customer,
- J, the NUMBER OF SERVICE STAGES TO BE COMPLETED by the clients seen in the system by the arriving customer

(Kleinrock, 1975, p. 126).

These two r.v. are obviously related. If the arriving customer finds:

- $L_s = 0$ customers in the system, then $J = 0$;
- $L_s = k$, $k \in \mathbb{N}$, and the customer being served is in her/his i^{th} stage of service, then

$$J = (k-1)\, r + (r - i + 1) \tag{6.127}$$

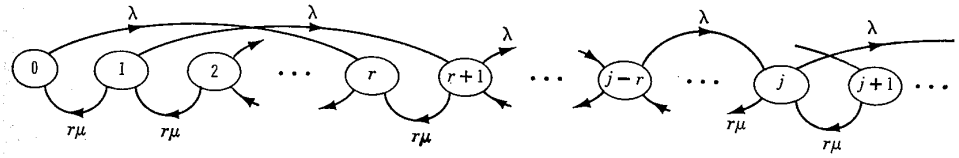
(Kleinrock, 1975, p. 126).

Consequently,

$$P(L_s = k) = \begin{cases} P(J=0), & k = 0 \\ \sum_{j=(k-1)r+1}^{kr} P(J=j), & k \in \mathbb{N} \end{cases} \tag{6.128}$$

(Kleinrock, 1975, p. 126).

Exercise 6.129 — M/E_r/1: rate diagram and balance equations
Considering the following rate diagram for J



(Kleinrock, 1975, p. 127), write the balance equations for $Q_j = P(J = j)$ by taking advantage of the previous rate diagram.[51] •

Exercise 6.130 — **M/E_r/1: solving the balance equations**
Let $Q_j = P(J = j)$.

(a) Use the $z-$transform method to prove that the p.g.f. of J is given by

$$\begin{aligned} Q(z) &= \sum_{j=0}^{+\infty} z^j Q_j \\ &= \frac{r\mu(1-z)Q_0}{r\mu + \lambda z^{r+1} - (\lambda + r\mu)z} \end{aligned} \tag{6.129}$$

(Kleinrock, 1975, p. 128).

(b) Prove that $Q_0 = 1 - \frac{\lambda}{\mu} = 1 - \rho$, by noting that $Q(1) = 1$ (Kleinrock, 1975, p. 128).

(c) Prove that, for $r = 1$, $Q(z) = \frac{1-\rho}{1-\rho^z}$ and $Q_k = \rho^k(1-\rho)$, $k \in \mathbb{N}_0$ (Kleinrock, 1975, pp. 128–129). •

Proposition 6.131 — **M/E_r/1: p.g.f. and p.f. of J; p.f. of L_s** (Kleinrock, 1975, p. 129)
Let:

- $\rho = \frac{\lambda}{\mu}$ be the traffic intensity;
- $Q(z)$ and $Q_j = P(J = j)$ be the p.g.f. and p.f. of J, respectively.

Then

$$Q(z) = \frac{1-\rho}{\prod_{i=1}^{r}\left(1 - \frac{z}{z_i}\right)} \tag{6.130}$$

$$= (1-\rho)\sum_{i=1}^{r}\frac{A_i}{1-\frac{z}{z_i}}, \tag{6.131}$$

where:

[51] They read as follows: $\lambda Q_0 = r\mu Q_1$; $(\lambda + r\mu)Q_j = \lambda Q_{j-r} + r\mu Q_{j+1}$, $j \in \mathbb{N}$ (Kleinrock, 1975, p. 127).

- $z_1, \ldots, z_r$ are the r unique roots (not equal to one!) of the denominator of $Q(z)$,

$$r\mu + \lambda z^{r+1} - (\lambda + r\mu)z = (1-z)\left[r\mu - \lambda \sum_{i=1} z^i\right]; \tag{6.132}$$

- $A_i = \prod_{n=1, n\neq i}^{r} \frac{1}{1-\frac{z}{z_i}}$, $i = 1, \ldots, r$.

Furthermore, the distribution of J is a weighted sum of Geometric* distributions,

$$Q_j = \begin{cases} 1-\rho, & j = 0 \\ (1-\rho)\sum_{i=1}^{r} A_i \times z_i^{-j}, & j \in \mathbb{N} \end{cases} \tag{6.133}$$

and the p.f. of L_s equals

$$P(L_s = k) = \begin{cases} P(J=0) = 1-\rho, & k = 0 \\ \sum_{j=(k-1)r+1}^{kr} P(J=j) & \\ = (1-\rho)\sum_{j=(k-1)r+1}^{kr} \left(\sum_{i=1}^{r} A_i \times z_i^{-j}\right), & k \in \mathbb{N}. \end{cases} \tag{6.134}$$

•

To provide an expression for $E(L_s)$ we have to invoke the Pollaczek-Khinchin mean-value formula (Kleinrock, 1975, p. 187):

$$E(L_s) = \rho + \rho^2 \frac{1 + CV^2(\text{service})}{2(1-\rho)}, \tag{6.135}$$

where $CV(\text{service}) = \frac{SD(\text{service})}{E(\text{service})}$ is the coefficient of variation of the service times.

After noting that $1 + CV^2(\text{service})$ is the sum of the squares of the coefficients of variation of the interarrival and service times, Kleinrock (1975, p. 191) adds that:

- $E(L_s)$ linearly increases with the square of the coefficient of variation of the service times;[52]
- $E(L_s)$ depends non linearly upon the traffic intensity.

Needless to say that the minimum value of $E(L_s)$ is achieved when $r = +\infty$, i.e., when the service times are constant and equal to μ^{-1}:

$$E(L_s^{M/D/1}) = \rho + \rho^2 \frac{1}{2(1-\rho)}. \tag{6.136}$$

Furthermore,

$$E(L_s^{M/D/1}) < E(L_s^{M/E_r/1}) \leq E(L_s^{M/M/1}),\ r \in \mathbb{N}. \tag{6.137}$$

[52]The more irregular are the services, the larger the expected number of customers found in the system.

Exercise 6.132 — M/E_r/1: expectations of L_s, L_q, W_s and W_q
Derive the following results by using the Pollaczek-Khinchin mean-value formula and Litte's law:

(a) $E(L_s) = \rho + \rho^2 \frac{1+\frac{1}{r}}{2(1-\rho)} = \rho + \frac{1+r}{2r} \frac{\rho^2}{1-\rho}$;

(b) $E(L_q) = E(L_s) - \rho = \frac{1+r}{2r} \frac{\rho^2}{1-\rho}$;

(c) $E(W_s) = \frac{1}{\lambda} E(L_s) = \frac{1}{\mu} + \frac{1+r}{2r} \frac{\rho}{\mu(1-\rho)}$;

(d) $E(W_q) = \frac{1}{\lambda} E(L_q) = \frac{1+r}{2r} \frac{\rho}{\mu(1-\rho)}$. •

Exercise 6.133 — M/E_r/1 system
Visitors arrive to an information desk according to a Poisson process having rate $\lambda = 0.5$ visitors per minute. Moreover, the times spent getting information have $Erlang(r, r\mu)$ distribution, with mean (resp. variance) equal to 1 minute (resp. 0.5 minute^2).

Consider from now on that this $M/E_r/1$ system is in equilibrium.

(a) Determine the value of r.

(b) Find the probability that the system is empty upon the arrival of a visitor.

(c) Obtain the probability that an arriving visitor finds 2 other customers in the system.

(d) Compute the probability that an arriving visitor finds 2 other customers waiting to be served.

(e) Calculate the following expectations: $E(L_s)$, $E(L_q)$, $E(W_s)$ and $E(W_q)$. •

Exercise 6.134 — M/E_r/1 system with no queue (Kleinrock and Gail, 1996, p. 99, Problem 4.3)
Consider a $M/E_r/1$ system in which NO QUEUE is allowed to form. Let J be the number of stages of service to be completed when one observes the system in equilibrium.

(a) Draw the rate diagram for J

(b) Find $P(J = j)$, $j = 0, 1, \ldots, r$ and the probability of a busy system.

(Kleinrock and Gail, 1996, p. 100.) •

6.6.2 E_r/M/1, the single-server system with Erlangian arrivals

To deal with the $E_r/M/1$ system we have to interchange the roles of interarrival times and service times (Kleinrock, 1975, p. 130), keeping in mind that this system does not have the PASTA property.

Expectedly, the description of the dynamics of a $M/E_r/1$ system in equilibrium relies on:

- L_s, the NUMBER OF CUSTOMERS actually in the system, AS SEEN BY AN EXTERNAL OBSERVER,[53]
- J, the TOTAL NUMBER OF COMPLETE ARRIVAL PHASES by the customers already in the system and the customer in the i^{th} phase of her/his arrival

(Kleinrock, 1975, p. 126).

These two r.v. are obviously related: if $L_s = k$, $k \in \mathbb{N}_0$, and the customer in the i^{th} phase of her/his arrival, then

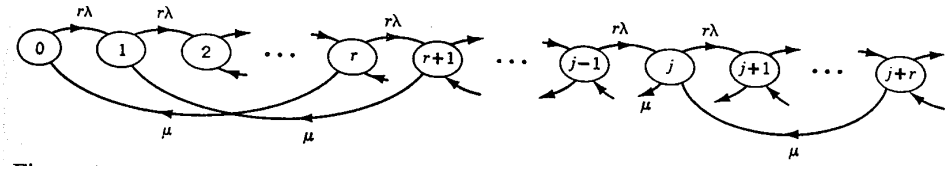
$$J = k\,r + (i-1) \tag{6.138}$$

(Kleinrock, 1975, p. 130). Thus,

$$P(L_s = k) = \sum_{j=kr}^{(k+1)r-1} P(J = j),\ k \in \mathbb{N}_0 \tag{6.139}$$

(Kleinrock, 1975, p. 131).

Exercise 6.135 — **E_r/M/1: rate diagram and balance equations**
Considering the rate diagram for J



(Kleinrock, 1975, p. 131), write the balance equations for $Q_j = P(J = j)$.[54] •

[53] NOT by an arriving or departing customer!

[54] They can be found in Kleinrock (1975, p. 131): $r\lambda\,Q_0 = \mu\,Q_r$; $r\lambda\,Q_j = r\lambda\,Q_{j-1} + \mu\,Q_{j+r}$, $j = 1,\ldots,r-1$; $(r\lambda+\mu)\,Q_j = r\lambda\,Q_{j-1} + \mu\,Q_{j+r}$, $j = r, r+1,\ldots$

Exercise 6.136 — $E_r/\mathbf{M}/1$**: solving the balance equations**
Let $Q_j = P(J = j)$ and $\rho = \frac{\lambda}{\mu}$. Use the $z-$transform method to prove that the p.g.f. of J is given by

$$\begin{aligned} Q(z) &= \sum_{j=0}^{+\infty} z^j Q_j \\ &= \frac{(1-z^r)\sum_{j=0}^{r-1} z^j Q_j}{r\rho z^{r+1} - (1+r\rho)z^r + 1} \\ &= \frac{(1-z^r)\left(1-\frac{1}{z_0}\right)}{r(1-z)\left(1-\frac{z}{z_0}\right)}, \end{aligned} \tag{6.140}$$

where z_0 is the root of $[r\rho z^{r+1} - (1+r\rho)z^r + 1]$ larger than 1 (Kleinrock, 1975, pp. 131–132). •

Proposition 6.137 — $\mathbf{M}/E_r/\mathbf{1}$**: p.f. of** J**,** L_s **and** L_q
Let: $\rho = \frac{\lambda}{\mu}$ be the traffic intensity; z_0 be the non-unit root of $[r\rho z^{r+1} - (1+r\rho)z^r + 1]$. Then the p.f. of J and L_s equal

$$P(J=j) = \begin{cases} \frac{1}{r}\left[1 - z_0^{-(j+1)}\right], & j = 0, 1, \ldots, r-1 \\ \rho(z_0 - 1)z_0^{r-(j+1)}, & j = r, r+1, \ldots \end{cases} \tag{6.141}$$

$$P(L_s = k) = \sum_{j=kr}^{(k+1)r-1} P(J=j) = \begin{cases} 1-\rho, & k = 0 \\ \rho(z_0^r - 1)\, z_0^{-rk}, & k \in \mathbb{N} \end{cases} \tag{6.142}$$

(Kleinrock, 1975, p. 133) and the p.f. of L_q is given by

$$P(L_q = k) = \begin{cases} P(L_s \leq 1) = 1 - \rho\, z_0^{-r}, & k = 0 \\ P(L_s = k+1) = \rho(z_0^r - 1)\, z_0^{-r(k+1)}, & k \in \mathbb{N}. \end{cases} \tag{6.143}$$

•

Exercise 6.138 — $E_r/\mathbf{M}/1$**: expectations of** L_s**,** L_q**,** W_s **and** W_q
Show that:

(a) $E(L_s) = \frac{\rho z_0^r}{z_0^r - 1}$;

(b) $E(L_q) = E(L_s) - \rho = \frac{\rho}{z_0^r - 1}$;

(c) $E(W_s) = \frac{1}{\lambda}E(L_s) = \frac{z_0^r}{\mu(z_0^r - 1)}$;

(d) $E(W_q) = \frac{1}{\lambda}E(L_q) = \frac{1}{\mu(z_0^r - 1)}$. •

6.6.3 (Random-sized) batch arrival systems

As put by Kleinrock (1975, p. 134), the system $M/E_r/1$ may be viewed as $M/M/1$ in which each "customer" arrival is in reality the arrival of a batch (or bulk) of exactly r customers, each requiring a single stage of service with exponential distribution. For convenience sake, we shall consider that the service times have mean μ^{-1}.[55]

It is unrealistic to consider batch of fixed size, thus we might as well allow the batch size at each arrival to be a r.v. X with p.f.

$$g_i = P(\text{batch size} = i),\ i \in \mathbb{N}. \tag{6.144}$$

We take

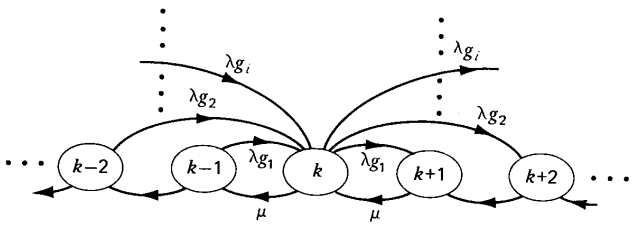
- L_s, the NUMBER OF CUSTOMERS IN THE SYSTEM AS VIEWED BY AN EXTERNAL OBSERVER,

as the state variable (Kleinrock, 1975, p. 134) of the system in equilibrium. The associated queueing system is represented by $M^X/M/1$ and its traffic intensity is equal to

$$\rho = \frac{\lambda E(X)}{\mu}. \tag{6.145}$$

Needless to say that $\rho < 1$ to reach the steady state.

Exercise 6.139 — **M^X/M/1: rate diagram and balance equations**
The rate diagram when the state variable is equal to k is



(Kleinrock, 1975, p. 135).

Write the balance equations based on this rate diagram.[56] •

Exercise 6.140 — **M^X/M/1: solving the balance equations**
Let $G(z) = \sum_{k=1}^{+\infty} z^k g_k$ be the p.g.f. of the batch size X and $\rho = \frac{\lambda E(X)}{\mu}$.

Use the $z-$transform method to prove that the p.g.f. of L_s is given by

[55] Instead of $(r\mu)^{-1}$ which would make the correspondence complete between the $M/E_r/1$ system and $M/M/1$ system with batch arrival of fixed size r.

[56] They are: $\lambda \left(\sum_{i=1}^{+\infty} g_i\right) P_0 = \mu P_1$; $\left[\lambda \left(\sum_{i=1}^{+\infty} g_i\right) + \mu\right] P_k = \mu P_{k+1} + \sum_{i=0}^{k-1} P_i \lambda g_{k-i}$, $k \in \mathbb{N}$ (Kleinrock, 1975, p. 135).

$$\begin{aligned} P(z) &= \sum_{k=0}^{+\infty} z^k P_k \\ &= \frac{\mu\, P_0\, (1-z)}{\mu(1-z) - \lambda z[1-G(z)]} \\ &= \frac{\mu\, (1-\rho)\, (1-z)}{\mu(1-z) - \lambda z[1-G(z)]} \end{aligned} \tag{6.146}$$

(Kleinrock, 1975, p. 136). •

Remark 6.141 — M^X**/M/1: f.p. and (factorial) moments of** L_s

Capitalizing on the p.g.f. of L_s in (6.146), we can get:

$$\begin{aligned} P(L_s = 0) &= P(0) \\ &= 1 - \rho \qquad (6.147) \\ P(L_s = k) &= \frac{1}{k!} \times \left. \frac{d^k\, P(z)}{dz^k} \right|_{z=0}, \; k \in \mathbb{N} \qquad (6.148) \\ E(L_s) &= \left. \frac{d\, P(z)}{dz} \right|_{z=1} \qquad (6.149) \\ E\left[L_s^{(k)}\right] &= E[L_s \times (L_s - 1) \times \ldots \times (L_s - k + 1)] \\ &= \left. \frac{d^k\, P(z)}{dz^k} \right|_{z=1}, \; k \in \mathbb{N}. \qquad (6.150) \end{aligned}$$

•

Exercise 6.142 — M^X**/M/1 system**

Consider a $M^X/M/1$ system with batch size constant and equal to 2 customers.

(a) Draw the corresponding rate diagram and use it to write the balance equations.

(b) Find $P(z)$ and determine $E(L_s)$ and $V(L_s)$.

(c) Derive the p.f. of L_s. •

Exercise 6.143 — M^X**/M/1 system (bis)** (Kleinrock and Gail, 1996, p. 106, Problem 4.6)

Consider a $M^X/M/1$ system with batch size $X \sim \text{Uniform}(\{1, 2\})$ distribution.

(a) Draw the its rate diagram and use it to write the equilibrium equations.

(b) Determine $P(z)$ and obtain $E(L_s)$.

(Kleinrock and Gail, 1996, pp. 106–107.) •

Exercise 6.144 — **M^X/M/1 system (bis, bis)** (Kleinrock and Gail, 1996, p. 108, Problem 4.7)
Find the p.f. of the number of customers in a $M^X/M/1$ system in equilibrium, with batch size $X \sim$ Geometric$(1-\alpha)$, i.e., $g_i = \alpha^{i-1}(1-\alpha)$, $i \in \mathbb{N}$ (Kleinrock and Gail, 1996, p. 108).[57] •

Exercise 6.145 — **M^X/M/1 system (bis, bis, bis)** (Adan and Resing, 2002, p. 40, Exercise 13)
Orders arrive to a work station according to a Poisson process with rate λ. An order consists of a random number N of independent jobs, with $N \sim$ Geometric$(1-p)$ and where $p \in [0,1)$. Each job requires an exponentially distributed amount of processing time with mean μ^{-1}.

(a) Derive the distribution of the total processing time of an order.

(b) Determine the distribution of the number of orders in the system.

(Adan and Resing, 2002, p. 135.) •

By using Remark 6.141 and Little's law, is certainly possible to add general expectations of L_s, L_q, W_s and W_q in terms of ρ and first and second moments of the batch size X.

Exercise 6.146 — **M^X/M/1: expectations of L_s, L_q, W_s and W_q**
Let:

- X be the batch size;
- $G(z)$, $E(X) = \frac{d\,G(z)}{dz}\Big|_{z=1}$ and $E(X^2) = \frac{d^2\,G(z)}{dz^2}\Big|_{z=1} + E(X)$ be the p.g.f., the expectation and the second moment of X, respectively;
- $\rho = \frac{\lambda E(X)}{\mu}$ the traffic intensity of the $M/M/1$ with batch arrivals of size X.

(a) Show that $E(L_s) = \frac{\rho}{1-\rho} \times \frac{E(X^2)+E(X)}{2E(X)}$ (Kleinrock and Gail, 1996, p. 110).[58]

(b) Derive expressions for: $E(L_q) = E(L_s) - \rho$; $E(W_s) = \frac{E(L_s)}{\lambda E(X)}$; $E(W_q) = \frac{E(L_q)}{\lambda E(X)}$.[59] •

[57] $P_0 = 1-\rho$; $P_k = (1-\rho)\,\frac{\lambda}{\lambda+\mu}\left(\frac{\lambda+\mu}{\lambda}\,\alpha\right)^k$, $k \in \mathbb{N}$.

