Proof of the Erlang C formula

Alexander Herzog
alexander.herzog@tu-clausthal.de
a-herzog.github.io/QueueCalc

1 The Erlang C model

The Erlang C queueing model describes a service system with exponentially distributed inter-arrival times with an arrival rate of $\lambda > 0$, exponentially distributed service times with a service rate of $\mu > 0$ and $c \in \mathbb{N}$ operators at the service desk. The arriving customers form a queue in front of the service desk. Each time an operator gets idle, the customer who has waited for the longest time will be routed to the free operator (FIFO queueing discipline; first-in-first-out), see figure 1.

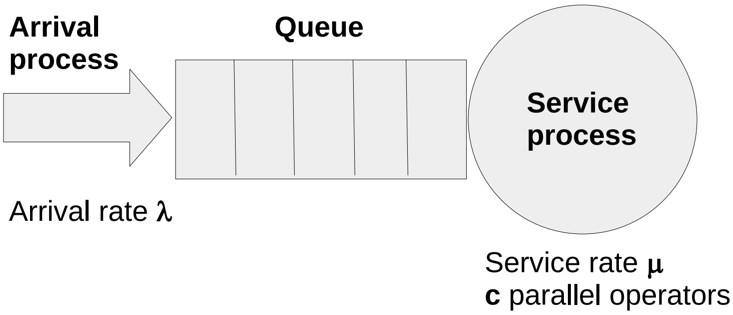

Figure 1: Erlang C queueing model

The long-run utilization of the system is

$$\rho := \frac{\lambda}{c\mu}\,.$$

The system can only operate stably in the long term if $\rho < 1$ (the long-run arrival rate is lower than the available workforce).

The following indicators are usually of interest:

- $P(W \leq t)$: probability for a new arriving customer to have to wait no longer than $t \geq 0$ seconds (Erlang C formula),
- $\mathbf{E}[W]$: average waiting time,
- $\mathbf{E}[V]$: average residence time,
- $\mathbf{E}[N_Q]$: average number of customers in the queue,
- $\mathbf{E}[N]$: average number of customers in the system.

2 Idea for calculating $P(W \leq t)$

The idea for calculating the customer's waiting time distribution $P(W \leq t)$ (the probability that the customer has to wait at most $t \geq 0$ seconds) is to determine the waiting time probability distribution for all possible states in which the system can be and to weight these times with the probability that the system is in the respective state. This leads to the formula:

$$P(W \leq t) = \sum_{n=0}^{\infty} P_{\mu,n-(c-1)}(W \leq t) \cdot p_n \tag{1}$$

where p_n in the probability that there are n customers in the system and $P_{\mu,m}(W \leq t)$ is the probability for a new customer to have to wait t or less seconds in the case the service rate is $\mu > 0$ and there are m customers that will have to be served to the end before the new customer's service process starts. For $m = 0$ the new service process will start immediately ($P_{\mu,0}(W \leq t) = 1$ for all $t \geq 0$) and for consistence we assume the same for $m < 0$.

Case $n \geq c$**:**
If there are $n < c$ customers in the system, a newly arriving customer can be served immediately, as in this case not all operators are currently busy. If $n \geq c$, one or more already running service processes has to end first, and then any customers who were already waiting before the newly arrived customer must be served before the newly arrived customer can be served. A distinction must therefore be made between the service times of customers whose service has yet to begin and the remaining service times of customers who are already in the service process. However, since the Erlang models uses the exponential distribution for the service times, these two time durations are subject to the same probability distribution with the same parameter. This is due to the memorylessness of the exponential distribution. (This property of the exponential distribution will be proofed below in section 3).

Case $n < c$**:**
For $n < c$ customers in the system, the probability that a newly arriving customer will have to wait less than $t \geq 0$ seconds is always 1 for all $t \geq 0$ (the newly arriving customer does not have to wait at all). This means we have $P_{\mu,n-(c-1)}(W \leq t) = 1$ for $n < c$ and for all $t \geq 0$ and formula (1) can be simplified to:

$$P(W \leq t) = \sum_{n=0}^{c-1} p_n + \sum_{n=c}^{\infty} P_{\mu,n-(c-1)}(W \leq t) \cdot p_n \,. \tag{2}$$

Since $P_{\mu,n-(c-1)}(W \leq t)$ represents the sequential execution of several exponential distributions (and thus is an Erlang distribution, see section 5) and p_n is the probability that the corresponding Markov chain is in state n, the following sections will present results on exponential distributions (section 3), Erlang distributions (section 5), and birth-and-death processes (section 6), which will ultimately allow us to calculate formula (2) in section 7.

3 Exponential distribution

Since the inter-arrival and the service times are distributed due to the exponential distribution in an Erlang model, we will need some results for the exponential distribution which will be presented and proofed here.

Definition 1 (Probability density function of the exponential distribution)**.** The function

$$f_\lambda(x) := \begin{cases} \lambda \mathrm{e}^{-\lambda x}, & x \geq 0\,, \\ 0, & x < 0 \end{cases}$$

with $\lambda > 0$ is called the probability density function of the exponential distribution.

Theorem 1 (Cumulative distribution function of the exponential distribution). *The function*

$$F_\lambda(x) := \begin{cases} 1 - \mathrm{e}^{-\lambda x}, & x \geq 0\,, \\ 0, & x < 0 \end{cases}$$

is the cumulative distribution function of the exponential distribution with probability density function $f_\lambda(x)$, $\lambda > 0$.

