Stationary Queueing Models
with Aspects of
Customer Impatience and Retrial Behaviour

DI Mag. Christian Dombacher (BDD)
Nikolaus Lenaugasse 8
A-2232 Deutsch-Wagram

20.02.2008, rev. 18.01.2009

Contents

Chapter 1

Introduction to Queueing Theory

1.1 History

Queueing theory as part of probability theory has evolved from classic teletraffic engineering in the last decades. In 1909 A.K. Erlang, a Danish teletraffic engineer published a paper called *The Theory of Probabilities and Telephone Conversations*. In the early 1920s he developed the famous *Erlang model* to evaluate loss probabilities of multi-channel point-to-point conversations. The Erlang model was extended to allow for calculation in finite source input situations by Engset several years later leading to the *Engset model*. In 1951 D.G. Kendall published his work about *embedded Markov chains*, which is the base for the calculation of queueing systems under fairly general input conditions. He also defined a naming convention for queueing systems which is still used. Nearly at the same time D.V. Lindley developed an equation allowing for results of a queueing system under fairly general input and service conditions. In 1957 J.R. Jackson started the investigation of networked queues thus leading to so called queueing network models. With the appearance of computers and computer networks, queueing systems and queueing networks have been identified as a powerful analysis and design tool for various applications.

1.2 Applications

As mentioned above, queueing theory allows for calculation of a broad spectrum of applications. These include

- In *manufacturing systems*, raw materials are transported from station to station using a conveyor belt. With each station having performed its task, the item is allowed to proceed to the next station. If processing times at all stations are equal and the conveyor belt is filled in the same frequency as items proceed from one station to the other, no waiting can occur, as the assembly line works in *synchronous* mode. In *asynchronous* mode, queueing for stations might occur and clearly has an impact on overall performance.

- *Computer systems* to perform real-time or high speed operations are often subject to bad performance due to a single bottleneck device such as CPU, disk drive, graphics card, communication ports or bus system. By the use of analytical models the bottleneck device may be detected and as a consequence upgraded.

- By nature of the protocols used in *computer networks*, delays occur due to congestion of the transport network. These delays may be seen as waiting time until the media becomes free again thus allowing for calculation of throughput, overall delay and other performance values.

- *Teletraffic engineering* deals with the availability of stations, trunks and interconnection lines. Although these systems are characterized by *blocking* more than by delay, they still belong to the world of queueing systems. With the introduction of new media in teletraffic engineering, the delay paradigm becomes more important again. Teletraffic engineering now also has to cover a broad spectrum of new units such as announcement boards, interactive voice response units, media servers, media and signaling gateways.

- *Workforce management* is concerned about the most efficient allocation of personell. The application of queueing theory in workforce management is most visible in call centres, where agents have to be allocated according to the call load. Relying on other techniques such as forecasting, queueing theory may be seen just as another brick in the wall

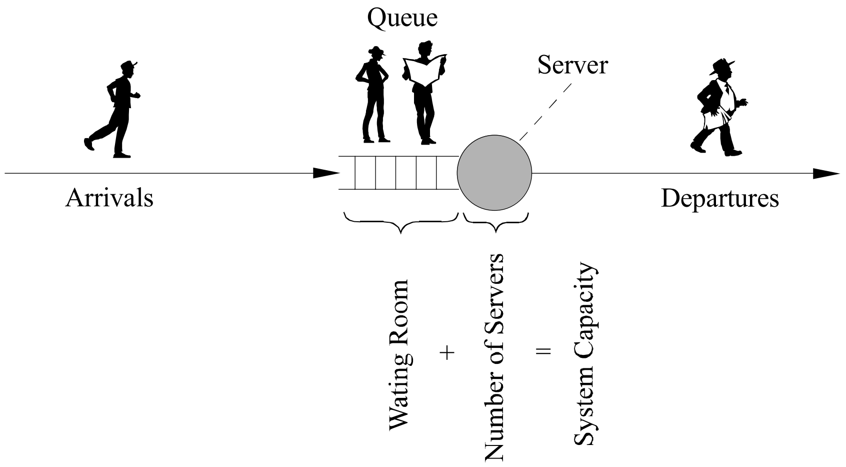


Figure 1.1: Schematic representation of a queueing system

in a wide range of solution methods to be applied to solve problems appearing in workforce management.

Obviously, the list above is far from being complete and may be extended further to other applications as well. For more information, the reader is referenced to publications such as *IEEE Communications Magazine*, *IEEE Computers*, *Bell Labs Technical System Journal* or similar.

1.3 Characterization

A queueing system may be described as a system, where customers arrive according to an *arrival process* to be serviced by a service facility according to a *service process.* Each service facility may contain one or more *servers.* It is generally assumed, that each server can only service one customer at a time. If all servers are busy, the customer has to queue for service. If a server becomes free again, the next customer is picked from the queue according to the rules given by the *queueing discipline.* During service, the customer might run through one or more *stages of service*, before departing from the system. A schematic representation of such a queueing system is given in figure 1.1. Before going into further detail, the most important aspects of queueing systems will be listed and briefly described.

- The *arrival process* is given by a statistical distribution and its parameters. Very often the exponential distribution is assumed resulting in the arrival pattern to be measured as the average number of arrivals per unit of time. When determining the trunk load in a PBX, the arrival pattern is often given in calls per busy hour. More general arrival processes are characterized by other pattern as well. These include batch arrivals and time dependence.

- The *service process* is described similar to the arrival process. Again, exponentiality is often assumed in practice due to intractabilities when releasing these assumptions. In opposite to the arrival process, the service process is highly dependent on the state of the system. In case, the queueing system is empty, the service facility is idle.

- The *queueing discipline* refers to the way, customers are selected for service under queueing conditions. Often used and most common is the *first come, first serve (FCFS)* discipline. Others include *last come, first serve (LCFS)*, random and priority service.

- The *departure process* is seldom used to describe a queueing system, as it can be seen as a result of queueing discipline, arrival and service process. Under certain conditions, arrival and departure process follow the same statistical distribution. This has become a very important fact in queueing network modeling.

- The *system capacity* introduces a natural boundary in queueing systems. In life systems, there are only limited number of resources such as trunks in a PBX, computer memory or network buffers. In queueing networks, nodes with finite system capacities may *block* customers from the previous node, when the node's capacity limit has been reached.

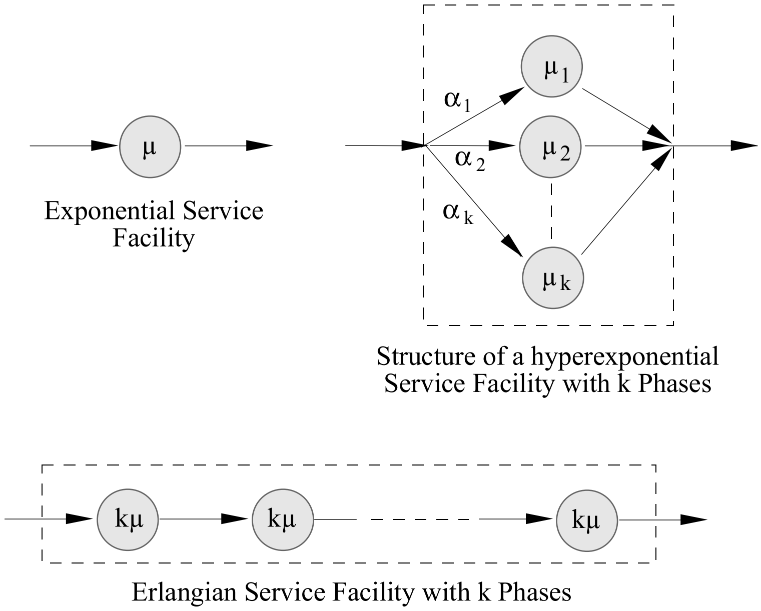
- The *number of servers* refers to the number of parallel nodes, which can service customers simultaneously. In telephone systems servers might describe trunks, tone detectors, tone generators and time slots.

- The number and structure of *service stages*, a customer might have to visit before departing the system. In a computer system, a job might have to visit the CPU twice and the I/O processor once during a single service. In practice, there exist a lot of situations, which can be modelled by complex queueing systems with service stages or simple computer networks.

1.4 Use of Statistical Distributions in Queueing Systems

As mentioned above, arrival, service and departure processes are described by means of *statistical distributions*. The most common distributions are the exponential and Poisson distributions. Statistical distributions are adjusted for life situations by customizing their parameters. Clearly, the more parameters are available for a certain distribution, the more flexible the distribution may be adjusted. On the other hand, estimating lots of parameters might become an infeasible task. It also turns out, that more complicated distributions result in almost intractable queueing models. Therefore the main target of selecting a proper distribution and estimating their parameters is to provide a tractable analytical model giving a close approximation to the life system under consideration. Sometimes the results are limited to a specific region only. One example are heavy load approximations, which fail to provide proper results for lightly loaded systems.

The *exponential* distribution with density $f(t) = \lambda e^{-\lambda t}$ posseses only one parameter $\lambda > 0$. Although severely limited, the exponential distribution is widely accepted, as queueing models based on the exponential distribution are very easy to handle. Some of shortcomings might be alleviated by creating mixtures of exponential distributions to define more complex distribution types such as the *Erlangian* (with density $f(t) = \frac{(\lambda k)^k}{(k-1)!} t^{k-1} e^{-k\lambda t}$) or the *hyperexponential* distribution (with density $f(t) = \sum_{i=1}^{k} \alpha_i \lambda_i e^{-\lambda_i t}$, $\lambda_i > 0$). Seen from the perspective of a service facility, one complex service facility is replaced by a certain arrangement of more simple service facilities each having an exponentially distributed service time. For a graphical representation, please refer to figure 1.2. The Erlangian distribution provides a good starting point for systems with *phases* or *stages* such as conveyor belts used in manufacturing systems. The design pattern is purely sequential, whereas the hyperexponential service facility follows a parallel arrangement. For the densities above the number of stages is denoted by k. More general arrangements may be achieved by mixing sequential and parallel arrangements resulting in so called *phase type distributions [42][59][53][15]*. This family of distributions



Exponential Service Facility

Structure of a hyperexponential Service Facility with k Phases

Erlangian Service Facility with k Phases

Figure 1.2: Simple and complex service facilities

includes all of the aforementioned as special cases. State transitions between phases comply to Markovian requirements thus allowing for generalization of analytic methods for memoryless systems.

Focusing on the parameter of the exponential distribution, it turns out, that λ describes the average rate. For service facilities, very often the average service time s is given, which may be easily converted to the average service rate μ by calculating the reciprocal, i.e. $\mu = \frac{1}{s}$. A similar description is available for the arrival and departure processes. Assuming an exponential distribution for the arrival process implicitely defines, that the times between subsequent arrivals, the so called *interarrival times* t, are exponentially distributed. This is graphically illustrated in figure 1.3.

Focusing on the exponential and the Poisson distribution, a useful equivalency may be derived. More formally, consider t_j as the time between two arrivals at T_j and T_{j-1}

$$t_j = T_j - T_{j-1}$$

assuming t_j for all j being exponentially distributed with parameter λ, i.e.

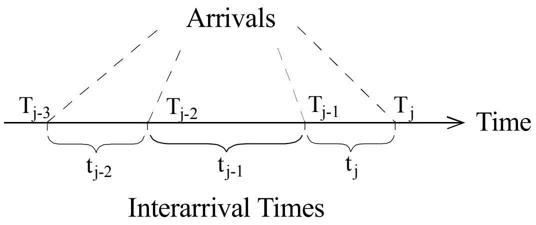


Figure 1.3: Interarrival times in an arrival process

$$\Pr\{t_j \geqq t\} = e^{-\lambda t}$$

then the number of arrivals N_t within $[0, t]$ follows a *Poisson* distribution:

$$\Pr\{N_t = j\} = f(j, \lambda) = \frac{(\lambda t)^j}{j!} e^{-\lambda t} \tag{1.1}$$

Without loss of generality, $t = 1$ might be assumed thus allowing for interpretation of λ as the average arrival rate. Thus an equivalent representation for arrival and departure processes has been found in terms of exponentially distributed interarrival times and Poisson distributed arrivals. A poisson probability mass function for $\lambda = 4.0$ is shown in figure 1.4. Spelled out, the graph describes the probability of N customers arriving at a queueing system, when the average rate of arrivals is four customers per unit time. The related probability density and cumulative distribution functions for $\lambda = 4.0$ are shown in figures 1.5 and 1.6.

In order to demonstrate several aspects of statistical distributions and their effect on queueing models, an analysis of a life system has been included as examples throughout the entire chapter. Instead of working through a large example at the end of the chapter, we have chosen to work out portions of the analysis where appropriate. In this subsection it will be shown, how sampled data can be matched with an exponential service time distribution.

Example 1 *Consider a call centre during the busy hour. Using the log of a CTI server, call holding times for each single call have been determined. In total 1707 calls have been measured. These calls have been arranged in groups with unit time of 15 sec, i.e. the first group includes calls with a holding time of 0-14 sec, the second group includes calls with holding time*

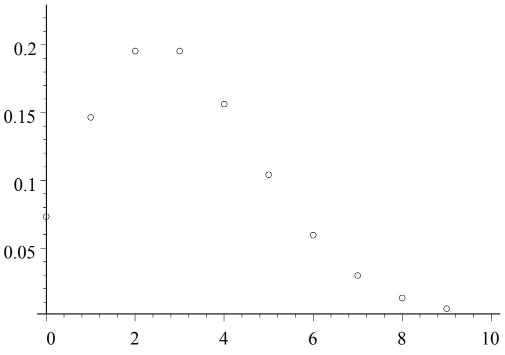


Figure 1.4: Poisson probability mass function with $\lambda = 4$

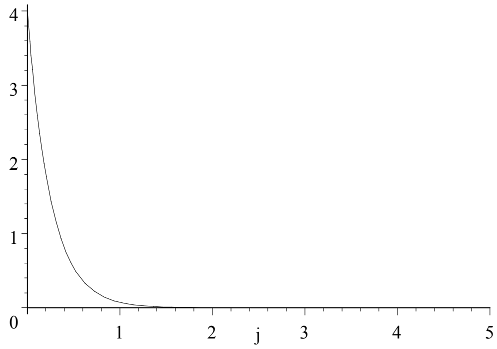


Figure 1.5: Exponential probabiliy density function with $\lambda = 4$

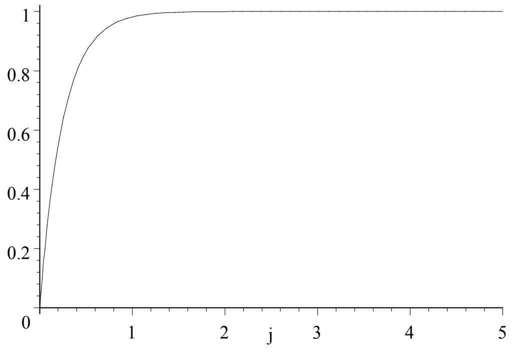


Figure 1.6: Exponential probability distribution function with $\lambda = 4$

between 15 and 29 seconds, etc. A call lasts 162 sec on the average, whereas the standard deviation σ of the holding time is 169, i.e. $\sigma = 169$. In order to visualize the distribution of calls, a histogram as shown in figure 1.7 has been created. Please note, that the wording distribution has not been used in the strong statistical sense.Giving a closer look to figure 1.7, the shape suggests an exponential distribution. Taking into account, that mean and standard deviation of the exponential distribution are both equal to $\frac{1}{\lambda}$, it can be seen, that measured data exhibit a similar value for sample mean and standard deviation. We therefore ignore the slight difference and attempt a so called two-moment approximation. As a grouping of data with interval length 15 secs has been introduced, the average holding time will be scaled as well, i.e. $\frac{1}{\lambda} = 162/15 = 10.8$. Plotting the formula for the exponential probability density function (PDF)

$$f(t) = \lambda e^{-\lambda t} \tag{1.2}$$

reveals figure 1.8.The same procedure has been applied to the cumulative distribution function (CDF)

$$F(t) = 1 - e^{-\lambda t} \tag{1.3}$$

to create figure 1.9.In order to compare the result with the histogram shown in figure 1.7, the probability density function has to be scaled by the number of calls 1707 on the y-axis and the interval length 15 on the x-axis. The resulting plot is shown in figure 1.10.By overlapping the two figures it turns out, that the fitted exponential distribution provides an acceptable approximation to the measured data. Thus we have justified the exponentiality assumption for

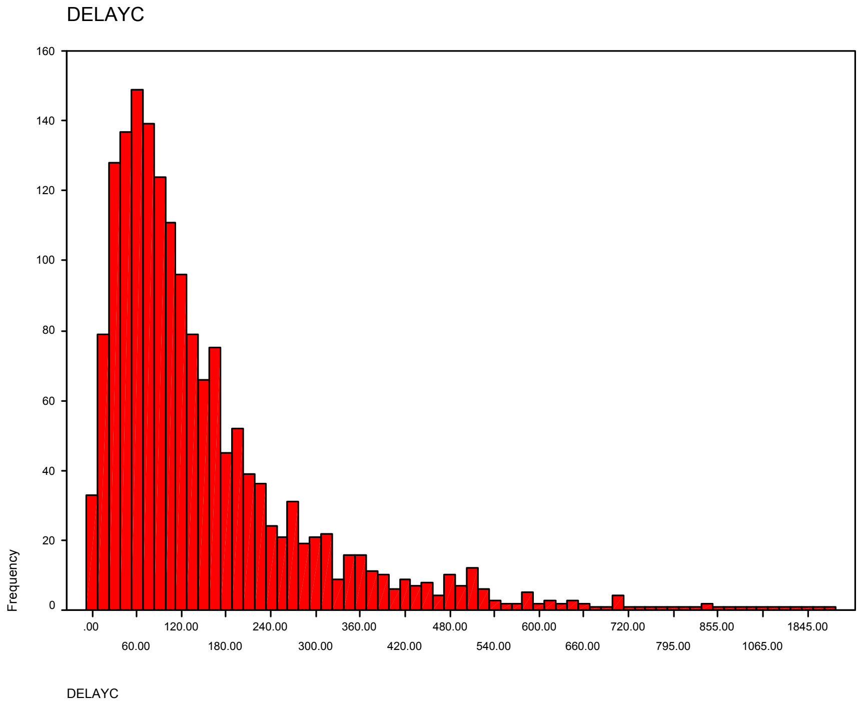


Figure 1.7: Histogram showing the distribution of calls in a call centre

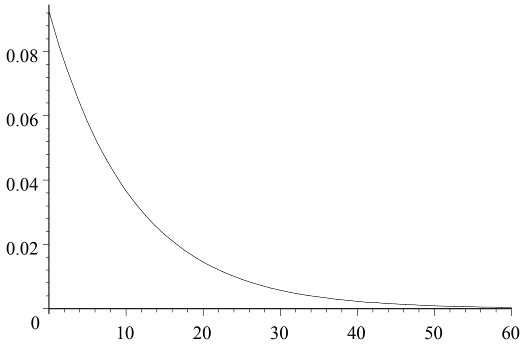


Figure 1.8: Fitted exponential PDF for $\frac{1}{\lambda} = 10.8$

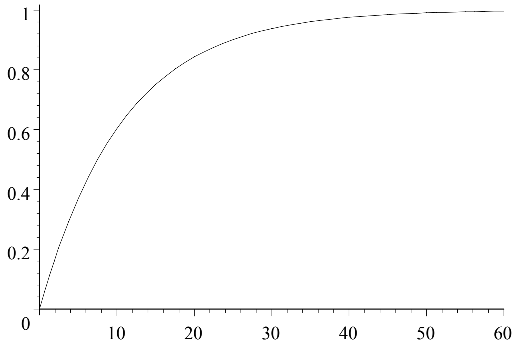


Figure 1.9: Fitted exponential CDF for $\frac{1}{\lambda} = 10.8$

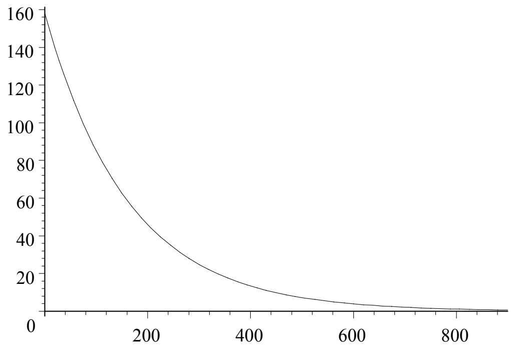


Figure 1.10: Scaled fitted exponential PDF for $\frac{1}{\lambda} = 10.8$

this set of data. Please note, that usually the match between measured data and the chosen statistical distribution is verified by a so called goodness-of-fit test. Using goodness-of-fit tests, a statistic is derived from the differences between the theoretical distribution and matched data to allow for acceptance or rejection of the fitted distribution.

One of the most appealing properties arising in queueing systems is the *memoryless or Markov property* of the exponential distribution. The memoryless property states, that the remaining (residual) time of an exponential process does not depend on the past. Consequences for the analysis of queueing systems include

- Given an exponentially distributed service time, a customer in service to be completed at some future time is independent of the time he has been in service so far. The remaining service time is still exponentially distributed. End of work can be seen as a sudden event, not as a result of work progress. The server simply forgets, how long it has been operating.
- Given Poisson distributed arrivals, the time to the next arrival at any point of time is exponentially distributed.

Theorem 2 *The exponential distribution is memoryless.*

Proof. The proof is based on the definition of conditional probability. A random variable T is said to be memoryless, if

$$\Pr\{T > t + t_0 | T > t_0\} = \Pr\{T > t\}$$

Now a random variable T is assumed to be exponentially distributed with parameter λ, i.e. $\Pr\{T \leqq t\} = 1 - e^{-\lambda t}$. Hence,

$$\begin{aligned} \Pr\{T &> t + t_0 | T > t_0\} = \frac{\Pr\{T > t + t_0\}}{\Pr\{T > t_0\}} \\ &= \frac{e^{-\lambda(t+t_0)}}{e^{-\lambda t_0}} = e^{-\lambda t} = \Pr\{T > t\} \end{aligned}$$

which completes the proof. ■

Furthermore, it can be shown, that the exponential distribution is the only continuous distribution exhibiting the memoryless property. For more

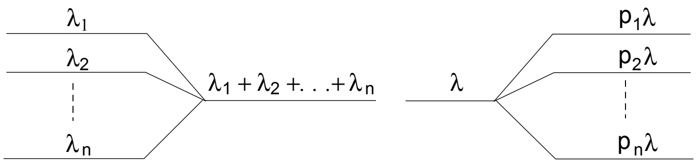


Figure 1.11: Merging and splitting of Poisson streams

information, refer to [29]. The discrete counterpart of the exponential distribution is the *geometric distribution*, which is commonly used to model cell based networks.

Without the need to handle the residual times, the use of the exponential distribution became very popular. In order to overcome the shortcomings of the exponential distribution in queueing systems, very often mixtures of exponential distributions are used in standard models instead of deriving a new model suitable for the distribution required. In fact, it turns out, that these mixture distributions are highly flexible due to their extensive sets of parameters.

Based on the memoryless property of the exponential distribution, additional useful relations for Poisson processes may be derived. Given n Poissonian streams, interarrival times are exponentially distributed with parameter λ_i, where $i = 1 \ldots n$, i.e. $F_i(t) = 1 - e^{-\lambda_i t}$, these streams may be merged to a single Poissonian stream, where interarrival times are distributed according to the distribution function $F(t) = 1 - e^{-(\lambda_1+\lambda_2+\ldots+\lambda_n)t}$. Consequently, a single Poisson stream may be splitted up still preserving the Poissonian nature of each substream. The related interarrival times are exponentially distributed with parameter $p_i\lambda$, where p_i denotes the propability, that a single customer joins substream i. For a graphical representation of these relations, refer to figure 1.11. As a consequence, multiple independent Poisson arrival streams may be seen as a single arrival stream. On the other hand, a single Poisson arrival stream presented to multiple servers, may be treated like multiple arrival streams. Special care has to be taken, if dependent streams are considered, e.g. those in feedback systems. The Poisson assumption does not necessarily hold under these circumstances.

1.5 Approximation of Arbitrary Distributions

In most practical situations one will rarely encounter distributions such as exponential and Erlang. This raises the question, which class of distributions might be sufficient to capture almost all situations in practice. Fortunately there is an answer to it. It turns out, that mixtures of *exponential distributions in serial and/or parallel* (also called generalized Erlang distributions) are capable of reasonably approximating any distribution. Before proceeding to the result the concept of *weak convergence* of probability distributions has to be introduced

Definition 3 *Given a series of distribution functions F_n and a distribution function F with $\lim_{n\to\infty} F_n(x) = F(x)$ for all continuity points x of F, then F_n is said to converge weakly (or in distribution). This is denoted by $F_n \rightharpoonup F$.*

The F_n above will become elements out of the class of exponential distributions in serial and/or paralell and F denotes the distribution to be approximated. In case of a continuous distribution F the limit is valid for all x.

Theorem 4 *Choose F to be an arbitrary distribution on the positive reals $(0, \infty)$ with finite k-th moment $\mu_F^{(k)}$. Then for each n there exists a F_n out of the class of exponential distributions in serial and/or parallel, which converges weakly to F. Furthermore the moments $\mu_{F_n}^{(l)}$ of F_n converge to $\mu_F^{(l)}$ for all $l \leq k$.*

The proof is omitted here, as it consults concepts such as completeness and denseness in probabilistic metric spaces. For the question raised above, it is interesting to note, that the class of exponential distributions in serial/parallel is equivalent to the family of *Cox distributions*, which is in turn part of the class of phase type distributions. As a consequence each of the stated distribution families sufficiently approximates the desired target distribution. For a mathematical treatment of the subject the reader is referred to [4].

1.6 Renewal Processes

In the previous sections we've learned that the Poisson process corresponds to exponential interarrival times. By relaxing the exponential assumption, one arrives at the so called *renewal process.* Renewal processes are characterized by independent interarrival times following a common distribution. They may be applied for the arrival as well as service processes, so in the following the event of an arrival will be called a *renewal.* Let T_n now denote the time between the n-1st and nth renewal, $S_n = \sum_{k=1}^{n} T_n$ with $S_0 = 0$ the time of the nth renewal and $N(t) = \sup\{n : S_n \leq t\}$ the total number of renewals in the interval $[0, t]$. Then $N(t)$ for all $t \geq 0$ will formally describe the renewal process. Taking expectation one arrives at the *renewal function* $m(t) = \mathbb{E}N(t)$.

For the Poisson process the T_n were independent identically distributed according to an exponential distribution. Consequently the distribution of S_n results from the n-fold convolution of the exponential distribution, that is an n-stage Erlang distribution. $N(t)$ counts the number of renewals up to the time t, which describes the Poisson process.

By denoting $s = \mathbb{E}T_n = \int_0^\infty t dF(t)$ to be the expected renewal time (e.g. interarrival or service time), where T_n is identically distributed with distribution function F for all $n \geq 1$, one arrives at certain interesting limits

Theorem 5 *Based on the notation above the following limits hold*

$$\lim_{t\to\infty} \frac{m(t)}{t} = \frac{1}{s}$$

and

$$\Pr\left\{\lim_{t\to\infty} \frac{N(t)}{t} = \frac{1}{s}\right\} = 1$$

The proof is based on the strong law of large numbers and is omitted here. The interested reader may consult [48] or [4]. The second limit holds only with probability one. That means, that there are exceptions to the rule, but these exceptions are negligible. In the context of arrival and service processes s simply describes the interarrival or service time. Consequently both limits converge to the arrival and service rates $\frac{1}{s}$. These results confirm our intuition: Observing a process for a very long time and dividing the number of occurences by the total time, one arrives at the rate of that process.

By utilizing the central limit theorem, we are also able to derive asymptotic results for sufficiently large t. As $t \longrightarrow \infty$, $N(t)$ is asymptotically normal distributed with mean $\frac{t}{s}$ and variance $\frac{t\sigma^2}{s^3}$ given the variance $\sigma^2 = \int_0^\infty (t-s)^2 \, dF(t)$ of the renewal distribution F exists. More details may be found in [4] and [14].

One could ask now, why the Poisson process plays such a prominent rule among the class of renewal processes. The answer lies in the fact of merging and splitting. It will turn out, that this feature is unique in the class of stationary renewal processes. A process $N(t)$ is called *stationary*, if a shift in time does not alter the distribution of the epochs, i.e. $N(t+s)-N(t)$ has the same distribution as $N(s)$.