[58] Since you shall deal with an indetermination, it may be useful to use L'Hôpital's rule.

[59] Check whether the expression you obtained for $E(W_q)$ coincides with the one in Ross (2003, p. 512), with $N \equiv X$, $E(S) = \frac{1}{\mu}$ and $E(S^2) = \frac{2}{\mu^2}$.

6.6.4 Batch service systems

Scenario 1

There is an equivalence between an $M/M/1$ system with services to groups of exactly r customers and a $E_r/M/1$ system. According to Kleinrock (1975, p. 137), in the former system:

- if the server becomes free then he/she will accept a batch (or bulk) of exactly r customers from the queue and serve them COLLECTIVELY; the service time for this group is exponentially distributed with mean μ^{-1};
- if, upon becoming free, the server finds less than r customers in the queue, he/she then waits to start service until a total of r customers accumulate and then accepts them and serve them as group.

Expectedly, the balance equations associated to the $E_r/M/1$ system lead to the same equilibrium distribution for the number of customers in the $M/M/1$ system with batch services of exactly r customers. However, this is not the batch service systems Kleinrock (1975, pp. 137-139) or Ross (2003, pp. 493–496) describe.

Exercise 6.147 — M/M^X/1 system in equilibrium — scenario 1 (Adan and Resing, 2002, p. 85, Exercise 55)
At a small river cars are brought from the left side to the right side of the river by a ferry. On average 15 cars per hour arrive according to a Poisson process. It takes the ferry an exponentially distributed time with a mean of 3 minutes to cross the river and return. The capacity of the ferry is equal to 2 cars. The ferry only takes off when there are two or more cars waiting.

(a) Draw the corresponding rate diagram.

(b) What is the fraction of time that the ferry is on its way between the two river sides?

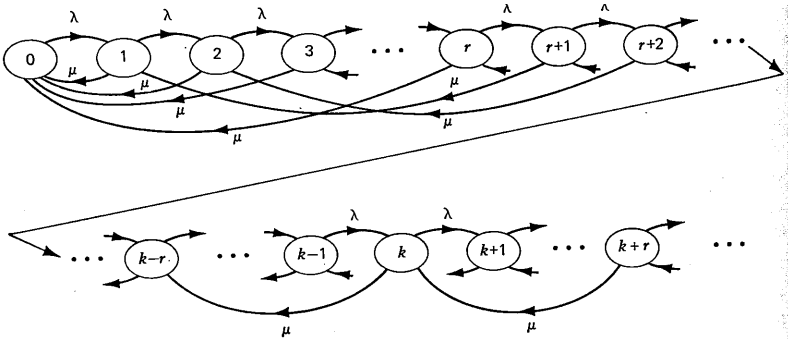
(c) Determine the distribution of the number of cars that are waiting for the ferry.

(d) Determine the mean waiting time of a car.

(Adan and Resing, 2002, p. 167.) •

Scenario 2

As Kleinrock (1975, p. 137) aptly refers, it seems more reasonable to avoid that the server remains idle when m ($m = 1, \ldots, r-1$) customers are available for a batch service. Thus, in scenario 2, we consider that the server, upon becoming free, will accept at most r customers for a bulk service. The associated rate diagram is:



(Kleinrock, 1975, p. 138).

Exercise 6.148 — M/M^X/1: solving the balance equations — scenario 2
Consider a $M/M^X/1$ system and $\rho = \frac{\lambda}{r\mu}$.

(a) Write the balance equations for L_s.[60]

(b) Apply the $z-$transform method to prove that the p.g.f. of L_s is given by

$$P(z) = \frac{\sum_{k=0}^{r-1} P_k (z^k - z^r)}{r\rho z^{r+1} - (1 + r\rho) z^r + 1} \tag{6.151}$$

(Kleinrock, 1975, p. 138).

(c) Prove that

$$P(z) = \frac{1 - \frac{1}{z_0}}{1 - \frac{z}{z_0}}, \tag{6.152}$$

where z_0 is the only root of the denominator of $P(z)$ larger than the unit (Kleinrock, 1975, p. 139). •

Proposition 6.149 — M/M^X/1: the p.f. of L_s and L_q — scenario 2
Let $\rho = \frac{\lambda}{r\mu}$ and z_0 be the only root of the denominator of $P(z)$, $r\rho z^{r+1} - (1 + r\rho) z^r + 1$, larger than the unit. Then, by inverting $P(z)$, we conclude that $L_s \sim \text{Geometric}^*(1 - \alpha)$, where $\alpha = z_0^{-1}$, i.e.,

[60] They are: $\lambda P_0 = \mu \sum_{i=1}^{r} P_i$; $(\lambda + \mu) P_k = \mu P_{k+r} + \lambda P_{k-1}$, $k \in \mathbb{N}$ (Kleinrock, 1975, p. 138).

$$P(L_s = k) = \alpha^k(1-\alpha),\ k \in \mathbb{N}_0 \tag{6.153}$$

(Kleinrock, 1975, p. 139). Furthermore, the p.f. of L_q is given by

$$P(L_q = k) = \begin{cases} P(L_s \leq 1) = 1-\alpha^2, & k = 0 \\ P(L_s = k+1) = \alpha^{k+1}(1-\alpha), & k \in \mathbb{N}. \end{cases} \tag{6.154}$$

•

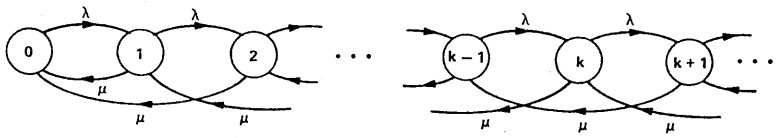
Exercise 6.150 — **M/M^X/1 — scenario 2**

(a) Obtain $E(L_s)$, $E(L_q)$, $E(W_s)$ and $E(W_q)$.

(b) Repeat Exercise 6.147. •

Exercise 6.151 — **M/M^X/1 — scenario 2** (Kleinrock and Gail, 1996, p. 111, Problem 4.9)

Consider a $M/M/1$ system with batch service with the following rate diagram



(Kleinrock and Gail, 1996, p. 111).

Let L_s represent the number of customers in the system in equilibrium and $P(z)$ the corresponding p.g.f.

(a) Write down the balance equations for $P_k = P(L_s = k)$.

(b) Show that $P(z)$, written in terms of P_0 and P_1, is equal to

$$P(z) = \frac{(2\mu-\lambda)z + \mu P_0(1-z)}{\mu(z+1) - \lambda z^2}.$$

(c) Verify that

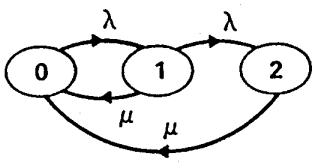
$$P(z) = \frac{(2\mu-\lambda)}{\lambda(1-z_1)z_2} \times \frac{1}{1-\frac{z}{z_2}},$$

where $z_1 = \frac{\mu - \sqrt{\mu^2+4\lambda\mu}}{2\lambda}$, $z_2 = \frac{\mu+\sqrt{\mu^2+4\lambda\mu}}{2\lambda}$ are the roots of the denominator of $P(z)$, $\mu(z+1) - \lambda z^2$, and $z_1 z_2 = -\frac{\mu}{\lambda}$.

(d) Verify that the p.f. $P_k = \frac{2\mu-\lambda}{\mu+\lambda z_2} z_2^{-k}$, $k \in \mathbb{N}_0$, is indeed associated to $P(z)$.

(Kleinrock and Gail, 1996, pp. 111–112.) •

Exercise 6.152 — **M/M^X/1 with finite capacity — scenario 2** (Kleinrock and Gail, 1996, p. 97, Exercise 4.1)
Consider the Markovian queueing system shown below. Branch labels are birth and death rates; node labels give the number of customers in the system.



(a) Obtain the p.f. and the expected value of the number of customers in this system in equilibrium.

(b) Repeat (a) for $\lambda = \mu$ and try to interpret the values you have obtained.

(Kleinrock and Gail, 1996, pp. 97–98.) •

6.7 G/M/1 systems in equilibrium

The single-server system $G/M/1$ is associated to:

- interarrival times that are i.i.d. r.v. with an arbitrary distribution G with mean λ^{-1};[61]
- i.i.d. service times with exponential distribution with rate μ.

Regretfully, the number of customers in the system at time t, $X(t)$ is not informative enough (Ross, 2007, p. 543) to serve as a state variable capable of describing the dynamics of this system in full.[62] Moreover, $\{X(t),\, t \geq 0\}$ is not a CTMC (Kulkarni, 2011, p. 216).

For a complete portrait of the $G/M/1$ system, we need the pair $(X(t), x)$ where x denotes the elapsed time since the last arrival (Adan and Resing, 2002, p. 79) in case G is not memoryless.

To get around this two-dimensional state description we shall only look at the system when a customer arrives (Ross, 2007, p. 544; Adan and Resing, 2002, p. 79).

Proposition 6.153 — G/M/1 embedded Markov chain (Adan and Resing, 2002, pp. 79–80)
Let $\hat{X}_n$ be the number of customers in the system, as seen by the n^{th} arrival. $\{\hat{X}_n,\, n \in \mathbb{N}\}$ forms a DTMC — usually called the $G/M/1$ *embedded Markov chain* — with transition probability matrix

$$\mathbf{P} = \begin{bmatrix} P_{00} & \beta_0 & 0 & \cdots & \cdots & \cdots \\ P_{10} & \beta_1 & \beta_0 & 0 & \cdots & \cdots \\ P_{20} & \beta_2 & \beta_1 & \beta_0 & 0 & \cdots \\ P_{30} & \beta_3 & \beta_2 & \beta_1 & \beta_0 & \ddots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots \end{bmatrix}, \tag{6.155}$$

where β_i denotes the probability of serving i ($i \in \mathbb{N}_0$) customers during an interarrival time[63] and

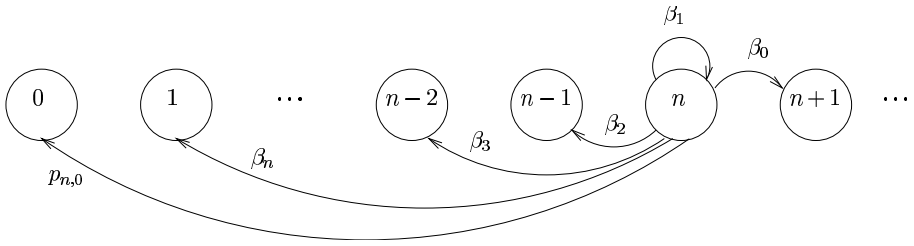
$$\beta_i = \int_0^{+\infty} e^{-\mu t} \frac{(\mu t)^i}{i!}\, dG(t). \tag{6.156}$$

•

[61] The interarrival times may not be exponentially distributed (Kulkarni, 2011, p. 216).
[62] Unless the interarrival times are exponentially distributed (Kulkarni, 2011, p. 216).
[63] Given that the server remains busy during this interval.

Exercise 6.154 — G/M/1 embedded Markov chain
Justify the entries of the TPM defined in Proposition 6.153 (Adan and Resing, 2002, p. 80).[64] •

Exercise 6.155 — G/M/1 embedded Markov chain: transition diagram
The following transition diagram for the G/M/1 embedded Markov chain can be found in Adan and Resing (2002, pp. 79–80):



Draw a more detailed transition diagram. •

For the sake of equilibrium, we require that the traffic intensity $\rho = \frac{\lambda}{\mu} < 1$ (Adan and Resing, 2002, p. 79),[65] as in the $M/M/1$ system.

Remark 6.156 — Limiting probabilities of the G/M/1 embedded Markov chain
The $G/M/1$ embedded DTMC has a limiting distribution

$$\lim_{n\to+\infty} P_{ij}^n = \hat{\pi}_j > 0,\ i, j \in \mathbb{N}_0, \tag{6.157}$$

where $\{\hat{\pi}_j : j \in \mathbb{N}_0\}$ satisfies the following system of equilibrium equations:

$$\begin{cases} \hat{\pi}_j = \sum_{i=0}^{+\infty} \hat{\pi}_i\, P_{ij} = \sum_{i=0}^{+\infty} \hat{\pi}_{j-1+i}\, \beta_i,\ j \in \mathbb{N} \\ \sum_{j=0}^{+\infty} \hat{\pi}_j = 1. \end{cases} \tag{6.158}$$

The $z-$transform method is pointless as an approach to find the solution of these equations (Adan and Resing, 2002, p. 81). By trying solutions of the form $\hat{\pi}_j = c\,\sigma^j$, $j \in \mathbb{N}_0$ (Ross, 2007, p. 545), plugging them in (6.158) and dividing by c and the common power σ^{j-1} (Adan and Resing, 2002, p. 81), we get

[64] Note that: $P_{ij} = 0$ for $j > i+1$; P_{ij}, for $j \leq i+1$ is equal to the probability that exactly $i+1-j$ customers are served during the interarrival time of the $(n+1)^{th}$ customer; $P_{i0} = 1 - \sum_{j=0}^{i} \beta_j$ because $\mathbf{P}$ is a stochastic matrix.

[65] We refer the reader to any advanced books on queueing for a proof of this result.

$$\begin{aligned}
\sigma &= \sum_{i=0}^{+\infty} \sigma^i \left[\int_0^{+\infty} e^{-\mu t} \frac{(\mu t)^i}{i!} \, dG(t) \right] \\
&= \int_0^{+\infty} e^{-\mu t} \left[\sum_{i=0}^{+\infty} \frac{(\sigma \mu t)^i}{i!} \right] dG(t) \\
&= \int_0^{+\infty} e^{-\mu(1-\sigma)t} \, dG(t) \\
&= \tilde{G}[\mu(1-\sigma)], \qquad (6.159)
\end{aligned}$$

where $\tilde{G}[\mu(1-\sigma)]$ represents the Laplace-Stieltjes transform of the common c.d.f. of the interarrival times evaluated at $\mu(1-\sigma)$.[66] $\sigma = 1$ is a root of equation (6.159) because $\tilde{G}(0) = 1$; however, it is a useless root. It can be shown that, as long as $\rho = \frac{\lambda}{\mu} < 1$, equation (6.159) has a unique root σ in the interval $(0, 1)$;[67] the value of σ is usually obtained by numerical methods (Adan and Resing, 2002, p. 81). •

Exercise 6.157 — G/M/1 system in equilibrium: obtaing σ
Evaluate σ explicitly for the following interarrival distributions:

(a) Exponential(λ) (Kleinrock, 1975, p. 252; Adan and Resing, 2002, p. 82);

(b) Erlang$(2, 3)$ and assume $\mu = 4$ (Adan and Resing, 2002, p. 82);

(c) Erlang$(2, 2\lambda)$ (Kleinrock and Gail, 1996, pp. 180–181);

(d) Exponential(μ)$\star$Exponential(2μ)[68] (Kleinrock, 1975, p. 253; Adan and Resing, 2002, pp. 82–83). •

We can finally state results concerning the distributions of:

- L_s, the number of customers in the system an ARRIVING CUSTOMER sees;
- L_q, the number of customers in the queue the ARRIVING CUSTOMER sees;
- W_s, the time an ARRIVING CUSTOMER will spend in the system;

[66] Please note that the Laplace-Stieltjes transform of G is equal to $\tilde{G}(s) = \int_0^{+\infty} e^{-st} \, dG(t)$, hence it coincides with the m.g.f. of the distribution of the service time evaluated at $-s$, $M_G(-s)$.

[67] According to Adan and Resing (2002, p. 85), this can be done by setting $f(\sigma) = \tilde{G}[\mu(1-\sigma)]$ and proving that: (i) $f(\sigma)$ is stricly convex over $[0, 1]$; (ii) $f(0) > 0$ and $f'(1) > 1$, provided that $\rho < 1$; (iii) conditions (i) and (ii) imply that the equation $\sigma = f(\sigma)$ has exactly one root in the interval (0, 1).

[68] That is, every interarrival time consists of two exponential and independent phases, the first phase with parameter μ and the second one with parameter 2μ.

- W_q, the time this ARRIVING CUSTOMER will spend in the queue waiting to be served.

Needless to say that these performance measures refer to the $G/M/1$ system in equilibrium and their distributions are similar to the ones of the $M/M/1$ system in equilibrium: ρ is simply replaced by σ.[69]

Proposition 6.158 — G/M/1: distribution of L_s (Kleinrock, 1975, p. 251; Adan and Resing, 2002, p. 81; Ross, 2007, p. 546)
An arriving customer will find k customers in the $G/M/1$ system in equilibrium with probability

$$P(L_s = k) = \sigma^k (1 - \sigma), \quad k \in \mathbb{N}_0, \tag{6.160}$$

i.e., $L_s \sim \text{Geometric}^*(1 - \sigma)$, where $\sigma \in (0, 1) : \sigma = \tilde{G}[\mu(1 - \sigma)]$. •

Proposition 6.159 — G/M/1: distribution of L_q
An arriving customer will find k customers waiting to be served in the $G/M/1$ system in equilibrium with probability

$$P(L_q = k) = \begin{cases} P(L_s \leq 1) = 1 - \sigma^2, & k = 0 \\ P(L_s = k + 1) = \sigma^{k+1} (1 - \sigma), & k \in \mathbb{N}. \end{cases} \tag{6.161}$$

•

Proposition 6.160 — G/M/1: distribution of W_s
Since the service times are memoryless in the $G/M/1$ system, the sojourn time of an arriving customer has the following properties in equilibrium:

$$(W_s \mid L_s = k) \sim \text{Gamma}(k + 1, \mu), \; k \in \mathbb{N}_0; \tag{6.162}$$

$$W_s \sim \text{Exponential}(\mu(1 - \sigma)). \tag{6.163}$$

•

Proposition 6.161 — G/M/1: distribution of W_q
In the $G/M/1$ queueing system, the time spent by an arriving customer waiting to be served is a MIXED r.v. with the following characteristics in equilibrium:

$$(W_q \mid L_s = 0) \overset{st}{=} 0; \tag{6.164}$$

$$(W_q \mid L_s = k) \sim \text{Gamma}(k, \mu), \; k \in \mathbb{N}; \tag{6.165}$$

$$(W_q \mid W_q > 0) \sim \text{Exponential}(\mu(1 - \sigma)); \tag{6.166}$$

$$F_{W_q}(t) = \begin{cases} 0, & t < 0 \\ (1 - \sigma) + \sigma \times F_{Exp(\mu(1-\sigma))}(t), & t \geq 0. \end{cases} \tag{6.167}$$

•

[69] But $\lim_{t \to +\infty} P[X(t) = j] \neq P(L_s = j)$, etc., unless the interarrival times are exponentially distributed.

Exercise 6.162 — G/M/1: distributions of L_s, L_q, W_s and W_q
Prove propositions 6.158–6.161. •

The following table condenses the distributions and expected values of the four performance measures of an $G/M/1$ in equilibrium.

G/M/1	
L_s	$P(L_s = k) = \sigma^k (1-\sigma)$, $k \in \mathbb{N}_0$ $\sigma \in (0,1) : \sigma = \tilde{G}[\mu(1-\sigma)]$ $E(L_s) = \frac{\sigma}{1-\sigma}$
L_q	$P(L_q = k) = \begin{cases} 1-\sigma^2, & k = 0 \\ \sigma^{k+1}(1-\sigma), & k \in \mathbb{N} \end{cases}$ $E(L_q) = \frac{\sigma^2}{1-\sigma}$
W_s	$(W_s \mid L_s = k) \sim \text{Gamma}(k+1, \mu)$, $k \in \mathbb{N}_0$ $W_s \sim \text{Exponential}(\mu(1-\sigma))$ $E(W_s) = \frac{1}{\mu(1-\sigma)}$
W_q	$(W_q \mid L_s = k) \sim \text{Gamma}(k, \mu)$, $k \in \mathbb{N}$ $F_{W_q}(t) = \begin{cases} 0, & t < 0 \\ 1-\sigma, & t = 0 \\ (1-\sigma) + \sigma \times F_{Exp(\mu(1-\sigma))}(t), & t > 0 \end{cases}$ $(W_q \mid W_q > 0) \sim \text{Exponential}(\mu(1-\sigma))$ $E(W_q) = \frac{\sigma}{\mu(1-\sigma)}$

Exercise 6.163 — D/M/1 system in equilibrium
Admit a customer arrives to a single-server system every minute and the service times are independent and exponentially distributed with rate equal to 90 customers per hour. Consider this system in equilibrium and find the expectations of:

(a) the number of customers in the system as seen by an arriving customer;

(b) the number of customers a customer sees upon her/his arrival waiting to be served;

(c) the time spent in system by an arriving customer;

(d) the time spent by an arriving customer in the waiting line. •

Exercise 6.164 — D/M/1 system in equilibrium (bis)
Repeat Exercise 6.163 considering now:

(a) $\lambda = 1$ and $\mu = 2$;

(b) $\lambda^{-1} = \frac{11}{16}$ and $\mu = 2$ (Kleinrock and Gail, 1996, pp. 183–184). •

Exercise 6.165 — D/M/1 system in equilibrium (bis, bis) (Adan and Resing, 2002, p. 85, Exercise 54)
Consider a queueing system where the interarrival times are exactly 4 minutes. The service times are i.i.d. and exponentially distributed with a mean of 2 minutes.