Theorem 6 *Given stationary renewal processes $N_1(t),...,N_n(t)$ and $N(t) = N_1(t)+...+N_n(t)$ each with common density function continues on the interval $(0,\infty)$ and right continous at 0, whereas the $N_1(t),...,N_n(t)$ are independent for all $t \geq 0$. Then $N_1(t),...,N_n(t)$ are all Poisson processes.*

Again the proof is omitted, because it requires the theory of point processes, which is not central to the current discussion. Also note the exact description of continuity above. It stems from the fact, that distributions defined for the positive reals can not be continous at 0 from the left. For more information on the superposition of point processes, consult [11].

1.7 Performance Characteristics of Queueing Systems

So far aspects of queueing models and statistical distributions have been discussed. As the usefulness of a model varies with its results, appropriate models and algorithms have to be selected. Another important factor is the point of view taken. Performance values calculated with respect to an arriving customer are not necessarily the same as those determined from a servers viewpoint. Again, the impact of statistical distributions is not negligible. However, it turns out, that these performance values are the same, when using models with exponentially distributed interarrival and service times. On the other hand, a lot of useful relations have been determined for more general cases as well. Although queueing models vary in application and complexity, a common set of performance characteristics may be determined as follows.

- The *state propability* p_n is described by the probability of n customers residing in the system, either being served or waiting. Thus,

$$p_n = \Pr\{n \text{ customers in system}\}$$

- The *traffic intensity* ρ is given by the ratio of arrival rate λ and service rate μ, i.e.

  $$\rho = \frac{\lambda}{\mu} \tag{1.4}$$

  Alternatively, the traffic intensity may also be seen as the ration of average service time $s = \frac{1}{\mu}$ and average interarrival time $t = \frac{1}{\lambda}$, i.e.

  $$\rho = \frac{s}{t} \tag{1.5}$$

  The traffic intensity is sometimes expressed in *erlangs* with respect to the Danish teletraffic engineer. In the United States very often *centum call seconds (CCS)* are used instead of erlang, as some manufacturers poll traffic sensitive equipment every 100 seconds [41]. In fact, a server being busy for an hour, carries a load of 36 CCS or equivalently 1 erlang. Expressed in a formula,

  $$\rho_{ccs} = 36\rho_{erl}$$

- The proportion of time, a server or a group of servers may be busy, is given by the *server utilization*

  $$u = \frac{\lambda}{m\mu} = \frac{\rho}{m},$$

  whereas m describes the number of servers in a queueing system. Please note, that a system with $u = 1$ is called a fully loaded system. Many common models are based on steady state concepts, which are comparable to the physical concept of equilibrium. As a consequence, they are not applicable to systems in overload, i.e. $u > 1$. Due to statistical effects, they don't provide proper results in fully loaded systems as well. Thus $u < 1$ defines a necessary stability condition for commonly used models.

- The *departure rate* or *throughput* X describes the average number of customers leaving the system. In a stable and work preserving system, the departure rate is usually equal to the arrival rate. The throughput is determined from the state propabilites and the service rate,

  $$X = \sum_{n=1}^{\infty} \mu_n p_n \tag{1.6}$$

Please note, that a load dependent service rate has been assumed. In systems with multiple servers, μ_n is different for each state. Take as an example a call centre with 3 agents assuming each agent with the same average call handle time. With one agent being engaged, the effective service rate is μ. The other two agents are still waiting for a call and this can be identified with an idle server. If the second agent receives a call with the first agent still talking, the effective service rate becomes 2μ. When three or more agents are serving an active call, the effective service rate is 3μ. Clearly the fourth call in the system experiences a waiting time as he has to queue for service. Thus a load dependent service rate has to be assumed.

- The *average queueing time* W_q defines the time a customer has to wait, until service begins.

- The *average time in system* W defines the time between arrival and departure of a customer. The average time in system is related to the average waiting time as follows

$$W = W_q + s = W_q + \frac{1}{\mu} \tag{1.7}$$

- The *average queue size* L_q defines the average number of customers in the queue.

- The average system size L defines the average number of customers in the system and may be determined as follows

$$L = \sum_{n=1}^{\infty} n p_n \tag{1.8}$$

Please note, that starting the summation from $n = 1$ delivers the same result as starting from $n = 0$.

A very useful relation between average queueing time and the average number of customers in the system has been determined by J. D. C. Little in the year 1961. He found out, that given the average queueing time, the average queue size may be determined by simply multiplying the former with the arrival rate, i.e.

$$L_q = \lambda W_q \tag{1.9}$$

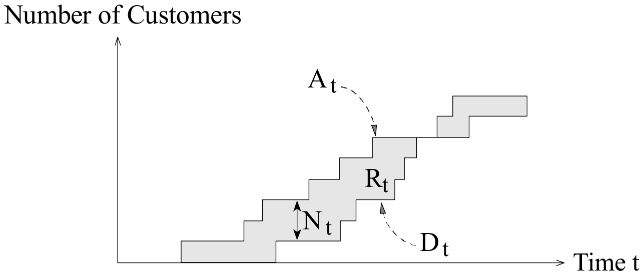


Figure 1.12: Little's Law

The same applies to the average system size and the average time in system

$$L = \lambda W \tag{1.10}$$

These relations are called *Little's Law.* Interestingly, Little's Law remains valid under very general assumptions. It does not assume any specific arrival distribution or service process, nor does it depend on the queueing discipline or the number of servers. With limited system capacity, Little's Law does still hold, but the arrival rate λ has to be redefined to exclude the number of customers lost due to blocking.

As shown in figure 1.12, Little's Law may also be derived graphically. By observing the number of customers entering and leaving a queueing system as functions of time in the interval $[0, t]$ denoted by A_t for the arrivals and D_t for the departures, the number of customers N_t in the system is given by

$$N_t = A_t - D_t$$

Defining arrival rate λ_t as

$$\lambda_t = \frac{A_t}{t}$$

Based on the area R_t between A_t and D_t, the average number of customers in the system L_t can be determined as follows

$$L_t = \frac{R_t}{t}$$

Please note, that R_t can be interpretated as the cumulated waiting time in interval $[0, t]$. The average waiting time W_t may now be calulated as the ratio between cumulated waiting time and the number of customers entering the system A_t, i.e.

$$W_t = \frac{R_t}{A_t}$$

Aggregating the last three formulas leads to

$$L_t = \frac{R_t}{t} = \frac{W_t A_t}{t} = W_t \lambda_t$$

Taking the limit as $t \longrightarrow \infty$ results in Little's well known formula. Please note, that Little's Law only applies to the average values, but not to the entire distribution. Many proofs have been presented in the literature since 1961, the original text *A Proof of the Queueing Formula* $L = \lambda W$ has been published in *Operations Research No. 9* by J. D. C. Little in the year 1961.

1.8 Little's Law for Distributions

It can be said, that Little's law is one of the most important rules in queueing theory. Easy to understand and simple to use it is commonly applied for theoretical and practical purposes. It turns out, that Little's law may be further extended to become a distributional statement. Assume a queueing system, where customers are served one at a time and leave the system (or the queue) in the order of arrival (FCFS). Furthermore, customers entering the system (or the queue) shall remain in the system (or the queue) until served, i.e. there is no form of customer impatience. Now define the random variables $\breve{L}_q$ (number of customers in the queue) and $\breve{W}_q$ (queueing time). Also note the relations to the performance indicators used in the classic version of Little's law

$$L_q = \mathbb{E}\breve{L}_q \qquad W_q = \mathbb{E}\breve{W}_q$$

Let the arrivals up to time t be described by a stationary renewal process $A(t)$. Then in steady state the average queue size $\breve{L}_q$ is distributed with $A(W)$, that is

$$\Pr\left\{\breve{L}_q \geq k\right\} = \Pr\left\{A(\breve{W}_q) \geq k\right\}$$

Given the time in system $\breve{W}$ and the number of customers in system $\breve{L}$, a similar relation also holds for single server systems. With multiple servers, overtaking customers would violate the assumption of customers leaving the system in order of their arrival. Our approach is based on [7] and [4], a thorough treatment based on deterministic processes appears in [19].

1.9 Notation

Due to the wide range of applications, statistical distributions, parameters and disciplines, the number of queueing system models steadily increases. As a consequence, D. G. Kendall developed a shorthand notation for queueing systems. According to that notation, a queueing system is described by the string $A/B/X/Y/Z$, where A indicates the arrival distribution, B the service pattern, X the number of servers, Y the system capacity and Z the queueing discipline. Standard symbols commonly used in queueing systems are presented in table 1.1.

For example, the shorthand $M/D/3/100/PRI$ describes a queueing system with exponential interarrival times, 3 servers each with deterministic service time, a system capacity of 100 places and a priority service discipline. Clearly the exponential interarrival times directly relate to Poisson arrivals. Also note, that a system capacity of 100 places in a system with 3 servers specify a maximum queue size of 97. Please note, that not all symbols are mandatory, as symbols Y and Z may be omitted thus resulting in an abbreviated string $A/B/X$. In that case, the system capacity is unlimited and the queueing discipline is *first come first served* per default. Thus a queueing system denoted by $M/M/1/\infty/FCFS$ is commonly abbreviated by $M/M/1$.

Kendall's notation has been extended in various ways. One such extension will be adopted to cover the description of impatient customers. Following Bacelli and Hebuterne [5], an impatience distribution I will be added to the standard string, both seperated by a plus sign. The distribution itself is defined similar to the first two elements A, B of the standard notation. The extended notation will then appear as $A/B/X/Y/Z + I$ in full length or as $A/B/X + I$ for the abbreviated notation.

The symbols mentioned in table 1.1 are not exclusive, as some characteristics can be seen as generalizations or specifications of other characteristics. So an exponential distribution may be seen as Erlang distribution with one phase, which in turn is a specialization of the phase type distribution. As a consequence, the capabilities of queueing models may be deducted from this short description. Formulas derived for $M/G/1$ are generalizations of the

Characteristics	**Symbol**	**Description**
A - Interarrival distribution	D	Deterministic
	C_k	Cox (k phases)
	E_k	Erlang (k phases)
	G	General
	GI	General independent
	GEO	Geometric (discrete)
	H_k	Hyperexponential
	M	Exponential (Markov)
	PH	Phase Type
B - Service time distribution	D	Deterministic
	C_k	Cox (k phases)
	E_k	Erlang (k phases)
	G	General
	GI	General independent
	GEO	Geometric (discrete)
	H_k	Hyperexponential
	M	Exponential (Markov)
	PH	Phase Type
X - Number of parallel servers	$1, 2, ..., \infty$	
Y - System capacity	$1, 2, ..., \infty$	
Z - Queueing discipline	$FCFS$	First come first serve
	RSS	Random selection for service
	PRI	Priority
	RR	Round Robin
	PS	Processor sharing
	GD	General

Table 1.1: Kendall notation for queueing systems

formulas used in $M/M/1$ systems.

As queueing theory originated from congestion theory in telephone systems, some application specific models survived over the years. The most common is the so called *lost calls cleared (LCC)* system, which can be expressed as $M/M/c/c$ model using Kendell notation. The LCC system does not have any waiting places, calls arriving to a system with all servers busy are *cleared*. As trunks in telephone systems usually do not have a queueing mechanism, the LCC model suits the need of calculating required trunk resources for a given offered load. The counterpart of the LCC system is the *lost calls held (LCH)* system, which relates to $M/M/c/K$ and $M/M/c$ models. Customers, which can not be immediately served on arrival are put in a queue. These models are commonly used to dimension the desired tone detector or tone generator resources in telephone systems given a certain waiting time objective.

Chapter 2

Simple Queueing Models

The most basic model in queueing theory is the $M/M/1$ model. In this section the mathematical derivation will be combined with intuitive insights to prepare the path for more complex models. In a $M/M/1$ model, random arrivals and exponentially distributed service times are assumed. Please note, that *random arrival*s is exactly defined to be Poisson in statistics. Furthermore, there is only a single server serving customers on a first come, first serve base. The population is infinite, so arriving customers are unaffected by the queue size. Parameters given for the $M/M/1$ model are λ, the average arrival rate, μ, the average service rate, which may be calculated from the average service time $\mu = \frac{1}{s}$. Bulk arrivals and group service are not allowed for this type of model.

By giving a closer look on the arrivals to a queueing system, it turns out, that time may be partitioned in slices with length $\triangle t$ such, that only one arrival per slice is allowed. By following a linear approach, it can be assumed, that the number of customers entering the system during interval $[t, t + \triangle t]$ are proportional to λ, i.e.

$$\begin{aligned} \Pr\{\text{single arrival in } [t, t + \triangle t]\} &= \lambda \triangle t \\ \Pr\{\text{no arrival in } [t, t + \triangle t]\} &= 1 - \lambda \triangle t \\ \Pr\{\text{more than one arrival in } [t, t + \triangle t]\} &= 0 \end{aligned}$$

Please note, that more arrivals during interval $[t, t + \triangle t]$ can be handled by introducing a function $o(\triangle t)$. Due to the fact, that this term will vanish in the subsequent derivation under very general assumptions, the presented results will still hold.

The same arguments may be applied to the service process as well, but it has to be taken into account, that departures can only occur, if the system is not empty, i.e.

$$\Pr\{\text{single departure in } [t, t + \triangle t] | \text{system not empty}\} = \mu \triangle t$$

By defining $p_n(t + \triangle t)$ as the propability of $n > 0$ customers residing in the system at time $t + \triangle t$, an expression may be found in terms of the propabilities at time t, i.e.

$$\begin{aligned} p_n(t + \triangle t) &= p_n(t) \Pr\{\text{no departure}\} \Pr\{\text{no arrival}\} \\ &\quad + p_{n-1}(t) \Pr\{1 \text{ arrival in } [t, t + \triangle t]\} \\ &\quad + p_{n+1}(t) \Pr\{1 \text{ departure in } [t, t + \triangle t]\} \end{aligned}$$

Due to the conditional nature of the departure process, a different equation has to be specified for state 0

$$\begin{aligned} p_0(t + \triangle t) &= p_0(t) \Pr\{\text{no arrival in } [t, t + \triangle t]\} \\ &\quad + p_1(t) \Pr\{1 \text{ departure in } [t, t + \triangle t]\} \end{aligned}$$

Expressed in formulas

$$\begin{array}{rlr} p_n(t + \triangle t) = & p_n(t)(1 - \lambda \triangle t)(1 - \mu \triangle t) & \\ & + p_{n-1}(t)\lambda \triangle t + p_{n+1}(t)\mu \triangle t & \text{for } n > 1 \\ p_0(t + \triangle t) = & p_0(t)(1 - \lambda \triangle t) + p_1(t)\mu \triangle t & \text{for } n = 0 \end{array}$$

Rearranging terms and taking the limit $\triangle t \longrightarrow 0$ gives

$$\begin{array}{rll} \frac{dp_n(t)}{dt} & = -(\lambda + \mu)p_n(t) + \lambda p_{n-1}(t) + \mu p_{n+1}(t) & \text{for } n > 1 \\ \frac{dp_0(t)}{dt} & = -\lambda p_0(t) + \mu p_1(t) & \text{for } n = 0 \end{array} \tag{2.1}$$

At this point it is guaranteed, that only one arrival per time slice $\triangle t$ can arrive. Please note, that the exception of bulk arrivals and group service has been excluded in the beginning of this section. Formula 2.1 may be used for a time dependent, so called *transient* analysis of the $M/M/1$ model. If someone is interested in the long term behaviour of the system, the so called *steady state equations* have to be determined. These steady state equations are related to the concepts of stochastic balance and furthermore physical

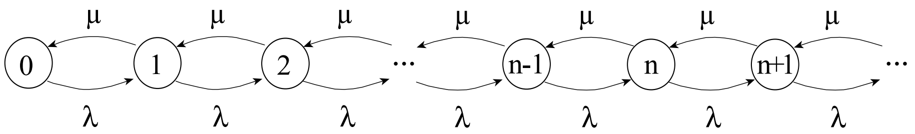


Figure 2.1: State transition diagram for the $M/M/1$ model

gravity, as will be shown below. Assuming a balanced state, changes over time are assumed to be negligible, i.e. $\frac{dp_n(t)}{dt} = 0$ leads to

$$\begin{array}{ll} 0 = -(\lambda+\mu)p_n + \lambda p_{n-1} + \mu p_{n+1} & \text{for } n > 1 \\ 0 = -\lambda p_0 + \mu p_1 & \text{for } n = 0 \end{array} \tag{2.2}$$

Queueing models are often analyzed by using so called *state transition diagrams*. Such a diagram is shown in figure 2.1 for the $M/M/1$ model. With a single look, flows into a state and flows out of a state may be determined. For the $M/M/1$ model it turns out, that the flow out of each state $n > 0$ is always $\lambda+\mu$, whereas the the flow into a state n are λ from the previous state $n-1$ and μ from the next state $n+1$. By following the intuitive concept of balance, i.e. by equating the rates into a state with the rates out of a state, the following equations may be written down immediately

$$\begin{array}{ll} (\lambda+\mu)p_n = \lambda p_{n-1} + \mu p_{n+1} & \text{for } n > 1 \\ \lambda p_0 = \mu p_1 & \text{for } n = 0 \end{array}$$

Comparing these equations with formula 2.2 above, they turn out to be the same. Thus analysis of a wide range of queueing models may be carried out based on concepts of balance. As a next step, the steady state propabilities of the $M/M/1$ model are to be derived. Rearranging equation 2.2 results in

$$\begin{array}{ll} p_{n+1} = \frac{\lambda+\mu}{\mu}p_n - \frac{\lambda}{\mu}p_{n-1} & \text{for } n > 1 \\ p_1 = \frac{\lambda}{\mu}p_0 & \text{for } n = 0 \end{array}$$

By using the definition of the traffic intensity $\rho = \frac{\lambda}{\mu}$, and continuously substituting, i.e.

$$\begin{aligned}
p_1 &= \rho p_0 \\
p_2 &= \rho p_1 = \rho^2 p_0 \\
p_3 &= \rho p_2 = \rho^3 p_0 \\
&\dots \\
p_n &= \rho p_{n-1} = \rho^n p_0
\end{aligned}$$

the state propabilities may be easily determined. As a last step, the propability of state 0 p_0 can be derived by recalling the fact, that propabilities always sum to 1, i.e.

$$1 = \sum_{n=1}^{\infty} p_n = \sum_{n=1}^{\infty} \rho^n p_0$$

Rewriting and utilizing the definition of a geometric series leads to

$$p_0 = \frac{1}{\sum_{n=1}^{\infty} \rho^n} = 1 - \rho$$

and finally to

$$p_n = \rho^n (1 - \rho) \tag{2.3}$$

Please note, that the arrival rate is not allowed to exceed the service rate due to stability of the system. This may be expressed by the stability conditions $\rho < 1$ or $\lambda < \mu$. In order to determine the performance characteristics of the $M/M/1$ system, one statistic has to be derived from the steady state propabilities. Most of the other statistics may be concluded from the relations presented in the previous section. Recalling the fact, that the average system size L is defined by

$$L = \sum_{n=0}^{\infty} n p_n$$

and substituting for p_n leads to

$$\begin{aligned} L &= (1-\rho)\sum_{n=0}^{\infty} n\rho^n \\ &= (1-\rho)\rho\sum_{n=1}^{\infty} n\rho^{n-1} \\ &= (1-\rho)\rho\frac{d(\sum_{n=0}^{\infty}\rho^n)}{d\rho} \\ &= (1-\rho)\rho\frac{d(\frac{1}{1-\rho})}{d\rho} \\ &= \frac{(1-\rho)\rho}{(1-\rho)^2} = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu-\lambda} \end{aligned} \tag{2.4}$$

Please note that $\sum_{n=1}^{\infty} np_n = \sum_{n=0}^{\infty} np_n$ and that $\sum_{n=1}^{\infty} n\rho^{n-1}$ is the first order derivative of $\sum_{n=0}^{\infty}\rho^n$. By using Little's Law, the average time in system may be calculated as follows

$$W = \frac{1}{\lambda}L = \frac{1}{\mu(1-\rho)} = \frac{1}{\mu-\lambda}$$

As a next step the average queueing time may be determined by simply subtracting the service time, i.e.

$$W_q = W - s = W - \frac{1}{\mu} = \frac{\rho}{\mu(1-\rho)} = \frac{\lambda}{\mu(\mu-\lambda)}$$

Applying Little's Law another time finally leads to the average queue size

$$L_q = \lambda W_q = \frac{\lambda^2}{\mu(\mu-\lambda)} = \frac{\rho^2}{1-\rho} \tag{2.5}$$

Based on the equations above, another useful relation may be obtained between the system size L and the queue size L_q

$$L = L_q + \rho \tag{2.6}$$

Due to the fact, that Little's Law holds under very general distribution assumptions, equation 2.6 is valid for almost any queueing system with a single server. In case of system capacity limitations, the law still holds, but only with slight modification of the arrival rate.

Example 7 *Consider a database system with an average service time of 450 msec. As database requests are inititated by a large number of clients, a random arrival pattern may be assumed. Thus the arrival process is assumed to be Poisson. On the average a new database query arrives every 500 msec. Service times are assumed to be exponentially distributed, the queueing discipline is assumed to follow a first come, first serve pattern. The numbers given for interarrival and service times lead to*

$$\begin{aligned}\lambda &= \frac{1}{0.5}\,\mathrm{sec}\\ \mu &= \frac{1}{0.45}\,\mathrm{sec}\\ u &= \rho = \frac{\lambda}{\mu} = 0.9\end{aligned}$$

With 90% utilization the database server can be deemed heavily loaded. This is also reflected in the average system size and the average answer times, i.e.

$$\begin{aligned}L &= \frac{\rho}{1-\rho} = \frac{0.9}{0.1} = 9\\ W &= \frac{1}{\lambda}L = \frac{1}{2}9 = 4.5\,\mathrm{sec}\end{aligned}$$

Due to the high frustration level, it was decided to replace the harddisc by a faster model. As a consequence service times were reduced to 350 msec. As a result, performance characteristics improved, i.e.

$$\begin{aligned}\lambda &= \frac{1}{0.5}\,\mathrm{sec}\\ \mu &= \frac{1}{0.35}\,\mathrm{sec}\\ u &= \rho = \frac{\lambda}{\mu} = 0.7\\ L &= \frac{\rho}{1-\rho} = \frac{0.7}{0.3} = 2.333\\ W &= \frac{1}{\lambda}L = \frac{1}{2}2.333 = 1.167\,\mathrm{sec}\end{aligned}$$

As can be seen from this example, heavy loaded systems are very sensible to small changes. In this case, the answer times have been reduced to a quarter of the original answer times only by a compareable small increase in service speed.

2.1 Idle and Busy Period

The work process of a queueing system may be splitted into idle and *busy periods*. The latter starts with an arrival at an empty system and ends, when the system becomes idle again. A *busy cycle* is the time between two successive arrivals at an empty system. It is also the sum of a busy and an adjacent idle period [27]. Let the random variable T_{busy} denote the busy time. As shown in [46], the density for a stable $M/M/1$ queue with one customer present in the system at time 0 is given by

$$f_{idle}(t) = \sqrt{\frac{\mu}{\lambda}} \frac{I_1\left(2\sqrt{\lambda\mu}t\right)}{t} e^{-(\lambda+\mu)t}, \quad t > 0 \tag{2.7}$$

where I_1 is the modified Bessel function of order 1,

$$I_1(x) = \sum_{k=0}^{\infty} \frac{(x/2)^{2k+1}}{k!(k+1)!}$$

Please note, that this argument is transient, i.e. depends on time t. Based on the density 2.7 or other means [27], one may derive mean and variance of T_{busy}:

$$\begin{aligned} \mathbb{E}T_{busy} &= \frac{1}{\mu - \lambda} \\ Var(T_{busy}) &= \frac{\mu + \lambda}{(\mu - \lambda)^3} \end{aligned}$$

Because of the memoryless property of the arrival process, the idle period T_{idle} is exponentially distributed with mean $\frac{1}{\lambda}$. The distribution of the busy cycle is simply the convolution of the idle time and busy period distribution. The analysis of the busy period is a standard concept in queueing theory, a classic reference is [27].

2.2 Capacity Constraints

Consequently, the next step is to extend the $M/M/1$ model to include a system capacity constraint thus becoming a $M/M/1/K$ model. A similar condition as introduced for state 0 before has to be applied to state K as well. This is also reflected in the state transition diagram in figure 2.2.Transition

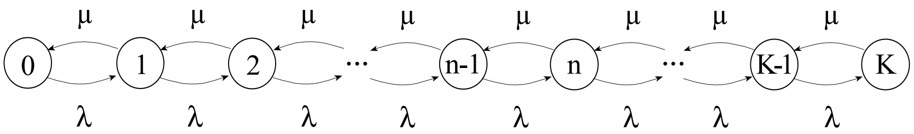


Figure 2.2: State transition diagram for the $M/M/1/K$ model

rates and state propabilities for states $1 \dots K$ remain the same as given by the $M/M/1$ model, i.e.

$$p_n = \rho^n p_0$$

Only p_0 has to be derived with respect to the limited system capacity. Following the same approach as before leads to

$$\begin{aligned}
1 &= \sum_{n=0}^{K} p_n = \sum_{n=0}^{K} \rho^n p_0 \\
p_0 &= \frac{1}{\sum_{n=0}^{K} \rho^n} = \begin{cases} \frac{1-\rho}{1-\rho^{K+1}} & \text{for } \rho \neq 1 \\ \frac{1}{K+1} & \text{for } \rho = 1 \end{cases} \\
p_n &= \begin{cases} \frac{(1-\rho)\rho^n}{1-\rho^{K+1}} & \text{for } \rho \neq 1 \\ \frac{1}{K+1} & \text{for } \rho = 1 \end{cases}
\end{aligned}$$

Due to the capacity constraint, the stability condition is no longer required. Customers arriving to a system with busy servers and full queue are lost. In order to determine the performance characteristics of the $M/M/1/K$ model, the same procedure is applied as has been done before, i.e.

$$\begin{aligned}
L &= \sum_{n=0}^{K} n p_n = p_0 \rho \sum_{n=1}^{K} n \rho^{n-1} \\
&= p_0 \rho \frac{d(\sum_{n=0}^{K} \rho^{n})}{d\rho} = p_0 \rho \frac{d}{d\rho} (\frac{1-\rho^{K+1}}{1-\rho}) \\
&= p_0 \rho \frac{1-(K+1)\rho^K + K\rho^{K+1}}{(1-\rho)^2} \\
&= \begin{cases} \frac{\rho(1-(K+1)\rho^K + K\rho^{K+1})}{(1-\rho^{K+1})(1-\rho)} & \text{for } \rho \neq 1 \\ \frac{\sum_{n=0}^{K} n}{K+1} = \frac{K}{2} & \text{for } \rho = 1 \end{cases}
\end{aligned}$$

Little's Law is still applicable, but based on the fact, that no customers are lost. Therefore, an *effective arrival rat*e has to be calculated. The effective arrival rate is dependent on the fraction of calls, which suceeded in entering the system. Giving a closer look on the state transition diagram in figure 2.2, it turns out, that up to state $K-1$, an arriving customer always can enter the system. But in state K, the call is *blocked* and as a consequence, lost. Thus $\sum_{n=0}^{K-1} p_n$ describes the propability, that customers can enter the system and p_K refers to the so called *blocking propability.* Considering the fact, that both propabilities sum to 1 leads to the following calculation for the effective arrival rate $\bar{\lambda}$

$$\bar{\lambda} = \lambda(1 - p_K)$$

By proceeding as before, all important performance characteristics may be obtained

$$\begin{aligned} W &= \frac{1}{\bar{\lambda}} L \\ W_q &= W - \frac{1}{\mu} \\ L_q &= \bar{\lambda} W_q \end{aligned}$$

Consequently, the relation between average queue size and average system size has to be reviewed. Equation 2.6 has to be modified to include the effective arrival rate $\bar{\lambda}$ instead of the arrival rate λ.

2.3 Queueing Disciplines

So far only first come, first serve has been assumed as a queueing discipline. As all performance characteristics mentioned are averages in a statistical sense, they are very insensible to changes of the queueing discipline. All results remain valid for *last come, first serve (LCFS)*, *round robin (RR)* and *processor sharing (PS)* disciplines. In priority systems, customers are grouped in classes and seperate characteristics are derived for each class. Although the average values remain the same, the underlying distributions change with the queueing discipline. This is a common feature in queueing theory. One often assumes a system or queueing discipline to be *work conserving*, that is [57]

- the server does not remain idle with customers waiting

- the queueing discipline does not affect the arrival time of any message
- the queueing discipline does not affect the amount of service time

If one works within the class of work conserving queueing disciplines, the performance key indicators such as average queueing time and average system size will remain untouched by the choice of a dedicated member. Note, that such an invariance property does not hold for the corresponding distributions. A very good discussion on the topics related to different queueing disciplines is found in [13].

Chapter 3

Birth-Death Process

Both, the $M/M/1$ and the $M/M/1/K$ model considered so far, are special cases of a more general system, which can be modelled by the so called *birth-death process*. The name is related with an application in biology, the birth-death process provides a simple frame to model populations of any sort. In view of technical systems, the birth-death process may be used to model *load dependent systems*. In such a system, arrival and service rates are dependent on the current state of the system. This is also reflected in the state transition diagram in figure 3.1. A wide range of queueing systems can be modelled by customizing the parameters of the birth-death process. Please note, that still exponential interarrival and service time distributions are assumed.

Although more complex, the balance approach still provides easy access

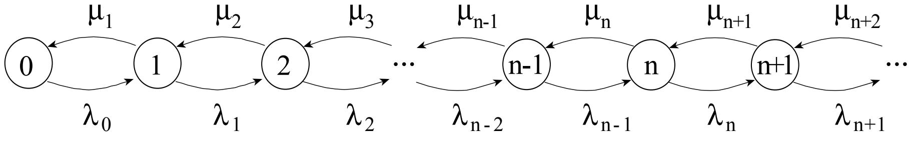


Figure 3.1: State transition diagram for a load dependent system

to the solution

$$\begin{aligned}
p_1 &= \frac{\lambda_0}{\mu_1} p_0 \\
p_2 &= \frac{\lambda_1}{\mu_2} p_1 = \frac{\lambda_1 \lambda_0}{\mu_2 \mu_1} p_0 \\
&\cdots \\
p_n &= \frac{\lambda_{n-1}}{\mu_n} p_{n-1} = \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} p_0 \qquad (3.1)
\end{aligned}$$

As all propabilities sum to 1, p_0 may be readily obtained as follows

$$\begin{aligned}
1 &= \sum_{n=0}^{\infty} p_n = p_0 + \sum_{n=1}^{\infty} p_n = p_0 + \sum_{n=1}^{\infty} \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} p_0 \\
p_0 &= \frac{1}{1 + \sum_{n=1}^{\infty} \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i}} \qquad (3.2)
\end{aligned}$$

The average number of customers in the system L may be determined as usual

$$L = \sum_{n=1}^{\infty} n p_n \qquad (3.3)$$

By defining the system throughput X as follows

$$X = \sum_{n=1}^{\infty} \mu_n p_n \qquad (3.4)$$

and using Little's Law leads to the average time in system W

$$W = \frac{1}{X} L = \frac{\sum_{n=1}^{\infty} n p_n}{\sum_{n=1}^{\infty} \mu_n p_n} \qquad (3.5)$$

The formulas of the birth-death process may applied to systems with limited capacity as well, but with a slight modification. The upper summation limit of each equation has to be replaced by the system capacity K. Before proceeding with more complex variants, the models considered so far are reviewed in terms of the birth-death process.

Example 8 *The following parametrization of the birth death process leads to the same results as the $M/M/1$ model mentioned above*

$$\begin{aligned}
\lambda_n &= \lambda \quad \textit{for all } n \\
\mu_n &= \mu \quad \textit{for all } n
\end{aligned}$$

There is also a parametrization available for the $M/M/1/K$ model

$$\begin{aligned}\lambda_n &= \begin{cases}\lambda & \textit{for } n = 0 \dots K-1 \\ 0 & \textit{for } n > K-1\end{cases} \\ \mu_n &= \begin{cases}\mu & \textit{for } n = 1 \dots K \\ 0 & \textit{for } n > K\end{cases}\end{aligned}$$

Please note, by introducing conditions to the parameters, interchanging the upper summation limit of the original formulae becomes obsolete.

3.1 Multiserver Systems

Queueing systems with multiple servers may be modelled by a single server system with state dependent service rate. Given n customers are in the system, work is processed n times as fast as a single server would need to do so. Given a limited supply of servers, the load dependent service rate remains the same, if the limit is exceeded. The related model is called $M/M/c$ in the limited case and $M/M/\infty$ in the umlimited case. The latter system is also called *delay server*, as the average answer time is insensible to the number of customers currently in the system. As a single system, the delay server is almost useless, but if combined with other systems to a *queueing network*, it plays an important role. The $M/M/c$ requires the following parametrization

$$\begin{array}{ll}\lambda_n = \lambda & \text{for all } n \\ \mu_n = n\mu & \text{for } 1 \leqq n \leqq c \\ \mu_n = c\mu & \text{for } n > c\end{array}$$

Substituting these parameters in equations 3.1 and 3.2 leads to

$$\begin{aligned}p_n &= \begin{cases}\prod_{i=0}^{n} \frac{\lambda}{i\mu} p_0 = \frac{1}{n!}\rho^n p_0 & \text{for } 1 \leqq n \leqq c \\ p_c \prod_{i=0}^{n} \frac{\lambda}{c\mu} = \frac{1}{c!c^{n-c}}\rho^n p_0 & \text{for } n > c\end{cases} \\ p_0 &= \left(1 + \sum_{n=1}^{c-1} \frac{1}{n!}\rho^n + \sum_{n=c}^{\infty} \frac{1}{c!c^{n-c}}\rho^n\right)^{-1} \\ &= \left(\sum_{n=0}^{c-1} \frac{1}{n!}\rho^n + \sum_{n=c}^{\infty} \frac{1}{c!c^{n-c}}\rho^n\right)^{-1} \\ &= \left(\sum_{n=0}^{c-1} \frac{1}{n!}\rho^n + \frac{1}{c!}\rho^c \frac{1}{1-\frac{\rho}{c}}\right)^{-1}\end{aligned}$$

As no system capacity constraint has been defined, a stability condition is required to preserve proper analytical results. Based on the intuitive argument, that not more customers should arrive than can be served, a stability condition may be written down immediately

$$\frac{\lambda}{c\mu} = \frac{\rho}{c} = u < 1$$

Please note, that the propability mass function p_n consists of two seperate functions. It turns out, that it is easier to determine the average queue size L_q first instead of the average system size L, as only one function is required. A very useful parameter in the derivation of L_q is the propability of delay p_d, which may be obtained as follows

$$\begin{aligned} p_d &= \sum_{n=c}^{\infty} p_n = \sum_{n=c}^{\infty} \frac{p_0 \rho^n}{c! c^{n-c}} \\ &= \frac{p_0 \rho^c}{c!} \sum_{n=c}^{\infty} \left(\frac{\rho}{c}\right)^{n-c} \\ &= \frac{p_0 \rho^c}{c!(1-\frac{\rho}{c})} \end{aligned} \tag{3.6}$$

Expression 3.6 is often referred to as *Erlang C formula* or *Erlang formula of the second kind.* The Erlang C formula has been derived for lost calls held systems (LCH) long before the $M/M/c$ model was developed. Most performance characteristics of interest may be expressed in terms of this expression. Consequently it will be used to derive the average queue size L_q

$$\begin{aligned} L_q &= \sum_{n=c}^{\infty} (n-c) p_n = \sum_{n=0}^{\infty} n p_{n+c} \\ &= \sum_{n=0}^{\infty} n \frac{p_0 \rho^{n+c}}{c! c^n} = \frac{p_0 \rho^c}{c!} \sum_{n=0}^{\infty} n \left(\frac{\rho}{c}\right)^n \\ &= \frac{p_0 \rho^c}{c!} \frac{\frac{\rho}{c}}{(1-\frac{\rho}{c})^2} = \frac{\frac{\rho}{c}}{1-\frac{\rho}{c}} p_d \\ &= \frac{\lambda}{c\mu - \lambda} p_d \end{aligned} \tag{3.7}$$

Proceeding as before leads to the other measures of effectiveness

$$
\begin{aligned}
W_q &= \frac{L_q}{\lambda} = \frac{1}{c\mu - \lambda} p_d \\
W &= W_q + \frac{1}{\mu} = \frac{1}{c\mu - \lambda} p_d + \frac{1}{\mu} \\
L &= \lambda W = \frac{\lambda}{c\mu - \lambda} p_d + \rho \qquad (3.8)
\end{aligned}
$$

Example 9 *Consider a supermarket with eight cashiers open and a common queue. Customers arrive approximately every 1.25 sec and take on average 5 minutes to be served. Assuming exponential arrival and service processes following a first come, first serve discipline, the model input is given as follows*

$$
\begin{aligned}
\lambda &= \frac{1}{1.25} = 0.8 \\
\mu &= \frac{1}{5} = 0.2 \\
c &= 8
\end{aligned}
$$

From these parameters the traffic intensity ρ and the server utilization u may be calculated

$$
\begin{aligned}
\rho &= 4 \\
u &= 0.5
\end{aligned}
$$

thus satisfying the stability condition $u < 1$. Substitution into the formulas for the $M/M/c$ model yield

$$
\begin{aligned}
p_d &= 0.059 \\
L &= 4.059 \\
L_q &= 0.059 \\
W &= 5.0738 \\
W_q &= 0.0738
\end{aligned}
$$

Thus the propability, that no customer has to wait is given by $1 - p_d = 0.941$. Also note, that the queues are very small in this 50% loaded system, less than one customer has to wait on the average. If three cashiers become ill, a rest *of five employees has to serve the entire customers. By setting* $c = 5$ *and reapplying the formulas, the following performance measures are obtained*

$$\begin{aligned}\rho &= 4\\ u &= 0.8\\ p_d &= 0.5541\\ L &= 6.2165\\ L_q &= 2.2165\\ W &= 7.7706\\ W_q &= 2.7706\end{aligned}$$

Now on the average more than two customers have to wait. The propability, that customers are served without having to wait has decreased to $1 - p_d = 0.4459$*. Although this might still be acceptable for a supermarket, this example shows, how sensible queueing systems might become with increased load. If an additonal cashier becomes unavailable or the average service time increases beyond 6.25 minutes, the supermarket can not serve their customers anymore. This can be seen from the stability condition.*

3.2 Capacity Constraints in Multiserver Systems

As already mentioned above, by customizing the parameters for the load dependent model, capacity constraints may be introduced to a multiserver system $M/M/c/K$, i.e.

$$\begin{array}{ll}\lambda_n = \lambda & \text{for } 0 \leqq n < K\\ \lambda_n = 0 & \text{for } n \geqq K\\ \mu_n = n\mu & \text{for } 1 \leqq n < c\\ \mu_n = c\mu & \text{for } c \leqq n \leqq K\\ \mu_n = 0 & \text{for } n > K\end{array}$$

Having identified the capacity limitations as the only difference between the limited and the unlimited model, the same propabilities for states $1 \ldots K$ can be assumed

$$p_n = \begin{cases} \prod_{i=0}^{n} \frac{\lambda}{i\mu} p_0 = \frac{1}{n!}\rho^n p_0 & \text{for } 1 \leqq n \leqq c\\ p_c \prod_{i=0}^{n} \frac{\lambda}{c\mu} = \frac{1}{c!c^{n-c}}\rho^n p_0 & \text{for } c < n \leqq K\\ 0 & \text{for } n > K\end{cases} \tag{3.9}$$

Proceeding as before for the $M/M/1/K$ model, the propability for state 0 remains to be determined

$$\begin{aligned}
1 &= \sum_{n=0}^{K} p_n = p_0 \sum_{n=0}^{c} \frac{1}{n!}\rho^n + p_0 \sum_{n=c+1}^{K} \frac{1}{c!c^{n-c}}\rho^n \\
p_0 &= \left(1 + \sum_{n=1}^{c} \frac{1}{n!}\rho^n + \sum_{n=c+1}^{K} \frac{1}{c!c^{n-c}}\rho^n\right)^{-1} \\
&= \left(\sum_{n=0}^{c-1} \frac{1}{n!}\rho^n + \sum_{n=c}^{K} \frac{1}{c!c^{n-c}}\rho^n\right)^{-1} \\
&= \begin{cases} \left(\sum_{n=0}^{c-1} \frac{1}{n!}\rho^n + \frac{\rho^c}{c!}\frac{1-(\frac{\rho}{c})^{K-c+1}}{1-\frac{\rho}{c}}\right)^{-1} & \text{for } \frac{\rho}{c} \neq 1 \\ \left(\sum_{n=0}^{c-1} \frac{1}{n!}c^n + \frac{c^c}{c!}(K-c+1)\right)^{-1} & \text{for } \frac{\rho}{c} = 1 \end{cases}
\end{aligned}$$

Following the same procedure as for the $M/M/c$ model, the propability of delay p_d is now derived

$$\begin{aligned}
p_d &= \sum_{n=c}^{K-1} p_n = \sum_{n=c}^{K-1} \frac{p_0\rho^n}{c!c^{n-c}} \\
&= \begin{cases} p_0\frac{\rho^c}{c!}\frac{1-(\frac{\rho}{c})^{K-c}}{1-\frac{\rho}{c}} & \text{for } \frac{\rho}{c} \neq 1 \\ p_0\frac{c^c}{c!}(K-c) & \text{for } \frac{\rho}{c} = 1 \end{cases}
\end{aligned}$$

Please note, that no delay can occur, if no waiting room exists, i.e. $K = c$. Then only blocking may occur, whereas the propability of blocking is given by p_K for all values of K. The $M/M/c/K$ model without waiting room also denoted by $M/M/c/c$ directly relates to the lost calls cleared (LCC) system. The blocking propability $p_b = p_c = p_K$ for the $M/M/c/c$ model is often referred to as *Erlang B formula, Erlang loss formula* or *Erlang formula of the first kind.* Substituting p_0 into equation 3.9 and simplifying yields

$$p_b = p_c = \frac{\frac{\rho^c}{c!}}{\sum_{n=0}^{c} \frac{\rho^n}{n!}} \tag{3.10}$$

The most appealing property of the Erlang loss formula lies in the fact, that its validity is not limited to exponential service times. It can be shown, that the Erlang loss formula still holds under very general conditions, i.e. for the

$M/G/c/c$ model. Thus any service time distribution dependency has been reduced to the mean service time only. As the proof is very extensive, it will be omitted here. A proof $M/M/1/1 = M/G/1/1$ for a single server model is presented in [27]. Turning attention back to the more general $M/M/c/K$ model, it remains to determine the performance characteristics. As before L_q provides the most convenient way to receive results

$$
\begin{aligned}
L_q &= \sum_{n=c}^{K}(n-c)p_n = \sum_{n=0}^{K-c} n p_{n+c} \\
&= \sum_{n=0}^{K-c} n \frac{p_0 \rho^{n+c}}{c! c^n} = \frac{p_0 \rho^c}{c!} \sum_{n=0}^{K-c} n \left(\frac{\rho}{c}\right)^n \\
&= \frac{p_0 \rho^c}{c!} \frac{\rho}{c} \sum_{n=1}^{K-c} n \left(\frac{\rho}{c}\right)^{n-1} \\
&= \begin{cases} \frac{p_0\rho^c}{c!}\frac{\rho}{c}\frac{d}{d\rho}\left(\frac{1-(\frac{\rho}{c})^{K-c+1}}{1-\frac{\rho}{c}}\right) & \text{for } \frac{\rho}{c} \neq 1 \\ \frac{p_0 c^c}{c!}\frac{(K-c)(K-c+1)}{2} & \text{for } \frac{\rho}{c} = 1 \end{cases} \\
&= \begin{cases} \frac{p_0\rho^{c+1}}{c!c(1-u)^2}(1-u^{K-c+1}-(1-u)(K-c+1)u^{K-c}) & \text{for } \frac{\rho}{c} \neq 1 \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{with } u = \frac{\rho}{c} & \\ \frac{p_0 c^c}{c!}\frac{(K-c)(K-c+1)}{2} & \text{for } \frac{\rho}{c} = 1 \end{cases}
\end{aligned}
$$

The other measures of effectiveness may be obtained by using Little's Law. Due to the system limitation, the arrival rate has to be modified to exclude lost customers. For telephony applications, one would say the *calls carried* have to be used instead of the *calls offered.* This may be expressed as follows

$$
\begin{aligned}
W_q &= \frac{1}{\lambda(1-p_K)} L_q \\
W &= W_q + \frac{1}{\mu} \\
L &= \lambda(1-p_K)W
\end{aligned}
$$

3.3 Erlang B revisited

Instead of directly deriving the result for the Erlang B formula 3.10, the following convenient recursion formula may be applied

$$p_c = \frac{\rho p_{c-1}}{c + \rho p_{c-1}}, \quad p_0 = 1$$

Substituting $\varepsilon_c = \frac{1}{p_c}$ provides an equivalent recurrence formula

$$\varepsilon_c = 1 + \frac{c}{\rho}\varepsilon_{c-1}, \quad \varepsilon_0 = 1 \tag{3.11}$$

Due to the waiting room limitation of the $M/M/c/c$ queue, a steady state distribution exists also for the case of *heavy traffic*, that is $u = \frac{\rho}{c} > 1$. Assuming an arrival rate of $c\lambda$, the utilization u exceeds 1. In fact, one can show [46], that the number of empty places converges weakly to a geometric distribution with parameter $\frac{1}{\rho}$. As a consequence the following relation for the blocking probability p_c holds

$$\lim_{c\to\infty} p_c = 1 - \frac{1}{\rho} = 1 - \frac{\mu}{\lambda} \tag{3.12}$$

With c sufficiently large, formula 3.12 provides a reasonable approximation in heavy traffic situations.

3.4 Customer Impatience

Modeling customer frustration may be achieved in different ways. In a *balking* scenario, customers are refusing to enter the queue given that it has reached a certain length. At its most extreme, such a system is described by a $M/M/c/K$ model. Alternatively, customer discouragement may be modeled by a monotonic decreasing function b_n. By carefully selecting a proper function b_n, one is able to express customer expectations in a nice and accurate way [27].

With respect to the birth-death model introduced in equations 3.1 and 3.2, balking affects the arrival rate, i.e.

$$\lambda_n = b_n \lambda \tag{3.13}$$

Please note, that the system arrival rate has been assumed to be constant λ.

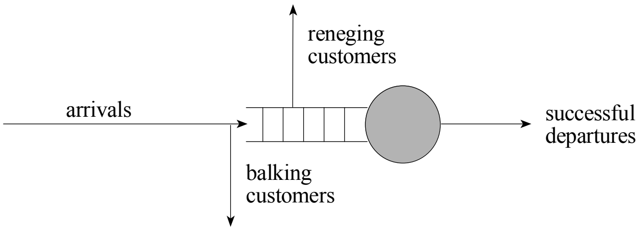


Figure 3.2: A queue with impatient customers

Another form of customer impatience is *reneging*. Other than the balking customer, a reneging customer joins the queue waiting for service. If the perceived waiting time exceeds customer expectations, the customer leaves the queue. Proceeding similar as above, a reneging function is introduced

$$r(n) = \lim_{\Delta t \longrightarrow 0} \Pr \left\{ \begin{array}{c} \text{customer reneges during } \Delta t \\ \text{given } n \text{ customers in the system} \end{array} \right\}$$

The reneging function clearly affects the service rate, as reneging customers may be seen as virtually serviced customers in addition to regularily service customers. Mathematically expressed

$$\bar{\mu}_n = \mu_n + r(n) \tag{3.14}$$

Both types of impatience may be combined in a single model as shown in figure 3.2. Application of the expressions 3.13 and 3.14 to the general birth-death equations 3.1 and 3.2 yields [27]

$$\begin{aligned} p_n &= \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} p_0 = \lambda^n p_0 \prod_{i=1}^{n} \frac{b_{i-1}}{\mu_n + r(i)} \\ p_0 &= \left(1 + \sum_{n=1}^{\infty} \lambda^n \prod_{i=1}^{n} \frac{b_{i-1}}{\mu_n + r(i)} \right)^{-1} \end{aligned}$$

For practical purposes, very often a more specific set of parameters is defined. Assuming c servers with constant service rate μ, i.e.

$$\mu_n = \left\{ \begin{array}{ll} n\mu & \text{for } 1 \leqq n \leqq c \\ c\mu & \text{for } c < n \leqq K \end{array} \right.$$

a system capacity of $K > c$, a constant balking rate in queueing situations, i.e.

$$b_n = \begin{cases} 1 & \text{for } n \leqq c \\ (1-\beta) & \text{for } c < n \leqq K \\ 0 & \text{for } n > K \end{cases} \tag{3.15}$$

and that customers don't have any knowledge about the system state [63], i.e.

$$r(n) = \begin{cases} 0 & \text{for } n \leqq c \\ (n-c)\delta & \text{for } c < n \leqq K \end{cases}$$

the model becomes

$$\begin{aligned} p_n &= \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} p_0 = \begin{cases} \frac{1}{n!}\rho^n p_0 & \text{for } 1 \leqq n \leqq c \\ \frac{\rho^c \lambda^{n-c}(1-\beta)^{n-c}}{c! \prod_{i=c+1}^{n} c\mu + (i-c)\delta} p_0 & \text{for } c < n \leqq K \\ 0 & \text{for } n > K \end{cases} \\ p_0 &= \left(1 + \sum_{n=1}^{c} \frac{1}{n!}\rho^n + \sum_{n=c+1}^{K} \frac{\rho^c \lambda^{n-c}(1-\beta)^{n-c}}{c! \prod_{i=c+1}^{n} c\mu + (i-c)\delta}\right)^{-1} \\ &= \left(\sum_{n=0}^{c} \frac{1}{n!}\rho^n + \sum_{n=c+1}^{K} \frac{\rho^c \lambda^{n-c}(1-\beta)^{n-c}}{c! \prod_{i=c+1}^{n} c\mu + (i-c)\delta}\right)^{-1} \end{aligned} \tag{3.16}$$

with ρ set to $\rho = \frac{\lambda}{\mu}$ as before. This limited capacity system covers balking as well as reneging behaviour. It is best solved by using numerical computation, as no closed form solution is known to the author. The performance characteristics W, L and X are calculated by substituting state probabilities 3.16 in formulas 3.3 to 3.5 for the birth-death model.

By omitting the balking behaviour one arrives at the $M/M/c+M$ model first introduced by C. Palm before 1960 [43]. It has also been given names such as *Erlang A* or *Palm/Erlang A*, because it provides a tradeoff between the Erlang C ($M/M/c$) queueing model and the Erlang B loss ($M/M/c/c$) system. Our treatment will be based on [40]. First note, that by eliminating the balking definition 3.15, the system becomes infinite. As a consequence the second sum in

$$p_0 = \left(\sum_{n=0}^{c} \frac{1}{n!}\rho^n + \sum_{n=c+1}^{\infty} \frac{\rho^c \lambda^{n-c}}{c! \prod_{i=c+1}^{n} c\mu + (i-c)\delta}\right)^{-1} \tag{3.17}$$

must converge to allow for meaningful results of p_n. In fact, convergence can be assured by deriving an upper bound:

$$\begin{aligned} p_0^{-1} &= \sum_{n=0}^{c} \frac{1}{n!}\rho^n + \sum_{n=c+1}^{\infty} \frac{\rho^c \lambda^{n-c}}{c! \prod_{i=c+1}^{n} c\mu + (i-c)\delta} \\ &= \sum_{n=0}^{c} \frac{1}{n!}\frac{\lambda^n}{\mu^n} + \frac{\lambda^c}{\mu^c c!} \sum_{n=c+1}^{\infty} \prod_{i=c+1}^{n} \frac{\lambda}{(c\mu + (i-c)\delta)} \\ &\leq \sum_{n=0}^{\infty} \frac{(\lambda / \min(\mu,\delta))^n}{n!} = e^{-\frac{\lambda}{\min(\mu,\delta)}} \end{aligned}$$

Hence the $M/M/c+M$ queue always remains stable. Once p_0 is known the entire steady state distribution may be derived from

$$p_n = \begin{cases} \frac{1}{n!}\rho^n p_0 & \text{for } 1 \leqq n \leqq c \\ \frac{\rho^c \lambda^{n-c}}{c! \prod_{i=c+1}^{n} c\mu + (i-c)\delta} p_0 & \text{for } n > c \end{cases} \tag{3.18}$$

To avoid numerical difficulties caused by the infinite sum in equation 3.17, Palm presented an ingenious derivation based on the Erlang loss formula and the incomplete gamma function. Rewriting expression 3.17 leads to

$$\begin{aligned} p_0^{-1} &= \sum_{n=0}^{c} \frac{\rho^n}{n!} + \frac{\rho^c}{c!} \sum_{n=c+1}^{\infty} \prod_{i=c+1}^{n} \frac{\lambda}{(c\mu + (i-c)\delta)} \\ &= \frac{\rho^c}{c!}\left[\frac{1}{p_b} + \sum_{n=1}^{\infty} \frac{(\lambda/\delta)^n}{\prod_{i=1}^{n} (c\mu/\delta + i)}\right] \\ &= \frac{\rho^c}{c!}\left[\frac{1}{p_b} + F\left(\frac{c\mu}{\delta}, \frac{\lambda}{\delta}\right) - 1\right] \end{aligned} \tag{3.19}$$

The auxiliary function F is defined as

$$F(x,y) = \frac{xe^y}{y^x}\gamma(x,y)$$

where $\gamma(x,y) = \int_0^y t^{x-1}e^{-t}dt$ denotes the *incomplete gamma function* [51]. Inverting expression 3.19 results in

$$p_0 = \frac{c!}{\rho^c} \frac{p_b}{1 + [F(c\mu/\delta, \lambda/\delta) - 1]\, p_b}$$

Inserting p_0 in the system of equation 3.18 and realizing, that

$$p_c = \frac{\rho^c}{c!} p_0$$

finally leads to

$$p_n = \begin{cases} \frac{c!}{n!\rho^{c-n}} p_c & \text{for } 0 \leqq n < c \\ \frac{p_b}{1+[F(c\mu/\delta,\lambda/\delta)-1]p_b} & \text{for } n = c \\ \frac{(\lambda/\delta)^{n-c}}{\prod_{i=1}^{n-c} c\mu/\delta+i} p_c & \text{for } n > c \end{cases} \tag{3.20}$$

The above set of formulas provides an easy way to calculate the steady state distribution by performing the following steps

1. Calculate the blocking probability for a c server loss system p_b by applying the Erlang B formula 3.10
2. Look up the value of the incomplete gamma function $\gamma\left(c\mu/\delta, \lambda/\delta\right)$
3. Insert both results in the expression for p_c
4. Use the remaining formulas to derive p_n, $n \neq c$ from p_c

Having derived a closed form solution for the equilibrium distribution, we are now able to derive various performance characteristics. Let $\breve{W}_q$ denote a random variable associated with the current queueing time. Then the probability of delay is given by

$$\begin{aligned} p_d &= \Pr\left\{\breve{W}_q > 0\right\} = \sum_{n=c}^{\infty} p_j \\ &= \frac{p_b}{1+\left[F\left(c\mu/\delta,\lambda/\delta\right)-1\right]p_b} \left[1 + \sum_{n=c+1}^{\infty} \frac{\left(\lambda/\delta\right)^{n-c}}{\prod_{i=1}^{n-c} c\mu/\delta + i}\right] \\ &= \frac{F\left(c\mu/\delta,\lambda/\delta\right)p_b}{1+\left[F\left(c\mu/\delta,\lambda/\delta\right)-1\right]p_b} \end{aligned} \tag{3.21}$$

In the $M/M/c+M$ queue a customer decides to leave the queue at an exponential rate. In determining the probability of getting ultimately served, one encounters, what has been called competition of exponentials in [40]:

$$\begin{aligned} p_0^s &= \frac{c\mu}{c\mu+\delta} \\ p_1^s &= \frac{c\mu+\delta}{c\mu+2\delta} p_0^* = \frac{c\mu}{c\mu+2\delta} \end{aligned}$$

Proceeding further one arrives at p_n^s the probability of the n-th customer getting served, that is

$$p_n^s = \frac{c\mu}{c\mu + (n+1)\,\delta}, \quad n \geq 1$$

The probability to abandon service and loosing the customer is given by

$$p_n^a = 1 - p_n^s = \frac{(n+1)\,\delta}{c\mu + (n+1)\,\delta}, \quad n \geq 0$$

One can now derive the conditional probability that a customer abandons given he does not receive immediate service

$$\begin{aligned} \Pr\left\{\text{Abandon}|\breve{W}_q > 0\right\} &= \sum_{n=c}^{\infty} \frac{p_n p_{c-n}^a}{p_d} \\ &= \frac{1}{\rho F\left(c\mu/\delta, \lambda/\delta\right)} + 1 - \frac{1}{\rho} \end{aligned} \tag{3.22}$$

The calculations have been omitted, because they are rather lenghty. They utilize an idendity derived by Palm for the function F based on properties of the incomplete gamma function [51]. A partial derivation is given in [40]. If required, one may consult the original paper [43] by Palm. Due to independence, the probability of an arbitrary customer abandoning the queue is given by the product of the expressions 3.21 and 3.22:

$$\begin{aligned} p_a &= \Pr\{\text{Abandon}\} = \Pr\left\{\text{Abandon}|\breve{W}_q > 0\right\} \Pr\left\{\breve{W}_q > 0\right\} \\ &= \left(\frac{1}{\rho F\left(c\mu/\delta, \lambda/\delta\right)} + 1 - \frac{1}{\rho}\right) p_d \end{aligned} \tag{3.23}$$

Based on the loss probability 3.23, the average queueing time W_q may be easily derived by an application of Little's law. First note that in equilibrium, the rate of customers abadoning the queue and the rate of customers entering the system have to be the same, i.e. $\delta L_q = \lambda p_a$. So the average queue size is given by

$$L_q = \frac{\lambda}{\delta} p_a$$

Applying Little's law $L_q = \lambda W_q$ leads to

$$W_q = \frac{p_a}{\delta}$$

The remaining measures of effectiveness are obtained in much the same way as has been done for the $M/M/c/K$ model, that is

$$\begin{aligned} W &= W_q + \frac{1}{\mu} = \frac{p_a}{\delta} + \frac{1}{\mu} \\ L &= \lambda W = \frac{\lambda p_a}{\delta} + \rho \end{aligned}$$

In comparison to the $M/M/c$ queueing system, performance is superior in terms of average waiting time and mean queue length. Additionally, the $M/M/c+M$ system is immune to any kind of congestion. This is also, what we encounter especially in real life situations concerned with human behaviour. Impatience becomes a mandatory assumption for the analysis of such models. This might be different for technical systems.