Compute σ and determine the distribution of the sojourn time (Adan and Resing, 2002, p. 166). •

Exercise 6.166 — G/M/1 system in equilibrium
Admit that the times (in minutes) between consecutive arrivals to the information desk of the Sainsbury Wing are i.i.d. r.v. with Uniform$(0, 4)$ distribution. The only server at this desk provides information for periods of time that are i.i.d. r.v. having exponential distribution with mean equal to $1m30s$.

After having identified the system and considering it in equilibrium:

(a) verify that the probability an arriving visitor finds at least one person at the information desk equals 0.650373;

(b) obtain the probability that an arriving visitor waits at least 10 minutes for her/his turn. •

Exercise 6.167 — E_n/M/1 system in equilibrium
Suppose the times between consecutive requests to land in an airport are i.i.d. r.v. with distribution with common p.d.f.

$$\frac{(n\lambda)^n t^{n-1} e^{-n\lambda t}}{(n-1)!},\ t > 0.$$

Assume that the times spent in landing manoeuvres are i.i.d. r.v. with exponential distribution with mean μ^{-1}.

(a) Verify that the probability that at least one airplane is involved in landing manoeuvres when a pilot makes a request to land satisfies

$$\sigma = \left(\frac{n\lambda}{n\lambda + \mu - \mu\sigma}\right)^n.$$

(b) Confirm that, in case $n = 3$, the mean time between consecutive requests is equal to 20 minutes and $\mu^{-1} = 10$ minutes, the probability that a pilot "sees" at least two other airplanes waiting to land is approximately 0.1093.

(c) The airport management is inquiring if there is a decrease of $E(L_q)$ if the requests to land arrive according to a Poisson process having rate equal to 3 per hour.

What can you tell to the airport management? •

The Hyper-exponential distribution is also used to model the interarrival times. Let us remind the reader that the r.v. X is said to have a Hyper-exponential distribution if its p.d.f. is given by

$$f_X(x) = \sum_{i=1}^{n} p_i \, f_{X_i}(x), \tag{6.168}$$

where: $X_i \overset{indep}{\sim}$ Exponential(λ_i), $i = 1, \ldots, n$, with $\lambda_i \neq \lambda_j$ whenever $i \neq j$; $p_i > 0$ and $\sum_{i=1}^{n} p_i = 1$.[70]

Exercise 6.168 — **H_2/M/1 system in equilibrium** (Kleinrock and Gail, 1996, p. 182, Problem 6.5)
Consider an $H_2/M/1$ system in which $\lambda_1 = 2$, $\lambda_2 = 1$, $p_1 = \frac{5}{8}$, $p_2 = \frac{3}{8}$ and $\mu = 2$.

(a) Find σ and write down $P(L_s = k)$.

(b) Obtain the c.d.f. and the expectation of W_q.

(Kleinrock and Gail, 1996, pp. 182–183.) •

Exercise 6.169 — **H_2/M/1 system in equilibrium (bis)** (Adan and Resing, 2002, p. 85, Exercise 51)
Customers arrive to a shop according to a Hyper-exponential arrival process; in this case the interarrival times are i.i.d. and exponentially distributed with a mean of 1 minute (resp. 3 minutes) with probability $\frac{1}{3}$ (resp. $\frac{2}{3}$). The service times are i.i.d. r.v. with exponential distribution and unitary mean.

[70] The Hyper-exponential distribution is an example of a *mixture density*. Its name is due to the fact that the coefficient of variation of this distribution is greater than the one of the Exponential distribution, whose coefficient of variation is 1 (http://en.wikipedia.org/wiki/Hyper-exponential_distribution). It is represent for short by $X \sim \text{Hyper-exponential}_n(\lambda_1, \ldots, \lambda_n; p_1, \ldots, p_n)$.

(a) Calculate the distribution of the number of customers found in the shop by an arriving customer.

(b) Calculate the mean number of customers in the record shop found upon arrival.

(c) Determine the mean time an arriving customer spends in the record shop.

(Adan and Resing, 2002, p. 163.) •

Exercise 6.170 — **H_2/M/1 system in equilibrium (bis, bis)** (Adan and Resing, 2002, p. 85, Exercise 52)
Consider a $G/M/1$ system, whose interarrival times have c.d.f. given by

$$G(t) = \frac{13}{24}\left(1 - e^{-3t}\right) + \frac{11}{24}\left(1 - e^{-2t}\right), \, t \geq 0,$$

and whose service times are exponentially distributed with a mean of $\frac{1}{6}$.

(a) Determine the distribution of the number of customers in the system just before an arrival.

(b) Determine the distribution of the waiting time of an arriving customer.

(Adan and Resing, 2002, p. 164.) •

Exercise 6.171 — **H_2/M/1 system in equilibrium (bis, bis, bis)** (Adan and Resing, 2002, p. 85, Exercise 53)
Determine the distribution of the sojourn time of an arriving customer to $G/M/1$ system in equilibrium, with Hyper-exponentially distributed interarrival times with c.d.f.

$$G(t) = \frac{13}{24}\left(1 - e^{-3t}\right) + \frac{11}{24}\left(1 - e{-}2t\right), \, t \geq 0,$$

and exponentially distributed service times with unit mean (Adan and Resing, 2002, p. 165). •

Exercise 6.172 — **G/M/1 system with finite storage** (Kleinrock and Gail, 1996, p. 184, Problem 6.7)
Consider a $G/M/1$ queueing system with room for at most two customers (one in service plus one waiting) and let $\tilde{G}(s) = \int_0^{+\infty} e^{-st}\, dG(t)$ be the Laplace-Stieltjes transform of the interarrival distribution.

(a) Prove that the TPM of the embedded DTMC is given by

$$\mathbf{P} = \begin{bmatrix} P_{00} & P_{01} & 0 \\ P_{10} & P_{11} & P_{12} \\ P_{10} & P_{11} & P_{12} \end{bmatrix},$$

where

$$\begin{aligned} P_{10} &= P_{12} \\ &= \tilde{G}(\mu) \\ P_{10} &= \int_0^{+\infty} \left(1 - e^{-\mu t} - \mu t e^{-\mu t}\right) dG(t) \\ &= 1 - \tilde{G}(\mu) + \mu \tilde{G}^{(1)}(\mu) \end{aligned}$$

(Kleinrock and Gail, 1996, p. 185).

(b) Let $\underline{\hat{\pi}} = [\hat{\pi}_0 \quad \hat{\pi}_1 \quad \hat{\pi}_2]$ be the vector with the p.f. of L_s.

By solving the system of equations $\underline{\hat{\pi}} = \underline{\hat{\pi}} \times \mathbf{P}$, write the p.f. of L_s in terms of μ and $\tilde{G}(s)$ (Kleinrock and Gail, 1996, pp. 185–186). •

Exercise 6.173 — G/M/1 system with costs (Kleinrock and Gail, 1996, p. 186, Problem 6.8)
Consider a $G/M/1$ system in which the cost of making a customer wait y seconds is $c(y) = ae^{by}$, $y > 0$ ($a > 0$).

(a) Find the average cost of queueing for a customer.

(b) Under what conditions will the average cost be finite?

(Kleinrock and Gail, 1996, pp. 186–187.) •

Since the arrival process in a $G/M/1$ system is not necessarily Poisson, the PASTA property is not applicable and therefore the long-run fraction of the time that the system has k customers,

$$p_k = \lim_{t \to +\infty} P[X(t) = k], \tag{6.169}$$

is DIFFERENT from the long-run fraction of arrivals that see k customers ahead of them in the system

$$\hat{\pi}_k = \lim_{n \to +\infty} P[\hat{X}_n = k]. \tag{6.170}$$

However, we can obtain p_k in terms of $\hat{\pi}_k$, as stated in the next proposition.

Proposition 6.174 — G/M/1: limiting distribution of $X(t)$ (Kulkarni, 2011, p. 221)
In a stable $G/M/1$ system,[71] the limiting distribution of the number of customers in the system is given by

$$\begin{aligned} p_0 &= 1-\rho && (6.171)\\ p_k &= \rho\,\hat{\pi}_{k-1},\ k\in\mathbb{N}, && (6.172) \end{aligned}$$

where $\hat{\pi}_k = \sigma^k\,(1-\sigma)$, $k\in\mathbb{N}_0$. •

Expectedly, $E(L_s)=\frac{\sigma}{1-\sigma}$ is different from the expectation of the r.v. $\lim_{t\to+\infty} X(t)$.

Exercise 6.175 — G/M/1: limiting distribution of $X(t)$ (Kulkarni, 2011, p. 221)
Prove that, by taking advantage of the limiting distribution of the number of customers in the system, in particular of result (6.172), we get:

$$\begin{aligned} E\left[\lim_{t\to+\infty} X(t)\right] &= \sum_{k=1}^{+\infty} k\times\rho\,\sigma^{k-1}(1-\sigma)\\ &= \frac{\rho}{1-\sigma} && (6.173)\\ &\neq E(L_s). \end{aligned}$$

•

Interestingly enough, by using $E\left[\lim_{t\to+\infty} X(t)\right]$, we can now apply Little's law and get the expected value of W_s:

$$\begin{aligned} E(W_s) &= \frac{1}{\lambda}\times E\left[\lim_{t\to+\infty} X(t)\right]\\ &= \frac{1}{\mu(1-\sigma)} && (6.174) \end{aligned}$$

(Adan and Resing, 2002, p. 82).

[71] With traffic intensity $\rho=\frac{\lambda}{\mu}$ and associated to a root $\sigma\in(0,1):\sigma=\tilde{G}[\mu(1-\sigma)]$.

6.8 M/G/1 systems in equilibrium

This section is devoted to another generalization of the $M/M/1$ system. In this case, the service times are still i.i.d. but are not necessarily exponentially distributed. Let:

- S and $F_S(t)$ be the service time and its c.d.f., respectively;
- $E(S) = \mu^{-1}$ and $CV(S) = \frac{SD(S)}{E(S)}$;
- $\tilde{F}_S(s) = \int_0^{+\infty} e^{-st}\, dF_S(t)$ the Laplace-Stieltjes transform of the c.d.f. of S.

In what follows we shall assume that: the service discipline is FCFS; $\rho = \frac{\lambda}{\mu} < 1$, for stability sake.

Due to the Poisson arrival process (with rate λ), the $M/G/1$ system has the PASTA property, thus, the limiting distributions of the number of customers in the system seen by

- an arriving customer or
- an outside observer

are the same. However, knowing the number of customers in the system at time t, $X(t)$, does not provide enough information about this system, unless the distribution of the service times is memoryless; in fact, $\{X(t),\, t \geq 0\}$ is not a CTMC (Kulkarni, 2011, p. 212), as in the $G/M/1$ system.

For a complete description of the $M/G/1$ system, we need to consider a two-dimensional state variable $(X(t), x)$, where x represents the additional time required to complete the service of the customer which is currently being served. But we can certainly avoid this two-dimensional state description by simply looking at the system just after the customers depart. In fact, the limiting distribution of

- the number of customers left behind in the system by a departing customer

coincides with the one of the number of customers in the system seen by an arriving customer or an outside observer (Adan and Resing, 2002, p. 60).

Proposition 6.176 — M/G/1 embedded Markov chain (Adan and Resing, 2002, p. 61)
Let X_n be the number of customers left behind by the n^{th} departure. $\{X_n,\, n \in \mathbb{N}\}$ forms a DTMC — termed the $M/G/1$ *embedded Markov chain* — with transition probability matrix

$$\mathbf{P} = \begin{bmatrix} \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 & \cdots \\ \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 & \cdots \\ 0 & \alpha_0 & \alpha_1 & \alpha_2 & \cdots \\ 0 & 0 & \alpha_0 & \alpha_1 & \ddots \\ 0 & 0 & 0 & \alpha_0 & \ddots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \tag{6.175}$$

where α_i denotes the probability that exactly i ($i \in \mathbb{N}_0$) customers arrived during a service time and is equal to

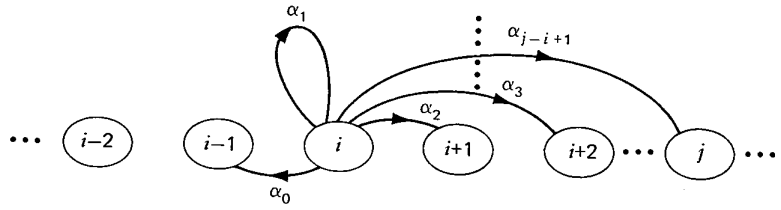
$$\alpha_i = \int_0^{+\infty} e^{-\lambda t} \frac{(\lambda t)^i}{i!} \, dF_S(t). \tag{6.176}$$

•

Exercise 6.177 — M/G/1 embedded Markov chain and its transition diagram

(a) Justify the entries of the TPM defined in Proposition 6.153 (Adan and Resing, 2002, p. 61).

(b) The following transition diagram for the M/G/1 embedded Markov chain can be found in Kleinrock (1975, p. 179):



Draw a more detailed transition diagram. •

Remark 6.178 — Limiting probabilities of the M/G/1 embedded Markov chain

The $M/G/1$ embedded DTMC has a limiting distribution

$$\lim_{n \to +\infty} P_{ij}^n = \pi_j > 0, \; i, j \in \mathbb{N}_0, \tag{6.177}$$

where $\{\pi_j : j \in \mathbb{N}_0\}$ satisfies the following system of equilibrium equations:

$$\begin{cases} \pi_j = \sum_{i=0}^{+\infty} \pi_i \, P_{ij}, \; j \in \mathbb{N} \\ \sum_{j=0}^{+\infty} \pi_j = 1. \end{cases} \tag{6.178}$$

Unlike the $G/M/1$ system, we can successfully use the $z-$transform to derive the solution of (6.178). The p.g.f. of the limiting distribution of X_n can be actually written in terms of the p.g.f. of the number, say A, of customers arrived during a service time

$$P_A(z) = \sum_{i=0}^{+\infty} z^i \, \alpha_i, \tag{6.179}$$

or alternatively as a function of the Laplace-Stieltjes of the c.d.f. of the service time because

$$\begin{aligned} P_A(z) &= \sum_{n=0}^{+\infty} z^i \left[\int_0^{+\infty} e^{-\lambda t} \, \frac{(\lambda t)^i}{i!} \, dF_S(t) \right] \\ &= \int_0^{+\infty} e^{-\lambda t} \left[\sum_{i=0}^{+\infty} \frac{(z\lambda t)^i}{i!} \right] dF_S(t) \\ &= \int_0^{+\infty} e^{-\lambda(1-z)t} \, dF_S(t) \\ &= \tilde{F}_S[\lambda(1-z)]. \end{aligned} \tag{6.180}$$

•

Proposition 6.179 — M/G/1: p.g.f. of L_s (Adan and Resing, 2002, pp. 62–63)
Let L_s be the NUMBER OF CUSTOMERS LEFT BEHIND BY A DEPARTING CUSTOMER in the $M/G/1$ system in equilibrium.

Capitalizing on the particular structure of the TPM of the $M/G/1$ embedded Markov chain leads to the conclusion that the p.f. of L_s satisfies

$$\begin{aligned} P(L_s = j) &= \pi_j \\ &= \pi_0 \, \alpha_j + \pi_1 \, \alpha_j + \ldots + \pi_j \, \alpha_1 + \pi_{j+1} \, \alpha_0 \\ &= \pi_0 \, \alpha_j + \sum_{i=0}^{j} \pi_{j-i+1} \, \alpha_i, \end{aligned} \tag{6.181}$$

for $j \in \mathbb{N}_0$. In addition, the p.g.f. of L_s is given by

$$\begin{aligned} P_{L_s}(z) &= \sum_{j=0}^{+\infty} z^j \, \pi_j \\ &= \sum_{j=0}^{+\infty} z^j \left(\pi_0 \, \alpha_j + \sum_{i=0}^{j} \pi_{j-i+1} \, \alpha_i \right) \\ &= \pi_0 \, P_A(z) + \frac{P_A(z) \times [P_{L_s}(z) - \pi_0]}{z}. \end{aligned} \tag{6.182}$$

Given that $\pi_0 = 1 - \rho$,[72] we get:

$$\begin{aligned} P_{L_s}(z) &= \frac{(1-\rho)\,P_A(z)\,(1-z)}{P_A(z)-z} \\ &= \frac{(1-\rho)\,\tilde{F}_S[\lambda(1-z)]\,(1-z)}{\tilde{F}_S[\lambda(1-z)]-z}. \end{aligned} \tag{6.183}$$

•

Remark 6.180 — M/G/1: Pollaczek-Khinchin transform equation; Laplace-Stieltjes transform of S; inverting the p.g.f. of L_s

(6.183) is one form of the Pollaczek-Khinchin transform equation (Kleinrock, 1975, p. 194; Adan and Resing, 2002, p. 63);[73] others will soon follow for W_s and W_q.

Note once again that the Laplace-Stieltjes transform of the c.d.f. F_S is equal to $\tilde{F}_S(s) = \int_0^{+\infty} e^{-st}\,dF_S(t)$, thus, it can be obtain by evaluating the m.g.f. of the r.v. S at $-s$, $M_S(-s)$, and the tables with expressions for the m.g.f. of popular distributions may come in handy.

As for the inversion of the p.g.f. of L_s, if we are led, for instance, to $P_{L_s}(z) = \frac{a}{1-bz}$, then recalling that $\sum_{j=0}^{+\infty} ab^j \times z^j = \frac{a}{1-bz}$ ($|bz| < 1$), we can certainly conclude that $P(L_s = j) = P_{L_s}^{-1}(z) = ab^j$, $j \in \mathbb{N}_0$. Moreover, we can always resort to

$$P(L_s = 0) = 1 - \rho \tag{6.184}$$

$$P(L_s = k) = \frac{1}{k!} \times \left.\frac{d^k\,P_{L_s}(z)}{dz^k}\right|_{z=0},\ k \in \mathbb{N} \tag{6.185}$$

to derive the p.f. of L_s. •

Exercise 6.181 — M/G/1: p.g.f. and p.f. of L_s

Evaluate the p.g.f. of L_s and invert it in order to provide an explicit expression for the p.f. of L_s associated to the following service time distributions and traffic intensities:

(a) Exponential(μ) and $\rho = \frac{\lambda}{\mu}$ (Kleinrock, 1975, p. 195; Adan and Resing, 2002, p. 63);

(b) Erlang($2, \mu$) and $\rho = \frac{1}{3}$ (Adan and Resing, 2002, pp. 63–64);

(c) HyperExponential($\mu_1 = 1, \mu_2 = 2, p_1 = \frac{1}{4}, p_2 = \frac{3}{4}$), $\lambda = 1$, $\mu^{-1} = \frac{5}{8}$ and $\rho = \frac{\lambda}{\mu}$ (Kleinrock, 1975, pp. 195–196; Adan and Resing, 2002, p. 64). •

[72] This result follows from the identity $P_{L_s}(1) = 1$.

[73] We ought to mention that Prabhu (1997, p. 11) notes that the mathematical theory of queues emerged through the work of Pollaczek (1957, 1961) and Khintchine (1960).