Another, although not well known form of customer impatience exists in multiqueue systems and is called *jockeying*. Customer dissatisfaction is expressed by simply joining another queue. Rather simple in description, these models are hard to solve and will not be covered in this text. For general information on customer impatience refer to [27]. A more specific model with limited sources, limited capacity and reneging is described in [1]. Equivalence relations between systems with customer impatience and machine inference problems are derived in [28]. Balking and reneging for birth-death processes has also been considered in [49].

3.5 Bounded Holding Times

In certain situations it becomes necessary to bound the time a customer resides in the system. As an example consider a call centre, where customers are rerouted to an IVR system, when a predefined waiting time limit has been reached. By expressing the waiting time limit in terms of an exponential distribution, the system fits nicely in the framework of birth-death processes. It has been introduced by Gnedenko and Kovalenko in their book [25]. Assume arrival rate, service rate, system capacity and the number of servers to be λ, μ, K and c similar to the $M/M/c/K$ model. Further let the waiting time boundary be exponentially distributed with rate δ. Then the corresponding average is given by $\frac{1}{\delta}$. Any customer having reached the waiting time limit will depart from the system. Combination of rates leads to

$$\mu_n = \begin{cases} n\mu + n\delta & \text{for } 1 \leqq n < c \\ c\mu + n\delta & \text{for } c \leqq n \leqq K \\ 0 & \text{for } n > K \end{cases}$$

Substitution of $\lambda_n = \lambda$ and μ_n into the birth-death equation 3.1 leads to expression 3.18 for the steady state distribution. It follows, that the model with exponentially bounded holding times is equivalent to the Erlang A $M/M/c + M$ queueing system.

3.6 Finite Population Models

The previous discussion focused on queueing problems with infinite customer population. Although mathematically convenient, such an assumption only serves well as an approximation to situations with a large population. One anticipates, that prediction errors become negligible. If this is not the case, then one has to take care about finiteness. This is best done by modifying the birth rate λ in the standard birth-death model as follows

$$\lambda_n = \begin{cases} (M-n)\,\lambda & \text{for } 0 \leqq n < M \\ 0 & \text{for } n \geqq M \end{cases}$$

Here M denotes the size of the population. Assuming a system with $c < M$ service units, i.e.

$$\mu_n = \begin{cases} n\mu & \text{for } 1 \leqq n < c \\ c\mu & \text{for } n \geqq c \end{cases}$$

and substituting in equation 3.1 leads to

$$p_n = \begin{cases} \binom{M}{n}\rho^n p_0 & \text{for } 0 \leqq n < c \\ \binom{M}{n}\frac{n!}{c^{n-c}c!}\rho^n p_0 & \text{for } c \leqq n \leqq M \end{cases} \tag{3.24}$$

with $\binom{M}{n} = \frac{M!}{(M-n)!n!}$ denoting the *binomial coefficient*. Applying $\sum_{n=0}^{M} p_n = 1$ and solving for p_0 gives

$$p_0 = \left[\sum_{n=0}^{c-1}\binom{M}{n}\rho^n + \sum_{n=c}^{M}\binom{M}{n}\frac{n!}{c^{n-c}c!}\rho^n\right]^{-1}$$

For efficient calculation of the steady state probabilities, Gross and Harris [27] suggest the following recursion based on the properties of the binomial coefficient:

$$f_n = \frac{p_{n+1}}{p_n} = \begin{cases} \frac{M-n}{n+1} & \text{for } 0 \leqq n < c \\ \frac{M-n}{c} & \text{for } c \leqq n \leqq M \end{cases}, \qquad p_n = \prod_{i=0}^{n-1} f_i p_0$$

Using the definition of the expected value, one is now able to derive the average system size

$$L = \sum_{n=0}^{M} np_n = \left[\sum_{n=0}^{c-1} n \binom{M}{n} \rho^n + \sum_{n=c}^{M} n \binom{M}{n} \frac{n!}{c^{n-c} c!} \rho^n \right] p_0$$

Following [27], the average queue size L_q may be derived from L as follows

$$\begin{aligned} L_q &= \sum_{n=c}^{M} (n-c)\, p_n = \sum_{n=c}^{M} np_n - c \sum_{n=c}^{M} p_n \\ &= L - \sum_{n=c}^{c-1} np_n - c \left(1 - \sum_{n=0}^{c-1} np_n \right) \\ &= L - c + \sum_{n=0}^{c-1} (c-n)\, p_n \\ &= L - c + p_0 \sum_{n=0}^{c-1} (c-n) \binom{M}{n} \rho^n \end{aligned}$$

In order to derive the waiting time indicators using Little's law, one first has to determine the mean arrival rate $\bar{\lambda}$. With n customers already in the system, a maximum of $M-n$ customers remain outside waiting for arrival. This results in a mean arrival rate of $(M-n)\,\lambda$. Averaging yields

$$\bar{\lambda} = \sum_{n=0}^{M} (M-n)\, \lambda p_n = \lambda \left(M \sum_{n=0}^{M} p_n - \sum_{n=0}^{M} np_n \right) = \lambda\,(M-L)$$

Using Little's law with the just derived mean arrival rate $\bar{\lambda}$ leads to

$$W = \frac{L}{\lambda\,(M-L)}, \qquad W_q = \frac{L_q}{\lambda\,(M-L)}$$

Assuming the size of the waiting room to be 0 results in a finite-source variation of the classic $M/M/c/c$ Erlang Loss system. This model is often used in telecommunication applications and is called the *Engset model*. The steady state distribution 3.24 simplifies to

$$p_n = \frac{\binom{M}{n}\rho^n}{\sum_{i=0}^{M}\binom{M}{i}\rho^i}, \qquad 0 \le n \le M$$

This is also known as the *Engset distribution*. The probability of a customer being blocked and getting lost due to call congestion is determined by a full system, that is $p_b = p_c$. Similar to the general finite population model, a recurrence relation may be deducted for easy calculation. In telephony applications, the recursion is usually defined for the blocking probability:

$$p_c = \frac{(M-c)\,\rho p_{c-1}}{c + (M-c)\,\rho p_{c-1}}$$

With M getting very large, the Engset distribution approaches the probabilities p_n given by the $M/M/c/c$ Erlang loss system. As a reference related to queueing theory consider any standard text book such as [27]. For telephony applications we refer to [6].

3.7 Relation to Markov Chains

Now an attempt will be made to relate birth-death processes to continous time Markov chains. For a short introduction please refer to appendix A.3. A birth-death process may be understood as a *skip-free* Markov chain, meaning that the process can only move to a neighbouring state in a single step. Combining birth and death rates

$$\begin{aligned} q_{n,n+1} &= \lambda_n \\ q_{n,n-1} &= \mu_n \\ q_{nn} &= -(\lambda_n + \mu_n) \\ q_{mn} &= 0 \text{ for } |m-n| > 1 \end{aligned}$$

leads to the infinitesimal generator

$$\mathbf{Q} = \begin{pmatrix} -\lambda_0 & \lambda_0 & 0 & 0 & \cdots \\ \mu_1 & -(\lambda_1+\mu_1) & \lambda_1 & 0 & \cdots \\ 0 & \mu_2 & -(\lambda_2+\mu_2) & \lambda_2 & \cdots \\ \vdots & & & \ddots & \end{pmatrix}$$

Alternatively one may address the discrete Markov chain embedded into the birth-death process. By chosing the occurences of the state transitions as regeneration points, the corresponding transition matrix becomes

$$\mathbf{P} = \begin{pmatrix} 0 & 1 & 0 & 0 & \cdots \\ \frac{\mu_1}{\lambda_1+\mu_1} & 0 & \frac{\lambda_1}{\lambda_1+\mu_1} & 0 & \cdots \\ 0 & \frac{\mu_2}{\lambda_2+\mu_2} & 0 & \frac{\lambda_2}{\lambda_2+\mu_2} & \cdots \\ \vdots & & & \ddots & \end{pmatrix}$$

Appearantly both matrices suggest irreducibility. From the conservation equation 3.2 it follows, that an equilibrium can only be assumed, if

$$\sum_{n=1}^{\infty} \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} < \infty \tag{3.25}$$

Due to the fact, that stationarity implies positive recurrence, equation 3.25 may be used as a criterion for positive recurrence. Focusing on the embedded chain, it can be shown [66], that a birth-death process is recurrent, if

$$\frac{1}{\lambda_0} + \sum_{n=1}^{\infty} \frac{\mu_1 \mu_2 \cdots \mu_n}{\lambda_0 \lambda_1 \cdots \lambda_n} = \infty \tag{3.26}$$

holds and vice versa. Multiplying the left hand side by λ_0 and omitting the first term simplifies formula 3.26 to

$$\sum_{n=1}^{\infty} \prod_{i=1}^{n} \frac{\mu_i}{\lambda_i} = \infty$$

The stationary distribution may also be calculated using Markov chain methods. One can either chose to solve the Chapman Kolmogorov equations for the embedded chain or apply Kolmogorov's differential systems directly. Either case leads to the same results. For further information please consult [4] and [66]. Especially the latter reference provides a rigorous treatment on the topic.

Chapter 4

Extended Markovian Models

Because of the memoryless property of the exponential distribution, Markovian models such as the birth-death process become analytically tractable. In most cases it is possible to derive a closed form solution as well. In the current section we enhance the models treated so far by features, which can not be classified as typically Markovian.

4.1 Some Useful Relations

Before getting hands on some rarities in queueing we will derive some useful tools. The first deals with an interesting property of Poisson arrivals. A Poisson stream is sometimes called purely random. Provided the state of the system changes at most by one, a customer arriving in the stream finds the same state distribution as an outside observer. One can say, that *Poisson arrivals see time averages* (PASTA). It turns out, that PASTA also applies to the transient case, which obviously includes the steady state version as special case.

Theorem 10 (PASTA) *Define $a_n(t)$ as the probability of n customers in the system seen by an arrival just after entering the system. Let $p_n(t)$ denote the distribution of n customers in the system at an arbitrary point in time. Then for Poisson arrivals*

$$a_n(t) = p_n(t) \quad \textit{for all } n \geq 0, t \geq 0$$

Proof. Define $N(t)$ as the number of customers in the system at time t. Now consider the number of arrivals $A(t, t+h)$ in an infinitesimal interval

$(t, t+h)$. Then $a_n(t)$ is defined as the limit $h \to 0$ of the probability, that the number of customers in the system is n given an arrival has occured just after t. In mathematical terms

$$\begin{aligned}
a_n(t) &= \lim_{h \to 0} \Pr\left\{N(t) = n | A\left(t, t+h\right) = 1\right\} \\
&= \lim_{h \to 0} \frac{\Pr\left\{N(t) = n, A\left(t, t+h\right) = 1\right\}}{\Pr\left\{A\left(t, t+h\right) = 1\right\}} \\
&= \lim_{h \to 0} \frac{\Pr\left\{A\left(t, t+h\right) = 1 | N(t) = n\right\} \Pr\left\{N(t) = n\right\}}{\Pr\left\{A\left(t, t+h\right) = 1\right\}} \\
&= \lim_{h \to 0} \frac{\Pr\left\{A\left(t, t+h\right) = 1\right\} \Pr\left\{N(t) = n\right\}}{\Pr\left\{A\left(t, t+h\right) = 1\right\}} \\
&= \lim_{h \to 0} \Pr\left\{N(t) = n\right\} = p_n\left(t\right)
\end{aligned}$$

Please note, that $\Pr\left\{A\left(t, t+h\right) = 1 | N(t) = n\right\} = \Pr\left\{A\left(t, t+h\right) = 1\right\}$ follows from the fact, that the number of arrivals occuring in two disjoint time intervals are independent. ■

Another proof based on the assumption, that future increments are independent of the past has been given by Wolff in [64]. A proof tailored to the requirements of the $M/G/1$ queue may be found in [27].

A similar result also holds for exponential service times. Assuming equilibrium, let p_n be the probability that n customers are in the system. The probability that n customers are in the system just prior to an arrival is denoted by $\tilde{p}_n$.

Theorem 11 (Rate Conservation Law) *Consider a queueing system with general arrivals, exponential service times, $c \leq \infty$ servers and system limit $K \leq \infty$. Furthermore assume a work conserving queueing discipline and no interruption of service. Then the following relation holds*

$$\min\left(c, n\right) p_n = \rho \tilde{p}_{n-1}$$

Rewriting the above equation to $\min\left(c, n\right) \mu p_n = \lambda \tilde{p}_{n-1}$ one may intuitively explain the result as follows. The left term represents a state transition from state n to state $n-1$, whereas the right term is just the opposite. Given a work conserving queueing discipline in accordance with the local balance principle, the rate downwards must equal the rate upwards [1]. For the proof we refer to theorem 6.4.3 of [8] or to page 154 of [56].

We will turn attention now to a theorem from complex analysis often employed in queueing theory. Its main use lies in assuring the existence of roots within a closed contour such as the unit circle $|z| = 1$. Usually a given function F is split into two parts, i.e. $F(z) = f(z) + g(z)$, where $f(z)$ has a known number of zeros inside a given domain.

Theorem 12 (Rouche) *If $f(z)$ and $g(z)$ are functions analytic inside and on a closed contour C and if $|g(z)| < |f(z)|$ on C, then both $f(z)$ and $f(z) + g(z)$ possess the same number of zeros inside C.*

A proof may be found in almost any standard textbook on complex analysis, for example see [52]. In some cases, one wants to extend the assumptions of Rouche's theorem to the boundary of C without invalidating the conclusion. This is indeed possible, as shown in a recent paper by Klimenok[38].

Theorem 13 (Extended Rouche) *Let $f(z)$ and $g(z)$ be analytic inside in the open unit disc $|z| < 1$, continous on the boundary $|z| = 1$ and differentiable at the point $z = 1$. Assume, that the following relations are satisfied*

$$\begin{gathered} |g(z)|\big|_{|z|=1, z\neq 1} < |f(z)|\big|_{|z|=1, z\neq 1} \\ f(1) = -g(1) \neq 0 \\ \frac{\frac{d}{dz} f(z)|_{z=1} + \frac{d}{dz} g(z)|_{z=1}}{f(1)} > 0 \end{gathered}$$

Then the number of zeros N_f and N_{f+g} of the functions $f(z)$ and $f(z)+g(z)$ in the unit disc $|z| < 1$ are related as follows:

$$N_{f+g} = N_f - 1$$

For the proof we refer to the paper of Klimenok [38]. As pointed out there, theorem 13 is in particular very useful for matrix geometric models of the $M/G/1$ type.

4.2 General impatience distribution

One possible generalization to Palm's $M/M/c + M$ model is to allow for a general impatience distribution. In doing so, one arrives at the $M/M/c + G$ model. A variant thereof has first been introduced by Bacelli and Hebuterne in 1981. Our treatment will be based on their paper [5] and the paper by Zeltyn and Mandelbaum [65]. Although the model assumes, that arriving customers are fully aware of the offered wait V and abandon service immediately, if their patience time is exceeded, the model coincides with the $M/M/c+G$ model in terms of all relevant stationary performance characteristics. The patience time is assumed to be distributed according the $G(.)$, in the latter more often referenced to by the *survival function* $\bar{G}(.) = 1 - G(.)$. Although the exponentiality assumption is violated, the model may be described by a Markov process $\{N(t), \eta(t) : t \geqq 0\}$, where $N(t)$ describes the number of customers in the system at time t and $\eta(t)$ denotes the virtual offered waiting time of a customer arriving at time t. As long as there are $c-1$ customers in the system, $\eta(t) = 0$, whereas from c customers on $\eta(t)$ becomes positive. In the latter case, it is only relevant to know, that there are c or more customers in the system, the exakt number is irrelevant. Therefore we choose $N(t) = c$ for $\eta(t) > 0$ and $\eta(t) = 0$ for $0 \leqq N(t) \leqq c-1$. With the system described that way, one preserves the Markov property. Let $v(x)$ denote the density of the virtual offered waiting time and define

$$\begin{array}{rlr} v(x) = & \lim_{t\to\infty} \lim_{h\to 0} \frac{\Pr\{N(t)=c, x<\eta(t)\leqq x+h\}}{h} & x \geqq 0 \\ p_n = & \lim_{t\to\infty} \Pr\{N(t) = n, \eta(t) = 0\} & 0 \leqq n \leqq c-1 \end{array}$$

Assign to λ, μ and $\rho = \frac{\lambda}{\mu}$ the usual meanings. Obviously the system behaves like a classical $M/M/c$ queue for $0 \leqq n \leqq c-1$, i.e.

$$p_n = \frac{\rho^n}{n!} p_0 \quad \text{for } 1 \leqq n \leqq c-1 \tag{4.1}$$

By realizing that state $c-1$ may only be entered from above, if an arriving customer would not have to wait for service, one arrives at

$$\lambda p_{c-1} = v(0) \tag{4.2}$$

The case of positive virtual wait may be obtained from

$$\begin{aligned} \Pr\{\eta(t+h)\} &= \Pr\{\eta(t) > x+h\} + \Pr\{\eta(t+h) > x, \eta(t) = 0\} \\ &\quad + \Pr\{\eta(t+h) > x, 0 < \eta(t) \leqq x+h\} \end{aligned}$$

The second term on the right side describes the increase of the virtual wait caused by an arrival occupying the last free server. This increase will exceed

x with probability $e^{-c\mu x}$, as the intervals between service departures are distributed according to an exponential distribution with rate $c\mu$. The third term corresponds to an arrival to a full system with positive virtual wait. The probability of receiving service is determined by the survival function $\bar{G}(x)$. Thus in steady state one can expect

$$\int_x^\infty v(y)dy = \int_{x+h}^\infty v(y)dy + \lambda h p_{n-1} e^{-c\mu x} + \lambda h \int_0^x e^{-c\mu(x-y)} v(y)\bar{G}(y)dy + o(h)$$

Differentiating with respect to h and letting $h \to 0$ yields

$$v(x) = \lambda p_{n-1} e^{-c\mu x} + \lambda e^{-c\mu x} \int_0^x e^{c\mu y} v(y)\bar{G}(y)dy \tag{4.3}$$

By observing, that $H(x) = e^{c\mu x}v(x)$ is a solution to the integral equation $H(x) = \lambda p_{n-1} + \lambda \int_0^x H(y)\bar{G}(y)dy$ and solving that integral equation directly one obtains $H(x) = \lambda p_{n-1} \exp\left\{\lambda \int_0^x \bar{G}(y)dy\right\}$ and

$$v(x) = \lambda p_{n-1} \exp\left\{\lambda \int_0^x \bar{G}(y)dy - c\mu x\right\} \tag{4.4}$$

Normalizing probabilities

$$\begin{aligned} 1 &= \sum_{n=0}^{c-1} p_n + \Pr\{V > 0\} = \sum_{n=0}^{c-1} p_n + \int_0^\infty v(x)dx \qquad (4.5) \\ &= p_0 \sum_{n=0}^{c-1} \frac{\rho^n}{n!} + \lambda \frac{\rho^{c-1}}{(c-1)!} \int_0^\infty \exp\left\{\lambda \int_0^x \bar{G}(y)dy - c\mu x\right\} dx \end{aligned}$$

results in

$$p_0 = \left[\sum_{n=0}^{c-1} \frac{\rho^n}{n!} + \frac{\rho^{c-1}\lambda J}{(c-1)!}\right]^{-1}, \quad J := \int_0^\infty \exp\left\{\lambda \int_0^x \bar{G}(y)dy - c\mu x\right\} dx \tag{4.6}$$

Please note, that we are now able to describe the steady state behaviour of the $M/M/c+G$ model in terms of equations 4.1, 4.2, 4.3 and 4.6. Furthermore, there is a straightforward generalization to include balking behaviour into the model. As before for the birth-death equations define a state dependent arrival rate

$$\lambda_n = \begin{cases} b_n\lambda & 0 \leqq n \leqq c-1 \\ b_{c-1}\lambda & \eta(t) > 0 \end{cases}$$

One may immediately write down the steady state equations

$$\begin{aligned}
p_n &= p_0 \prod_{i=1}^{n} \frac{\lambda_{i-1}}{i\mu} = \lambda^n p_0 \prod_{i=1}^{n} \frac{b_{i-1}}{i\mu} \quad \text{for } 1 \leqq n \leqq c-1 \\
\lambda_{c-1} p_{c-1} &= \lambda b_{c-1} = v(0) \\
v(x) &= \lambda_{c-1} \left(p_{n-1} e^{-c\mu x} + e^{-c\mu x} \int_0^x e^{c\mu y} v(y) \bar{G}(y) dy \right) \qquad (4.7) \\
&= \lambda b_{c-1} \left(p_{n-1} e^{-c\mu x} + e^{-c\mu x} \int_0^x e^{c\mu y} v(y) \bar{G}(y) dy \right) \\
p_0 &= \left[\sum_{n=0}^{c-1} \prod_{i=1}^{n} \frac{\lambda_{i-1}}{i\mu} + \lambda_{c-1} J \prod_{i=1}^{c-1} \frac{\lambda_{i-1}}{i\mu} \right]^{-1} \\
&= \left[\sum_{n=0}^{c-1} \lambda^n \prod_{i=1}^{n} \frac{b_{i-1}}{i\mu} + \lambda^c b_{c-1} J \prod_{i=1}^{c-1} \frac{b_{i-1}}{i\mu} \right]^{-1}
\end{aligned}$$

For both models, the system may be assumed to be stable, if the integral in the expression for J in equation 4.6 converges. This in turn is equivalent to the condition $\lambda \bar{G}(\infty) < c\mu$ or $u\bar{G}(\infty) < 1$ with u the utilization. For a proper probability distribution $G(.)$, i.e. $\lim_{x \to \infty} \bar{G}(x) = 0$, the system will not become unstable and show behaviour similar to the $M/M/c/K$ queueing system. Otherwise $G(.)$ is called *defective* and the above mentioned condition has to be considered.

Considering the model without balking, let p_a denote the probability, that an arriving customer refrains from being serviced because of excessive wait. Then one may express the server occupancy as $\bar{u} = u(1-p_a) = \frac{\rho}{c}(1-p_a)$. Alternatively it may be calculated from $\bar{u} = \frac{1}{c}\sum_{n=0}^{c-1} np_n + \Pr\{V > 0\} = \frac{1}{c}\sum_{n=0}^{c-1} np_n + (1 - \sum_{n=0}^{c-1} p_n)$. The last substitution is evident from the normalization condition 4.5. Combining expressions leads to

$$\begin{aligned}
p_a &= 1 - \frac{c}{\rho}\bar{u} = 1 - \frac{c}{\rho}\left(\frac{1}{c}\sum_{n=0}^{c-1} np_n + 1 - \sum_{n=0}^{c-1} p_n \right) \\
&= 1 - \frac{c}{\rho}\left(\sum_{n=1}^{c-1} \frac{np_n}{c} + 1 - \sum_{n=0}^{c-1} p_n \right) = 1 - \sum_{n=1}^{c-1} \frac{\rho^{n-1} p_0}{(n-1)!} + \frac{c}{\rho}\left(1 - \sum_{n=0}^{c-1} p_n \right) \\
&= 1 - \sum_{n=0}^{c-2} p_n + \frac{c}{\rho}\left(1 - \sum_{n=0}^{c-1} p_n \right) = \left(1 - \frac{c}{\rho} \right)\left(1 - \sum_{n=0}^{c-1} p_n \right) + p_{c-1}
\end{aligned}$$

For the average waiting time one starts with equation 4.4 to derive

$$\Pr\{V > t\} = \lambda p_{n-1} \int_t^\infty \exp\left\{\lambda \int_0^x \bar{G}(y)dy - c\mu x\right\} dx$$

which is just the survival function of V. With V defined only for non-negative values, the average queueing time may be calculated as follows

$$\begin{aligned} W_q &= \int_0^\infty \bar{G}(t)\Pr\{V > t\}\,dt \\ &= \lambda p_{c-1} \int_0^\infty \bar{G}(t)\exp\left\{\lambda \int_0^x \bar{G}(y)dy - c\mu x\right\} dxdt \qquad (4.8) \\ &= \lambda p_{c-1} \int_0^\infty \int_0^t \bar{G}(y)dy \exp\left\{\lambda \int_0^t \bar{G}(y)dy - c\mu t\right\} dt \qquad (4.9) \end{aligned}$$

whereas the last expression has been derived integrating by parts. Note, that by combining expression 4.1 and 4.6 and using Erlang's loss formula 3.10 for $c-1$ servers, one derives p_{c-1} as follows

$$p_{c-1} = \frac{\frac{\rho^{c-1}}{(c-1)!}}{\sum_{n=0}^{c-1}\frac{\rho^n}{n!} + \rho^{c-1}\frac{\lambda J}{(c-1)!}} = \left[\frac{1}{p_b} + \lambda J\right]^{-1}$$

The remaining performance characteristics may be determined by an application of Little's Law

$$L_q = \lambda W_q, \quad W = W_q + \frac{1}{\mu}, \quad L = \lambda W$$

With balking included into the model, the average queueing time 4.9 adapts well to the modification, one has only to replace λ by λb_{c-1} and calculate p_{c-1} according to the first equation of 4.7. Giving a closer look to the effective arrival rate and the expression for p_a, the results become quite cumbersome. It turns out, that an explicit inclusion of states above c provides a feasible solution. This will be one of the main ideas of the model by Brandt and Brandt introduced next.

The $M/M/c+G$ model is a rather general one, as it includes the $M/M/c$ queueing system and Palm's $M/M/c+M$ model as special cases. For the latter let the patience times follow an exponential distribution with parameter δ, whereas for the former assume infinite patience, i.e. $\bar{G}(x) = 1$. Note, that

$\bar{G}(x) = 1$ is indeed a defective distribution putting the stability condition $1 > u\bar{G}(\infty) = u$ in effect. Obviously we arrived at the stability condition for the $M/M/c$ model. Another important special case is the $M/M/c + D$ queueing system as introduced by Gnedenko and Kovalenko in their book [25], which is based on Barrer's derivations [49]. In practice, it applies well to computer networks with deterministic timeouts.

As mentioned above, Brandt and Brandt derived a generalization to the $M/M/c + G$ queueing system by allowing for state dependent arrival and service rates [9]. We will only provide the results here, as the derivations are rather lengthy but in concept similar to the classic model. The main difference lies in the fact, that steady state probabilities are now defined for c or more customers and that the residual patience time has been taken into account. As before, as long as there are servers available, the system follows the well-known birth-death approach. It assumes a bounded sequence of arrival rates λ_n, i.e. there is only a finite number of $\lambda_n > 0$. This leads to the steady-state distribution

$$p_n = \begin{cases} p_0 \left(\prod_{i=0}^{n-1} \lambda_i\right) \left(\prod_{i=n+1}^{c} \mu_i\right) & 0 < n \leqq c \\ p_0 \left(\prod_{i=0}^{n-1} \lambda_i\right) \frac{\mu_c}{(n-c)!} \int_0^\infty \left(\int_0^y \bar{G}(z)dz\right)^{n-c} e^{-\mu_c y} dy & n > c+1 \\ \left[\sum_{j=0}^{c-1} p_j + \sum_{j=0}^{\infty} \left(\prod_{i=0}^{c+j-1} \lambda_i\right) \frac{\mu_c}{j!} \int_0^\infty \left(\int_0^y \bar{G}(z)dz\right)^j e^{-\mu_c y} dy\right]^{-1} & n = 0 \end{cases}$$

The system can be considered stable, if one is able to calculate a non-trivial p_0. Isolating the relevant part yields the stability condition

$$\sum_{j=0}^{\infty} \left(\prod_{i=0}^{c+j-1} \lambda_i\right) \frac{1}{j!} \int_0^\infty \left(\int_0^y \bar{G}(z)dz\right)^j e^{-\mu_c y} dy < \infty$$

Based on the effective arrival rate $\bar{\lambda} = \sum_{n=0}^{\infty} \lambda_n p_n$, the probability, that a customer has to wait, i.e.

$$p_d = 1 - \frac{1}{\bar{\lambda}} \sum_{n=0}^{c} \lambda_n p_n$$

and the probability, that an arriving customer will leave the system later due to impatience

$$p_a = 1 - \frac{1}{\bar{\lambda}} \left(\mu_c + \sum_{n=0}^{c-1} (\mu_n - \mu_c) p_n\right)$$

The mean queueing time may be derived by using Little's law

$$W_q = \frac{1}{\bar{\lambda}} \sum_{n=c+1}^{\infty} (n-c)\, p_n$$

One may split the queueing time into two parts, representing the wait an arriving customer is exposed to in case of being served or lost due to impatience,

$$\begin{aligned} W_q^s &= \frac{p_s}{(1-p_a)\,\bar{\lambda}} \sum_{j=1}^{\infty} \left(\prod_{i=0}^{c+j-1} \lambda_i \right) \frac{\mu_c}{j!} \int_0^{\infty} \left(\int_0^y \bar{G}(z)dz \right)^j (\mu_c y - 1)\, e^{-\mu_c y} dy \\ W_q^a &= \frac{p_s}{p_a \bar{\lambda}} \sum_{j=1}^{\infty} \left(\prod_{i=0}^{c+j-1} \lambda_i \right) \frac{\mu_c}{j!} \int_0^{\infty} \left(\int_0^y \bar{G}(z)dz \right)^j (j+1-\mu_c y)\, e^{-\mu_c y} dy \end{aligned}$$

For the proofs we refer to the paper of Brandt and Brandt [9]. Some of them rely on Palm distributions and stationary point processes, especially those, which are concerned with the relation between distributions at arrival epochs and their general counterpart. For more information on these topics, please consult [8]. In their paper, Brandt and Brandt also consider the special case of an impatience time defined as the minimum of a constant and an exponentially distributed random variable. In the extreme, one arrives either at Palm's $M/M/c + M$ model or Gnedenko's $M/M/c + D$ queueing system.

4.3 Retrials and the Orbit Model

Up to now it has been assumed for systems with limited capacity, that blocked customers are lost. In the following we will consider these customers to retry for service after some period of time. Obviously there is some dependency introduced in the model, which violates the memoryless property of the arrival stream. By describing the system as a two dimensional Markov process $\{C(t), N(t) : t \geqq 0\}$, one restores the desired features. Here $C(t)$ denotes the number of busy servers at time t and $N(t)$ describes the number of retrying sources. One can think of blocked customers beeing redirected to an *orbit* instead of getting lost. For clarification, this situation is shown in figure 4.1. From a different viewpoint, a retrial system forms a queueing network

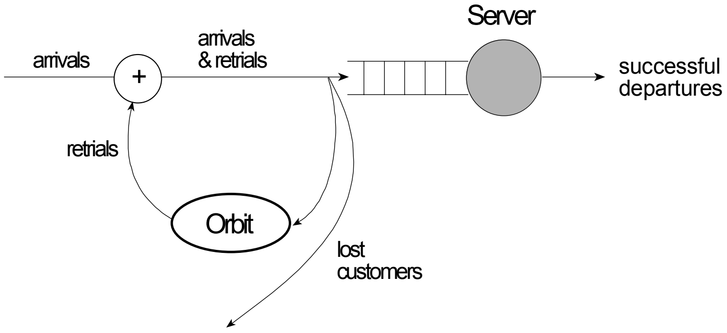


Figure 4.1: Orbit Model

consisting of loss and infinite server nodes. Retrial systems have become important in telephony applications, as a typical caller retries after some time, if he does not reach the desired target. For this reason, the service facility is often modeled as loss system, not as queueing system with limited capacity. We will follow this convention and introduce a $M/M/c/c$ queue as service facility. If the service distribution is not exponential, we'll loose the Markov property of the above mentioned process again. Then a supplementary variable describing the elapsed service time has to be introduced to preserve it.

Adhering to the usual notation we'll turn attention to the single server case now, i.e. assume a $M/M/1/1$ service facility. Consequently $C(t)$ can only take the values 0 and 1. Assume that the time lengths between the retrials are independent and follow an exponential distribution with parameter η. Thus on the average every $\frac{1}{\eta}$ seconds (or any other preferred time unit) a retrial occurs. Introducing $p_{m,n} := \Pr\{C(t) = m, N(t) = n\}$, we may proceed as usual and equate the flow in with the flow out. Hence,

$$\begin{aligned}(\lambda + n\eta)\, p_{0,n} &= \mu p_{1,n} \\ (n+1)\, \eta p_{0,n+1} &= \lambda p_{1,n}\end{aligned}$$

Following the treatment of [30], both expressions may be combined to

$$p_{1,n+1} = \rho \left(1 + \frac{\lambda}{\eta\,(n+1)}\right) p_{1,n}$$

This leads to

$$p_{1,n} = \begin{cases} \rho^n \prod_{i=1}^n \left(1 + \frac{\lambda}{\eta i}\right) p_{1,0} & n \geqq 1 \\ p_{1,0} & n = 0 \end{cases} \tag{4.10}$$

For a single server system the server utilization $u = \rho$ may be written alternatively as

$$\begin{aligned} \rho &= \sum_{n=0}^{\infty} p_{1,n} = p_{1,0} \left[1 + \sum_{n=1}^{\infty} \rho^n \prod_{i=1}^{n} \left(1 + \frac{\lambda}{\eta i}\right)\right] \\ &= p_{1,0} \left[1 + \sum_{n=1}^{\infty} \rho^n \prod_{i=1}^{n} \left(\frac{i + \lambda/\eta}{i}\right)\right] \\ &= p_{1,0} \left[1 + \sum_{n=1}^{\infty} \rho^n \frac{(1 + \lambda/\eta) \cdots (n + \lambda/\eta)}{n!}\right] \\ &= p_{1,0} \left[\sum_{n=0}^{\infty} \rho^n \binom{n + \lambda/\eta}{n}\right] \\ &= p_{1,0} \left[(1 - \rho)^{-1-\lambda/\eta}\right] \end{aligned}$$

where for the last step the binomial theorem has been applied. A simple multiplication yields

$$p_{1,0} = \left[(1 - \rho)^{-1-\lambda/\eta}\right]^{-1} = \rho\,(1 - \rho)^{1+\lambda/\eta} \tag{4.11}$$

So far the steady state distribution for the single server retrial model has been derived. The computation of the expected number of customers in orbit will be performed in two steps. By calculating

$$\begin{aligned}
\sum_{n=1}^{\infty} n p_{1,n} &= p_{1,0} \sum_{n=1}^{\infty} n \rho^n \prod_{i=1}^{n} \left(1 + \frac{\lambda}{\eta i}\right) \\
&= p_{1,0} \left(1 + \frac{\lambda}{\eta}\right) \rho \frac{d}{d\rho} \sum_{n=1}^{\infty} \rho^n \prod_{i=2}^{n} \left(1 + \frac{\lambda}{\eta i}\right) \\
&= p_{1,0} \left(1 + \frac{\lambda}{\eta}\right) \rho \frac{d}{d\rho} \sum_{n=1}^{\infty} \rho^{n-1} \prod_{i=1}^{n-1} \left(1 + \frac{\lambda}{\eta i}\right) \\
&= p_{1,0} \left(1 + \frac{\lambda}{\eta}\right) \rho \frac{d}{d\rho} (1-\rho)^{-1-\lambda/\eta} \\
&= p_{1,0} \left(1 + \frac{\lambda}{\eta}\right) \rho (1-\rho)^{-2-\lambda/\eta} \\
&= \frac{\rho^2}{1-\rho} \left(1 + \frac{\lambda}{\eta}\right)
\end{aligned}$$

and

$$\begin{aligned}
\sum_{n=1}^{\infty} n p_{0,n} &= \sum_{n=1}^{\infty} \frac{n\mu}{(\lambda + n\eta)} p_{1,n} = \sum_{n=1}^{\infty} \frac{\mu}{\eta} \left(\frac{\lambda + n\eta}{n\eta}\right)^{-1} p_{1,n} \\
&= \frac{\mu}{\eta} p_{1,0} \sum_{n=1}^{\infty} \left(1 + \frac{\lambda}{n\eta}\right)^{-1} \rho^n \prod_{i=1}^{n} \left(1 + \frac{\lambda}{\eta i}\right) \\
&= p_{1,0} \frac{\mu}{\eta} \rho \left(1 + \sum_{n=2}^{\infty} \rho^{n-1} \prod_{i=1}^{n-1} \left(1 + \frac{\lambda}{\eta i}\right)\right) \\
&= p_{1,0} \frac{\lambda}{\eta} \left(1 + \sum_{n=1}^{\infty} \rho^{n} \prod_{i=1}^{n} \left(1 + \frac{\lambda}{\eta i}\right)\right) \\
&= p_{1,0} \frac{\lambda}{\eta} (1-\rho)^{-1-\lambda/\eta} = \frac{\lambda}{\eta} \rho
\end{aligned}$$

one easily derives the expected number of customers in orbit

$$\begin{aligned}
L_q &= \sum_{n=1}^{\infty} n \left(p_{0,n} + p_{1,n}\right) = \frac{\rho^2}{1-\rho} \left(1 + \frac{\lambda}{\eta}\right) + \frac{\lambda}{\eta} \rho \\
&= \frac{\rho (\rho\eta + \lambda)}{\eta (1-\rho)} = \frac{\rho^2}{1-\rho} + \frac{\lambda\rho}{\eta (1-\rho)} = \frac{\lambda\rho}{\eta (1-\rho)} + L_q^{M/M/1} \qquad (4.12)
\end{aligned}$$

By comparing the result with expression 2.5 for the $M/M/1$ model, one identifies the second term as some kind of *expected excess* in the number of customers [30]. Letting η approach infinity, the excess vanishes and we arrive at an ordinary $M/M/1$ queueing model. There is no delay between subsequent retries and so the orbit attaches as queue to the $M/M/1/1$ loss model. The remaining performance characteristics are determined by an application of Little's law

$$\begin{aligned} W_q &= \frac{1}{\lambda} L_q = \frac{\lambda(\rho\eta + \lambda)}{\eta(1-\rho)} \\ L &= L_q + \rho = \frac{\rho(\rho\eta + \lambda) + \rho\eta(1-\rho)}{\eta(1-\rho)} = \frac{\rho(\eta+\lambda)}{\eta(1-\rho)} \\ W &= \frac{1}{\lambda} L = \frac{\eta + \lambda}{\eta\mu(1-\rho)} \end{aligned}$$

It turns out, that the single server retrial system is stable for $\rho < 1$ [20]. One may also derive the conditional average waiting time in orbit for an arriving customer given a busy server. By realizing, that the arrival rate to the orbit is $\lambda\rho$ and using Little's law yields

$$\mathbb{E}\left\{\breve{W}_q | \breve{W}_q > 0\right\} = \frac{1}{\lambda\rho} L_q = \frac{1}{1-\rho}\left(\frac{1}{\mu} + \frac{1}{\eta}\right)$$

Again we detect an excess to the average queueing time of the $M/M/1$ model. A slightly different approach to the one presented here is given in [20] by using a generating function approach to derive the main performance characteristics. Falin and Templeton also present results for the variance of the average number of customers in orbit and in system. Their results are stated here for completeness without proof

$$\begin{aligned} \sigma_L^2 &= \frac{\rho(\eta+\lambda)}{\eta(1-\rho)^2} \\ \sigma_{L_q}^2 &= \frac{\rho(\rho\eta + \rho^2\eta - \rho^3\eta + \lambda)}{\eta(1-\rho)^2} \end{aligned}$$

The calculation of the variance of the average waiting time is not straightforward, as customers may overtake each other randomly in orbit. For a detailed analysis on the waiting time distribution we refer to their book [20].

The multiserver case may be approached by letting $C(t)$ assume values between 0 and c. Proceeding as usual leads to the following system of equations for the steady state probabilities

$$\begin{aligned}\left(\rho + m + \frac{\eta}{\mu}\right) p_{m,n} &= \rho p_{m-1,n} + (m+1)\, p_{m+1,n} + \frac{\eta}{\mu}(n+1)\, p_{m-1,n+1} \\ (\rho + c)\, p_{c,n} &= \rho p_{c-1,n} + \rho p_{c,n-1} + \frac{\eta}{\mu}(n+1)\, p_{c-1,n+1}\end{aligned}$$

Please note, that we have divided the equations by μ for convinience. Introducing the partial generating function $P_m(z) = \sum_{n=0}^{\infty} z^n p_{m,n}$ for $0 \leqq m \leqq c$ and $|z| < 1$, the above system of equation may be written as [23]

$$\begin{aligned}(\rho + m)\, P_m(z) + \frac{\eta}{\mu} z \frac{d}{dz} P_m(z) &= \rho P_{m-1}(z) + (m+1)\, P_{m+1}(z) + \frac{\eta}{\mu} z \frac{d}{dz} P_{m-1}(z) \\ (\rho + c)\, P_c(z) &= \rho P_{c-1}(z) + \rho z P_c(z) + \frac{\eta}{\mu} \frac{d}{dz} P_{c-1}(z)\end{aligned}$$

Introducing the bivariate generating function $P(y,z) = \sum_{m=0}^{c} y^m P_m(z)$ and repeating the step above yields

$$\begin{aligned}&\rho\,(1-y)\, P(y,z) + \tfrac{\eta}{\mu}(z-y)\, \tfrac{\partial}{\partial z} P(y,z) + (y-1)\, \tfrac{\partial}{\partial y} P(y,z) \\ &+ \rho y^c\,(y-z)\, P_c(z) + \tfrac{\eta}{\mu}(y-z)\, \tfrac{d}{dz} P_{c-1}(z) = 0\end{aligned}$$

Differentiating with respect to z, y, yy, yz, zz at the point $y = 1$, $z = 1$ leads to a system of equations, which can be solved in terms of $L_q = \frac{d}{dz} P(1,1)$, the blocking probability $p_b = P_c(1)$, the utilization $u = \mathbb{E}C(t) = \frac{d}{dy} P(1,1)$ and other variables. Simplification yields

$$\begin{aligned}u &= \rho \\ L_q &= \left(1 + \frac{\eta}{\mu}\right) \frac{\rho - \sigma_C^2}{c - \rho}\end{aligned} \tag{4.13}$$

The detailed calculations have been omitted for sake of readability, they do not provide any further insight into the problem. The interested reader is referred to the book of Falin and Templeton [20]. It can be shown, that the multiserver retrial system is stable for $u < 1$. As will be shown next, the model is bounded from above by the classic $M/M/c$ queueing system and so the stability condition adheres to intuition.

This is as far as one can get with exact techiques. Closed form solutions only exist in the case of one or two servers [20], for $c \geqq 3$ the average number of customers in orbit depends on the variance of the number of busy servers σ_C^2. In the extreme for $\eta \to \infty$ the retrial model approaches the classic $M/M/c$ queueing system, whereas for $\eta = 0$ it reduces to an Erlang loss system. This allows for an approximation for high and low retrial rates η. In the former case the blocking probability p_b is approximated by the probability of delay $p_d^{(M/M/c)}$ for the $M/M/c$ queue given by expression 3.6. The same applies to the average queue length, i.e. $L_q \approx L_q^{(M/M/c)}$.

For η small the Erlang loss formula 3.10 with traffic intensity $\bar{\rho} = \frac{\lambda+r}{\mu}$ and c servers provides a starting point for an approximation. Hereby we assume, that the unknown retrial arrival rate r does not depend on the number of busy servers. It is easy to verify, that the Erlang loss formula constitutes a distribution allowing us to calculate mean and variance. For the purposes of the current section we will denote it by $E(\bar{\rho}, c)$, where $\bar{\rho}$ and c are the parameters. Keeping in mind, that this distribution describes the random variable *busy servers*, its expectation must equal the utilization of the retrial system, i.e. $u = \rho = \bar{\rho}\,(1 - E(\bar{\rho}, c))$ leading to $E(\bar{\rho}, c) = 1 - \rho/\bar{\rho} = \frac{r}{\lambda+r}$.

Similar considerations yield $\sigma_b^2 = \rho - (c - \rho)\,(\bar{\rho} - \rho)$, the variance of the random variable busy servers. Returning to the high rate approximation, the same idea may be applied to the Erlang delay formula leading to an approximation of the variance $\sigma_d^2 = \rho \left(1 - p_d^{(M/M/c)}\right)$. Although both variances are related to the number of busy servers, we kept the suffixes to show the origin of the formulas.

For intermediate values of η, the most straightforward way to provide an approximation is via interpolation, i.e.

$$\begin{aligned} p_b &\approx \frac{1}{1+\frac{\eta}{\mu}} E(\bar{\rho}, c) + \frac{\frac{\eta}{\mu}}{1+\frac{\eta}{\mu}} p_d^{(M/M/c)} \\ \sigma_C^2 &\approx \frac{1}{1+\frac{\eta}{\mu}} \sigma_b^2 + \frac{\frac{\eta}{\mu}}{1+\frac{\eta}{\mu}} \sigma_d^2 \end{aligned}$$

Inserting the expression for σ_C^2 into formula 4.13 yields

$$\begin{aligned}
L_q &= \left(1+\frac{\eta}{\mu}\right)\frac{\rho-\sigma_C^2}{c-\rho} = \left(1+\frac{\eta}{\mu}\right)\frac{\left(1+\frac{\eta}{\mu}\right)\rho-\sigma_b^2-\frac{\eta}{\mu}\sigma_d^2}{(c-\rho)\left(1+\frac{\eta}{\mu}\right)} \\
&\approx \frac{\left(1+\frac{\eta}{\mu}\right)\rho-\rho+(c-\rho)\left(\bar{\rho}-\rho\right)-\frac{\eta}{\mu}\rho\left(1-p_d^{(M/M/c)}\right)}{c-\rho} \qquad (4.14)\\
&= \frac{(c-\rho)\frac{r}{\mu}+\frac{\eta}{\mu}\rho p_d^{(M/M/c)}}{c-\rho} = \frac{r}{\mu}+\frac{\eta}{\mu}\frac{\rho}{c-\rho}p_d^{(M/M/c)} \\
&= \frac{r}{\mu}+\frac{\eta}{\mu}L_q^{(M/M/c)}
\end{aligned}$$

where the unknown quantity r is calulated from $E(\frac{\lambda+r}{\mu},c)=\frac{r}{\lambda+r}$ for given values of λ, μ and c. Although there is an appealing relation to the $M/M/c$ queue, it is in general not additive as one would expect from the single server case. The remaining performance characteristics may be determined from an application of Little's law, i.e.

$$W_q=\frac{1}{\lambda}L_q, \quad W=W_q+\frac{1}{\mu}, \quad L=\lambda W=L_q+\rho$$

Some of the ideas presented here have to be attributed to R.I. Wilkinson [45], but the most complete reference in the field is the book by Falin and Templeton [20]. Fayolle and Brun have treated a model with customer impatience and repeated calls in their paper [21]. Their model is rather cumbersome and difficult to analyze.

Chapter 5

Arbitrary Service Processes

In certain cases the Poisson assumption is not appropriate. One has to consider non-Poisson arrival and service processes. The current section will provide some results on queueing systems with rather arbitrary distributions. If either one of the distributions is Markovian, the corresponding queue may be analyzed by embedding a discrete time Markov chain (see appendix A.3). This method introduced by Kendall suggests the system to be modeled by viewing it only at times where the Markov property holds. By careful selection of the regeneration points, one is able to nullify the impact of residual service and residual interarrival time.

5.1 Single Server Systems

Consider a single server queue with Poissonian arrivals at rate λ and arbitrary (absolute continuous) service distribution $B(.)$ with average service time $\frac{1}{\mu}$ and finite variance. As regeneration points choose the instance at which customers complete service and depart from the system. At that time either a waiting customer commences service or the system becomes idle. More exact, the residual life time is zero, but the customer has not left the system yet. Define $\bar{b}(s) = \int_0^\infty b(x)e^{-sx}dx$ as the Laplace transform of the service density $b(.)$ corresponding to $B(.)$. Due to the Poissonian nature of the arrival process, the probability of n arrivals between two successive departures within an interval of length t is given as

$$q_{n|t} = \frac{e^{-\lambda t}\,(\lambda t)^n}{n!} \tag{5.1}$$

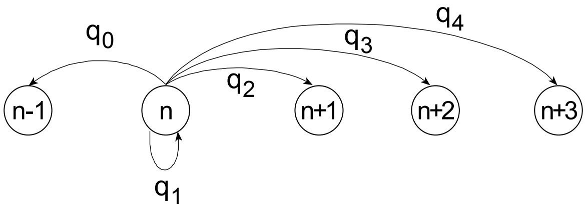


Figure 5.1: Embedded Markov chain for the M/G/1 queue

Averaging yields the probability of n arrivals within an interval of length t, that is

$$q_n = \int_0^\infty q_{n|t} b(t) dt$$

We are now able to assemble the matrix of transition probabilities

$$\mathbf{Q} := \begin{pmatrix} q_0 & q_1 & q_2 & \cdots \\ q_0 & q_1 & q_2 & \cdots \\ 0 & q_0 & q_1 & \cdots \\ 0 & 0 & q_0 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

The number of customers in the queue is decreased by one, if no customer arrives in the observed interval. If only one customer arrives, the system remains in state n. Otherwise the process can reach any state $k > n$. We say, the embedded Markov chain is *skip-free to the left*. Also refer to figure 5.1 for a graphical representation of the possible state transitions out of state n. As shown in appendix A.3, the equilibrium distribution $\mathbf{p}$ is found by solving $\mathbf{pQ} = \mathbf{p}$ and normalizing. Reverting to classic notation leads to

$$p_n = p_0 q_n + \sum_{i=0}^{n+1} p_i q_{n-i+1} = p_0 q_n + \sum_{i=0}^{n} p_{i+1} q_{n-i}, \qquad n \geq 0 \qquad (5.2)$$

By the use of generating functions

$$\begin{aligned}
P(z) &= \sum_{i=0}^{\infty} p_i z^i, \qquad |z| < 1 \\
P_+(z) &= \sum_{i=0}^{\infty} p_{i+1} z^i = z^{-1} \sum_{i=0}^{\infty} p_{i+1} z^{i+1} \\
&= -z^{-1} p_0 + z^{-1} \sum_{i=-1}^{\infty} p_{i+1} z^{i+1} \\
&= -z^{-1} p_0 + z^{-1} \sum_{i=0}^{\infty} p_i z^i = z^{-1} \left(P(z) - p_0\right) \\
Q(z) &= \sum_{i=0}^{\infty} q_i z^i = \sum_{i=0}^{\infty} z^i \int_0^{\infty} q_{i|t} b(t) dt \\
&= \sum_{i=0}^{\infty} z^i \int_0^{\infty} \frac{e^{-\lambda t} \left(\lambda t\right)^i}{i!} b(t) dt = \bar{b} \left(\lambda \left(1 - z\right)\right) \qquad (5.3)
\end{aligned}$$

the system of equations 5.2 may be written as

$$P(z) = p_0 Q(z) + P_+(z) Q(z) = p_0 Q(z) + z^{-1} \left(P(z) - p_0\right) Q(z)$$

Rearranging

$$P(z) \left(1 - z^{-1} Q(z)\right) = p_0 \left(1 - z^{-1}\right) Q(z)$$

finally leads to

$$P(z) = \frac{p_0 \left(1 - z^{-1}\right) Q(z)}{1 - z^{-1} Q(z)} = \frac{p_0 \left(1 - z\right) Q(z)}{Q(z) - z} \qquad (5.4)$$

To determine p_0 one needs to apply the properties of moment generating functions [23], that is

$$P(1) = 1, \quad Q(1) = 1, \quad \frac{d}{dz} Q(z)|_{z=1} = \frac{\lambda}{\mu} = \rho$$

The last equation stems from the fact, that ρ is the expected value for a transition to occur. Rearranging expression 5.4 and letting z approach 1 yields

$$p_0 = \lim_{z \to 1} \frac{P(z) \left(Q(z) - z\right)}{\left(1 - z\right) Q(z)}$$

Applying L'Hospital's rule provides us with the desired result

$$p_0 = 1 - \rho \tag{5.5}$$

Hence equation 5.4 now becomes

$$P(z) = \frac{(1-\rho)\,(1-z)\,Q(z)}{Q(z) - z} \tag{5.6}$$

To obtain a solution to equation 5.6 and derive the steady state distribution one needs to invert the transforms involved. This is not possible without assuming a specific service distribution $B(.)$. Even then, it is questionable, if the desired inversion can be carried out. From a theoretical standpoint, one has only to isolate the coefficients of z^i in the series $P(z)$. Strictly speaking, we have only determined the system size distribution as seen by arrivals. Note, that in a stable system, the system size distribution seen by departures equals the one seen by arrivals. We are now able to apply theorem 10 (PASTA) to see, that the latter are equal to the system size distribution [64]. Following that daisy chain, we conclude, that expression 5.6 sufficiently describes the distribution of the number of customers in system. To determine the average system size L, one once again makes use of the properties of generating functions [23]:

$$L = \frac{d}{dz} P(z)|_{z=1}$$

It turns out, that this operation also might become rather cumbersome. Fortunately a more direct approach exists and will be presented next.