Proposition 6.182 — M/G/1: the Pollaczek-Khinchin mean formula for $E(L_s)$ (Kleinrock, 1975, p. 187)
Differentiating the p.g.f. of L_s, we obtain the Pollaczek-Khinchin mean formula providing an expression for $E(L_s)$ in terms of the traffic intensity $\rho = \frac{\lambda}{\mu}$ and the square of the coefficient of variation of the service time:

$$\begin{aligned} E(L_s) &= \left. \frac{d\, P_{L_s}(z)}{dz} \right|_{z=1} \\ &= \rho + \rho^2 \, \frac{1 + CV^2(S)}{2(1-\rho)}. \end{aligned} \tag{6.186}$$

•

Exercise 6.183 — M/G/1: the Pollaczek-Khinchin mean formula for $E(L_s)$
Prove Proposition 6.186 (Kleinrock, 1975, pp. 186–187). •

Proposition 6.184 — M/G/1: p.f. and expectation of L_q
Let L_q be the number of customers left behind by a departing customer waiting to be served in the $M/G/1$ system in equilibrium. Expectedly:

$$P(L_q = k) = \begin{cases} P(L_s \leq 1), & k = 0 \\ P(L_s = k+1), & k \in \mathbb{N}; \end{cases} \tag{6.187}$$

$$\begin{aligned} E(L_q) &= E(L_s) - \rho \\ &= \rho^2 \, \frac{1 + CV^2(S)}{2(1-\rho)} \end{aligned} \tag{6.188}$$

(Kleinrock, 1975, p. 188). •

Exercise 6.185 — M/G/1: expectations of L_q and L_s (Kulkarni, 2011, p. 215, Example 6.12)
Packets arrive at an infinite-capacity buffer according to a Poisson process at a rate of 400 per second. All packets are exactly 512 bytes long.[74] The buffer is emptied at a rate of 2 megabits per second.

Compute:

(a) the expected number of packets waiting for transmission;

(b) the expected number of packets in the buffer (including any in transmission). •

[74] A byte equals eight bits.

There are other Pollaczek-Khinchin formulae, namely relating the Laplace-Stieltjes transform of the service time $\tilde{F}_S(s)$ and the ones of

- W_s, the TIME AN ARRIVING CUSTOMER WILL SPEND IN THE SYSTEM (sojourn time)
- W_q, the TIME THIS ARRIVING CUSTOMER WILL SPEND IN THE QUEUE waiting to be served (waiting time)

in a $M/G/1$ system in equilibrium. To prove the results we should remind the reader that:

- the limiting distribution of the number of customers left behind by a departing customer coincides with the one of an arriving customer in a $M/G/1$ system in equilibrium;
- all customers left behind by a departing customer are precisely those who arrived during her/his sojourn in the system.

Consequently,

$$\begin{aligned} \pi_j &= \hat{\pi}_j \\ &= \int_0^{+\infty} e^{-\lambda t} \frac{(\lambda t)^j}{j!} \, dF_{W_s}(t), \; j \in \mathbb{N}_0 \\ \sum_{j=0}^{+\infty} z^j \, \pi_j &= \sum_{j=0}^{+\infty} z^j \, \hat{\pi}_j \\ \frac{(1-\rho)\, \tilde{F}_S[\lambda(1-z)]\,(1-z)}{\tilde{F}_S[\lambda(1-z)] - z} &= \tilde{F}_{W_s}[\lambda(1-z)]. \end{aligned}$$

Proposition 6.186 — M/G/1 system: Laplace-Stieltjes transform of W_s
The Laplace-Stieltjes transform of the c.d.f. of the sojourn time of an arriving customer in the $M/G/1$ system in equilibrium equals

$$\tilde{F}_{W_s}[\lambda(1-z)] = \frac{(1-\rho)\, \tilde{F}_S[\lambda(1-z)]\,(1-z)}{\tilde{F}_S[\lambda(1-z)] - z}.$$

By making the change of variable $s = \lambda(1-z)$, we obtain

$$\tilde{F}_{W_s}(s) = \frac{(1-\rho)\, \tilde{F}_S(s)\, s}{\lambda \tilde{F}_S(s) + s - \lambda}. \tag{6.189}$$

•

Remark 6.187 — M/G/1 system: Laplace-Stieltjes transform of W_s
Admitting that W_s is an absolutely continuous r.v. and capitalizing on the following facts

- $\tilde{F}_{W_s}(s) = \int_0^{+\infty} e^{-st}\, dF_{W_s}(t) = \int_0^{+\infty} e^{-st} f_{W_s}(t)\, dt$, which is the Laplace transform of the p.d.f. of W_s,
- *Mathematica* inverts Laplace transforms,

we are capable of identifying the p.d.f. of W_s in various cases. •

Exercise 6.188 — M/G/1: p.d.f. of W_s
Use *Mathematica* to obtain the p.d.f. of W_s associated to the following service time distributions and traffic intensities:

(a) Exponential(μ) and $\rho = \frac{\lambda}{\mu}$

(b) Erlang($2, \mu$) and $\rho = \frac{1}{3}$

(c) HyperExponential($\mu_1 = 1, \mu_2 = 2, p_1 = \frac{1}{4}, p_2 = \frac{3}{4}$) and $\rho = \frac{\lambda}{\mu} = \frac{1}{5/8}$

(Adan and Resing, 2002, p. 65). •

Proposition 6.189 — M/G/1: Laplace-Stieltjes transform of W_q (Adan and Resing, 2002, p. 66)
Having in mind that

- the sojourn time is the sum of two independent r.v., the service and waiting times,
- the Laplace-Stieltjes transform of a convolution, such as $W_s = S + W_q$, is the product of the Laplace-Stieltjes transforms of the random and independent summands,

we can determine the Laplace-Stieltjes transform of W_q:

$$\begin{aligned} \tilde{F}_{W_q}(z) &= \frac{\tilde{F}_{W_s}(z)}{\tilde{F}_S(z)} \\ &= \frac{(1-\rho)s}{\lambda \tilde{F}_S(s) + s - \lambda}. \end{aligned} \tag{6.190}$$

•

Proposition 6.190 — M/G/1: expectations of W_s and W_q
By applying Little's law, we can conclude that:

$$\begin{aligned} E(W_s) &= \frac{E(L_s)}{\lambda} \\ &= \frac{1}{\mu} + \frac{\rho}{\mu}\frac{1+CV^2(S)}{2(1-\rho)}; \qquad (6.191) \\ E(W_q) &= \frac{E(L_q)}{\lambda} \\ &= \frac{\rho}{\mu}\frac{1+CV^2(S)}{2(1-\rho)}. \qquad (6.192) \end{aligned}$$

•

Remark 6.191 — M/G/1: expectation of W_q (Adan and Resing, 2002, p. 68)
An arriving customer has to wait for the residual service time R of the customer being served (if and) and then continues to wait for an amount of time corresponding to the service times of all customers who were already waiting in the queue upon her/his arrival. Moreover, the single-server is busy with probability ρ. Hence

$$\begin{aligned} E(W_q) &= E(L_q) \times E(S) + \rho \times E(R) \\ &= \lambda E(W_q) \times \frac{1}{\mu} + \rho \times E(R) \\ &= \frac{\rho E(R)}{1-\rho} \\ &= \frac{\rho \frac{E(S^2)}{2E(S)}}{1-\rho} \\ &= \frac{\lambda^2 E(S^2)}{2(1-\rho)} \qquad (6.193) \\ &= \frac{\rho}{\mu}\frac{1+CV^2(S)}{2(1-\rho)}. \qquad (6.194) \end{aligned}$$

•

Exercise 6.192 — M/G/1: expectation of W_q (Prabhu, 1997, p. 148, Problem 6)
Patients arrive in a Poisson fashion at the ER of a local hospital at a rate of 4 per hour. The time (in minutes) required to admit an emergency patient is uniformly distributed over the interval $(5, 15)$.

Admitting there is only one admission clerk at the ER and the system is equilibrium, what is the expected amount of time an ER patient spends to get admitted?[75] •

[75] Solution: 20.83 minutes.

Exercise 6.193 — M/G/1: expectation of W_q (bis) (Adan and Resing, 2002, p. 75, Exercise 41)

Consider a machine where jobs are being processed. The mean production time is 4 minutes and the standard deviation is 3 minutes. The mean number of jobs arriving per hour is 10. Suppose that the interarrival times are exponentially distributed.

Determine the mean waiting time of the jobs (Adan and Resing, 2002, p. 159). •

M/G/1	
L_s	$P_{L_s}(z) = \frac{(1-\rho)\,\tilde{F}_S[\lambda(1-z)]\,(1-z)}{\tilde{F}_S[\lambda(1-z)]-z}$
	$P(L_s = k) = \begin{cases} 1-\rho, & k = 0 \\ \frac{1}{k!} \times \frac{d^k\, P_{L_s}(z)}{dz^k}\Big\|_{z=0}, & k \in \mathbb{N} \end{cases}$
	$E(L_s) = \rho + \rho^2\, \frac{1+CV^2(S)}{2(1-\rho)}$
L_q	$P(L_q = k) = \begin{cases} P(L_s \leq 1), & k = 0 \\ P(L_s = k+1), & k \in \mathbb{N} \end{cases}$
	$E(L_q) = \rho^2\, \frac{1+CV^2(S)}{2(1-\rho)}$
W_s	$\tilde{F}_{W_s}(s) = \frac{(1-\rho)\,\tilde{F}_S(s)\,s}{\lambda\tilde{F}_S(s)+s-\lambda}$
	$E(W_s) = \frac{1}{\mu} + \frac{\rho}{\mu}\, \frac{1+CV^2(S)}{2(1-\rho)}$
W_q	$\tilde{F}_{W_q}(s) = \frac{(1-\rho)s}{\lambda\tilde{F}_S(s)+s-\lambda}$
	$E(W_q) = \frac{\rho}{\mu}\, \frac{1+CV^2(S)}{2(1-\rho)}$

Exercise 6.194 — M/G/1: expectations of L_s, L_q, W_s and W_q

Calculate the expected values of L_s, L_q, W_s and W_q for the following $M/G/1$ systems in equilibrium:

(a) Exponential(μ) and $\rho = \frac{\lambda}{\mu}$;

(b) Erlang($2, \mu$) and $\rho = \frac{1}{3}$;

(c) HyperExponential($\mu_1 = 1, \mu_2 = 2, p_1 = \frac{1}{4}, p_2 = \frac{3}{4}$) and $\rho = \frac{\lambda}{\mu} = \frac{1}{5/8}$. •

Exercise 6.195 — M/G/1 system (Prabhu, 1997, p. 148, Problem 7)

It has been found that students arrive to submit or pick up jobs in a computer center, in a manner that may be regarded as a Poisson process having rate equal to 4 jobs per

minute. Furthermore, there is only one person available to handle the students' requests and this server is found to spend T minutes in servicing a student, where T is a discrete r.v. with p.f.

$$P(T=t) = \begin{cases} 0.2, & t = 0.1 \\ 0.4, & t = 0.2 \\ 0.3, & t = 0.3 \\ 0.1, & t = 0.4 \end{cases}$$

.

Consider that this system is in equilibrium and find:

(a) the average time a student spends in the system;[76]

(b) average length of the waiting line.[77] •

Exercise 6.196 — M/G/1 system (bis) (Prabhu, 1997, p. 148, Problem 8)
Customers arrive at a cashier's window of a local bank according to a Poisson process with mean interarrival time equal to 15 minutes. Service times have a Gamma distribution with a mean of 6 minutes and a variance of 5 minutes^2.

During steady state, determine the expected:

(a) time a customer waiting in line;[78]

(b) number of customers in this system.[79] •

Exercise 6.197 — M/G/1 system (bis, bis) (Prabhu, 1997, p. 148, Problem 9)
Arrivals at a truck-weighing station occur in accordance to a Poisson process having rate of 5 per hour. The service time is constant and equal to 6 minutes per truck.

Find:

(a) the expected number of trucks at the weighing station;[80]

(b) probability that a truck will have to wait for service.[81] •

[76] Solution: 1.755 minutes.
[77] Solution: 7.02 jobs.
[78] Solution: 2.28 minutes.
[79] Solution: 0.55 customers.
[80] Solution: 0.75 trucks.
[81] Solution: $\frac{1}{2}$.

Exercise 6.198 — M/G/1 system (bis, bis, bis) (Prabhu, 1997, p. 148, Problem 10)
Parts arrive at an inspection station at a Poisson rate of 20 per hour. Currently inspection is performed manually, with inspection time distributed with a mean of 2 minutes and a standard deviation of 0.5 minutes. It has been proposed that inspection be performed automatically. With automatic inspection, the time required will be constant and equal to b minutes.

Show that the mean waiting time is reduced iff $b < 2.03$. •

Exercise 6.199 — M/E_2/1 system (Adan and Resing, 2002, p. 74, Exercise 38)
Consider a single machine where jobs arrive according to a Poisson stream with a rate of 10 jobs per hour. The processing time of a job consists of two phases — each phase takes an exponential time with a mean of 1 minute.

(a) Determine the Laplace-Stieltjes transform of the processing time.

(b) Find the distribution of the number of jobs in the system.

(c) Obtain the mean number of jobs in the system and the mean production lead time (waiting time plus processing time).

(Adan and Resing, 2002, p. 156.) •

Exercise 6.200 — M/H_2/1 system (Adan and Resing, 2002, p. 74, Exercise 39)
At a post office customers arrive according to a Poisson process with a rate of 60 customers per hour. Half of the customers have a service time that is the sum of a fixed time of 15 seconds and an exponentially distributed time with a mean of 15 seconds. The other half have an exponentially distributed service time with a mean of 1 minute.

Determine the mean waiting time and the mean number of customers waiting in the queue (Adan and Resing, 2002, p. 157). •

Exercise 6.201 — Another M/G/1 system (Adan and Resing, 2002, pp. 74–75, Exercise 40)
A machine produces products in two phases. The first phase is standard and the same for all products. The second phase is customer specific (the finishing touch). The first (resp. 74 second) phase takes an exponential time with a mean of 10 (resp. 2) minutes. Orders for the production of one product arrive according to a Poisson stream with a rate of 3 orders per hour. Orders are processed in order of arrival.

Determine the mean production lead time of an order (Adan and Resing, 2002, p. 158). •

6.9 The busy period

Suppose a single-server system is free initially, then a customer arrives, is served and during her/his service more customers arrive and will be served, and the process will continue until no customer is left and the server becomes free again. When this happens, a *busy period* has just ended (Prabhu, 1997, p. 7).

If the system has more than one server, the busy period is naturally defined as the time during which at least one of the servers is busy (Prabhu, 1997, p. 7).

Definition 6.202 — **Busy period** (http://en.wikipedia.org/wiki/M/M/1_queue#Busy_period_of_server)
The busy period is the time period measured between the instant a customer arrives to an empty system until the instant a customer departs leaving behind an empty system. •

The busy period is indeed the time spent in states $1, 2, 3, \ldots$ between visits to the state 0 (http://en.wikipedia.org/wiki/M/G/1_queue). It is followed by an *idle period* during which the system is empty (Adan and Resing, 2002, p. 37).

Prabhu (1997, p. 7) also defines the *system busy period* and *slack period* for systems with more than one server. The former is the period of time during which all the m servers are busy. A system busy period is followed by a slack period, during which less than m servers are busy.

Since the busy period is an important performance measure of queueing systems,[82] we proceed with a brief characterization of the busy periods of the following single-server systems:

- $M/M/1$
- $G/M/1$
- $M/G/1$.

M/M/1 system

If we assume that a *busy cycle* (Cohen, 1972) starts with a busy period (BP) and ends with the subsequent idle period (IP), we can

- apply the key renewal theorem to alternating renewal processes and

[82] The literature on the busy period is vast; the reader is encouraged to use *MathSciNet* to find papers on this relevant performance measure.

- use the fact that the fraction of time the single-server is busy is equal to $\rho = \frac{\lambda}{\mu}$

to derive an expression of the expected value of the busy period of a $M/M/1$ system.

Proposition 6.203 — M/M/1: expectation of the idle/busy period (Adan and Resing, 2002, pp. 37–38)
Due to the lack of memory of the exponential distribution, $IP \sim$ Exponential(λ). Moreover,

$$E(IP) = \frac{1}{\lambda} \tag{6.195}$$

and

$$\begin{aligned} \frac{E(BP)}{E(BP) + E(IP)} &= \rho & (6.196) \\ E(BP) &= \frac{1}{\mu(1-\rho)}. & (6.197) \end{aligned}$$

•

An expression for the p.d.f. of BP can be also provided.

Proposition 6.204 — M/M/1: Laplace-Stieltjes transform and p.d.f. of the busy period (Adan and Resing, 2002, p. 39)
The Laplace-Stieltjes transform of the c.d.f. of the BP of a $M/M/1$ system in equilibrium is given by

$$\begin{aligned} \tilde{F}_{BP}(s) &= \int_0^{+\infty} e^{-st}\, dF_{BP}(t) \\ &= \frac{1}{2\lambda}\left[\lambda + \mu + s - \sqrt{(\lambda+\mu+s)^2 - 4\lambda\mu}\right]. & (6.198) \end{aligned}$$

The inversion of this transforms leads to the p.d.f. of the BP:

$$f_{BP}(t) = \frac{e^{-(\lambda+\mu)\,t}}{t\sqrt{\rho}} \times I_1\left(2t\sqrt{\lambda\mu}\right),\ t > 0, \tag{6.199}$$

where

$$I_1(x) = \sum_{k=0}^{+\infty} \frac{(x/2)^{2k+1}}{k!\,(k+1)!} \tag{6.200}$$

denotes the modified Bessel function of the first kind of order one. •

Exercise 6.205 — **M/M/1: Laplace-Stieltjes transform and p.d.f. of the busy period**

Prove Proposition 6.204 (Adan and Resing, 2002, pp. 38–39). •

Exercise 6.206 — **M/M/1: variance of the busy period**

Prove that

$$V(BP) = \frac{1+\rho}{\mu^2(1-\rho)^3}.$$

•

Exercise 6.207 — **M/M/1: survival function of the busy period**

Use the *Mathematica* function **BesselI**$[n, x]$[83] to obtain values of the survival function of the busy period, $P(BP > t)$, when

- $\mu = 1$
- $\rho = 0.8, 0.9, 0.95$
- $t = 1, 2, 4, 8, 16, 40, 80$

(Adan and Resing, 2002, p. 39). •

G/M/1 system

Due to the similarities between the $G/M/1$ and $M/M/1$ systems when it comes to the probabilistic behavior of the performance measures L_s, L_q, W_s and W_q, the expected value of the BP can be obtained by simply replacing ρ with σ in (6.197).

Proposition 6.208 — **G/M/1: expectation of the busy/idle period** (Ross, 2007, p. 548)

Let $\tilde{G}$ be the Laplace-Stieltjes transform of the c.d.f. of the interarrival times of a $G/M/1$ system. Then $\frac{E(BP)}{E(BP)+E(IP)} = \lim_{t\to+\infty} P[X(t) > 0] = \rho$, where $\rho = \frac{\lambda}{\mu} = \frac{E(Service\,Time)}{E(Interarrival\,Time)}$ (see (6.171)). Consequently,

$$E(BP) = \frac{1}{\mu(1-\sigma)} \tag{6.201}$$

$$E(IP) = \frac{\mu-\lambda}{\lambda\mu(1-\sigma)}, \tag{6.202}$$

where $\sigma \in (0,1) : \sigma = \tilde{G}[\mu(1-\sigma)]$. •

[83] It gives the modified Bessel function of the first kind $I_n(x)$.

Exercise 6.209 — G/M/1: expectation of the busy/idle period
Prove Propositon 6.208 (Ross, 2007, p. 548). •

Proposition 6.210 — G/M/1: Laplace-Stieltjes transform of the idle period (Adan *et al.*, 2005)
The Laplace-Stieltjes transform of the c.d.f. of the idle period of a $G/M/1$ system is given by:

$$\begin{aligned}\tilde{F}_{IP}(s) &= \frac{\mu(1-\sigma)-\mu[1-\tilde{G}(s)]}{\mu(1-\sigma)-s} \\ &= \mu \times \frac{\tilde{G}(s)-\sigma}{\mu(1-\sigma)-s}. \end{aligned} \tag{6.203}$$

•

Needless to say that, after having obtained σ, we can use *Mathematica* to invert this transform and identify the p.d.f. of IP.

Exercise 6.211 — G/M/1: expectation of the idle period
Use Equation (6.203) to check that $E(IP) = \frac{\mu-\lambda}{\lambda\mu(1-\sigma)}$. •

Proposition 6.212 — G/M/1: Laplace-Stieltjes transform and c.d.f. of the busy period (Cohen, 1982, p. 226)
The Laplace-Stieltjes transform of the c.d.f. of the busy period of the $G/M/1$ system is equal to:

$$\tilde{F}_{BP}(s) = \frac{1-\xi(s)}{\mu^{-1}\, s + 1 - \xi(s)},\ Re(s) \geq 0, \tag{6.204}$$

where $\xi(s)$ is the root with the smallest absolute value of $z - \tilde{G}\left[s + \mu(1-z)\right] = 0$.

The inversion of this Laplace-Stieltjes transform leads to

$$F_{BP}(t) = \sum_{n=1}^{+\infty} \left\{ \frac{e^{-\mu\, t}\, (\mu\, t)^{n-1}}{n!} \times \int_0^t \mu[1 - G^{n\star}(u)]\, du \right\},\ t > 0, \tag{6.205}$$

where $G^{n\star}$ is the $n-$fold convolution of the c.d.f. of the interarrival times.[84] •

Adan *et al.* (2005) also add that the busy and idle periods of the $G/M/1$ queue are not necessarily independent r.v.

[84] That is, $G^{n\star}$ is the c.d.f. of a sum on n i.i.d. r.v. with common c.d.f. G. $G^{1\star}(u) = G(u)$; $G^{(n+1)\star}(u) = \int_{\mathbb{R}} G^{n\star}(u-x)\, dG(x)$, $n \in \mathbb{N}$.

Remark 6.213 — G/M/1: Laplace-Stieltjes transform of the busy period
After reading Ross (2007, p. 548), we are led to believe that the busy cycle $BC = BP+IP$ can be written as $BC = \sum_{i=1}^{N} T_i$, where T_i is the i^{th} interarrival time after the busy period begins and $N \sim$ Geometric$(1-\sigma)$. Consequently, the Laplace-Stieltjes transform of the c.d.f. of the busy cycle can be written in terms of the p.g.f. of N:

$$\begin{aligned}
\tilde{F}_{BC}(s) &= E\left[\left(e^{-s\sum_{i=1}^{N} T_i}\middle| N\right)\right] \\
&\overset{\tilde{F}_{T_i}(s)=\tilde{G}(s)}{=} E\left\{\left[\tilde{G}(s)\right]^N\right\} \\
&= P_N\left[\tilde{G}(s)\right] \\
&= \frac{(1-\sigma)\tilde{G}(s)}{1-\sigma\,\tilde{G}(s)}
\end{aligned} \tag{6.206}$$

We are certainly tempted to admit that BP and IP are independent r.v. and write:

$$\begin{aligned}
\tilde{F}_{BP}(s) &= \frac{\tilde{F}_{BP+IP}(s)}{\tilde{F}_{IP}(s)} \\
&= \frac{\frac{(1-\sigma)\tilde{G}(s)}{1-\sigma\,\tilde{G}(s)}}{\mu \times \frac{\tilde{G}(s)-\sigma}{\mu(1-\sigma)-s}} \\
&= \frac{1-\sigma}{\mu} \times \frac{\tilde{G}(s) \times [\mu(1-\sigma)-s]}{\left[1-\sigma\,\tilde{G}(s)\right] \times \left[\tilde{G}(s)-\sigma\right]}.
\end{aligned} \tag{6.207}$$

Once again, we would use *Mathematica* to invert this transform and identify the p.d.f. of BP. •

Exercise 6.214 — G/M/1: expectation of the busy period
Check that, oddly enough, the use of Equation (6.207) leads to $E(BP) = \frac{1}{\mu(1-\sigma)}$. •

M/G/1 system

The expected length of a busy period is $\frac{1}{\mu-\lambda}$ where μ^{-1} is the expected length of service time and λ the rate of the Poisson process governing the arrivals (http://en.wikipedia.org/wiki/M/G/1_queue#Busy_period). Furthermore, $\frac{E(BP)}{E(BP)+E(IP)} = \lim_{t\to+\infty} P[X(t)>0] = \rho = \frac{\lambda}{\mu}$ (Ross, 2007, p. 530).

Proposition 6.215 — M/G/1: expectation of the idle/busy period (Ross, 2007, pp. 530–531)
Let S represent once more the service time in this queueing system and $E(S) = \mu^{-1}$. From the Poisson arrivals, it follows that $IP \sim$ Exponential(λ) and

$$E(IP) = \frac{1}{\lambda} \tag{6.208}$$

$$E(BP) = \frac{1}{\mu(1-\rho)}, \tag{6.209}$$

as in the $M/M/1$ system. •

The Laplace-Stieltjes of the c.d.f. of the busy period can be shown to obey a functional relationship (http://en.wikipedia.org/wiki/M/G/1_queue#Busy_period), as stated in the next proposition.

Proposition 6.216 — M/G/1 system: Laplace-Stieltjes of the c.d.f. busy period (Adan and Resing, 2002, p. 72)
The Laplace-Stieltjes of the c.d.f. of the busy period is such that

$$\tilde{F}_{BP}(s) = \tilde{G}[s + \lambda - \lambda \tilde{F}_{BP}(s)], \tag{6.210}$$

where $\tilde{G}$ denotes the Laplace-Stieltjes transform of c.d.f. of the service time. •

Exercise 6.217 — M/G/1 system: Laplace-Stieltjes of the c.d.f. busy period
Prove Proposition 6.216 (Adan and Resing, 2002, pp. 71-72). •

Equation (6.210) can only be solved exactly in special cases such as the $M/M/1$ queue.

Exercise 6.218 — M/M/1 system: Laplace-Stieltjes of the c.d.f. busy period
Use Proposition 6.216 to check that

$$\tilde{F}_{BP}(s) = \frac{1}{2\lambda}\left[\lambda + \mu + s - \sqrt{(\lambda + \mu + s)^2 - 4\lambda\mu}\right]$$

(Adan and Resing, 2002, pp. 72–72). •

Proposition 6.219 — M/G/1: the spread of the busy period (Adan and Resing, 2002, p. 73)
Let S denote the service time. Then the second moment and the square of the coefficient of variation of the busy period of the $M/G/1$ system are given by

$$E(BP^2) = \frac{E(S^2)}{(1-\rho)^3} \tag{6.211}$$

$$CV^2(BP) = \frac{CV^2(S)+\rho}{1-\rho}, \tag{6.212}$$

respectively.[85] •

Exercise 6.220 — M/G/1: the spread of the busy period
Use (6.210) to show Proposition 6.219 (Adan and Resing, 2002, p. 73). •

Exercise 6.221 — M/H_2/1: the mean busy period duration and much more (Adan and Resing, 2002, p. 74, Exercise 37)
Customers arrive at a post office — with only one server — according to a Poisson process with a rate of 30 customers per hour.

A quarter of the customers wants to cash a cheque; the associated service time is exponentially distributed with a mean of 2 minutes. The remaining customers want to buy stamps and their service times are exponentially distributed with a mean of 1 minute.

Determine:

(a) the p.g.f., p.f. and expected value of the number of customers in the system;

(b) the Laplace-Stieltjes transform of the c.d.f., the c.d.f. and the expectation of the sojourn time;

(c) the mean busy period duration.

(Adan and Resing, 2002, p. 155.) •

Exercise 6.222 — M/M/1: the mean busy period duration and much more
Resume Exercise 6.221 and admit that all customers have an exponentially distributed service time with a mean of 75 seconds (Adan and Resing, 2002, p. 74, Exercise 37).

(a) Obtain the mean number of customers in the system and the mean sojourn time (Adan and Resing, 2002, p. 155).

(b) Calculate $E(BP)$, $V(BP)$ and $CV^2(BP)$. •

[85] $CV(BP)$ increases as ρ tends to one.

6.10 Networks of Markovian queues

This section is an introduction to the very important subject of queueing networks (or networks of queues), a current area of great interest and application, namely to manufacturing facilities and computer/communication networks (Gross and Harris, 1998, p. 165), as illustrated by Koole (2010).

We begin with a few definitions and remarks on queueing networks, systems in which a number of queues are connected by customer routing (http://en.wikipedia.org/wiki/Queueing_theory#Queueing_networks).

Definition 6.223 — **Queueing network** (Gross and Harris, 1998, p. 165; http://en.wikipedia.org/wiki/Queueing_theory#Queueing_networks)
A system consisting of a group of say k stations (or nodes), where each station i represents a service facility of some kind with m_i servers ($m_i \in \overline{\mathbb{N}}$ and $i = 1, \ldots, k$) and the stations are connected by customer routing, is said to be a queueing network. •

Remark 6.224 — **Queueing network** (Gross and Harris, 1998, p. 165)
In the most general case, customers may:

- arrive from *outside* the system at any station;
- depart from the system from any station;
- traverse from station to station;
- enter and leave the system at different stations;
- return to stations previously visited, skip some stations or choose to remain in the system indefinitely. •

Definition 6.225 — **Jackson network** (Gross and Harris, 1998, pp. 165–166; Kulkarni, 1995, p. 359)
A queueing network with the following characteristics is said to be a Jackson network:

- it consists of k service stations;
- there are m_i ($m_i \in \overline{\mathbb{N}}$) servers at station i ($i = 1, \ldots, k$);
- there is infinite waiting room at every station;
- customers may arrive from *outside* to station i ($i = 1, \ldots, k$) according to a Poisson process, with — *external input rates* r_i —, which is independent of the external arrival processes at other stations and also independent of any service times;

- the durations of the services, provided by each server at station i ($i = 1, \ldots, k$), are not only exponentially distributed i.i.d. r.v. with (*service*) *rate* μ_i, but also independent of the service times at other stations;
- upon completion of her/his service at station i ($i = 1, \ldots, k$), the customer
  - goes to station j ($j = 1, \ldots, k$), with *routing probability* P_{ij}, where $\sum_{j=1}^{k} P_{ij} \leq 1$, for all $i = 1, \ldots, k$, or
  - leaves the system from station i, with probability $1 - \sum_{j=1}^{k} P_{ij}$;
- the routing probabilities do not depend on the state of the queueing network;
- the *routing matrix* $\mathbf{P} = [P_{ij}]_{i,j=1,\ldots,k}$ is such that $\mathbf{I} - \mathbf{P}$ is invertible.[86] •

Definition 6.226 — **Closed Jackson network** (Gross and Harris, 1998, p. 166)
If

- $r_i = 0$, for all $i = 1, \ldots, k$, that is, no customer may enter the system from *outside*, and
- $\sum_{j=1}^{k} P_{ij} = 1$, for all $i = 1, \ldots, k$,[87] i.e, no customer may leave the system,

then we are dealing with a closed Jackson network. •

Definition 6.227 — **Open Jackson network** (Gross and Harris, 1998, p. 166)
If

- $r_i \neq 0$, for some i ($i = 1, \ldots, k$), or
- $\sum_{j=1}^{k} P_{ij} \neq 1$, for some i ($i = 1, \ldots, k$),

then the system is referred to as an open Jackson network. •

From now on, we shall restrict ourselves to Jackson networks. These systems can be described as (multivariate) birth and death Markov processes (Gross and Harris, 1998, p. 168).

[86] This assumption will become clear shortly.
[87] This means that the routing matrix is stochastic.

Remark 6.228 — State vector of a Jackson network (Gross and Harris, 1998, p. 166)
Since Jackson networks are Markovian systems, they can be described by a $k-$dimensional state vector, $\underline{N}(t) = (N_1(t), \ldots, N_k(t))$, where $N_i(t)$ represents the number of customers at station i $(i = 1, \ldots, k)$ at time t. Furthermore, $\underline{N}(t)$ is a CMTC. •

However, to describe the steady state behaviour of $\underline{N}(t)$ we have to briefly address the issue of the output of $M/M/m$ $(m \in \overline{\mathbb{N}})$ queues.

6.10.1 M/M/m queue output

Let us remind the reader that by Kolmogorov's criterion for reversibility, any birth and death process is a reversible continuous time Markov chain (http://en.wikipedia.org/wiki/Burke's_theorem). Thus, if we start at time t and look backward, the departures become the arrivals and vice-versa, and the probabilistic behaviour of the backward and forward paths are the same (Gross and Harris, 1998, p. 168).

Burke's Theorem (Burke, 1956) states for instance that, in the steady state, the output (or departure) process of $M/M/1$, $M/M/m$ or $M/M/\infty$ queues follows a Poisson process; moreover, because no traffic is lost in such queues, the departure rate must be equal to the arrival rate in steady state (Zukerman, 2010–2012, p. 208).

Theorem 6.229 — Burke's Theorem (Gross and Harris, 1998, p. 170; http://en.wikipedia.org/wiki/Burke's_theorem)
Let:

- $\{T_1, T_2, \ldots\}$ represent the sequence of times between consecutive departures — the departure process — from a $M/M/m$ $(m \in \overline{\mathbb{N}})$ queue in the steady state with arrival rate λ;
- $N(t)$ be the number of customers in the system at time t.

Then

- $N(t)$ is independent of the departure process prior to time t;[88]

[88] This follows from the fact that the departures prior to t in the reversed process coincide with the arrival process after t in the reversed process and it is clear that the number in the system at time t is independent of the arrivals after that time point in a Poisson system.

- the arrival and departure processes are unaffected by the exponential service mechanism and are both Poisson processes with rate λ.[89] •

Burke's startling result proves to be extremely useful in the analysis of Jackson networks:

- the steady state joint probability distribution of $(N_1, N_2, \ldots, N_m)$ has a product-form

(Gross and Harris, 1998, p. 166) and average metrics can be computed, as shown for instance for open Jackson networks, the topic of our next subsection.

[89]For a constructive proof of Burke's Theorem, which shows that, indeed, the inter departure times are i.i.d. r.v. with Exponential distribution with parameter λ when we deal with the $M/M/1$, $M/M/m$ or $M/M/\infty$ queues in steady state, please refer to Gross and Haris (1998, pp. 168–170). Burke first published this theorem along with a proof in 1956; however, the theorem was anticipated but not proved by O'Brien (1954) and Morse (1955), according to http://en.wikipedia.org/wiki/Burke's_theorem

6.10.2 Open Jackson networks

Before we proceed it is crucial to identify the *total input rates* of the k stations of an open Jackson network.

Proposition 6.230 — Total input rates; traffic equations (Kulkarni, 1995, p. 360; Gross and Harris, 1998, p. 166; Ross, 2003, p. 499)
The input to station j consists of two parts:

- external input, at rate r_j;
- internal input due to routing from stations $i = 1, \ldots, j$, at rate $\sum_{i=1}^{k} \lambda_i P_{ij}$.

If we let λ_j denote the *total input rate of customers to station* j, then in the steady state we must deal with the following *traffic equations*:

$$\lambda_j = r_j + \sum_{i=1}^{k} \lambda_i P_{ij},\ j = 1, \ldots, k. \tag{6.213}$$

•

Corollary 6.231 — Total input rates in matrix form (Kulkarni, 1995, pp. 360–361)
Let:

- $\underline{r} = [r_j]_{j=1,\ldots,k}$ represent the row vector of the external input rates;
- $\mathbf{P} = [P_{ij}]_{i,j=1,\ldots,k}$ be the routing matrix of the Jackson network;
- $\underline{\lambda} = [\lambda_j]_{j=1,\ldots,k}$ denote the row vector of the total input rates;
- $\mathbf{I}$ is the identity matrix with rank k.

Then $\underline{\lambda}$ can be obtained as the solution of

$$\underline{\lambda} \times (\mathbf{I} - \mathbf{P}) = \underline{r}, \tag{6.214}$$

i.e.,

$$\underline{\lambda} = \underline{r} \times (\mathbf{I} - \mathbf{P})^{-1}. \tag{6.215}$$

•

The invertibility of $(\mathbf{I} - \mathbf{P})$ is obviously needed to obtain $\underline{\lambda}$ and fundamentally means that no customer stays in the queueing network indefinitely (Kulkarni, 1995, p. 361).

Theorem 6.232 — Transient behaviour of $\underline{N}(t)$ (Kulkarni, 1995, p. 361)
$\{\underline{N}(t) : t \geq 0\}$ is a positive recurrent CTMC iff $\rho_i = \frac{\lambda_i}{m_i \mu_i} < 1$, for all $i = 1, \ldots, k$, where $\underline{\lambda} = [\lambda_i]_{i=1,\ldots,k}$ satisfies (6.215). •

Theorem 6.233 — Jackson's theorem: steady state behaviour of $\underline{N}(t)$ in an open Jackson network (Kulkarni, 1995, p. 361; Gordon and Harris, 1998, p. 178)
Let:

- $\underline{N} = (N_1, \ldots, N_k)$ be the state vector of the open Jackson network in the steady state;
- $L_s^{(i)}$ be the number of customers in a $M(\lambda_i)/M(\mu_i)/m_i$ system in steady state.

If $\{\underline{N}(t) : t \geq 0\}$ is a positive recurrent CTMC then

$$\begin{aligned} P(\underline{N} = \underline{n}) &= P(N_1 = n_1, \ldots, N_k = n_k) \\ &= \lim_{t \to +\infty} P[N_1(t) = n_1, \ldots, N_k(t) = n_k] \\ &= \prod_{i=1}^{k} P[L_s^{(i)} = n_i], \end{aligned} \tag{6.216}$$

i.e., in the steady state the open Jackson network *acts as if* the stations were k independent $M(\lambda_i)/M(\mu_i)/m_i$ systems $(i = 1, \ldots, k)$. •

Remark 6.234 — Jackson's theorem: steady state behaviour of $\underline{N}(t)$ in an open Jackson network (Kulkarni, 1995, p. 363; Gordon and Harris, 1998, pp. 175–176; http://en.wikipedia.org/wiki/Arrival_theorem)

- The joint probability function in (6.216) is called the product-form, for obvious reasons.
- The term *acts as if* is extremely important because the individual processes $\{N_i(t) : t \leq 0\}$ $(i = 1, \ldots, k)$ are not themselves birth and death processes arising from the $M(\lambda_i)/M(\mu_i)/m_i$ $(i = 1, \ldots, k)$ systems.

  Indeed, the reader should not be misled into believing that the open Jackson network actually decomposes into k individual $M(\lambda_i)/M(\mu_i)/m_i$ systems $(i = 1, \ldots, k)$ systems, with the flow into each a true Poisson process with rate λ_i.

  As a matter of fact, as long as customers can return to previously visited stations (i.e., as long as there is any kind of feedback) the internal flows are not Poisson.
- In an open Jackson network with m stations, the probability that the system is in state $\underline{n}$ immediately before an arrival to any node is also $P(\underline{N} = \underline{n})$.

  This result does not follow from Jackson's theorem and was first stated by Sevcik and Mitrani (1981); furthermore, it refers to particular points in time, namely arrival times. •

By plugging in the expressions of $P[L_s^{(i)} = n_i]$ to (6.216) we obtain the results stated in the next corollary.

Corollary 6.235 — Steady state behaviour of $\underline{N}(t)$ in an open Jackson network (Gross and Harris, 1998, p. 175, 178; Kulkarni, 1995, p. 361)
The product-form solution for $\lim_{t\to+\infty} P[N_1(t) = n_1, \ldots, N_k(t) = n_k]$ leads to the following results concerning an open Jackson network in steady state, with k:

- single-server stations (with $\rho_i = \frac{\lambda_i}{\mu_i} < 1$, $i = 1, \ldots, k$)

$$P(\underline{N} = \underline{n}) = \prod_{i=1}^{k} \left[(1-\rho_i)\,\rho_i^{n_i}\right],\ n_i \in \mathbb{N}_0; \tag{6.217}$$

- m_i-server stations (with $\rho_i = \frac{\lambda_i}{m_i\,\mu_i} < 1$, $i = 1, \ldots, k$)

$$P(\underline{N} = \underline{n}) = \begin{cases} \prod_{i=1}^{k} \left[\frac{m_i!}{n_i!}\,(1-\rho)(m_i\rho_i)^{n_i-m_i} C(m_i, m_i\rho_i)\right], \\ \qquad\qquad\qquad\qquad n_i = 0, 1, \ldots, m_i - 1 \\ \prod_{i=1}^{k} \left[(1-\rho_i)\,\rho_i^{n_i-m_i}\, C(m_i, m_i\rho_i)\right], \\ \qquad\qquad\qquad\qquad n_i = m_i, m_i+1, \ldots, \end{cases} \tag{6.218}$$

  where $C(m_i, m_i\rho_i) = P\left(L_s^{(i)} \geq m_i\right) = \frac{\frac{(m_i\rho_i)^{m_i}}{m_i!(1-\rho_i)}}{\sum_{j=0}^{m_i-1} \frac{(m_i\rho_i)^j}{j!} + \frac{(m_i\rho_i)^{m_i}}{m_i!(1-\rho_i)}}$;

- infinite-server stations (with $\rho_i = \frac{\lambda_i}{\mu_i}$, $i = 1, \ldots, k$)

$$P(\underline{N} = \underline{n}) = \prod_{i=1}^{k} \left(e^{-\rho_i}\,\frac{\rho_i^{n_i}}{n_i!}\right),\ n_i \in \mathbb{N}_0. \tag{6.219}$$

•

Capitalizing once again on the product-form we can compute the mean total number of customers in an open Jackson network in equilibrium. In addition, we can use Little's law to obtain the average time a customer spends in the open Jackson network in steady state.