Let A_k and D_k denote the number of customers entering and leaving the system immediately after the k-th departure has occured. Then the following recurrence relation for the number of customers in the system N_k may be observed

$$N_{k+1} = \begin{cases} N_k - 1 + A_{k+1} & \text{for } N_k > 0 \\ A_{k+1} & \text{for } N_k = 0 \end{cases} \tag{5.7}$$

By introducing an auxillary function

$$U\left(N_k\right) = \begin{cases} 1 & \text{for } N_k > 0 \\ 0 & \text{for } N_k = 0 \end{cases}$$

relation 5.7 may be simplified to

$$N_{k+1} = N_k - U\left(N_k\right) + A_{k+1} \tag{5.8}$$

Taking expectation gives

$$\mathbb{E}N_{k+1} = \mathbb{E}N_k - \mathbb{E}U\left(N_k\right) + \mathbb{E}A_{k+1}$$

By realizing, that the number of customers after the k-th departure N_k does not depend on k, both terms may be eliminated from the above equation. This leads to

$$\mathbb{E}U\left(N_k\right) = \mathbb{E}A_{k+1}$$

Another expression for $U\left(N_k\right)$ may be derived by averaging over all possible values, i.e.

$$\mathbb{E}U\left(N_k\right) = \sum_{n=0}^{\infty} U\left(N_k\right) \Pr\left\{N_k = n\right\} = \sum_{n=1}^{\infty} \Pr\left\{N_k = n\right\}$$

The last term on the right is simply the probability of the server being busy. Recall from equation 5.5, that the probability of an idle system is given by $p_0 = 1 - \rho$. Putting it all together results in

$$\mathbb{E}U\left(N_k\right) = \mathbb{E}A_{k+1} = 1 - p_0 = \rho \tag{5.9}$$

As a next step, equation 5.8 has to be squared

$$N_{k+1}^2 = N_k^2 + U^2\left(N_k\right) + A_{k+1}^2 - 2N_kU\left(N_k\right) - 2A_{k+1}U\left(N_k\right) + 2N_kA_{k+1}$$

Taking expectation and eliminating terms $\mathbb{E}N_{k+1}^2 = \mathbb{E}N_k^2$ yields

$$0 = \mathbb{E}U^2\left(N_k\right) + \mathbb{E}A_{k+1}^2 - 2\mathbb{E}\left(N_kU\left(N_k\right)\right) - 2\mathbb{E}\left(A_{k+1}U\left(N_k\right)\right) + 2\mathbb{E}\left(N_kA_{k+1}\right) \tag{5.10}$$

Notice the two obvious relations $U^2\left(N_k\right) = U\left(N_k\right)$ and $N_kU\left(N_k\right) = N_k$, which follow immediately from the definition. By independence of A_{k+1} and N_k, one may write $\mathbb{E}\left(N_kA_{k+1}\right) = \mathbb{E}N_k\mathbb{E}A_{k+1} = \rho^2$. Substitution of these idendities in expression 5.10 gives

$$0 = \rho + \mathbb{E}A_{k+1}^2 - 2L - 2\rho^2 + 2\rho L \tag{5.11}$$

As the arrivals follow a Poisson process, the number of arrivals occuring in the interval between two subsequent departures depends only on the length of the interval, not on the interval itself. Consequently the index may be omitted, i.e. $\mathbb{E}A^2 := \mathbb{E}A_{k+1}^2$. Next, the second moment is expressed in terms of expectation and variance, that is $\mathbb{E}A^2 = Var(A) + \left(\mathbb{E}A\right)^2 = Var(A) +$

ρ^2. With S denoting the service time (with $\mathbb{E}S = \frac{1}{\mu}$), one may split the variance to $Var(A) = \mathbb{E}\left(Var(A|S)\right) + Var\left(\mathbb{E}\left(A|S\right)\right) = \mathbb{E}\left(\lambda S\right) + Var\left(\lambda S\right)$. By denoting the variance of S as σ_S^2 and recalling $\mathbb{E}\left(\lambda S\right) = \lambda \mathbb{E}\left(S\right) = \frac{\lambda}{\mu} = \rho$ one arrives at the expression $Var\left(A\right) = \rho + \lambda^2 Var(S) = \rho + \lambda^2 \sigma_S^2$ leading to

$$\mathbb{E}A^2 = \rho + \lambda^2 \sigma_S^2 + \rho^2$$

Substituting in equation 5.11 yields

$$0 = 2\rho - \rho^2 + \lambda^2 \sigma_S^2 + 2\left(1-\rho\right)L$$

Rearranging terms finally results in

$$L = \frac{2\rho - \rho^2 + \lambda^2 \sigma_S^2}{2\left(1-\rho\right)} = \rho + \frac{\rho^2 + \lambda^2 \sigma_S^2}{2\left(1-\rho\right)} \tag{5.12}$$

The above result 5.12 is often refered to as *Pollaczek-Khintchine Formula* and enables us to derive the remaining performance characteristics by applying Little's law. This leads to

$$\begin{aligned} W &= \frac{1}{\lambda}L = \frac{1}{\mu} + \frac{\rho^2 + \lambda^2 \sigma_S^2}{2\lambda\left(1-\rho\right)} \\ W_q &= W - \frac{1}{\mu} = \frac{\rho^2 + \lambda^2 \sigma_S^2}{2\lambda\left(1-\rho\right)} \\ L_q &= \lambda W_q = \frac{\rho^2 + \lambda^2 \sigma_S^2}{2\left(1-\rho\right)} \end{aligned}$$

For further reading, we again refer to classic textbooks on queueing theory. We were mainly led by [27]. Similar derivations may be found also in [25] and [64]. For more advanced approaches consider [4] and [46]. Due to the vast amount of results on the $M/G/1$ model, we had to omit some of them. These include for example waiting time distributions, batch arrivals and priorities. Much of the material not covered may be found in the given references.

Example 14 *Assume an exponential distribution for the service times, i.e. we specialize on the $M/M/1$ model. Applying the Laplace transform of the* *exponential density* $\bar{b}(s) = \frac{\mu}{\mu+s}$ *to expression 5.3 and 5.6 results in*

$$\begin{aligned} P(z) &= \frac{(1-\rho)(1-z)\frac{\mu}{\mu+\lambda(1-z)}}{\frac{\mu}{\mu+\lambda(1-z)} - z} \\ &= \frac{(1-\rho)(1-z)\mu}{\mu - z\mu - z\lambda(1-z)} \\ &= \frac{(1-\rho)\mu}{\mu - z\lambda} = \frac{(1-\rho)}{1-z\rho} \end{aligned}$$

Expanding the result in a geometric series, i.e.

$$(1-\rho)\frac{1}{1-z\rho} = (1-\rho)\sum_{n=0}^{\infty} z^n \rho^n$$

and isolating the coefficients of z^n *leads to the steady state distribution*

$$p_n = (1-\rho)\rho^n$$

which equals expression 2.3. We have shown, that the $M/M/1$ *model fits perfectly in the framework of the more general* $M/G/1$ *queueing system. To derive a result for the average system size we first note, that exponential distributed service times* S *with rate* μ *have mean* $\mathbb{E}S = \frac{1}{\mu}$ *and variance* $\sigma_S^2 = \frac{1}{\mu^2}$. *Substituting these values in the Pollaczek-Khintchine formula 5.12 leads to*

$$L = \frac{2\rho - \rho^2 + \lambda^2\frac{1}{\mu^2}}{2(1-\rho)} = \frac{\rho}{1-\rho}$$

Comparing the result to expression 2.4 shows the expected result.

Example 15 *Now consider deterministic service times, i.e. we specialize to the* $M/D/1$ *model. To gain results one usually has to employ integro-differential equations. By applying the results for the* $M/G/1$ *model we are able to significantly reduce the mathematical effort necessary. The deterministic distribution is in some sense malformed, as there is only a single point with mass 1, i.e.*

$$b(x) = \delta\left(x - \frac{1}{\mu}\right)$$

Here the function $\delta(z)$ *describes the Kronecker function*

$$\delta(z) = \begin{cases} 0 & z \neq 0 \\ 1 & z = 0 \end{cases}$$

The Laplace transform of the density is given by

$$b(s) = e^{-\frac{1}{\mu}}$$

Applying to expression 5.3 and 5.6 leads to

$$\begin{aligned} P(z) &= \frac{(1-\rho)\,(1-z)\,e^{-\lambda(1-z)/\mu}}{e^{-\lambda(1-z)/\mu}-z} \\ &= \frac{(1-\rho)\,(1-z)}{1-ze^{\rho(1-z)}} \end{aligned}$$

Expanding in a geometric series

$$P(z) = (1-\rho)\,(1-z)\sum_{n=0}^{\infty} z^n e^{n\rho(1-z)}$$

and expressing the exponential function in an exponential series allows one to isolate the coefficients of z^n:

$$\begin{aligned} p_0 &= 1-\rho \\ p_1 &= (1-\rho)\,(e^{\rho}-1) \\ p_n &= (1-\rho)\sum_{i=0}^{n}\frac{(-i\rho)^{n-i}\,e^{i\rho}}{(n-i)!}-\sum_{i=0}^{n-1}\frac{(-i\rho)^{n-i-1}\,e^{i\rho}}{(n-i-1)!} \end{aligned} \tag{5.13}$$

Compared to the calculation so far, the derivation of the average system size gets even simpler than for the $M/M/1$ model. As there is no variation in the model, i.e. $\sigma_S^2 = 0$, the Pollaczek-Khintchine formula 5.12 immediately becomes

$$L = \rho + \frac{\rho^2}{2\,(1-\rho)} \tag{5.14}$$

Rewriting expression 5.14 reveals an interesting relation between the $M/D/1$ and the $M/M/1$ model:

$$L = \frac{\rho}{1-\rho} - \frac{\rho^2}{2\,(1-\rho)} = L^{(M/M/1)} - \frac{\rho^2}{2\,(1-\rho)}$$

It turns out, that given the same parameters the number of customers is always smaller for systems with deterministic service times. In case of heavy *traffic, i.e. $\rho \to 1$, the system size of the $M/M/1$ model is twice the size of the $M/D/1$ queueing system:*

$$\lim_{\rho \to 1} L = \frac{1}{2} \lim_{\rho \to 1} L^{(M/M/1)}$$

For sake of readability, some calculations have been omitted. The detailed calculations may be found in [49].

By introducing the coefficient of variation $c_S = \frac{\sqrt{Var(S)}}{\mathbb{E}S} = \mu\sigma_S$, one may think of using it as a control parameter to interpolate between deterministic ($c_S = 0$) and exponential service times ($c_S = 1$) . In fact, this is possible by reinterpreting the Pollaczek-Khintchine formula 5.12 as follows

$$\begin{aligned}
L &= \frac{2\rho - \rho^2 + \rho^2 c_S^2}{2(1-\rho)} = \frac{2\rho - \rho^2 - \rho^2 c_S^2 + 2\rho c_S^2 - 2\rho c_S^2}{2(1-\rho)} \\
&= c_S^2 \frac{2\rho}{2(1-\rho)} + \left(1 - c_S^2\right) \frac{2\rho - \rho^2}{2(1-\rho)} \\
&= c_S^2 \frac{\rho}{1-\rho} + \left(1 - c_S^2\right) \frac{2\rho - \rho^2}{2(1-\rho)} \\
&= c_S^2 L^{(M/M/1)} + \left(1 - c_S^2\right) L^{(M/D/1)} \qquad (5.15)
\end{aligned}$$

By Little's law the same convex combination may also be applied to the other performance characteristics. Although this interpretation is not very appealing in its own sense, it becomes of great interest for the approximation of multiserver limited capacity systems. In fact, it will turn out, that the idea extends to the most general models.

5.2 Finite Single Server Systems

In this section we will get in touch with finite source and limited capacity systems. We will provide some ideas and summarize the major results. For sake of readability lengthy discussions will be omitted, the details missing may be found in the given literature. First consider the $M/G/1/K$ model. As before for the exponential version a limit of K customers is allowed and customers arriving at a full system are turned away. The service time S follows an arbitrary distribution $B(.)$ with expectation $\frac{1}{\mu}$ and finite variance.

No stable system needs to be assumed, as the finite waiting room provides an upper limit for the number of customers in the system. It can be shown [25][58], that the steady state distribution of $M/G/1/K$ system is proportional to the stationary solution of a stable $M/G/1$ queue given the same parameter. The latter is sufficiently described by expression 5.6 and will be denoted as $p_n^{(M/G/1)}$. Following Gnedenko and Kovalenko [25], the first K probabilities are given by

$$\begin{aligned} p_n &= \kappa p_n^{(M/G/1)}, \qquad 0 \leq n < K \\ \kappa &= \left[1 - \rho + \rho \sum_{n=0}^{K-1} p_n^{(M/G/1)} \right]^{-1} \end{aligned} \tag{5.16}$$

Applying the usual normalization condition $\sum_{n=0}^{K-1} p_n$ leads to the expression for p_K, which is also the probability of being blocked

$$p_b = p_K = \kappa \left[1 - \rho + (\rho - 1) \sum_{n=0}^{K-1} p_n^{(M/G/1)} \right] \tag{5.17}$$

It turns out, that a similar relation also exists, when the infinite system becomes unstable. More details may be found in [25]. Having calculated the blocking probability we are now in the position to derive an expression for the effective arrival rate,

$$\bar{\lambda} = \lambda \left(1 - p_K\right)$$

From the steady state distribution, the average number of customers in the system may be determined by

$$L = \sum_{n=0}^{K} n p_n$$

Application of Little's law yields the remaining performance characteristics

$$\begin{aligned} W &= \frac{L}{\bar{\lambda}} = \frac{L}{\lambda \left(1 - p_K\right)} \\ W_q &= W - \frac{1}{\mu} \\ L_q &= \bar{\lambda} W_q = \lambda \left(1 - p_K\right) W_q \end{aligned}$$

Alternatively one may follow the embedded Markov chain approach as has been done before for the $M/G/1$ queueing system. This leads to a finite state Markov chain in discrete time. Each of the results given above may then be determined in terms of the corresponding stationary solution. For further details we refer to one of the most complete references on the $M/G/1/K$ model available, that is [58].

We now proceed to the finite population $M/G/1$ queueing system, also referred to as *machine-repairman system*. The working machines are associated with the source and the broken machines form a queue waiting for repair. In our case a single repairman is available to perform the job. The average time in system then becomes the expected machine outage time. Based on such key indicators the cost of operating a production business may be inferred. We will now present some results derived by Takagi in [58] without proof. Consider a system with population size N and the other parameters defined as in the $M/G/1$ model. Given the Laplace transform $\bar{b}(s)$ of the service time density, the mean arrival rate is given by

$$\bar{\lambda} = \frac{N\lambda\left[1 + \sum_{n=1}^{N-1}\binom{N-1}{n}\prod_{i=1}^{n}\left(\bar{b}^{-1}(i\lambda) - 1\right)\right]}{1 + N\rho\left[1 + \sum_{n=1}^{N-1}\binom{N-1}{n}\prod_{i=1}^{n}\left(\bar{b}^{-1}(i\lambda) - 1\right)\right]}$$

From the expression for the average time in system

$$W = \left[1 + \sum_{n=1}^{N-1}\binom{N-1}{n}\prod_{i=1}^{n}\left(\bar{b}^{-1}(i\lambda) - 1\right)\right]^{-1} + \frac{N}{\mu} - \frac{1}{\lambda}$$

the remaining performance characteristics may be derived by applying Little's law

$$\begin{aligned} L &= \bar{\lambda}W = N - \frac{\bar{\lambda}}{\lambda} \\ W_q &= W - \frac{1}{\mu} \\ L_q &= \bar{\lambda}W_q \end{aligned}$$

From the results of the two above models one can image, that the corresponding calculations quickly become cumbersome. Both models occur relatively rare in queueing literature and are completely omitted in standard queueing theory textbooks. An exception to the rule is [58], where all the necessary details are to be found.

5.3 Multiserver Systems

In generalizing to more servers, we loose all the powerful tools used so far. In fact, the $M/G/c$ queueing system does not permit a simple analytical solution. We can not apply the method of embedded Markov chains the usual way and there is no such relation as the Pollaczek-Khintchine formula. The only item left in our toolbox is Little's law. Before obtaining some approximations for the $M/G/c$ model, we note, that rather simple solutions exist for two special cases. The first is the $M/G/c/c$ model already discussed on page 3.2. The second is the so called infinite server queue $M/G/\infty$, which derives from the $M/G/c/c$ model by allowing c to become infinite. This immediately leads to the stationary distribution

$$p_n = \frac{e^{-\rho}\rho^n}{n!}$$

The remaining results may be determined in the same fashion [27].

One of the simplest approximations to be obtained is based on a generalization of the idea, which led us to expression 5.15. By considering the result to be valid for multiple servers as well, one arrives at

$$L \approx c_S^2 L^{(M/M/c)} + \left(1 - c_S^2\right) L^{(M/D/c)} \tag{5.18}$$

The expression for $L^{(M/M/c)}$ may be determined from formula 3.8. As shown in [49] and [44], the generating function for system size distribution of the $M/D/c$ model is given by

$$P(z) = \frac{\sum_{n=0}^{c} p_n \left(z^n - z^c\right)}{1 - z^c e^{\rho(1-z)}}, \quad |z| < 1 \tag{5.19}$$

By realizing, that the numerator is a polynomial of degree c and rewriting expression 5.19 after some manipulations yields [49]

$$P(z) = -\frac{c-\rho}{(1-z_1)\cdots(1-z_{c-1})} \frac{(z-1)(z-z_1)\cdots(z-z_{c-1})}{1 - z^c e^{\rho(1-z)}}$$

The $z_1, \ldots, z_c$ are the zeros of the numerator within the unit circle, that is $z_n = \{z : |z| < 1 \text{ and } P(z) = 0\}$, $0 < n \leq c$. The last zero is always given by $z_c = 1$. Note, that Rouche's theorem 12 assures, that $z_1, \ldots, z_c$ are found within the unit circle. The probabilities p_n may now be obtained as the coefficient of z^n in the power series expansion of $P(z)$ for a given number of servers c. The average system size may be determined by applying the well known property of generating functions $L = \frac{d}{dz}P(z)|_{z=1}$ to expression 5.19. Some further algebra finally leads to

$$L^{(M/D/c)} = \frac{\rho^2 - c\,(c-1) + \sum_{n=0}^{c-1} \left[c(c-1) - n(n-1)\right] p_n^{(M/D/c)}}{2c\,(1-\rho/c)} + \rho$$

Due to the fact, that the $p_n^{(M/D/c)}$ have not vanished from the expression above, the derivation of an exact solution still remains a rather tedious task. Fortunately an approximation developed by Cosmetatos has been suggested in [60]:

$$L^{(M/D/c)} \approx \frac{L_q^{(M/M/c)}}{2}\left[1 + \left(1 - \frac{\rho}{c}\right)(c-1)\,\frac{\sqrt{4+5c}-2}{16\rho}\right] + \rho \tag{5.20}$$

By applying a regenerative approach and including information about the elapsed service time into the model, van Hoorn was able to deduce another approximation. Consider the following assumptions

- The residual service times are independent random variables each with residual life distribution
$$B_r\,(t) = \mu \int_0^t \left(1 - B(x)\right) dx$$
- Given a full system, the time until the next departure has distribution function $B\,(ct)$. Thus the $M/G/c$ queue is treated as $M/G/1$ queue with rate $c\mu$.

By applying the following recursion scheme, one arrives at an approximation for the steady state distribution

$$\begin{aligned}
p_n &\approx \begin{cases} \left[\sum_{n=0}^{c-1} \frac{\rho^n}{n!} + \frac{\rho^c}{c!(1-\rho/c)}\right]^{-1} & n = 0 \\ \frac{\rho^n}{n!} p_0 & n < 0 < c \\ \lambda\left(\alpha_{n-c} p_{c-1} + \sum_{i=c}^{n} \beta_{n-i} p_i\right) & n \geq c \end{cases} \qquad (5.21) \\
\alpha_n &= \int_0^\infty \left(1 - B_r(x)\right)^{c-1} \left(1 - B(x)\right) e^{-\lambda x} \frac{(\lambda x)^n}{n!} dx, \quad n \geq 0 \\
\beta_n &= \int_0^\infty \left(1 - B(cx)\right) e^{-\lambda x} \frac{(\lambda x)^n}{n!} dx, \quad n \geq 0
\end{aligned}$$

Proceeding further, van Hoorn was able to determine the expected queue length

$$L_q = \left[(c\mu - \lambda) \int_0^\infty (1 - B_r(x))^c \, dx + \frac{\lambda\mu}{c} \mathbb{E}S^2 \right] L_q^{(M/M/c)} \tag{5.22}$$

For all approximation presented above, the remaining performance indicators may be determined by the help of Little's law and the usual relations

$$L = L_q + \rho, \quad W_q = \lambda L_q = L + \frac{1}{\mu}, \quad W = \lambda L$$

Now consider a system with multiple servers and waiting room limitation K. By combining the results for the $M/G/1/K$ model with the approximation above, van Hoorn was able to derive reasonable approximations for the $M/G/c/K$ queueing system. Denoting the probabilities for the infinite server system given by 5.21 with $p_n^{(M/G/c)}$, the corresponding probabilities for the limited capacity system are given by

$$\begin{aligned} p_n &\approx \kappa p_n^{(M/G/c)}, \qquad 0 \le n < K \\ \kappa &= \left[1 - \rho + \rho \sum_{n=0}^{K-1} p_n^{(M/G/c)} \right]^{-1} \end{aligned}$$

Please note the similarity to expression 5.16 obtained for the single server system. The probability for an arriving customer being blocked from entering the system and getting lost is

$$p_d = p_K \approx \rho p_{c-1} - (1 - \rho) \sum_{n=0}^{K-1} p_n$$

By noting, that the effective arrival rate $\bar{\lambda} = (1 - p_K)\,\lambda$, one may now determine the performance characteristics the same way as has been several times before,

$$\begin{aligned} L &= \sum_{n=0}^{K} n p_n, \quad W = \frac{L}{\lambda(1 - p_K)} \\ W_q &= W - \frac{1}{\mu}, \quad L_q = \lambda(1 - p_K) W_q \end{aligned}$$

Classic queueing literature provides a wealth of information on bounds and approximation, although much of it is devoted to the more general $G/G/1$ and $G/G/c$ queues. For example, see [27]. A very simple relation between the $M/M/c$ model and the more general $M/G/c$ queueing system with processor sharing discipline has been derived by Wolff in [64]. Some exact results in terms of generating functions may be found in [49].

5.4 Customer Impatience

Balking may be introduced to the $M/G/1$ model in a straightforward manner by prescribing a probability b that a customer enters the system on arrival. In modifying equation 5.1 accordingly, i.e. writing instead

$$q_{n|t} = \frac{e^{-b\lambda t} \left(b\lambda t\right)^n}{n!}$$

one is able to carry out the entire set of calculations as given for the classic $M/G/1$ model [27]. Obviously it is much more difficult to attach more than a constant balking rate to the mode. Even for single server systems, the inclusion of customer impatience effects into the model becomes a tedious task. There exist some solutions in the literature. Most of them are based on the refinement of a $G/G/1$ queueing system with impatient customers to the case of Poissonian arrivals. Bacelli and Hebuterne [5] have shown, that the distribution for the virtual offered waiting time and the distribution for the waiting time coincide for the extended $M/G/1+G$ model with impatient customers. Consider a single server queue with Poissonian arrivals and arbitrary service distribution $B(.)$ as before for the classic $M/G/1$ queueing system. Let $V(.)$ denote the (absolutely continous) distribution of the virtual offered waiting time. Define the survival function $\bar{G}(.) = 1 - G(.)$ to the impatience distribution $G(.)$. Note, that $V(.)$ can be interpreted as a mixed distribution, as there is a positive probability for an arriving customer to join service immediately. It splits in a discrete part $V(0)$ and a (absolutely) continous part with density $v(.)$. Bacelli and Hebuterne have shown, that $v(.)$ is the solution of the following system of integral equations

$$\begin{aligned} v(t) &= \lambda V(0)(1 - B(t)) + \int_0^t v(s)G(s)(1 - B(t-s))ds \\ 1 &= V(0) + \int_0^\infty v(s)ds \end{aligned} \tag{5.23}$$

By substitution, they were able to identify expression 5.23 as Fredholm integral equation of the second kind. The solution allows $v(.)$ to be represented as integral series. For details we refer to the paper by Bacelli and Hebuterne [5]. Another approach is to generalize the equations derived by Takacs [56] for the classic $M/G/1$ model to consider forms of customer impatience. This has been carried out by Gnedenko and is shown in [49]. Transient solutions to $M/G/1$ queueing systems with balking and reneging have been investigated by Subba Rao in his papers [54] and [55]. It should be noted, that Subba Rao assumes the service distribution to belong to a certain class of distributions following an exponential pattern. He does not consider arbitrary service distributions.

5.5 Retrials

Under the usual assumptions for a $M/G/1$ queue, we will now attempt to analyze such a system with retrying customers. Time periods between retrials are assumed to follow an exponential distribution with mean $\frac{1}{\eta}$. The system state will be described by a Markov process $\{C(t), \xi(t) < x, N(t) : t \geqq 0\}$, where $C(t)$ denotes the number of busy servers, $N(t)$ represents the number of retrials and $\xi(t)$ describes the elapsed service time. To introduce $\xi(t)$ into the model preserves the Markov property and is called the technique of *supplementary variables*. Please note, that for $C(t) = 0$ there is no need to define an elapsed service time, as no customer is present in the system. The relevant states are collapsed into a single simpler state. Putting it together the equilibrium probabilities are defined as

$$\begin{aligned} p_{0,n} &= \Pr\{C(t), N(t)\} \\ p_{1,n}(x) &= \Pr\{C(t), \xi(t) < x, N(t)\} \end{aligned}$$

Following the approach by Falin and Templeton [20], the steady state equations are given by

$$\begin{aligned} (\lambda + n\eta)\, p_{0,n} &= \int_0^\infty p_{1,n}(x) b(x) dx \\ \frac{d}{dx} p_{1,n}(x) &= -(\lambda + b(x))\, p_{1,n}(x) + \lambda p_{1,n-1}(x) \\ p_{1,n}(0) &= \lambda p_{0,n}(x) + (n+1)\, \eta p_{1,n+1}(x) \end{aligned}$$

recalling that $b(x)$ describes the density of service times. Introducing the generating functions $P_0(z) = \sum_{n=0}^{\infty} z^n p_{0,n}$ and $P_1(z,x) = \sum_{n=0}^{\infty} z^n p_{1,n}(x)$ the above system of equations may be written as

$$\begin{aligned} \lambda P_0(z) + \eta z \frac{d}{dz} P_0(z) &= \int_0^{\infty} P_1(z,x) b(x) dx && (5.24) \\ \frac{\partial}{\partial x} P_1(z,x) &= -\left(\lambda - \lambda z + b(x)\right) P_1(z,x) && (5.25) \\ P_1(z,0) &= \lambda P_0(z) + \eta \frac{d}{dz} P_0(z) && (5.26) \end{aligned}$$

Solving the ordinary differential equation 5.25 yields

$$P_1(z,x) = P_1(z,0) \left(1 - b(x)\right) e^{-(\lambda - \lambda z)x} \tag{5.27}$$

which allows one to rewrite 5.24 as

$$\lambda P_0(z) + \eta z \frac{d}{dz} P_0(z) = \bar{b}\left(\lambda - \lambda z\right) P_1(z,0) \tag{5.28}$$

Expressing in terms $\eta \frac{d}{dz} P_0(z)$ and inserting in expression 5.26 leads to

$$P_1(z,0) = \lambda \frac{1-z}{\bar{b}\left(\lambda - \lambda z\right) - z} P_0(z)$$

Substituting in equation 5.27 gives

$$P_1(z,x) = \lambda \frac{1-z}{\bar{b}\left(\lambda - \lambda z\right) - z} P_0(z) \left(1 - b(x)\right) e^{-(\lambda - \lambda z)x} \tag{5.29}$$

In order to derive an expression for $P_0(z)$ one inserts $P_1(z,0)$ into equation 5.26 and 5.28, i.e.

$$\eta \left[\bar{b}\left(\lambda - \lambda z\right) - z\right] \frac{d}{dz} P_0(z) = \lambda \left[1 - \bar{b}\left(\lambda - \lambda z\right)\right] P_0(z) \tag{5.30}$$

By careful inspection of $\bar{b}\left(\lambda - \lambda z\right) - z$ leads to $z < \bar{b}\left(\lambda - \lambda z\right) \leqq 1$. By considering the limit of $\frac{1-\bar{b}(\lambda-\lambda z)}{\bar{b}(\lambda-\lambda z)-z}$ from below, one can show, that the convergence radius determined by the generating function may be extended to the boundary $0 \leqq z \leqq 1$. This in turn enables one to state expression 5.30 as ordinary differential equation

$$\frac{d}{dz} P_0(z) = \frac{\lambda}{\eta} \frac{1 - \bar{b}\left(\lambda - \lambda z\right)}{\bar{b}\left(\lambda - \lambda z\right) - z} P_0(z)$$

with solution

$$P_0(z) = P_0(1) \exp\left\{ \frac{\lambda}{\eta} \int_0^z \frac{1 - \bar{b}\,(\lambda - \lambda y)}{\bar{b}\,(\lambda - \lambda y) - y} dy \right\}$$

By noting, that $1 - P_0(1) = u = \rho := -\lambda \frac{d}{dz} \bar{b}(z)\big|_{z=0}$, one immediately arrives at

$$P_0(z) = (1 - \rho) \exp\left\{ \frac{\lambda}{\eta} \int_0^z \frac{1 - \bar{b}\,(\lambda - \lambda y)}{\bar{b}\,(\lambda - \lambda y) - y} dy \right\} \tag{5.31}$$

Thus the steady state solution for the single server retrial system with arbitrary service times is completely characterized in terms of generating functions by expression 5.29 and 5.31. It can be shown, that the system remains stable for $u = \rho < 1$, for details refer to [20]. The distribution of the number of repeating customers is given by

$$\begin{aligned} P(z) &= P_0(z) + \int_0^\infty P_1(z, x) dx \\ &= (1 - \rho) \frac{1 - \bar{b}\,(\lambda - \lambda z)}{\bar{b}\,(\lambda - \lambda z) - z} \exp\left\{ \frac{\lambda}{\eta} \int_0^z \frac{1 - \bar{b}\,(\lambda - \lambda y)}{\bar{b}\,(\lambda - \lambda y) - y} dy \right\} \end{aligned}$$

By using the properties of the Laplace transform one arrives at the average number of customers in orbit, i.e.

$$L_q = \frac{d}{dz} P(z)\bigg|_{z=1} = \frac{2\lambda\rho/\eta + \rho^2 + \lambda^2 \sigma_S^2}{2\,(1 - \rho)} = \frac{\lambda\rho}{\eta\,(1 - \rho)} + L_q^{(M/G/1)}$$

As highlighted above, the average queue length exceeds the one from the classic $M/G/1$ model by an additive factor. Considering, that the retrial system is not work conserving, i.e. there is a positive probability, that waiting customers do not immediately receive service, this factor becomes intuitively clear. Furthermore by letting η approach infinity, one arrives at the classic $M/G/1$ model. Please note, that we encountered the same effect for the average queue length of the exponential retrial system given in expression 4.12. The remaining performance characteristics may be determined from an application of Little's law, i.e.

$$W_q = \frac{1}{\lambda} L_q, \quad W = W_q + \frac{1}{\mu}, \quad L = \lambda W = L_q + \rho$$

We have already seen for the classic $M/G/c$ queue, that no simple analytical solution is available for more than one server. The same is true for the multiserver system with retrials. Fortunately the approach taken for the classic multiserver queue also works for the retrial version. Recall from equation 5.18 the approximation for the average system size:

$$L \approx c_S^2 L^{(M/M/c)} + \left(1 - c_S^2\right) L^{(M/D/c)}$$

Considering equation 5.20, the average number of customers in system for the $M/D/c$ queue may roughly approximated by $L^{(M/D/c)} \approx \frac{1}{2} L^{(M/M/c)}$ leading to

$$L \approx \frac{1 + c_S^2}{2} L^{(M/M/c)}$$

Replacing the $M/M/c$ model with its retrial counterpart results in

$$L \approx \frac{1 + c_S^2}{2} \left(L_q^{(rM/M/c)} + \rho\right)$$

where $L_q^{(rM/M/c)}$ is given by equation 4.14 for the multiserver retrial system with exponential service times. A slightly different argument for its derivation called the *processor sharing method* is given in [64]. Wolff also discusses an approximation deploying the *retrials see time averages* property for retrial models. He makes a difference between a customers initial entry and retry and assigns different probabilities to each event. With finite probability on retries the case of a finite (geometric) orbit is also covered by the model. Up to now we have only considered retrial models with exponential retry times. Note, that in relaxing this assumption, the system under consideration may become instable even with $\rho < 1$. For more details please refer to [3].