Proposition 6.236 — Mean number of customers and mean time spent in an open Jackson network in steady state
Consider an open Jackson network in steady state and let:

- $N = \sum_{i=1}^{k} N_i = \sum_{i=1}^{k} L_s^{(i)}$ be the total number of customers in this system;

- W_s the total time a customer spends in this queueing network.[90]

Then

$$E(N) = \sum_{i=1}^{k} E\left[L_s^{(i)}\right] \tag{6.220}$$

$$E(W_s) = \frac{E(N)}{\sum_{i=1}^{k} r_i}. \tag{6.221}$$

We have for an open Jackson network in steady state with k:

- single-server stations (with $\rho_i = \frac{\lambda_i}{\mu_i} < 1$, $i = 1, \ldots, k$),

$$\begin{aligned} E(N) &= \sum_{i=1}^{k} \frac{\rho_i}{1-\rho_i} \\ &= \sum_{i=1}^{k} \frac{\lambda_i}{\mu_i - \lambda_i} \end{aligned} \tag{6.222}$$

  (Ross, 2003, pp. 499–500);

- m_i-server stations (with $\rho_i = \frac{\lambda_i}{m_i\,\mu_i} < 1$, $i = 1, \ldots, k$),

$$E(N) = \sum_{i=1}^{k} \left[m_i\,\rho_i + \frac{\rho_i}{1-\rho_i}\, C(m_i, m_i\rho_i) \right]; \tag{6.223}$$

- infinite-server stations (with $\rho_i = \frac{\lambda_i}{\mu_i}$, $i = 1, \ldots, k$),

$$E(N) = \sum_{i=1}^{k} \rho_i. \tag{6.224}$$

-

Let us now consider an example and a few exercises to illustrate and practice the application of all these results. They refer, for instance, to:

- a network of three single-server stations (Ross, 2003, Chap. 8, Exercise 31);
- a single station queue with Bernoulli feedback (Kulkarni, 1995, pp. 363–364);
- single-server queues in tandem (Kulkarni, 1995, pp. 364–365).

[90] Ross (2003, p. 500) stresses that the denominator of $E(W_s) = \frac{E(N)}{\lambda}$ is equal to $\sum_{i=1}^{k} r_i$ and not $\sum_{i=1}^{k} \lambda_i$.

Example 6.237 — A network of three single-server stations (Ross, 2003, Chap. 8, Exercise 31)
Customers arrive at stations 1, 2, 3 in accordance with Poisson processes having respective rates 5, 10, 15. The service times at the three stations are exponential with rates 10, 50, 100, respectively. Moreover:

- a customer completing service at station 1 is equally likely to

  (i) go to station 2,

  (ii) go to station 3 or

  (iii) leave the system;

- a customer departing service at station 2 always goes to station 3;

- a departure from service at station 3 is equally likely to either go to station 2 or leave the system.

What is the average number of customers in the system (consisting of all three single-server stations) in equilibrium?

- **Jackson network**

  Since there is the possibility of entering and exiting the system, we are dealing with an open Jackson network. Moreover, it consists of $k = 3$ single-server queues such that:

  - the arrival rates of customers coming from outside to these 3 single-server stations are

    $$\begin{aligned} r_1 &= 5 \\ r_2 &= 10 \\ r_3 &= 15; \end{aligned}$$

  - the service rates of these 3 single-server stations are equal to

    $$\begin{aligned} \mu_1 &= 10 \\ \mu_2 &= 50 \\ \mu_3 &= 100. \end{aligned}$$

- **Routing probabilities** (P_{ij})

  Let P_{ij} represent the probability that a customer leaves station i $(i = 1, \ldots, k)$ and joins station j $(j = 1, \ldots, k)$.[91] Since

  - a customer completing service at station 1 is equally likely to
    (i) go to station 2,
    (ii) go to station 3, or
    (iii) leave the system,
  - a customer departing service at station 2 always goes to station 3,
  - a departure from service at station 3 is equally likely to either go to station 2 or leave the system,

  we obtain

$$\mathbf{P} = [P_{ij}]_{i,j=1,2,3} = \begin{bmatrix} 0 & \frac{1}{3} & \frac{1}{3} \\ 0 & 0 & 1 \\ 0 & \frac{1}{2} & 0 \end{bmatrix}.$$

- **Total input rate of customers to server** j (λ_j)

  Since $\rho_j = \frac{\lambda_j}{\mu_j} < 1$, $j = 1, 2, 3$, these rates are given by the traffic equations:

$$\lambda_j = r_j + \sum_{i=1}^{k} \lambda_i P_{ij},\ j = 1, \ldots, k.$$

  In this case,

$$\begin{aligned} \underline{\lambda} &= \underline{r} \times (\mathbf{I} - \mathbf{P})^{-1} \\ &= [r_1 \quad r_2 \quad r_3] \times \left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} 0 & \frac{1}{3} & \frac{1}{3} \\ 0 & 0 & 1 \\ 0 & \frac{1}{2} & 0 \end{bmatrix} \right)^{-1} \\ &= [5 \quad 10 \quad 15] \times \begin{bmatrix} 1 & -\frac{1}{3} & -\frac{1}{3} \\ 0 & 1 & -1 \\ 0 & -\frac{1}{2} & 1 \end{bmatrix}^{-1} \end{aligned}$$

[91] Recall that $1 - \sum_{j=1}^{k} P_{ij}$ represents the probability that a customer departs the system after being served by server i.

$$
\begin{aligned}
&= [5 \quad 10 \quad 15] \times \begin{bmatrix} 1 & 1 & \frac{4}{3} \\ 0 & 2 & 2 \\ 0 & 1 & 2 \end{bmatrix} \\
&= [5 \quad 40 \quad 170/3].
\end{aligned}
$$

- **Requested expectation**

  Recall that in the steady state the 3 single-server stations *act as if* they were 3 independent $M(\lambda_j)/M(\mu_j)/1$ $(j = 1, 2, 3)$ queues. Thus, the average number of customers in this queueing network in equilibrium, $E(N)$, is given by

$$
\begin{aligned}
\sum_{j=1}^{k} \frac{\lambda_j}{\mu_j - \lambda_j} &= \frac{5}{10-5} + \frac{40}{50-40} + \frac{\frac{170}{3}}{100 - \frac{170}{3}} \\
&= 1 + 4 + \frac{17}{13} \\
&\simeq 6.307692.
\end{aligned}
$$

•

Exercise 6.238 — A single station queue with Bernoulli feedback (Kulkarni, 1995, pp. 363–364, Example 7.5)
The simplest queueing network is a single station $M/M/m$ queue with feedback:

- customers arrive at the service station according to a Poisson process with rate r;
- the service station has m servers, each working at rate μ;
- upon completion of her/his service, a customer leaves the system with probability α or rejoins the system and behaves as a new arrival with probability $1 - \alpha$.

Derive the limiting probabilities. •

Exercise 6.239 — Open Jackson network with two single-server stations (Ross, 2003, pp. 500-501, Example 8.5)
Consider a system of two servers where:

- customers from outside the system arrive at server 1 at a Poisson rate 4 and at server 2 at a Poisson rate 5;
- the service rates of 1 and 2 are respectively 8 and 10;

- a customer, upon completion of service at server 1, is equally likely to go to server 2 or to leave the system (i.e., $P_{11} = 0$, $P_{12} = 1$);
- a departure from server 2 will go 25 percent of the time to server 1 and will depart the system otherwise (i.e., $P_{21} = 1$, $P_{22} = 0$).

Determine:

(a) the limiting probabilities;

(b) $E(N)$;

(c) $E(W_s)$. •

Exercise 6.240 — Open Jackson network with three stations (Gross and Harris, 1998, pp. 181–182, Example 4.2)
A company has a three-node telephone system with the following characteristics:

- calls coming into the 800 number are Poisson, with mean of 35/hr;
- the caller gets one of two options — press 1 for claims service and press 2 for policy service;
- it is estimated that the caller's listening, decision and button-pushing time is exponential with mean equal to 30 seconds;
- only one call at a time can be processed and all calls put on hold (with nice background music!) are not lost;
- approximately 55% of the calls go to claims and the remainder to policy service;
- the claims processing station has three parallel servers with exponential service times with mean equal to 20 minutes;
- about 2% of the customers finishing at claims then go to policy service;
- approximately 1% of the customers finishing at police service go to claims;

Obtain:

(a) the average queue sizes in front of each station;

(b) the average number of customers in the system;

(c) the total average time a customer spends in the system. •

Exercise 6.241 — Open Jackson network with six single-server stations (Zuckerman, 2000–2012, p. 211, Homework 19.1)
Consider a six station Jackson network of $M/M/1$ queues, such that:

- the service rate of all the queues is equal to one, i.e., $\mu_i = 1$ for $i = 1, \ldots, 6$;
- the arrival rates from the outside into the different stations is given by $r_1 = 0.6$, $r_2 = 0.5$, and $r_i = 0$ for $i = 3, \ldots, 6$;
- the routing matrix is given by

$$\mathbf{P} = \begin{bmatrix} 0 & 0.4 & 0.6 & 0 & 0 & 0 \\ 0 & 0.1 & 0 & 0.7 & 0.2 & 0 \\ 0 & 0 & 0 & 0.3 & 0.7 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.6 \\ 0 & 0 & 0 & 0.3 & 0 & 0.2 \\ 0 & 0 & 0.3 & 0 & 0 & 0 \end{bmatrix}.$$

Use *Mathematica* to find:

(a) the probability that the entire network is empty;

(b) the mean time a packet spends in the network from the moment it enters the network until it leaves the network. •

Exercise 6.242 — Open Jackson network with seven single-server stations (Gross and Harris, 1998, p. 205, Exercise 4.10C)
Consider a seven station, open single-server Jackson network, where:

- only stations 2 and 4 get input from the outside at rate 5/min;
- stations 1 and 2 have service rates of 85/min; stations 3 and 4 have service rates of 120/min;
- station 5 has a service rate of 70/min;
- stations 6 and 7 have service rates of 20/min;
- the routing matrix is given by

$$\mathbf{P} = \begin{bmatrix} \frac{1}{3} & \frac{1}{4} & 0 & \frac{1}{4} & 0 & \frac{1}{6} & 0 \\ \frac{1}{3} & \frac{1}{4} & 0 & \frac{1}{3} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 \\ 0 & 0 & 0 & \frac{4}{5} & 0 & 0 & \frac{1}{6} \\ \frac{1}{6} & 0 & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & 0 \\ 0 & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & 0 & \frac{1}{6} \end{bmatrix}.$$

Use *Mathematica* to find:

(a) the average number of customers in the system;

(b) the average time a customer spends in this queueing network.[92] •

Exercise 6.243 — Open Jackson network with six stations with different numbers of servers: amusement park (Kulkarni, 2011, pp. 224-225, 231, examples 6.19 and 6.24)
An amusement park has six rides:

- a roller coaster (station 1);
- a merry-go-round (station 2);
- a water-tube ride (station 3);
- a fantasy ride (station 4);
- a ghost-mountain ride (station 5);
- a journey to the moon ride (station 6).

Visitors arrive at the park according to a Poisson process at a rate of 600 per hour and immediately fan out to the six rides. A newly arriving visitor is equally likely to go to any of the six rides first. From then on, we assume that, after each ride, the riders choose to join one of the remaining five rides or leave in a completely random fashion. Moreover:

- the roller coaster ride lasts for an exponentially distributed time with mean equal to 2 minutes (including the loading and unloading times);
- two roller coaster cars, each carrying 12 riders, are on the track at any one time;

[92] The original question was: find the mean line delay at each node.

- the merry-go-round can handle 35 riders at one time, and the ride lasts for an exponentially distributed time with mean equal to 3 minutes;
- the water-tube ride lasts for an exponentially distributed time with mean equal to 1.5 minutes, and there can be ten tubes on the waterway at any one time, each carrying two riders;
- the fantasy ride is takes an exponentially distributed time with mean equal to 5-minute that can carry 60 persons at any one time;
- the ghost-mountain ride can handle 16 persons at a time and lasts for an exponentially distributed time with mean equal to 90 seconds;
- the journey to the moon ride takes an exponentially distributed time with mean equal to 100 seconds, and 20 riders can be on it at any one time.

If a ride can handle m riders simultaneously, we shall assume the corresponding service station has m servers. Furthermore, although the service times of all riders sharing a ride must be the same, we shall assume the service times are i.i.d.

(a) Use *Mathematica* to compute the expected number of visitors in the park in steady state.

(b) Which ride has the longest queue?[93] •

We treat now open Jackson networks whose

- stations can be viewed as forming a series system with flow always in a single direction from station to station and
- customers may enter the system from *outside* only at station 1 and depart only from station k (Gross and Harris, 1998, p. 166).

Definition 6.244 — **Tandem queue** (Gross and Harris, 1998, p. 166)
An open Jackson network associated with

- $r_i = \begin{cases} 1, & i = 1 \\ 0, & \text{otherwise} \end{cases}$

[93] We need to compute $E[L_q^{(i)}]$, the expected number of customers in the queue (not including those in service) which is given by $E[L_q^{(i)}] = E[L_s^{(i)}] - \frac{\lambda_i}{\mu_i}$.

- $P_{ij} = \begin{cases} 1, & i = 1, \ldots, k-1, \quad j = i+1 \\ 0, & \text{otherwise,} \end{cases}$

  that is, the routing matrix is given by

$$\mathbf{P} = [P_{ij}]_{i,j=1,\ldots,k} = \begin{bmatrix} 0 & 1 & 0 & \cdots & \cdots & 0 \\ 0 & 0 & 1 & \ddots & \vdots & 0 \\ 0 & 0 & 0 & \ddots & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & \cdots & 0 & 1 \\ 0 & 0 & \cdots & \cdots & 0 & 0 \end{bmatrix}, \tag{6.225}$$

is said to be a tandem (or series) queue. •

Remark 6.245 — Applications of tandem queues (Gross and Harris, 1998, p. 167) Tandem queues arise in:

- manufacturing and assembly-line processes in which units must proceed through a series of work stations, each station performing a given task;
- registration processes, when a registrant must visit a series of desks;
- clinic physical examination procedures, where a patient has a consult, is submitted to lab tests, etc. •

Exercise 6.246 — Two single-server queues in tandem (Ross, 2003, pp. 496–497) Consider a tandem queue with two single-server stations described as follows:

- customers arrive at a Poisson rate λ at server 1;
- after being served by server 1, customers then join the queue in front of server 2;[94]
- each server serves one customer at a time with server i taking an exponential time with rate μ_i $(i = 1, 2)$ for a service.

[94] We suppose there is infinite waiting space at both servers.

Find:

(a) the average number of customers in this system in equilibrium (Ross, 2003, pp. 497–498);

(b) the probability that the network is empty (Kulkarni, 2011, p. 230, Example 6.22). •

Exercise 6.247 — Three single-server queues in tandem (Jain, 2008, Homework 32)
In a series network of three routers, the packets arrive at the first router according to a Poisson process at the rate of 100 packets/s. The service times are exponentially distributed in any of the three routers with rates equal to 250 packets/s, 150 packets/s and 200 packets/s, respectively.

(a) Write an expression for the steady state probabilities.

(b) Calculate the probability of having 2 packets at each of the three routers. •

Exercise 6.248 — k single-server queues in tandem (Kulkarni, 1995, pp. 364–365, Example 7.6)
Admit k single-server queues are set in tandem:

- customers arrive from outside only at station 1 according to a Poisson process with rate $r_1 = \lambda$;
- station i provide service at rate μ_i;
- customers leaving station i join station $i + 1$ $(i = 1, \ldots, k - 1)$;
- customers leaving station k depart the system.

Derive:

(a) a condition that ensures this system reaches equilibrium;[95]

(b) the steady-state distribution. •

Exercise 6.249 — Tandem queue with four stations with different numbers of servers (Kulkarni, 2011, pp. 224, 230–231, examples 6.18 and 6.23)
Consider a queueing network with four service stations with the following characteristics:

- customers arrive at station 1 according to a Poisson process at a rate of 10 per hour and get served at stations 1, 2, 3, and 4 in that order;

[95] $r_1 = \lambda < \min_{i=1,\ldots,k} \mu_i$, i.e., the slowest server determines the traffic handling capacity of this tandem system.

- customers depart after getting served at station 4;
- the mean service time is 10 minutes at station 1, 15 minutes at station 2, 5 minutes at station 3, and 20 minutes at station 4;
- the first station has two servers, the second has three servers, the third has a single-server, and the last has four servers.

Compute the expected number of customers in this network in equilibrium. •

Exercise 6.250 — Optimizing the distribution of six servers in an open Jackson network with two stations in tandem (Kulkarni, 2011, p. 244, Exercise 6.41)
Consider a tandem queue with two stations characterised as follows:

- customers arrive at the first station according to a Poisson process at a rate of 24 per hour;
- the service times at the first station are i.i.d. exponential r.v. with mean 4 minutes;
- after service completion at the first station, the customers move to station 2, where the service times are i.i.d. exponential with mean 3 minutes.

The network manager has six servers at her disposal. How many servers should be stationed at each station so that the expected number of customers in the network is minimized in the long run? •

6.10.3 Closed Jackson networks

Closed queueing networks are suitable models for population studies, multiprogrammed computer systems,[96] etc. (Kulkarni, 1995, p. 369).

Let us remind the reader that in a closed Jackson network there are

- no external arrivals to the system and
- no departures from the system,

therefore the total number of customers remains constant (Kulkarni, 1995, p. 369). Furthermore, $r_i = 0$ and $\sum_{i=1}^{k} P_{ij} = 1$, for all $i = 1, \ldots, k$ (Prabhu, 1997, p. 103).

We suppose, from now on, that:

- there is a finite number of customers, say N, who continuously travel inside the network from station to station (Prabhu, 1997, p. 103);
- when there are n_i customers in station i ($i = 1, \ldots, k$), the service rate at this station is equal to $\mu_i \equiv \mu_i(n_i)$, with $\mu_i(0) = 0$ and $\mu_i(n_i) > 0$, $n_i = 1, \ldots, N$ (Kulkarni, 1995, p. 369);
- when a customer completes service at station i, he/she joins station j with probability P_{ij} (Kulkarni, 1995, p. 369);
- the routine matrix $\mathbf{P} = [P_{ij}]_{i,j=1,\ldots,k}$ is irreducible and aperiodic (Prabhu, 1997, p. 103).[97]

Proposition 6.251 — Characterizing the state vector of a closed Jackson network (Kulkarni, 1995, p. 369)
Let $\underline{N}(t) = (N_1(t), \ldots, N_k(t))$ represent the state vector of the closed Jackson network at time t.

Then $\{\underline{N}(t) : t \geq 0\}$ is CMTC with state space given by $\mathcal{S} = \{\underline{n} = (n_1, \ldots, n_k) \in \mathbb{N}_0^k : \sum_{i=1}^{k} n_i = N\}$. •

[96] *In the early days of computing, central processing unit (CPU) time was expensive, and peripherals were very slow. When the computer ran a program that needed access to a peripheral, the CPU would have to stop executing program instructions while the peripheral processed the data. This was deemed very inefficient. [...] The use of multiprogramming was enhanced by the arrival of virtual memory and virtual machine technology, which enabled individual programs to make use of memory and operating system resources as if other concurrently running programs were, for all practical purposes, non-existent and invisible to them.* (http://en.wikipedia.org/wiki/Computer_multitasking#Multiprogramming)

[97] Irreducible and non absorbing, according Gross and Harris (1998, p. 184). But note that if k is finite and larger than 1 and the DTMC is irreducible then there cannot be an absorbing state.

Proposition 6.252 — Traffic equations for a closed Jackson network
In a closed Jackson network,

- the input to station j is only due to routing from stations $i = 1, \ldots, k$, at rate $\sum_{i=1}^{k} \lambda_i P_{ij}$.

Therefore, in the steady state, λ_j can be obtained as the solution of the following traffic equations:

$$\lambda_j = \sum_{i=1}^{k} \lambda_i P_{ij},\ j = 1, \ldots, k \tag{6.226}$$

(Gross and Harris, 1998, p. 183). These equations have a positive solution, unique up to a multiplicative constant (Prabhu, 1997, p. 101) and can be written in matrix form,

$$\underline{\lambda} = \underline{\lambda}\,\mathbf{P} \quad \Leftrightarrow \quad \underline{\lambda} \times (\mathbf{I} - \mathbf{P}) = \underline{\mathbf{0}}, \tag{6.227}$$

where:

(i) $\underline{\lambda} = [\lambda_j]_{j=1,\ldots,k}$ denotes the row vector with the input rates;

(ii) $\mathbf{I}$ is the identity matrix with rank k;

(iii) $\underline{\mathbf{0}} = [0 \quad \cdots \quad 0]$ is a row vector with k zeroes. •

Remark 6.253 — Solving the traffic equations for a closed queueing network

- (6.227) remind us of the matrix representation of the equations that define the unique stationary distribution of a irreducible and positive recurrent DTMC. In fact, if we let

  (i) $\mathbf{P} = [P_{ij}]_{i,j=1,\ldots,k}$ be a $k \times k$ TPM of an irreducible DTMC with finite state space $\{1, \ldots, k\}$ (thus, a positive recurrent DTMC) and

  (ii) $\underline{\pi} = [\pi_j]_{j=1,\ldots,k}$ be the row vector denoting its stationary distribution,

  then

  $$\pi_j = \sum_{i=1,\ldots,k} \pi_i P_{ij},\ j = 1, \ldots, k \quad \Leftrightarrow \quad \underline{\pi} \times (\mathbf{I} - \mathbf{P}) = \underline{\mathbf{0}}. \tag{6.228}$$

- Recall that the proviso that $\sum_{j=1,\ldots,k} \pi_j = 1$, enables us to provide a solution in matrix form to (6.228). It is given by

  $$\underline{\pi} = \underline{\mathbf{1}} \times (\mathbf{I} - \mathbf{P} + \mathbf{ONE})^{-1}, \tag{6.229}$$

  where:

(i) $\underline{\mathbf{1}} = [1 \quad \cdots \quad 1]$ is a row vector with k ones;

(ii) **ONE** is the $k \times k$ matrix all of whose entries are equal to 1.

- Technically, $\underline{\lambda}$ is the left eigenvector of $\mathbf{P}$ that corresponds to the eigenvalue 1 (Gautam, 2012, p. 341).

  Understandably, being an eigenvector, $\underline{\lambda}$ is not unique, however, we have $\frac{\lambda_i}{\lambda_j}$ a constant for every pair i and j, regardless of the solution $\underline{\lambda}$ we have considered, and for that reason the $\underline{\lambda}$ values are also called visit ratios (Gautam, 2012, p. 341). Consequently, it is wiser to alternatively consider

$$\begin{aligned} \underline{\lambda} &= \underline{\pi} \\ &= \underline{\mathbf{1}} \times (\mathbf{I} - \mathbf{P} + \mathbf{ONE})^{-1}. \end{aligned} \tag{6.230}$$

•

The closed Jackson networks we are dealing with have a product-form stationary distribution (http://en.wikipedia.org/wiki/Queueing_theory#Queueing_networks) for the state vector $\underline{N}(t) = (N_1(t), \ldots, N_k(t))$.

Proposition 6.254 — Steady state behaviour of $\underline{N}(t)$, closed Jackson network (Kulkarni, 1995, p. 370; Gross and Harris, 1998, pp. 184, 178)
This is the product-form solution for

$$\lim_{t \to +\infty} P[\underline{N}(t) = \underline{n}] = P(\underline{N} = \underline{n}), \quad \underline{n} \in \mathcal{S}, \tag{6.231}$$

where

$$\mathcal{S} = \left\{ \underline{n} = (n_1, \ldots, n_k) \in \mathbb{N}_0^k : \sum_{i=1}^{k} n_i = N \right\}, \tag{6.232}$$

while dealing with a closed Jackson network in steady state, with k:

- single-server stations

$$P(\underline{N} = \underline{n}) = \frac{1}{G(N)} \times \prod_{i=1}^{k} \left(\frac{\pi_i}{\mu_i} \right)^{n_i}, \quad \underline{n} \in \mathcal{S}, \tag{6.233}$$

  where the normalising constant $G(N)$ is chosen such that $\sum_{\underline{n} \in \mathcal{S}} P(\underline{N} = \underline{n}) = 1$, that is, $G(N) = \sum_{\underline{n} \in \mathcal{S}} \prod_{i=1}^{k} \left(\frac{\pi_i}{\mu_i} \right)^{n_i}$;

- m_i server stations

$$P(\underline{N} = \underline{n}) = \frac{1}{G(N)} \times \prod_{i=1}^{k} \frac{\pi_i^{n_i}}{\prod_{j=1}^{n_i} \mu_i(j)}, \quad \underline{n} \in \mathcal{S}, \tag{6.234}$$

where $G(N) = \sum_{\underline{n} \in \mathcal{S}} \left[\prod_{i=1}^{k} \frac{\pi_i^{n_i}}{\prod_{j=1}^{n_i} \mu_i(j)} \right]$. •

Remark 6.255 — State space and the calculation of $G(N)$

- For arbitrary N and k, there are $\binom{N+k-1}{N}$ possible ways to allocate the N customers among the k stations (Gross and Harris, 1998, p. 186).
- It is essential to calculate $G(N)$ to determine the steady state probabilities. This is only feasible with a calculator for relatively small values of N and k (Ross, 2003, p. 502), as illustrated by Example 6.256.
- An efficient algorithm to calculate $G(N)$, for large N and k, was developed by Buzen (1973). This algorithm can also be found in Gross and Harris (1998, pp. 186–188) and http://en.wikipedia.org/wiki/Buzen's_algorithm. •

Example 6.256 — Closed Jackson network with 2 customers and 3 stations (Gross and Harris, 1998, pp. 184–186, Example 4.4)
Consider a closed Jackson network with:

- $N = 2$ customers;
- $k = 3$ stations (station 1 has 2 servers, whereas 2 and 3 are single-server stations);
- exponentially distributed services provided with rates $\mu_1 = 2$, $\mu_2 = 1$ and $\mu_3 = 3$ by the servers from stations 1, 2 and 3, respectively;
- routing matrix

$$\mathbf{P} = \begin{bmatrix} 0 & \frac{3}{4} & \frac{1}{4} \\ \frac{2}{3} & 0 & \frac{1}{3} \\ 1 & 0 & 0 \end{bmatrix}.$$

(a) Identify the state space.

- **State space**

  Since we are dealing with $N = 2$ customers and $k = 3$ stations, the state space has $\binom{2+3-1}{2} = 6$ elements and

$$\begin{aligned}\mathcal{S} &= \left\{\underline{n}=(n_1,n_2,n_3)\in\mathbb{N}_0^3 : \sum_{i=1}^{3} n_i = 2\right\}\\ &= \{(2,0,0),(0,2,0),(0,0,2),(1,1,0),(1,0,1),(0,1,1)\}.\end{aligned}$$

(b) Write and solve the traffic equations.

- **Traffic equations**

  Having in mind that the routing matrix is equal to

$$\mathbf{P} = \begin{bmatrix} 0 & \frac{3}{4} & \frac{1}{4} \\ \frac{2}{3} & 0 & \frac{1}{3} \\ 1 & 0 & 0 \end{bmatrix},$$

  the traffic equations $\lambda_j = \sum_{i=1}^{k} \lambda_i P_{ij}$, $j = 1,\ldots,k$, read as follows:

$$\begin{cases} \lambda_1 = \lambda_2 \times \frac{2}{3} + \lambda_3 \\ \lambda_2 = \lambda_1 \times \frac{3}{4} \\ \lambda_3 = \lambda_1 \times \frac{1}{4} + \lambda_2 \times \frac{1}{3}. \end{cases}$$

- **Solving the traffic equations**

$$\begin{aligned}\underline{\lambda} &= \underline{\mathbf{1}} \times (\mathbf{I} - \mathbf{P} + \mathbf{ONE})^{-1}\\ &= [1 \quad 1 \quad 1] \times \left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} 0 & \frac{3}{4} & \frac{1}{4} \\ \frac{2}{3} & 0 & \frac{1}{3} \\ 1 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \right)^{-1}\\ &= [1 \quad 1 \quad 1] \times \begin{bmatrix} 2 & \frac{1}{4} & \frac{3}{4} \\ \frac{1}{3} & 2 & \frac{2}{3} \\ 0 & 1 & 2 \end{bmatrix}^{-1}\\ &\overset{Mathematica}{=} [1 \quad 1 \quad 1] \times \begin{bmatrix} \frac{40}{81} & \frac{1}{27} & -\frac{16}{81} \\ -\frac{8}{81} & \frac{16}{27} & -\frac{13}{81} \\ \frac{4}{81} & -\frac{8}{27} & \frac{47}{81} \end{bmatrix}\\ &= \begin{bmatrix} \frac{4}{9} & \frac{1}{3} & \frac{2}{9} \end{bmatrix}.\end{aligned}$$

(c) Check that the steady state probabilities are:

(i) $p_{2,0,0} \simeq 0.0962$, $p_{0,2,0} \simeq 0.4332$, $p_{0,0,2} \simeq 0.0214$,

(ii) $p_{1,1,0} \simeq 0.2888$, $p_{1,0,1} \simeq 0.0642$, $p_{0,1,1} \simeq 0.0962$.

- **Steady state probabilities**

  Considering that

  (i) station 1 has 2 servers, whereas 2 and 3 are single-server stations, and

  (ii) the services last for exponentially distributed times with rates $\mu_1 = 2$, $\mu_2 = 1$ and $\mu_3 = 3$ by the servers from stations 1, 2 and 3,

  we can add that

$$
\begin{aligned}
[\mu_1(1) \quad \mu_1(2)] &= [\mu_1 \quad 2\,\mu_1] = [2 \quad 4] \\
[\mu_2(1) \quad \mu_2(2)] &= [\mu_2 \quad \mu_2] = [1 \quad 1] \\
[\mu_3(1) \quad \mu_3(2)] &= [\mu_3 \quad \mu_3] = [3 \quad 3]\,; \\
G(N) &= \sum_{\underline{n}\in\mathcal{S}} \left[\prod_{i=1}^{k} \frac{\pi_i^{n_i}}{\prod_{j=1}^{n_i} \mu_i(j)} \right] \\
&= \frac{\left(\frac{4}{9}\right)^2}{\mu_1(1) \times \mu_1(2)} \times 1 \times 1 + 1 \times \frac{\left(\frac{1}{3}\right)^2}{\mu_2(1) \times \mu_2(2)} \times 1 \\
&\quad +1 \times 1 \times \frac{\left(\frac{2}{9}\right)^2}{\mu_3(1) \times \mu_3(2)} + \frac{\frac{4}{9}}{\mu_1(1)} \times \frac{\frac{1}{3}}{\mu_2(1)} \times 1 \\
&\quad +\frac{\frac{4}{9}}{\mu_1(1)} \times 1 \times \frac{\frac{2}{9}}{\mu_3(1)} + 1 \times \frac{\frac{1}{3}}{\mu_2(1)} \times \frac{\frac{2}{9}}{\mu_3(1)} \\
&= \frac{\frac{16}{81}}{8} + \frac{1}{9} + \frac{\frac{4}{81}}{9} + \frac{\frac{4}{27}}{2} + \frac{\frac{8}{81}}{6} + \frac{\frac{2}{27}}{3} \\
&= \frac{2}{81} + \frac{1}{9} + \frac{4}{729} + \frac{2}{27} + \frac{4}{243} + \frac{2}{81} \\
&= \frac{187}{729}; \\
P(\underline{N} = \underline{n}) &= \frac{1}{G(N)} \times \left[\prod_{i=1}^{k} \frac{\pi_i^{n_i}}{\prod_{j=1}^{n_i} \mu_i(j)} \right] \\
&= \begin{cases}
\frac{\frac{2}{81}}{\frac{187}{729}} = \frac{18}{187} \simeq 0.0962, & \underline{n} = (2,0,0) \\
\frac{\frac{1}{9}}{\frac{187}{729}} = \frac{81}{187} \simeq 0.4332, & \underline{n} = (0,2,0) \\
\frac{\frac{4}{729}}{\frac{187}{729}} = \frac{4}{187} \simeq 0.0214, & \underline{n} = (0,0,2) \\
\frac{\frac{2}{27}}{\frac{187}{729}} = \frac{54}{187} \simeq 0.2888, & \underline{n} = (1,1,0) \\
\frac{\frac{4}{243}}{\frac{187}{729}} = \frac{12}{187} \simeq 0.0642, & \underline{n} = (1,0,1) \\
\frac{\frac{2}{81}}{\frac{187}{729}} = \frac{18}{187} \simeq 0.0962, & \underline{n} = (0,1,1).
\end{cases}
\end{aligned}
$$

  •

Exercise 6.257 — Closed Jackson network with 3 stations and 5 customers (Gautam, 2012, p. 343, Problem 57)
Consider a closed Jackson network with 3 stations and 5 customers who behave in the following fashion.

- Upon completing service at:
  - station 1, a customer rejoins station 1 with probability 0.5, or joins station 2 (resp. 3) with probability 0.1 (resp. 0.4);
  - stations 2 and 3, a customer always joins station 1.
- When it comes to the nodes:
  - station 1 has a single server that serves at rate i if there are i customers at the station;
  - station 2 has two servers each with service rate 1;
  - station 3 has one server with service rate 2.

(a) Identify the routing matrix.

(b) Write and solve the traffic equations.[98]

(c) Compute the joint and the marginal probabilities of the number of customers at each station in steady state.[99]

(d) Find the average number of customers in each station in steady state.

(Gautam, 2012, pp. 343–346, Problem 57.) •

Exercise 6.258 — Closed Jackson network model of a computer system with multiprogramming (Kulkarni, 1995, pp. 372–373, Example 7.9)
Consider the following model of a multiprogramming computer system consisting of 3 stations:

(i) a CPU (main computing device, station 1);

[98] $\underline{\lambda} = [\lambda_1 \quad \lambda_2 \quad \lambda_3] = \left[\frac{10}{15} \quad \frac{1}{15} \quad \frac{4}{15}\right]$.

[99] Marginal probabilities in steady state for: station 1, [0.0129 0.0642 0.1597 0.2611 0.3013 0.2009]; station 2, [0.7051 0.2521 0.0379 0.0045 0.0004 0.000015]; station 3, [0.3247 0.2945 0.2140 0.1167 0.0424 0.0077].

(ii) a printer (station 2);

(iii) a disc drive (station 3).

A program starts in the CPU. Furthermore:

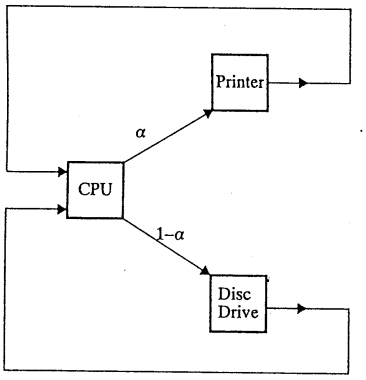
(i) when its computing part is done it goes to the printer (resp. disc drive) with probability α (resp. $1-\alpha$);

(ii) from the printer the program terminates (resp. goes back to the computing phase) with probability β (resp. $1-\beta$);

(iii) similarly, after the disc drive, the program goes back to the computing phase with probability 1;

(iv) when a program terminates, a new program is instantaneously admitted to the CPU queue, so that the total number of programs in the system remains constant, equal to N.[100]

Model this system as a closed Jackson shown in the figure below:



(a) Obtain the steady state probabilities.

(b) Compute these probabilities when $\alpha = \beta = \frac{1}{2}$ and $N = 2$. •

Definition 6.259 — **Cyclic queue** (Gross and Harris, 1998, p. 166)
Consider a closed Jackson network such that

$$P_{ij} = \begin{cases} 1, & i = 1, \ldots, k-1, \quad j = i+1 \\ 1, & i = k, \quad j = 1 \\ 0, & \text{otherwise,} \end{cases} \tag{6.235}$$

[100] N is usually called the *level of multiprogramming.*

that is, the routing matrix equals

$$\mathbf{P} = [P_{ij}]_{i,j=1,\ldots,k} = \begin{bmatrix} 0 & 1 & 0 & \cdots & \cdots & 0 \\ 0 & 0 & 1 & \ddots & \vdots & 0 \\ 0 & 0 & 0 & \ddots & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & \cdots & 0 & 1 \\ 1 & 0 & \cdots & \cdots & 0 & 0 \end{bmatrix}. \tag{6.236}$$

Then this closed Jackson network is referred to as a cyclic queue, essentially because customers flow in a *circle* always from station i to station $i+1$ ($i = 1, \ldots, k-1$), and then back to station 1, and so on. •

Remark 6.260 — Applications of cyclic queues (Koenigsberg, 1982)

- The concept of a cyclic queue was introduced by Taylor and Jackson in the *Operational Research Quarterly* in 1954. The paper dealt with the "flow" of aircraft engines from operation to maintenance to available for operation.
- Since then cyclic queue models have been applied to many other production and service industry problems, such as: computer design and control; ship operations; production processes; communications flow; etc. •

Exercise 6.261 — Cyclic queue with k single-server stations and one job (Kulkarni, 1995, pp. 370–372, Example 7.8)
Consider a cyclic queue with a single job/customer (i.e., $N = 1$) and k single-server stations with service rate μ_i, $i = 1, \ldots, k$.

(a) Write and solve the traffic equations.

(b) After having identified the state space, obtain the steady state probabilities. •

Exercise 6.262 — Cyclic queue with k single-server stations and 2 customers (Ross, 2003, pp. 504–506, Example 8.6)
Consider a cyclic queue with $N = 2$ customers and k single-server stations with service rate μ_i, $i = 1, \ldots, k$.

After having identified the state space, and written and solved the traffic equations, obtain the steady state probabilities. •

Exercise 6.263 — Cyclic queue with k single-server stations and N customers (Gross and Harris, 1998, pp. 199-200)
Consider a cyclic queue with N customers and k single-server stations with service rate μ_i, $i = 1, \ldots, k$.

(a) Write and solve the traffic equations.

(b) Obtain the steady state probabilities. •

Exercise 6.264 — Cyclic queue with two single-server stations and N customers (Jain, 2008, Example 32.1)
Consider a closed queueing network with two single-server queues and N jobs circulating among the queues. Admit servers 1 and 2 provide exponentially distributed service times with means 2 and 3, respectively.

Show that, in the steady state, the probability of having n_1 jobs in the first queue and $n_2 = N - n_1$ jobs in the second queue is equal to

$$P(N_1 = n_1, N_2 = N - n_1) = \frac{1}{G(N)} \times 2^{n_1} \times 3^{N-n_1},\ n_1 = 0, 1, \ldots, N,$$

where the normalizing constant $G(N)$ can be shown to be $3^{N+1} - 2^{N+1}$. •

Exercise 6.265 — Cyclic queue with two single-server stations and N customers (Kleinrock and Gail, 1996, p. 91, Problem 3.12)
Consider a cyclic queue in which N customers circulate around through two single-server stations. Servers 1 and 2 are of exponential type with rates μ_1 and μ_2, respectively.

Let $p_i = \lim_{t \to +\infty} P(i \text{ customers at station 1 and } N - i \text{ at station 2 at time } t)$.

(a) Draw the associated rate diagram.[101]

(b) Write down the balance equations for p_i, $i = 0, 1, \ldots, N$.