Chapter 6

Arbitrary Arrival Processes

In the the preceding section the $M/G/1$ queue was modeled as an embedded Markov chain with regeneration points taken at the instances of service departure. Using the same technique, we are able to analyze systems with arbitrary arrivals and exponential service distribution even for multiple servers. For this non-Poisson system, the regeneration points occur at the epochs of arrival. As there is only a slight difference in derivation of the $G/M/1$ and the $G/M/c$ model, we will focus on the latter. The interarrival times are assumed to follow a general distribution $A(.)$ with expectation $\frac{1}{\lambda}$.

6.1 Multiserver Systems

To obtain the matrix of transition probabilities $\mathbf{Q} = (q_{ij})$ for the embedded chain, one has to partition the (i, j) plane into four parts and consider each case seperately.

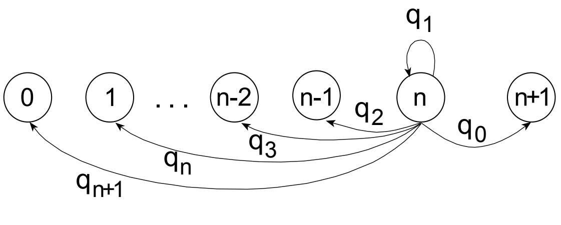


Figure 6.1: Some state transition for the G/M/c queue

- Customers do not arrive in batches, so there is obviously no transition from state $i+1$ to state j for $i+1 < j$, i.e

$$q_{ij} = 0, \qquad i+1 < j$$

- Let $a(t)$ denote the density of the interarrival times. With all servers being busy, the system serves at rate $c\mu$. Defining q_n as the probability of n customers getting served within two subsequent arrivals, i.e.

$$q_n = \int_0^\infty \frac{e^{-c\mu t}\left(c\mu t\right)^n}{n!} a(t)dt$$

yields

$$q_{ij} = q_{i+1-j}, \qquad c \le j \le i+1$$

For a graphical representation refer to figure 6.1.

- Now consider the case, when no customers are waiting. During an interarrival period, $i+1-j$ service completions must occur at time t. Because of the exponentially distributed service time, the probability of a customer to depart is given by $1 - e^{-\mu t}$. Thus the probability, that a customer will remain in service is $e^{-\mu t}$. Employing the binomial distribution leads to

$$q_{ij} = \int_0^\infty \binom{i+1}{i+1-j} \left(1 - e^{-\mu t}\right)^{i+1-j} e^{-\mu t j} a(t)dt, \qquad c \le i+1 \le j$$

By noting, that $\binom{i+1}{i+1-j} = \binom{i+1}{j}$, one may also write [37]

$$q_{ij} = \int_0^\infty \binom{i+1}{j} \left(1 - e^{-\mu t}\right)^{i+1-j} e^{-\mu t j} a(t)dt, \qquad c \le i+1 \le j$$

- It remains to discuss the case $j < c < i+1$. Following [27], assume, that arrival goes into service after some time V, when all prior customers have left. Then V is the time until $i-c+1$ customers have been served by a full system at rate $c\mu$. This V is distributed according to a $(i-c+1)$-stage Erlang distribution. To get from state i to j between two subsequent arrivals, $c-j$ service departures must occur in the interval from V to the end of the interarrival period. Again utilizing the binomial distribution and using the memoryless property of the service distribution, one arrives at

$$q_{ij} = \binom{c}{c-j} \frac{(c\mu)^{i-c+1}}{(i-c)!} \int_0^\infty \int_0^t \left(1 - e^{-\mu(t-v)}\right)^{c-j} e^{-\mu(t-v)j} v^{j-c} e^{-c\mu v} dva(t)dt$$

for $c < j < i+1$.

Assuming a stable system, the stationary solution for the embedded Markov chain may be obtained by solving the linear system of equations

$$\tilde{\mathbf{p}}\mathbf{Q} = \tilde{\mathbf{p}}, \qquad \sum_{i=0}^{\infty} \tilde{p}_i = 1 \tag{6.1}$$

Please note, that we have denoted the stationary solution by $\tilde{\mathbf{p}}$ to reflect the dependence on the arrival epochs. For $j \geq c$ this leads to

$$\tilde{p}_j = \sum_{i=0}^{j-2} 0\tilde{p}_i + \sum_{i=j-1}^{\infty} q_{i+1-j}\tilde{p}_i = \sum_{i=j-1}^{\infty} q_{i+1-j}\tilde{p}_i$$

Being familiar with the solution of the $M/M/1$ queueing model, this suggests a solution of the form

$$\tilde{p}_i = C\omega^i, \qquad i \geq c \tag{6.2}$$

where ω is the root of the equation

$$z = \sum_{n=0}^{\infty} q_n z^n = \bar{a}\left(c\mu - c\mu z\right) \tag{6.3}$$

with $\bar{a}$ denoting the Laplace transform of the interarrival density. Due to the assumption of a stable system, we only accept solutions from inside the unit disc, that is $|\omega| < 1$. In fact, in can be shown [56], that a unique solution exists for the case $\rho < 1$. Please note, that for Poissonian arrivals ω equals the utilization $u = \frac{\rho}{c}$ and that $u = \omega$ does not hold in general. For this reason, ω is often referred to as *generalized server occupancy* [13].It remains to determine the constant C from the normalization condition and the first $c-1$ equations of the system 6.1. Although possible, this may not be an easy task due to the fact, that $c+1$ equations in $c+1$ unknowns containing infinite sums have to be solved. However, Gross & Harris [27] suggest to modify the term to depend on $\tilde{p}_0, \dots \tilde{p}_c$ and ω only by the following procedure. By noting, that $1 = \sum_{i=0}^{\infty} \tilde{p}_i = \sum_{i=0}^{c-1} \tilde{p}_i + C \sum_{i=c}^{\infty} \omega^i$ one arrives at

$$C = \frac{1 - \sum_{i=0}^{c-1} \tilde{p}_i}{\sum_{i=c}^{\infty} \omega^i} = \frac{1 - \sum_{i=0}^{c-1} \tilde{p}_i}{\omega^c \left(1 - \omega\right)^{-1}}$$

Based on the above calculations one may easily obtain the average queueing time as follows. As queueing occurs only in a full system, customers are served at a constant rate $c\mu$. An arriving customer $j \geq c$ has to wait for the departure of the $j - c + 1$ preceding customers before entering service. Averaging over all possible cases yields [13][37]

$$\begin{aligned} W_q &= \frac{1}{c\mu} \sum_{i=c}^{\infty} \left(i - c + 1\right) \tilde{p}_i = \frac{C}{c\mu} \sum_{i=c}^{\infty} \left(i - c + 1\right) \omega^i \\ &= \frac{C\omega^c}{c\mu \left(1 - \omega\right)^2} \end{aligned} \tag{6.4}$$

Please note, that neither the result for W_q nor the queueing time distribution itself do depend on the arrival epochs [37]. Let the random variable $\breve{W}_q$ denote the queueing time and define A, A^c as the events {arrival queues} and {arrival does not queue}. The corresponding distribution may be written as

$$\begin{aligned} \Pr\left\{\breve{W}_q \leq w\right\} &= 1 - \Pr\left\{\breve{W}_q > w\right\} \\ &= 1 - \Pr\left\{\breve{W}_q > w|A\right\} - \Pr\left\{\breve{W}_q > w|A^c\right\} \\ &= 1 - \Pr\left\{\breve{W}_q > w|A\right\} = \Pr\left\{\breve{W}_q \leq w|A\right\} \end{aligned}$$

By summing the relevant terms, we obtain the probability of an arriving customer being delayed

$$\tilde{p}_d = \sum_{i=c}^{\infty} \tilde{p}_i = C \sum_{i=c}^{\infty} \omega^i = \frac{C\omega^c}{1 - \omega}$$

From Little's law we may readily compute the average queue size from expression 6.4:

$$L_q = \lambda W_q = \frac{C\rho\omega^c}{c\left(1 - \omega\right)^2}$$

Please note, that a similar result would have been obtained by applying the rate conservation law 11 to the formula $L_q = \sum_{i=c}^{\infty} (i-s)\, p_j$ [1]. Adding the average service time $\frac{1}{\mu}$ yields

$$W = W_q + \frac{1}{\mu}, \qquad L = L_q + \rho$$

Next, consider the queue length distribution given that all servers are busy, i.e. that an arrival is delayed. Let the random variable $\breve{L}_q$ denote the queue size. Following [37],

$$\begin{aligned} \Pr\left\{\breve{L}_q = n | \text{arrival delayed}\right\} &= \frac{\tilde{p}_{c+n}}{\tilde{p}_d} = \frac{C\omega^{c+n}}{C\omega^c / (1-\omega)} \\ &= (1-\omega)\,\omega^n \end{aligned} \tag{6.5}$$

one has to conclude, that the conditional queue length distribution is geometric. Proceeding further, one may also determine the distribution of the queueing time [27] and the distribution of the queueing time given an arrival is delayed [37]. As shown by Kleinrock, the latter is an exponential distribution. Loosely speaking, we encountered $M/M/1$ behaviour in the conditional distributions of the $G/M/c$ queueing system. In [13] some results on the waiting time of queues with disciplines other than FCFS are presented.

6.2 Single Server System

We may now readily apply the preceding results to the single server case $c = 1$. Similar to the calculations for the $M/M/1$ queueing model one may readily obtain from expression 6.2 the probabilities

$$\tilde{p}_n = (1-\omega)\,\omega^n \tag{6.6}$$

The root ω is calculated from equation 6.3 given $c = 1$. As stated in [37], the number of customers found in the system by an arriving customer is geometric. Thus we find resemblence with the system size distribution of the corresponding system with exponential service times. Furthermore it can be shown [37], that the queueing time distribution has the same form as for the

$M/M/1$ model. The performance characteristics are now easily calculated by substituting $C = 1 - \omega$ and $c = 1$ to the relevant expressions:

$$\begin{aligned} L_q &= \frac{\rho\omega}{(1-\omega)}, \qquad W_q = \frac{\omega}{\mu(1-\omega)} \\ L &= \frac{\rho}{(1-\omega)}, \qquad W = \frac{1}{\mu(1-\omega)} \end{aligned}$$

Please note, that the average time in system and the queueing time are the same for a random observer and an arriving customer. This is not true for the average system size and the average queue size, as can be seen by applying the rate conservation law. Rewriting the expression in theorem 11 yields $\min(c,n)\,p_n = \omega p_{n-1} = \rho\tilde{p}_{n-1}$. By generalizing to a random observer, we have to scale each probability by a factor $\frac{\omega}{\rho}$. Fortunately this factor carries over to the expressions for L and L_q by linearity. So a simple multiplication with $\frac{\omega}{\rho}$ yields the desired result

$$\tilde{L}_q = \frac{\omega}{\rho}L_q = \frac{\omega^2}{(1-\omega)}, \qquad \tilde{L} = \frac{\omega}{\rho}L = \frac{\omega}{(1-\omega)}$$

For a detailed derivation of the $G/M/1$ queue please consult [27]. Although the models for single and multiple servers are similar in analysis, only some queueing theory textbooks cover the more general case. For further interest we refer to the literature stated in the text.

Example 16 *First consider the case of exponential service times. Then the Laplace transform of the density is given by*

$$\bar{a}(s) = \frac{\lambda}{\lambda + s}$$

Substituting $\bar{a}(s)$ and $c = 1$ in equation 6.3 leads to

$$\omega = \bar{a}(\mu - \mu\omega) = \frac{\lambda}{\lambda + \mu - \mu\omega}$$

Rearranging

$$\mu\omega^2 - (\lambda + \mu)\,\omega + \lambda = 0$$

and factoring

$$(\omega - 1)(\mu\omega - \lambda) = 0$$

suggests two solutions $\omega = 1$ and $\omega = \frac{\lambda}{\mu} = \rho$. We omit the first solution as it does not lie inside the unit disc. Applying the result to equation 6.6 yields the well known steady state distribution

$$\tilde{p}_n = (1 - \rho)\, \rho^n$$

By either applying PASTA or the rate conservation law, one immediately proves the idendity of $\tilde{p}_n$ and p_n.

Example 17 *Assuming deterministic arrivals the corresponding Laplace transform of the arrival process is given by*

$$\bar{a}(s) = e^{-s/\lambda}$$

Substituting $\bar{a}(s)$ and $c = 1$ in equation 6.3 leads to

$$\omega = \bar{a}\,(\mu - \mu\omega) = e^{-\frac{\mu(1-\omega)}{\lambda}} = e^{-(1-\omega)/\rho}$$

which can be numerically solved for predetermined values of ρ.

As the second example shows, even for such a intuitively simple model as the $D/M/1$ model an analytic solution becomes intractable. In most cases one resorts to the use of numeric procedures to obtain the desired result. Fortunately some approximations devoted to more general models are available and indeed suitable to approximate queues like $G/M/c$ and $D/M/1$. We shall discuss these topics below in the context of arbitrary arrivals and departures.

6.3 Capacity Constraints

We will now state some results on the multiserver queue with limited capacity first derived by Takacs as presented in [24]. Denoting the Laplace transform of the interarrival density with $\bar{a}(s)$ as before the steady state probabilities seen by an arriving customer are given by

$$
\begin{aligned}
\tilde{p}_n &= \begin{cases} \sum_{i=0}^{c-1} (-1)^{i-n} \binom{i}{n} v a_*(i) \sum_{j=i+1}^{\infty} \frac{w_j a_*(j)}{1-\bar{a}(j\mu)} & 0 \le n \le c \\ v g_{K-n} & c < n \le K \end{cases} \\
v &:= \left[\sum_{j=0}^{K-c} g_j + \sum_{j=1}^{c} \frac{w_j a_*(j)}{1-\bar{a}(j\mu)} \right]^{-1} \\
g_n &:= \frac{1}{n!} \frac{d^n}{dz^n} \left(\frac{(1-z)\,\bar{a}(c\mu - c\mu z)}{\bar{a}(c\mu - c\mu z) - z} \right) \Bigg|_{z=0} \\
w_n &:= \begin{cases} \binom{c}{n} \Big[\sum_{i=1}^{K-c+1} g_i a_{K-c+1-i} & \\ -\bar{a}(n\mu) \sum_{i=1}^{K-c} g_i \left(\frac{c}{c-n}\right)^{K-c+1-i} & \\ +a_n + \sum_{i=1}^{K-c} g_i \sum_{k=1}^{K-c} a_k \left(\frac{c}{c-n}\right)^{K-c+1-i-k} & \\ + \sum_{k=1}^{K-c} a_k \left(\frac{c}{c-n}\right)^{K-c-k} - \bar{a}(n\mu) \left(\frac{c}{c-n}\right)^{K-c} \Big] & 0 \le n < c \\ \bar{a}(c\mu) g_{K-c+1} & n = c \end{cases} \\
a_*(n) &:= \prod_{k=1}^{n} \frac{\bar{a}(k\mu)}{1-\bar{a}(k\mu)}, \quad a_n := \int e^{-c\mu x} \frac{(c\mu x)^n}{n!} a(x) dx
\end{aligned}
$$

From these the unconditional equilibrium distribution is easily determined by applying the rate conservation law. Using some of the functions introduced above, the average queueing time may be calculated from

$$W_q = \frac{v}{c\mu\,(1-v)} \sum_{n=1}^{K-c} (K-c+1-n)\, g_n$$

An arriving customer is turned away from the system and gets lost with probability $p_b = \tilde{p}_K$. As a consequence the effective arrival rate is readily obtained, i.e.

$$\bar{\lambda} = \lambda\,(1 - \tilde{p}_K)$$

Through the application of Little's law we arrive at the remaining performance key indicators in terms of W_q and $\bar{\lambda}$:

$$W = W_q + \frac{1}{\mu}, \quad L = \bar{\lambda} W, \quad L_q = \bar{\lambda} W_q$$

Another useful reference for the $G/M/m/K$ model is the research paper [31] by Per Hokstad. It also covers the connection between the time-continous and the embedded process, which is just another name for the rate conservation law given by theorem 11. Hokstad managed to reduce the derivation of the steady state solution to a linear system of equations. Finally he determines performance key indicators and presents some examples.

Chapter 7

Arbitrary Arrival and Service Processes

Arriving at the most general queueing model, it is remarkable, that still some results may be found. In the following discussion we will first focus on the single server model and then provide some approximations for the multiserver case.

7.1 Single Server System

The single server system has first been analyzed by Lindley in 1952. Our approach will follow the presentation in [27]. Let the random variables $\breve{W}_q^{(n)}$, $\breve{S}^{(n)}$, $\breve{T}^{(n)}$ denote the queueing time, the service time and the interarrival time of the n-th customer. Assume, that $\breve{S}^{(n)}$ and $\breve{T}^{(n)}$ are mutually independent and independently indentically distributed according to $B(.)$ and $A(.)$, respectively. Furthermore both distribution functions shall be non negative and absolutely continous. Consequently the densities exist and we may denote their Laplace transforms with $\bar{b}(s)$ and $\bar{a}(s)$. Introducing the random variable $\breve{U}^{(n)} = \breve{S}^{(n)} - \breve{T}^{(n)}$ as the difference between service time and interarrival time, the Laplace transform of the density of $\breve{U}^{(n)}$ may be determined as the convolution

$$\bar{u}(s) = \bar{a}(-s)\bar{b}(s) \tag{7.1}$$

Following the given notation, we may describe the queueing time for the $(n+1)$-th customer as

$$\breve{W}_q^{(n+1)} = \begin{cases} \breve{W}_q^{(n)} + \breve{U}^{(n)} & \breve{W}_q^{(n)} + \breve{U}^{(n)} > 0 \\ 0 & \breve{W}_q^{(n)} + \breve{U}^{(n)} \leq 0 \end{cases} \tag{7.2}$$

Let $W_q^{(n)}(t)$ describe the distribution of the queueing time of the n-th customer. Then the waiting time distribution for the subsequent customer may be readily obtained from relation 7.2, i.e.

$$W_q^{(n+1)}(t) = \begin{cases} \int_{-\infty}^{t} W_q^{(n)}(t-x)u(x)dx & t \geq 0 \\ 0 & t < 0 \end{cases} \tag{7.3}$$

Here $u(.)$ denotes the density function of the random variable $\breve{U}^{(n)}$. Due to the fact, that arrivals and services are independently and indentically distributed, we may write down the average arrival and service rates as $\lambda = 1/\mathbb{E}\breve{T}^{(1)}$ and $\mu = 1/\breve{S}^{(1)}$. Assuming $\rho = \lambda/\mu < 1$, it may be shown [8], that a steady state solution $\lim_{n\to\infty} W_q^{(n)}(t) = W_q(t)$ for the queueing time exists. Consequently, the two waiting time distributions in equation 7.3 must be identical, i.e.

$$W_q(t) = \begin{cases} \int_{-\infty}^{t} W_q(t-x)u(x)dx & t \geq 0 \\ 0 & t < 0 \end{cases} \tag{7.4}$$

The above equation is often refered to as *Lindley Integral Equation* and belongs to the class of Wiener-Hopf Integral Equations. Introducing

$$W_q^-(t) = \begin{cases} 0 & t \geq 0 \\ \int_{-\infty}^{t} W_q(t-x)u(x)dx & t < 0 \end{cases}$$

equation 7.4 may be expressed as follows

$$W_q^-(t) + W_q(t) = \int_{-\infty}^{t} W_q(t-x)u(x)dx$$

Taking the Laplace transform of both sides and substituting equation 7.1 results in

$$\bar{W}_q^-(s) + \bar{W}_q(s) = \bar{W}_q(s)\bar{u}(s) = \bar{W}_q(s)\bar{a}(-s)\bar{b}(s) \tag{7.5}$$

By applying the properties of Laplace transforms [23], the Laplace transform of the unknown queueing time density $\bar{w}_q(s)$ may be expressed in terms of its distribution functions as $\bar{w}_q(s) = s\bar{W}_q(s)$. Rewriting equation 7.5

$$\bar{W}_q^-(s) + \frac{1}{s}\bar{w}_q(s) = \frac{1}{s}\bar{w}_q(s)\bar{a}(-s)\bar{b}(s)$$

finally leads to

$$\bar{w}_q(s) = \frac{s\bar{W}_q^-(s)}{\bar{a}(-s)\bar{b}(s) - 1} \tag{7.6}$$

The result stated above can not provide complete satisfaction, as it still remains to determine $\bar{W}_q^-(s)$. There exist various approaches to the solution of the $G/G/1$ queueing system, but no exact closed form solution is provided. For example, refer to the textbooks [37], [13], [4] and [64]. If a closed form solution is desired, one has to consider approximative methods. As discussed in section 1.4, the arrival and service distribution may be arbitrarily well approximated by the class of exponential distributions in serial and parallel, which is in turn part of the family of phase type distributions. This approach has been studied in detail by Schassberger in [50].

7.2 Multiserver systems

The analysis of the multiserver case poses some technical difficulties, e.g. in certain situations, a stable system does not become empty. Additional assumptions are required to deal with questions like which server will serve an arriving customer in an underload situation. Main results have already been provided by Kiefer and Wolfowitz in 1955, but we will turn attention to approximations for the $G/G/c$ queueing system. One such approximation for the average queueing time is the Allen-Cuneen formula

$$W_q \approx \frac{p_d^{(M/M/c)}}{\mu\,(c - \rho)} \left(\frac{c_T^2 + c_S^2}{2}\right) \tag{7.7}$$

Here c_T^2 and c_S^2 describe the coefficient of variation of the interarrival and the service time. The probability $p_d^{(M/M/c)}$ is given by the Erlang Delay formula 3.6. As shown in [2], formula 7.7 is exact for the queueing systems $M/G/1$ and $M/M/c$. Another reasonable approximation may be obtained by extending the concepts introduced in section 5 for the $M/G/c$ queueing system. Following the paper [36] by Kimura, the average queueing time is given by

$$W_q \approx \left(c_T^2 + c_S^2\right) \left[\frac{1 - c_T^2}{W_q^{(D/M/c)}} + \frac{1 - c_S^2}{W_q^{(M/D/c)}} + \frac{2\,(c_T^2 + c_S^2 - 1)}{W_q^{(M/M/c)}}\right]^{-1} \tag{7.8}$$

where $W_q^{(M/D/c)}$ may be derived from expression 5.20 through an application of Little's law. For $W_q^{(M/M/c)}$ one may substitute the exact result provided by formula 3.8. The remaining term is best approximated by a relation due to Cosmetatos, Krämer and Langenbach-Belz

$$W_q^{(D/M/c)} \approx \left(\frac{1}{2} - \left(1 - \frac{\rho}{c}\right)(c-1)\frac{\sqrt{4+5c}-2}{8\rho}\right) e^{-2(c-\rho)/3\rho} W_q^{(M/M/c)}$$

By an application of Little's law, the performance characteristics L, L_q and W are readily derived from expression 7.8. For some notes on the $G/G/c$ model, refer to the book [64]. The phase-type approach has been investigated by Schassberger in [50]. An advanced mathematical discussion covering relations between queueing processes and concepts of convergence may be found in [8].

7.3 Customer Impatience

The most straightforward way to build customer impatience into the $G/G/1$ model is to adapt expression 7.2 and derive a generalization of Lindley's integral equation. Without consideration of balking behaviour this leads to the so called $G/G/1+G$ model. As before, let the random variables $\breve{W}_q^{(n)}$, $\breve{S}^{(n)}$, $\breve{T}^{(n)}$ denote the queueing time, the service time and the interarrival time of the n-th customer. Also define $\breve{U}^{(n)} = \breve{S}^{(n)} - \breve{T}^{(n)}$ with the usual meaning. A new random variable $\breve{G}^{(n)}$ shall describe the patience time of the n-th customer. In order to model impatience effects, a potential customer refuses to join the queue, if $\breve{G}^{(n)} \leqq \breve{W}_q^{(n)}$. On the other hand, for $\breve{G}^{(n)} > \breve{W}_q^{(n)}$ he enters the system to get served. Taking this into account, one arrives at

$$\breve{W}_q^{(n+1)} = \begin{cases} \breve{W}_q^{(n)} + \breve{U}^{(n)} & \breve{W}_q^{(n)} + \breve{U}^{(n)} > 0,\ \breve{G}^{(n)} > \breve{W}_q^{(n)} \\ 0 & \breve{W}_q^{(n)} + \breve{U}^{(n)} \le 0,\ \breve{G}^{(n)} > \breve{W}_q^{(n)} \\ \breve{W}_q^{(n)} - \breve{T}^{(n)} & \breve{W}_q^{(n)} - \breve{T}^{(n)} > 0,\ \breve{G}^{(n)} \leqq \breve{W}_q^{(n)} \\ 0 & \breve{W}_q^{(n)} - \breve{T}^{(n)} \le 0,\ \breve{G}^{(n)} \leqq \breve{W}_q^{(n)} \end{cases}$$

Introducing the survival function $\bar{G}(.) = 1 - G(.)$ for the distribution $G(.)$ of the impatience time and proceeding as before leads to

$$W_q(t) = \begin{cases} \int_{-\infty}^{t} W_q(t-x)\left[\bar{G}(x)u(x) + \left(1 - \bar{G}(x)\right)a(-x)\right]dx & t \ge 0 \\ 0 & t < 0 \end{cases}$$

Here $u(.)$ and $a(.)$ denote the density functions of the random variables $\breve{T}^{(n)}$ and $\breve{U}^{(n)}$, respectively. If the distribution of impatience times $G(.)$ is non-defective, that is for $\lim_{x\to\infty} G(x) = 1$ the system remains stable for $\rho < 1$. For details on a similar derivation we refer to the paper of Bacelli and Hebuterne [5]. An approximative solution to a $G/G/1$ queueing system with balking and reneging is given in [62].

Appendix A

Stochastic Processes

A.1 Introduction

If one wishes to model real world phenomena and gain deeper insight into them, an adequate set of mathematical and probabilistic tools is required. One such tool is the theory of stochastic processes concerned with the abstraction of empirical processes. Examples include the flows of events in time and evolutionary models in biological science. Associated with the concept of a stochastic process are the *state space* and the *parameter space*. Whereas the latter is often identified with time, the former contains all values the process can assume. If the parameter space is not limited to time only, i.e. possesses a higher dimension, stochastic processes are also called *random fields*. In the following discussion, the parameter space is limited to the one dimensional case and treated as time.