(c) Find the p.g.f. $P(z) = \sum_{i=0}^{N} z^i\, p_i$.

(d) Obtain p_i.

(Kleinrock and Gail, 1996, pp. 92–93.) •

[101] Note that this turns out to be exactly the rate diagram associated to a $M/M/1/N$ system, where $\lambda = \mu_1$ and $\mu = \mu_2$. Use this fact throughout this exercise.

Now, it is time to provide results for

- $E[L_s^{(i)}(N)]$, the average number of customers at station i, and
- $E[W_s^{(i)}(N)]$, the average time spent at station i,

in a closed Jackson network with N customers and k **single-server stations**.

The *mean value analysis* (MVA) enable us to determine recursively these two performance measures without computing the normalizing constant $G(N)$ (Ross, 2003, p. 502).

This approach relies on the Arrival Theorem stated next. This theorem is due to Reiser and Lavenberg (1980), according to http://en.wikipedia.org/wiki/Arrival_theorem, and its proof can be also found in Ross (2003, pp. 502–503).

Proposition 6.266 — Arrival theorem (Ross, 2003, p. 503)
In a closed Jackson network with N customers, the system in steady state as seen by an arriving customer to station j is distributed as the stationary distribution in a similar closed Jackson network with only $(N-1)$ customers. •

We now proceed with the derivation of the MVA.

Firstly, recall that the average time an arriving customer spends in a single-server system, with exponentially distributed interarrival and service times, is equal to the sum between the expected service time of the arriving customer, say $\frac{1}{\mu_i}$, and the sum of the expected service times of all the customers found in the system by this arriving customer. Consequently, upon conditioning on the number of customers found at server i by an arrival to that station i $(i = 1, \ldots, k)$, it follows that

$$\begin{aligned} E[W_s^{(i)}(N)] &= \frac{1}{\mu_i} \\ &\quad + \frac{1}{\mu_i} \times E(\text{no. of customers at station } i \text{ as seen by an arrival}) \\ &= \frac{1}{\mu_i} + \frac{1}{\mu_i} \times E[L_s^{(i)}(N-1)], \end{aligned} \tag{6.237}$$

where the last equality follows from the arrival theorem (Ross, 2003, p. 5003–504).

Secondly, the MVA also relies on:

- the principle that Little's formula is applicable throughout the queueing network (Gross and Harris, 1998, p. 189), therefore

$$E[L_s^{(i)}(N-1)] = \lambda_i(N-1) \times E[W_s^{(i)}(N-1)], \tag{6.238}$$

where $\lambda_i(N-1)$ denotes the arrival rate for station i in a closed Jackson network with $N-1$ customers (Gross and Harris, 1998, p. 190);

- the fact that the arrival rate to station i, $\lambda_i(N-1)$, can be obtained by multiplying the throughput rate[102]

$$\lambda^\star(N-1) = \sum_{j=1}^{k} \lambda_j(N-1) \tag{6.239}$$

and the steady state probability π_i, that is,

$$\lambda_i(N-1) = \lambda^\star(N-1) \times \pi_i. \tag{6.240}$$

Thirdly, plugging equation (6.238) and (6.240) in (6.237) leads to:

$$E[W_s^{(i)}(N)] = \frac{1}{\mu_i} + \frac{1}{\mu_i} \times \left\{\lambda^\star(N-1) \times \pi_i \times E[W_s^{(i)}(N-1)]\right\}, \tag{6.241}$$

where the throughput rate of the closed Jackson network with $N-1$ customers can be obtained by recalling that the number of customers is fixed in a closed Jackson network, i.e.,

$$\begin{aligned}
\sum_{i=1}^{k} E[L_s^{(i)}(N-1)] &= N-1 \\
\sum_{i=1}^{k} \lambda_i(N-1) \times E[W_s^{(i)}(N-1)] &= N-1 \\
\sum_{i=1}^{k} \lambda^\star(N-1) \times \pi_i \times E[W_s^{(i)}(N-1)] &= N-1 \\
\lambda^\star(N-1) &= \frac{N-1}{\sum_{i=1}^{k} \pi_i \times E[W_s^{(i)}(N-1)]}.
\end{aligned} \tag{6.242}$$

Finally, we get the recursion:

$$E[W_s^{(i)}(N)] = \frac{1}{\mu_i} + \frac{1}{\mu_i} \times \frac{(N-1) \times \pi_i \times E[W_s^{(i)}(N-1)]}{\sum_{j=1}^{k} \pi_j \times E[W_s^{(j)}(N-1)]}. \tag{6.243}$$

(Ross, 2003, p. 504).

The next proposition is actually an algorithm to obtain $E[W_s^{(i)}(N)]$ and $E[L_s^{(i)}(N)]$ via MVA.

[102]The throughput rate is the average service completion rate of the entire system, also known as the throughput rate of the closed Jackson network (Ross, 2003, p. 502).

Proposition 6.267 — Calculating $E[W_s^{(i)}(N)]$ and $E[L_s^{(i)}(N)]$ via MVA
$E[W_s^{(i)}(N)]$ and $E[L_s^{(i)}(N)]$ can be obtained as follows:

- **Step 1** — Obtain the vector of the steady state probabilities using the equality

$$\underline{\pi} = [\pi_i]_{i=1,\ldots,k} = \underline{\mathbf{1}} \times (\mathbf{I} - \mathbf{P} + \mathbf{ONE})^{-1}.$$

- **Step 2** — Set

$$E[W_s^{(i)}(1)] = \frac{1}{\mu_i},\ i = 1, \ldots, k.$$

- **Step 3** — Determine $E[W_s^{(i)}(2)]$, $E[W_s^{(i)}(3)]$, ..., $E[W_s^{(i)}(N)]$ by using the recursion

$$E[W_s^{(i)}(N)] = \frac{1}{\mu_i} + \frac{1}{\mu_i} \times \frac{(N-1) \times \pi_i \times E[W_s^{(i)}(N-1)]}{\sum_{j=1}^{k} \pi_j \times E[W_s^{(j)}(N-1)]}.$$

- **Step 4** — Set

$$\lambda^{\star}(N) = \frac{N}{\sum_{i=1}^{k} \pi_i \times E[W_s^{(i)}(N)]}.$$

- **Step 5** — Set

$$E[L_s^{(i)}(N)] = \lambda^{\star}(N) \times \pi_i \times E[W_s^{(i)}(N)].$$

•

Exercise 6.268 — Calculating $E[W_s^{(i)}(N)]$ and $E[L_s^{(i)}(N)]$ via MVA, for a cyclic queue with $N = 2$ customers
Consider a cyclic queue with k single-server stations and $N = 2$ customers.

Determine $E[W_s^{(i)}(N)]$ and $E[L_s^{(i)}(N)]$, for $i = 1, \ldots, k$, by using the MVA (Ross, 2003, pp. 504–505, Example 8.6). •

Exercise 6.269 — Calculating $E[W_s^{(i)}(N)]$ and $E[L_s^{(i)}(N)]$ via MVA, for a closed Jackson network with $N = 2$ customers and 3 stations
Consider the closed Jackson network with $N = 2$ customers and $k = 3$ stations from Example 6.256, but now assume all the stations have just one server.

Determine $E[W_s^{(i)}(N)]$ and $E[L_s^{(i)}(N)]$, for $i = 1, \ldots, k$, by using the MVA (Gross and Harris, 1998, pp. 191–192, Example 4.5). •

Exercise 6.270 — Closed Jackson network with seven single-server stations and 35 customers (Gross and Harris, 1998, p. 207, Exercise 4.17C)
Write a program in *Mathematica* to find the average number of customers and the average time spent at each station, for a closed Jackson network with $N = 35$ customers circulating between $k = 7$ single-server stations according to the routing matrix

$$\mathbf{P} = \begin{bmatrix} \frac{1}{3} & \frac{1}{4} & 0 & \frac{1}{4} & 0 & \frac{1}{6} & 0 \\ \frac{1}{3} & \frac{1}{4} & 0 & \frac{1}{4} & \frac{1}{6} & 0 & 0 \\ 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 \\ 0 & 0 & 0 & \frac{5}{6} & 0 & 0 & \frac{1}{6} \\ \frac{1}{6} & 0 & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & 0 \\ 0 & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{bmatrix}$$

and assuming the same service rates as in Exercise 6.242, that is:

- stations 1 and 2 have service rates of 85/min;
- stations 3 and 4 have service rates of 120/min;
- station 5 has a service rate of 70/min;
- stations 6 and 7 have service rates of 20/min. •

For a modification of the MVA algorithm for multiple-server stations, the reader is referred to Gross and Harris (1998, pp. 193–194).

6.10.4 Tandem queues with blocking and a few other networks

The main purpose of this very short section is to briefly introduce the reader to other (not necessarily Markovian) queueing networks.

Let us start by addressing tandem queues with blocking (Gross and Harris, 1998, pp. 172-174). The probabilistic analysis of this sort of queue is rather complex, essentially because of the blocking effect. These are tandem queues with the following characteristics:

- no queue is allowed to form in any of the stations (Gross and Harris, 1998, pp. 172), that is, none of the stations has a waiting area;
- if a station downstream comes up to capacity then any further processing at upstream stations, which feeds it, is prevented (Gross and Harris, 1998, pp. 172).

The next exercise is actually an unnumbered example taken from Gross and Harris (1998, pp. 172–174); it details how a tandem series with two single-server stations operates.

Exercise 6.271 — Tandem queues with blocking (and no waiting room)
Suppose two $M/M/1$ stations are set in series and interact as follows:

- all customers require service at both stations 1 and 2, in this specific order;
- if a customer is being served at station 2 and a service is completed at station 1, the station 1 customer must wait there until the service of the station 2 customer is completed; moreover, in this case the system is blocked and any arrivals to station 1 are rejected (namely because this station is not empty and it has no waiting area);
- if a customer is being served at station 1 then, even if station 2 is empty, any arriving customers to station 1 are turned away;
- arrivals to the system (i.e., to station 1) are governed by a Poisson process with parameter λ.

Finally, services at stations 1 and 2 are independent and exponentially distributed with parameters μ_1 and μ_2, respectively.

After having considered the following possible system states, essentially referring to the number of customers in each station,

(n_1, n_2)	Description
$(0,0)$	empty system
$(1,0)$	customer in process at station 1; no customer in station 2
$(0,1)$	no customer in station 1; customer in process at station 2
$(1,1)$	two customers in process, one at each station
$(b,1)$	customer finished at station 1 but waiting for the completion of the service of the customer at station 2 (system is blocked)

(a) draw the associated rate diagram;

(b) write the balance equations for this multidimensional CTMC;[103]

(c) obtain the steady state probabilities when $\mu_1 = \mu_2 = \mu$.[104]

(Gross and Harris, 1998, pp. 172–174.) •

Tackling tandem queues, with a (positive and) finite waiting room and blocking, is far more complex than the system described in Exercise 6.271. The complexity results from having to write a balance equation for each possible state (Gross and Harris, 1998, p. 173).

For more detailed accounts on this sort of tandem queues, the reader is referred to:

- Hunt (1956), who treated a two station tandem queue in which the waiting room of station 1 is unlimited but no waiting is allowed between stations (Gross and Harris, 1998, p. 173);
- Perros (1994), a good general reference on queueing networks with blocking (Gross and Harris, 1998, p. 173).

As for other types of Jackson networks, open Jackson networks with multiple customer classes are briefly addressed by Gross and Harris (1998, pp. 182-183).

For extensions of open Jackson networks with state-dependent service[105] and state-dependent arrival rate, the reader is referred to Kulkarni (1995, pp. 367–369, sections 4.1

[103] According to Gross and Harris (1998, p. 173) they are: $0 = -\lambda\, p_{(0,0)} + \mu_2\, p_{(0,1)}$; $0 = -\mu_1\, p_{(1,0)} + \mu_2\, p_{(1,1)} + \lambda\, p_{(0,0)}$; $0 = -(\lambda + \mu_2)\, p_{(0,1)} + \mu_1\, p_{(1,0)} + \mu_2\, p_{(b,1)}$; $0 = -(\mu_1 + \mu_2)\, p_{(1,1)} + \lambda\, p_{(0,1)}$; $0 = -\mu_2\, p_{(b,1)} + \mu_1\, p_{(1,1)}$.

[104] They can be found in Gross and Harris (1998, p. 173): $p_{(0,0)} = \frac{2\mu^2}{3\lambda^2+4\lambda\mu+2\mu^2}$; $p_{(1,0)} = \frac{\lambda(\lambda+2\mu)}{2\mu^2} \times p_{(0,0)}$; $p_{(0,1)} = \frac{\lambda}{\mu} \times p_{(0,0)}$; $p_{(1,1)} = \frac{\lambda^2}{2\mu^2} \times p_{(0,0)}$; $p_{(b,1)} = \frac{\lambda}{2\mu^2} \times p_{(0,0)}$.

[105] Actually, when we are dealing with a $M(\lambda_i)/M(\mu_i)/m_i$ station with $m_i > 1$, the service rate at this station is equal to $\min\{m_i, n\} \times \mu_i$ when there are n customers at that node, thus state dependent. Thus, what Kulkarni (1995, p. 367, Section 4.1) proposes is to further extend this model.

and 4.2). However, it is worth mentioning that these two extensions of Jackson networks have been originally treated by Jackson himself in his 1963 paper (Gross and Harris, 1998, p. 200). Interestingly, the open Jackson networks with state-dependent service considered by Kulkarni (1995, p. 367) admit a product-form solution involving independent r.v., and so does an open Jackson networks with state-dependent arrival rate; however, the queues at various nodes are not associated to independent r.v. (Kulkarni, 1995, p. 368).

Finally, Gross and Harris (1998, pp. 200-204) has an interesting discussion on these and other queueing networks.

References

- Adan, I., Boxma, O. and Perry, D. (2005). The G/M/1 queue revisited. *Mathematical Methods in Operations Research* **62**, 437–452.
- Adan, I. and Resing, J. (2002). *Queueing Theory.* (http://www.win.tue.nl/∼iadan/queueing.pdf)
- Asmussen, S. (2003). *Applied Probability and Queues* (2nd. edition). Springer-Verlag.
(QA274.12-.76.ASM.59288)
- Burke, P.J. (1956). The output of a queuing system. *Operations Research* **4**, 699–704.
- Buzen, J.P. (1973). Computational algorithms for closed queueing networks with exponential servers. *Communications of the ACM* **16**, 527–531.
- Cohen, J.W. (1972). The suprema of the actual and virtual waiting times during a busy cycle of the $K_m/K_n/1$ queueing system. *Advances in Applied Probability* **4**, 339–356.
- Cohen, J.W. (1982). *The single-server Queue (2nd., revised ed.).* North-Holland.
(QA274.12-.76.COH.30384)
- Cooper, R.B. (1981). *Introduction to Queueing Theory (2nd. edition).* North-Holland.
- Gautam, N. (2012). *Analysis of Queues: Methods and Applications.* CRC Press.
- Gross, D. and Harris C.H. (1998). *Fundamentals on Queueing Theory (3rd. edition).* John Wiley & Sons.
- Erlang, A.K. (1909). The theory of probability and telephone conversations. *Nyt Tidsskrift for Matematik B.*
- Hunt, G.C. (1956). Sequential arrays of waiting lines. *Operations Research* **4**, 674-683.
- Isaacson, D.L. and Madsen, R.W. (1976). *Markov Chains: Theory and Applications.* John Wiley & Sons.
(QA274.12-.76.ISA.28858)

- Jain, R. (2008). Queueing Networks. (www.cse.wustl.edu/∼jain/cse567-08/ftp/k_32qn.pdf)

- Kendall, D.G. (1951). Some problems in the theory of queues. *Journal of the Royal Statistical Society – Series B. Methodological* **13**, 151–173.

- Kendall, D.G. (1953). Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain. *The Annals of Mathematical Statistics* **24**, 338–354.

- Kleinrock, L. (1975). *Queueing Systems, Volume I: Theory.* John Wiley & Sons. (T57.9.KLE)

- Kleinrock, L. and Gail, R. (1996). *Queueing Systems: Problems and Solutions.* John Wiley & Sons.
(T57.92.KLE.49916)

- Koenigsberg, E. (1982). Twenty five years of cyclic queue and closed queue networks: a review. *The Journal of the Operational Research Society* **33**, 605–619.

- Koole, G. (2010). Optimization of Business Processes: An Introduction to Applied Stochastic Modeling. Department of Mathematics, VU University Amsterdam (www.math.vu.nl/∼koole/obp/obp.pdf?)

- Kulkarni, V.G. (1995). *Modeling and Analysis of Stochastic Systems.* Chapman & Hall.
(QA274.12-.76.KUL.59065, QA274.12-.76.KUL.45259)

- Kulkarni, V.G. (2011). *Introduction to Modeling and Analysis of Stochastic Systems* (Second Edition). Springer.

- Lee, A.M. (1966). *Applied Queueing Theory.* MacMillan.
(IST – Biblioteca de Civil, IO-04.7043)

- Morais, M.C. (2013). *Lecture Notes — Stochastic Processes* (Caps. 0–4), 236 pp. (https://fenix.ist.utl.pt/disciplinas/ipe64/2012-2013/2-semestre/material-didactico/lecture-notes)

- Morse, P.M. (1955). Stochastic Properties of Waiting Lines. *Journal of the Operations Research Society of America* **3**, 255–261.

- O'Brien, G.G. (1954). The Solution of Some Queueing Problems. *Journal of the Society for Industrial and Applied Mathematics* **2**, 133–142.

- Pires, A.M.P. (1990). *Sistemas M/M/r/n: comparação de sistemas com iguais taxas de chegadas e de serviço*. M.Sc. Thesis, IST, Universidade Técnica de Lisboa. (17-11.36399)

- Pacheco, A. (2002). *Class Notes – Stochastic Manufacturing and Service Systems*. Georgia Institute of Technology, Atlanta, USA. (https://fenix.ist.utl.pt/disciplinas/ipe64/2012-2013/2-semestre/material-didactico)

- Perros, H.G. (1994). *Queueing networks with blocking — Exact and approximate solutions*. Oxford University Press

- Pollaczek, F. (1957). *Problèmes Stochastiques Posés par le Phénomène de Formation d'une Queue d'Attente un Guichet et par des Phénomànes Apparantés*. Gauthier Villars.

- Pollaczek, F. (1961). *Théorie Analytique des Problèmes Stochastiques Relatif un Group de Lignes Téléphoniques avec Dispositif d'Attente*. Gauthier Villars.

- Prabhu, N.U. (1997). *Foundations of Queueing Theory*. Kluwer Academic Publishers. (QA274.12-.76.PRA.48373)

- Reiser, M. and Lavenberg, S.S. (1980). Mean-value analysis of closed multichain queuing networks. *Journal of the ACM* **27**, 313–***.

- Reynolds, J.F. (1968). The stationary solution of a multi server queueing model with discouragement. *Operations Research* **16**, 64–71. Birkhauser.

- Resnick, S. (1992). *Adventures in Stochastic Processes*. Birkhauser. (QA274.12-.76.RES.43493)

- Ross, S.M. (1983). *Stochastic Processes*. John Wiley & Sons. (QA274.12-.76.ROS.36921, QA274.12-.76.ROS.37578)

- Ross, S.M. (1989). *Introduction to Probability Models* (4th. edition). Academic Press.

- Ross, S.M. (2003). *Introduction to Probability Models (8th. edition).* Academic Press.
(QA273.ROS.62694)

- Ross, S.M. (2007). *Introduction to Probability Models (9th. edition).* Academic Press.

- Serfozo, R. (1999). *Introduction to Stochastic Networks.* Springer-Verlag.
(QA274.8.SER.51450)

- Sevcik, K.C. and Mitrani, I. (1981). The distribution of queuing network states at input and output instants. *Journal of the ACM* **28** 358–371.

- Srivastava, H.M. and Kashyap, B.R.K. (1982). *Special Functions in Queueing Theory and Related Stochastic Processes.* Academic Press.
(QA274.12-.76.SRI.34351)

- Wolff, R.W. (1989). *Stochastic modeling and the theory of queues.* Prentice-Hall.
(QA274.12-.76.WOL.35949)

- Zuckerman, M. (2000–2012). *Introduction to Queueing Theory and Stochastic Teletraffic Models.* (arxiv.org/pdf/1307.2968)