Definition 18 *Let T denote the parameter space. A stochastic process is a family of random variables $\{X(t) : t \in T\}$ defined on the same probability space.*

Definition 19 *A stochastic chain is a stochastic process with countable (discrete) state space. It will be denoted by $\{X_t : t \in T\}$.*

As stated above, it is very important, that each random variable $X(t)$ assumes the same state space. Flipping a coin and throwing a dice are independent experiments. Their combination alone does not constitute a stochastic process.

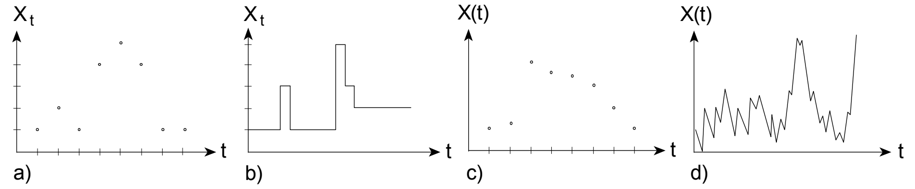


Figure A.1: Examples of stochastic processes

Figure A.1 shows examples for each combination of state and parameter space. Notation has chosen according to above definitions specifying two stochastic chains on the left and the more general processes with continous state space on the right. These graphs are called sample graphs, because they reflect one possible realization of a stochastic process over time. By keeping the time fixed and considering the possible realizations one arrives at another view, that is the process as a random variable for a certain point in time. These dualistic perspectives are central to the theory of stochastic processes and give rise to concepts such as *ergodicity.* Ergodicity deals with the problem of determining measures for stochastic processes. For example, the ergodic theorem states, that under certain conditions, the time average equals the ensemble average (almost sure). The time average is deducted from a single realization over infinite time and the ensemble average is the mean over all possible realizations for a certain point in time.

A central concept to the theory of stochastic processes is stationarity. It releases the requirement of time dependence allowing for a steady state view of certain processes.

Definition 20 *A stochastic process $X(t)$ will be called stationary, if its joint distributions are left unchanged by shifts in time, i.e. $(X(t_1),...,X(t_n))$ and $(X(t_1+h),...,X(t_n+h))$ have the same distribution for all h and $t_1,...,t_n$.*

The general discussion will end here and we will turn to examples important to queueing theory. Proceeding further would require an introduction to measure theory. For readable accounts refer to [33][34][12][32]. The standard reference in the field is [16].

A.2 Markov Processes

General stochastic processes may exhibit a complicated dependence structure. By restricting the dependence of the future to the present, not allowing for any influence from the past, one arrives at what is called a *Markov process.* At first sight such a restriction seems to be a serious one, but this is not necessarily the case. Considering effects of births and deaths to determine tomorrows distribution of a population is only one example. Compared to general stochastic processes, Markov processes are mathematically more tractable. Furthermore they may serve as approximations to more elaborate models.

Definition 21 *A stochastic process $\{X(t) : t \in T\}$ is called Markov process, if for any set of n time points $t_1 < ... < t_n$ the conditional distribution of $X(t_n)$, given the values $X(t_1), ..., X(t_{n-1})$, depends only on $X(t_{n-1})$, i.e. for any $x_1, ..., x_n$*

$$\begin{aligned} & \Pr\{X(t_n) \leq x_n | X(t_1) = x_1, ..., X(t_{n-1}) = x_{n-1}\} \\ = \ & \Pr\{X(t_n) \leq x_n | X(t_{n-1}) = x_{n-1}\} \end{aligned} \tag{A.1}$$

According to the type of parameter space, Markov processes are classified in *discrete parameter Markov processes* and *continous parameter Markov processe*s. They are determined by the *transition probabilities* $P(s, t, x, A)$ and an initial distribution. Here $P(s, t, x, A)$ describes the probability of the transition from state x to a state $y \in A$ within time $|t - s|$. In case A is finite or at most countable, it can be calculated by

$$P(s, t, x, A) = \sum_{y \in A} \Pr\{X(t) = y | X(s) = x\} \tag{A.2}$$

If the transition probability depends only on the difference of s and t, i.e. $P(s, t, x, A) = P(|t - s|, x, A)$, the Markov process is called a *(time-)homogenous Markov process.* The case of discrete parameter space will be discussed in more detail in the next section. The literature focusing on Markov processes is highly based on measure theory and functional analytic concepts, one classical reference is [47].

A.3 Markov Chains

Markov chains provide the means to calculate limiting distributions under relatively mild conditions. As such they find wide applications in modeling real world phenomena met in engineering and science. The ease in calculation is reached by restricting the state space.

Definition 22 *A Markov process with countable (discrete) state space is called Markov chain. Alternatively it can be seen as stochastic chain, which satisfies equation A.1.*

Markov chains are classified by their parameter space thus appearing either as *discrete time* or *continous time* variants. Based on the definition for Markov processes, a Markov chain will be called *(time-)homogenous*, if the transition probabilities do not depend on time. In the following sections only homogenous Markov chains will be discussed.

A.3.1 Homogenous Markov Chains in Discrete Time

As both state space S and parameter space T are now discrete, the transition probabilities given in A.2 may be simplified to $p_{ij} = \Pr\{X_n = j | X_{n-1} = i\}$. These probabilities may be assembled to the *matrix of transition probabilities of stage 1*:

$$\mathbf{P} = (p_{ij})_{i,j \in T} = \begin{pmatrix} p_{11} & p_{12} & \cdots \\ p_{21} & p_{22} & \cdots \\ \vdots & \vdots & \end{pmatrix}$$

Please note, that the term matrix is used in the wide sense. It includes the possibility of infinite dimension. A short introduction to the algebra of denumerable matrices is given in [35]. The probability of reaching state j starting from state i within m steps is denoted by $p_{ij}^{(m)}$ and calculated using the *Chapman Kolmogorov equations*

$$p_{ij}^{(m)} = \sum_{k \in T} p_{ik}^{(m-1)} p_{kj}$$

for $m \geq 2$. In other words, the m-step transition probabilities are recursively defined by the single step transition probabilities. The idea behind the Chapman Kolmogorov equations is, that before proceeding to state j within

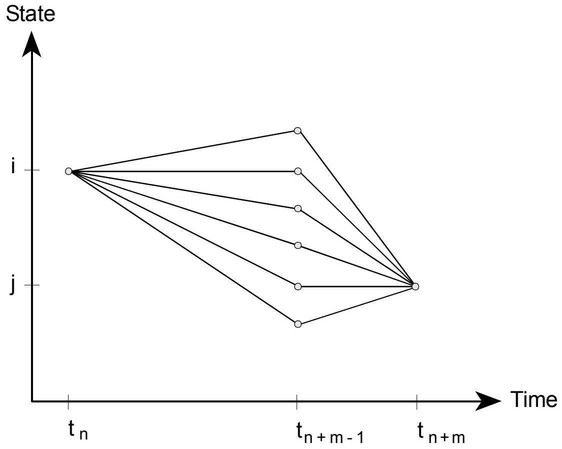


Figure A.2: Chapman Kolmogorov Equations

a single step, state k has to be reached from state i within $m-1$ steps. As state k can be any state in state space, one has to sum up all probabilities. This is also shown in figure A.2. In matrix notation, the system of Chapman Kolmogorov equations is reduced to the power operation, that is $\mathbf{P}^{(m)} = \mathbf{P}^m$. Given a *vector of initial distributions* $\mathbf{a}$, the state of the process after m steps is given by $\mathbf{aP}^m$. So a Markov chain is fully determined by an initial distribution and the transition probabilities. This we had expected already from the more general framework of Markov processes.

Based on the experience with empirical processes one may ask for a steady state following a startup phase. In mathematical terms steady state is described by a *limiting distribution*, which is independent of any initial distribution. For such a limit to exist, certain restrictions become necessary. One of the more obvious conditions is, that the Markov chain under analysis shall not consist of independent subchains. Visualized as graph, there must be a path of positive probability between each pair of states.

Definition 23 *A Markov chain is irreducible, if all its states communicate, i.e. any $i \neq j \in S$ satisfies*

$$p_{ij}^{(m)} > 0$$

Otherwise it is called reducible.

If a chain is reducible, it can be splitted up into subchains with each of them analyzed seperately. Mathematically expressed irreducibility defines an equivalence relation resulting in a certain grouping of states. It turns out, that the properties discussed below are shared by all members of an irreducible Markov chain. One of them is periodicity. As an example consider a bistable flip-flop toggling between states 0 and 1 at every time instant. This behaviour is clearly periodic, as the process returns to the starting state every second step. The initial distribution is preserved to infinity and no steady state can be achieved.

Definition 24 *A state is called aperiodic, if the greatest common divisor of*

$$\left\{ m | p_{ii}^{(m)} > 0 \right\}$$

equals 1*. If not, the state is called periodic. For an irreducibility class, the period is the same for all class members.*

In order to assume a steady state, we have to assure, that the Markov chain does not drift to infinity or get stuck in a group of states. The visiting behaviour of each state has to be inspected more closely.

Definition 25 *Let $f_i^{(m)}$ describe the probability to return to state i within m steps, i.e.*

$$f_i^{(m)} = \Pr\left\{ X_m = i, X_k \neq i : k = 1, 2, ..., m-1 | X_0 = i \right\}$$

and define f_i to be the probability to return to state i in a finite or infinite number of steps, that is

$$f_i = \sum_{m=1}^{\infty} f_i^{(m)}$$

then state i is called recurrent, if $f_i = 1$. For $f_i < 1$ it is called transient. In case of a recurrent state

$$m_i = \sum_{m=1}^{\infty} m f_i^{(m)}$$

is called the mean recurrence time. For $m_i < \infty$ state i is called positive recurrent, otherwise it is called null recurrent. For an irreducibility class, these properties extend to all members.

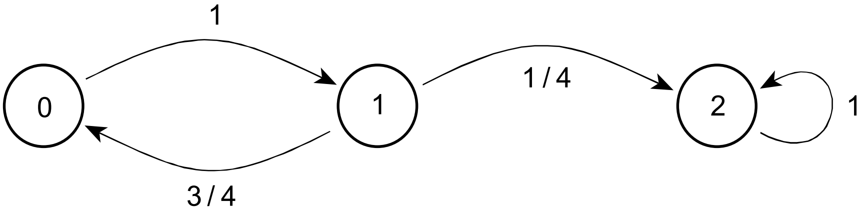


Figure A.3: A simple Markov chain example

Instead of assuming an almost sure return to state i for recurrence, we could also ask for an almost sure infinite number of visits to state i in the above definition. Indeed, it turns out that both conditions are equivalent:

Theorem 26 *Given $\nu_i = \#\left(n > 0 : X_n = i\right)$ the number of visits to state i, the following conditions are equivalent: $f_i = 1 \Leftrightarrow \Pr\left\{\nu_i = \infty\right\} = 1 \Leftrightarrow \mathbb{E}\nu_i = \infty$.*

For a proof, please refer to [12]. Combining all the properties discussed so far leads to the concept of *ergodicity.*

Definition 27 *An irreducible, aperiodic and positive recurrent Markov chain is called ergodic.*

Before proceeding to the existence of the limiting distribution, consider the example shown in figure A.3. State 0 may be reached from state 1 and vice versa. These states communicate and are periodic. State 2 acts like a trap - once entered the process can not escape. Such a state is called an *absorbing state.* A little calculation shows that $f_0^{(2)} = f_1^{(2)} = \frac{3}{4}$and $f_0^{(m)} = f_0^{(m)} = 0$ for $m \neq 2$. Hence $f_0 = f_1 = \frac{3}{4} < 1$. Both states are transient, whereas state 2 is positive recurrent. The properties are not shared among states, so the chain is not irreducible. This can also be seen from the fact, that there is no communication from state 2 to state 1.

The limiting distribution assumed by a Markov chain in steady state will be denoted by $\pi_i = \lim_{n\to\infty} \Pr\left\{X_n = i\right\}$. These probabilities are sometimes called *stationary.* If they exist, no transition of the underlying Markov chain affects the probability vector $\boldsymbol{\pi} = (\pi_i)_{i\in S}$. In matrix form, the following equilibrium conditions hold

$$\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}, \qquad \sum_{i\in S} \pi_i = 1 \tag{A.3}$$

We are now ready to summarize the main results on the existence of the limiting distribution $\boldsymbol{\pi}$:

Theorem 28 *Given an aperiodic Markov chain in discrete time, the limits* $\pi_i = \lim_{n\to\infty} \Pr\{X_n = i\}$ *for all* $i \in S$ *exist. For an irreducible and aperiodic Markov chain the following expression holds*

$$\pi_i = \frac{1}{m_i}$$

These limits are independent of the initial distribution but do not necessarily constitute a probability distribution, because m_i *might become infinite. In case the underlying Markov chain is ergodic, the vector* $\boldsymbol{\pi} = (\pi_i)_{i\in S}$ *represents a valid probability distribution.*

Turning attention to the requirement of ergodicity for the existence of a stationary distribution it turns out, that irreducibility and aperiodicity are easy to verify. Proving recurrence often becomes cumbersome. So one often starts from the opposite direction, that is calculating the solution to A.3 first. By the existence of the stationary distribution vector $\boldsymbol{\pi}$, the underlying discrete time Markov chain may be assumed to be positive recurrent. Another useful fact is, that every irreducible Markov chain with a finite number of states is also positive recurrent. Furthermore standard matrix calculus may be applied to derive the solutions.

Although a Markov chain is per definition memoryless, it may be applied to a wider class of models. Therefore a stochastic process is observed only, when state transitions occur. These occurences are called *regeneration points.* The resulting process satisfies the definition of a discrete time Markov chain. It is called an *embedded Markov chain* and proves to be useful especially in the theory of queues.

Markov chains in discrete time have been widely explored, so there exists a vast amount of literature. Classics include [33], [34], [12] and [16]. Numerical aspects relevant for applied Markov chains are discussed in [53]. A very detailed treatment is found in [35].

A.3.2 Homogenous Markov Chains in Continous Time

There are several approaches to the analysis of Markov chains in continous time. We will follow the traditional approach, because it nicely relates to the methods used for discrete time Markov chains. With some slight modifications in notation to reflect the continuity of the parameter the *transition probabilities* are defined as $p_{ij}(s,t) = \Pr\{X(t) = j|X(s) = i\}$. Time-homogenity allows us to write $p_{ij}(s,t) = p_{ij}(0,t-s) =: p_{ij}(t-s)$. Please note, there is no such concept like a single step transition probability, because a dedicated time unit does not exist. Consequently infitesimal calculus has to be applied to gain results in continous time. This in turn requires additional restrictions to be imposed on the *transition rate matrix* $\mathbf{P}(t)$:

Definition 29 *A matrix* $\mathbf{P} = (p_{ij})_{i,j\in S}$ *is called stochastic, if* $p_{ij} > 0$ *for all* $i,j \in S$, $\sum_j p_{ij} = 1$ *for all* $i \in S$ *and at least one element in each column differs from zero.*

Definition 30 $\mathbf{P}(t)$ *is called a transition semigroup on state space* S*, if* $\mathbf{P}(t)$ *is a stochastic matrix,* $\mathbf{P}(0) = \mathbf{I}$ *and* $\mathbf{P}(t+s) = \mathbf{P}(t)\mathbf{P}(s)$.

The last condition assumed for $\mathbf{P}(t)$ to be a transition semigroup is the continous time equivalent of the system of *Chapman Kolmogorov equations.* Furthermore assume, that the transition probabilities are continous at 0, that is $\lim_{t\to 0}\mathbf{P}(t) = \mathbf{P}(0) = \mathbf{I}$. This in turn implies

$$\lim_{t\to 0} p_{ij}(t) = p_{ij}(0)$$

and

$$q_{ij} := \lim_{t\to 0}\frac{p_{ij}(t) - p_{ij}(0)}{t} \tag{A.4}$$

with $0 \le q_{ij} < \infty$ for $i \neq j$ and $q_{ii} \le 0$. Rewritten in matrix notation $\mathbf{Q} := (q_{ij})_{i,j\in S}$ one arrives at the *infinitesimal generator.* The matrix equivalent of A.4 is

$$\mathbf{Q} = \lim_{t\to 0}\frac{\mathbf{P}(t) - \mathbf{P}(0)}{t} = \lim_{t\to 0}\frac{\mathbf{P}(t) - \mathbf{I}}{t}$$

Based on the inifinitesimal generator we are able to define further properties for the transition semigroup $\mathbf{P}(t)$:

Definition 31 $\mathbf{P}(t)$ *is called stable, if* $-q_{ii} < \infty$ *for all* $i \in S$. $\mathbf{P}(t)$ *is called conservative, if* $-q_{ii} = \sum_{j\in S, j\neq i} q_{ij}$ *for all* $i \in S$.

The latter probability derives from the *conservation equality* $\sum_{j\in S} p_{ij}(t) = 1$ for fixed t. In other words, any work performed by the process is preserved.

Rewriting the system of Chapman Kolmogorov equations as $\mathbf{P}(t+s)-\mathbf{P}(t) = \mathbf{P}(t)\mathbf{P}(s) - \mathbf{P}(t)$, dividing by s

$$\frac{\mathbf{P}(t+s)-\mathbf{P}(t)}{s} = \frac{\mathbf{P}(t)\mathbf{P}(s)-\mathbf{P}(t)}{s} = \mathbf{P}(t)\frac{\mathbf{P}(s)-\mathbf{I}}{s} \tag{A.5}$$

and passing to the limit $s \to \infty$ one arrives at *Kolmogorov's forward differential system*

$$\frac{d}{dt}\mathbf{P}(t) = \mathbf{P}(t)\mathbf{Q}$$

Extracting $\mathbf{P}(t)$ in A.5 to the right side results in *Kolmogorov's backward differential system*

$$\frac{d}{dt}\mathbf{P}(t) = \mathbf{Q}\mathbf{P}(t)$$

In traditional notation, these systems may be written as

$$\begin{aligned} \frac{d}{dt}p_{ij}(t) &= p_{ij}(t)q_{jj} + \sum_{k\in S, k\neq j} p_{ik}(t)q_{kj} \qquad (A.6) \\ \frac{d}{dt}p_{ij}(t) &= q_{ii}p_{ij}(t) + \sum_{k\in S, k\neq i} q_{ik}p_{kj}(t) \end{aligned}$$

By embedding a discrete time Markov chain the concepts of irreducibility, communication, transience, recurrence and positive recurrence are inherited. Therefore one has to note, that for each $c > 0$, $Y_n = X(t)$ with $t = cn$ describes a Markov chain in discrete time [26]. Obviously there is no aperiodicity, as we miss a dedicated time unit for continous time Markov chains.

Now we are in the position to calculate the steady state distribution of a Markov chain in continous time. Let $p_j(0) = \Pr\{X(0) = j\}$ denote the *initial probability* for state j and define $\mathbf{p} = (p_j(0))_{j\in S}$ as the *initial probability vector.* Choosing equation A.6 and applying the law of total probability $p_j(t) = \sum_{i\in S} p_{ij}(t)p_i(0)$ results in

$$\frac{d}{dt}p_j(t) = q_{jj}p_j(t) + \sum_{k\in S, k\neq i} p_k(t)q_{kj} \tag{A.7}$$

Irreducibility now assures the existence of the *limiting probabilities* $p_j = \lim_{t\to\infty} p_j(t)$. Assuming an equilibrium, there is no variation in $p_j(t)$, that is $\frac{d}{dt}p_j(t) = 0$. Equation A.7 now becomes

$$0 = q_{jj}p_j(t) + \sum_{k\in S, k\neq i} p_k(t)q_{kj}$$

or in matrix notation

$$\mathbf{0} = \mathbf{pQ} \tag{A.8}$$

This system of equations is often associated with the concept of *global balance*. Forcing the p_j to form a valid probability distribution by imposing the additional restriction

$$\sum_{j \in S} p_j = 1$$

one has successfully derived the *stationary probabilities* with $\mathbf{p}$ the *stationary probability vector.* A similar derivation also exists for Kolmogorov's backward differential system. Knowing how to calculate steady state distributions, one may ask, under what circumstances such solutions remain valid. Based on the definitions of stability and conservativity we are able to state two simple conditions:

Theorem 32 *Given a conservative continous time Markov chain, Kolmogorov's backward differential system is valid. Kolmogorov's forward differential system applies for a stable Markov chain in continous time.*

Please note, that the global balance equations also remain valid for the discrete case, as the infinitesimal generator may be constructed as $\mathbf{Q} = \mathbf{P} - \mathbf{I}$. In either case they have an intuitive interpretation. The flow out of a certain state has to equal the flow into that state. Clearly this concept is related to conservativity and stability. For further reading on the topics discussed we recommend [10] and [53]. Proofs, which were skipped, are found in the former reference.

Index

Bibliography

[1] *Teletraffic Theory and Applications* - H. Akimaru, K. Kawashima - Springer Verlag 1999

[2] *Probability, Statistics and Queueing Theory* - A.O. Allen - Academic Press 1978

[3] On the stability of retrial queues - E. Altmann, A.A. Borovkov - published in *Queueing Systems* 26 (1997), pages 343-363

[4] *Applied Probability and Queues* - S. Asmussen, Springer 2003

[5] *On Queues with Impatient Customers* - F. Bacelli, G. Hebuterne - published in *Performance* (1981) pages 159-179

[6] *Principles of Telecommunication Traffic Enginerring 3rd Edition* - D. Bear - IEE Communications 1988

[7] *The general distributional Little's law and its applications* - D. Bertsimas, D. Nakazato - March 1991

[8] *Stationary Stochastic Models* - A. Frank, B. Lisek, P. Franken - Akademie Verlag 1990

[9] On the $M(n)/M(m)/s$ Queue with impatient Calls - A. Brandt, M.Brandt - published in *Performance Evaluation* 35, pages 1-18

[10] *Markov Chains* - Pierre Brémaud - Springer Verlag 1999

[11] *Superposition of Point Processes* - E. Cinlar 1972 - published in *Stationary Point Processes* from Wiley (1972) pages 549-606

[12] *Introduction to Stochastic Processes* - E. Cinlar - Prentice Hall 1975

[13] *Introduction to Queueing Theory* - R.B. Cooper - Macmillan Publishing 1972

[14] *Renewal Theory* - D. R. Cox - Methuen 1962

[15] *Queueing Theory for Telecommunications* - John N. Daigle - Addison Wesley 1992

[16] *Stochastic Processes* - J. L. Doob - Wiley 1953

[17] *IP Telephony* - Bill Douskalis - Prentice Hall 2000

[18] *Putting VoIP to Work* - Bill Douskalis - Prentice Hall 2002

[19] *Sample-Path Analysis Of Queueing Systems* - M. El-Taha, S. Stidham - Kluwer 1999

[20] *Retrial Queues* - G.I. Falin, J.G.C. Templeton - Chapman & Hall 1997

[21] On a System with Impatience and Repeated Calls - G. Fayolle, M.A. Brun - published in *Queueing Theory and its Applications* by North-Holland (1988), pages 283-305

[22] *Telecommunications Switching, Traffic and Networks* - J.E. Flood - Prentice Hall 1995

[23] *Transform Techniques for Probability Modeling* - W.C. Giffin - Academic Press 1975

[24] *Handbuch der Bedienungstheorie 2* - Autorenkollektiv - Akademie Verlag 1984

[25] *Introduction to Queueing Theory, 2nd Edition* - B.V. Gnedenko, I.N. Kovalenko - Birkhäuser Verlag 1989

[26] *Theorie Stochastischer Prozesse* - Karl Grill - TU Wien 2007

[27] *Fundamentals of Queueing Theory* - D. Gross, C. M. Harris - Wiley 1985

[28] *Queueing Model with State Dependent Balking and Reneging: Its Complementary and Equivalence* - Surenda M. Gupta - published in *Performance Evaluation Review Vol. 22, No. 2-4*

[29] *Performance Modelling of Communication Networks and Computer Architectures* - Peter G. Harrison, Naresh M. Patel - Addison Wesley 1993

[30] *To Queue or Not To Queue, Equilibrium Behavior in Queueing Systems* - R. Hassin, M. Haviv - Kluwer Academic Publishers 2003

[31] *The $G/M/m$ Queue with Finite Waiting Room* - P. Hokstad - published in *Journal of Applied Probability (1975)* pages 779-792

[32] *Basic Stochastic Processes* - R. Iranpour, F. Chagon - Macmillan Publishing 1988

[33] *A First Course in Stochastic Processes* - S. Karlin, H. Taylor - Academic Press 1975

[34] *A Second Course in Stochastic Processes* - S. Karlin, H. Taylor - Academic Press 1981

[35] *Denumerable Markov Chains* - J. Kemeny, J. Snell, A. Knapp - Springer Verlag 1976

[36] *Heuristic approximation for the mean waiting time in the $GI/G/s$ queue* - T. Kimura - published in *Economic Journal of Hokkaido University (1987)*, pages 87-98

[37] *Queueing Systems: Theory* - L. Kleinrock - Wiley 1975

[38] *On the Modification of Rouche's Theorem for Queueing Theory Problems* - V. Klimenok - published in *Queueing Systems 38 (2001)*

[39] *IP Telephony with H.323* - Vineet Kumar, Markku Kopi, Senthil Sengodan - Wiley 2001

[40] *The Palm/Erlang-A Queue, with Applications to Call Centers* - A. Mandelbaum, S. Zeltyn - Technion (Israel) 2005

[41] *Basic Traffic Analysis* - Roberta Martine - AT&T Bell Labs 1994

[42] *Matrix-Geometric Solutions in Stochastic Models* - Marcel F. Neuts - Dover 1994

[43] *Contributions to the Theory on Delay Systems* - C. Palm - published in *Tele (1957)* pages 37-67

[44] *Queues and Inventories, A Study of their Stochastic Processes* - N.U. Prabhu - Wiley 1965

[45] *Stochastic Service Systems* - J. Riordan - Wiley 1962

[46] *Stochastic Networks and Queues* - Philippe Robert - Springer Verlag 2003

[47] *Diffusions, Markov Processes and Martingales Volume 1+2* - L. Rogers, D. Williams - Cambridge University Press 2000

[48] *Applied Probability Models With Optimization Applications* - Sheldon M. Ross - Holden Day 1970

[49] *Elements of Queueing Theory with Applications* - T.L. Saaty - McGraw Hill 1961

[50] *Warteschlangen* - R. Schassberger - Springer Verlag 1973

[51] *Special Functions in Queueing Theory* - H.M. Srivastava, B.R.K. Kashyap - Academic Press 1982

[52] *Complex Analysis* - I. Stewart, D. Tall - Cambridge University Press 1983

[53] *Introduction to the Numerical Solution of Markov Chains* - William J. Stewart - Princeton University Press 1994

[54] Queueing with Balking and Reneging in $M/G/1$ Systems - S. Subba Rao - published in *Metrika* 12 (1967)

[55] Balking and Reneging in $M/G/1$ Systems with Post–Ponable Interruptions - S. Subba Rao - published in *Metrika* 14 (1969)

[56] *Introduction to the Theory of Queues* - L. Takacs - Oxford University Press 1962

[57] *Queueing Analysis Volume 1 (Vacation and Priority Systems)* - H. Takagi - North Holland 1991

[58] *Queueing Analysis Volume 2 (Finite Systems)* - H. Takagi - North Holland 1993

[59] *Analytische Leistungsbewertung verteilter Systeme* - Phuoc Tran Gia - Springer Verlag 1996

[60] *Algorithms and Approximations for Queueing Systems* - M.H. van Hoorn - Mathematisch Centrum 1984

[61] *Internet QoS* - Zheng Wang - Morgan Kaufmann Publishers 2001

[62] A Diffusion Approximation for a $GI/GI/1$ Queue with Balking and Reneging - A. Ward, P. Glynn - published in *Queueing Systems* 50 (2005) pages 371-400

[63] *Improving Service by Informing Customers about Anticipated Delays* - Ward Whitt 1998 - published in *Management Science 45 (1999)* pages 192-207

[64] *Stochastic Modeling and the Theory of Queues* - R.W. Wolff - Prentice Hall 1989

[65] Call Centers with Impatient Customers: Many-Server Asymptotics of the $M/M/n+G$ Queue - A. Mandelbaum, S. Zeltyn 2005 - published in *Queueing Systems* 51 (2005) pages 361-402

[66] *Birth and Death Processes and Markov Chains* - W. Zikun, Y. Yiangqun - Springer Verlag 1992