Proof. According to the definition of the cumulative distribution function $F(x) := \int_{-\infty}^{x} f(t)\,\mathrm{d}t$ holds:

$$F_\lambda(x) = \int_0^x \lambda \mathrm{e}^{-\lambda t}\,\mathrm{d}t = \left[-\mathrm{e}^{-\lambda t}\right]_{t=0}^{t=x} = -\mathrm{e}^{-\lambda x} - (-1) = 1 - \mathrm{e}^{-\lambda x}$$

for $x \geq 0$ and $F_\lambda(x) = 0$ else. □

Theorem 2. *The function $f_\lambda(x)$, $\lambda > 0$, is the probability density function of a probability distribution, i. e. $f_\lambda(x) \geq 0$ and $\int_{-\infty}^{\infty} f_\lambda(x)\,dx = 1$.*

Proof. The proposition $f_\lambda(x) \geq 0$ follows directly from the definition of $f_\lambda(x)$. The previous theorem gives us:

$$\int_{-\infty}^{\infty} f_\lambda(t)\,\mathrm{d}t = \lim_{x\to\infty} F_\lambda(x) = \lim_{x\to\infty} 1 - \mathrm{e}^{-\lambda x} = 1\,.$$

□

Theorem 3 (Indicators of the exponential distribution). *For an exponential distribution with parameter $\lambda > 0$, the following holds:*

1. $\mathbf{E}[X] = \frac{1}{\lambda}$,
2. $\mathbf{Std}[X] = \frac{1}{\lambda}$ *and*
3. $\mathbf{CV}[X] = \mathbf{SCV}[X] = 1$.

Proof. 1. With the definition of $\mathbf{E}[X]$ follows:

$$\begin{aligned}\mathbf{E}[X] &= \int_{-\infty}^{\infty} x \cdot f_\lambda(x)\,\mathrm{d}x = \int_0^\infty \lambda x \mathrm{e}^{-\lambda x}\,\mathrm{d}x \\ &= \left[\lambda x \cdot \left(-\frac{1}{\lambda}\right)\mathrm{e}^{-\lambda x}\right]_{x=0}^{x=\infty} - \int_0^\infty \lambda\left(-\frac{1}{\lambda}\right)\mathrm{e}^{-\lambda x}\,\mathrm{d}x = 0 + \int_0^\infty \mathrm{e}^{-\lambda x}\,\mathrm{d}x \\ &= \left[\left(-\frac{1}{\lambda}\right)\mathrm{e}^{-\lambda x}\right]_{x=0}^{x=\infty} = 0 - \left(-\frac{1}{\lambda}\right) = \frac{1}{\lambda}\,.\end{aligned}$$

2. With the definition of $\mathbf{E}[X^2]$ follows initially:

$$\begin{aligned}\mathbf{E}[X^2] &= \int_{-\infty}^{\infty} x^2 \cdot f_\lambda(x)\,\mathrm{d}x = \int_0^\infty \lambda x^2 \mathrm{e}^{-\lambda x}\,\mathrm{d}x \\ &= \left[\lambda x^2\left(-\frac{1}{\lambda}\right)\mathrm{e}^{-\lambda x}\right]_{x=0}^{x=\infty} - \int_0^\infty 2\lambda x\left(-\frac{1}{\lambda}\right)\mathrm{e}^{-\lambda x}\,\mathrm{d}x = 0 + \int_0^\infty 2x\mathrm{e}^{-\lambda x}\,\mathrm{d}x \\ &= \left[2x\left(-\frac{1}{\lambda}\right)\mathrm{e}^{-\lambda x}\right]_{x=0}^{x=\infty} - \int_0^\infty 2\left(-\frac{1}{\lambda}\right)\mathrm{e}^{-\lambda x}\,\mathrm{d}x = 0 + \frac{2}{\lambda}\left[\left(-\frac{1}{\lambda}\right)\mathrm{e}^{-\lambda x}\right]_{x=0}^{x=\infty} \\ &= 0 - \frac{2}{\lambda}\left(-\frac{1}{\lambda}\right) = \frac{2}{\lambda^2}\,.\end{aligned}$$

Using the shift theorem we get:

$$\mathbf{Var}[X] = \mathbf{E}[X^2] - (\mathbf{E}[X])^2 = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}\,.$$

And so, ultimately, the following applies to the standard deviation:

$$\mathbf{Std}[X] = \sqrt{\mathbf{Var}[X]} = \frac{1}{\lambda}\,.$$

3. From 1 and 2 follows directly:

$$\mathbf{CV}[X] = \frac{\mathbf{Std}[X]}{|\mathbf{E}[X]|} = \frac{\frac{1}{\lambda}}{\frac{1}{\lambda}} = 1\,.$$

Furthermore, $\mathbf{SCV}[X] = (\mathbf{CV}[X])^2 = 1$.

□

3.1 Memorylessness of the exponential distribution

Theorem 4 (Memorylessness of the exponential distribution)**.** *If a time interval is exponentially distributed and the event has not yet occurred at time $s \geq 0$, then the time distribution that the event will occur within the next $t \geq 0$ time units (i. e. up to time $s + t$), is again exponentially distributed with the same parameter $\lambda > 0$, i. e. the following applies:*

$$P(X \leq t + s | X \geq s) = P(X \leq t)\,, \tag{3}$$

where $P(X \leq a | X \geq b)$ is the conditional probability of the event "$X \leq a$" knowing in advance that $X \geq b$.

Proof. Using the calculation rules for conditional probabilities, the following holds:

$$\begin{aligned} P(X \leq t + s | X \geq s) &= \frac{P(\{X \leq t + s\} \cap \{X \geq s\})}{P(X \geq s)} = \frac{F_\lambda(t+s) - F_\lambda(s)}{1 - F_\lambda(s)} \\ &= \frac{1 - \mathrm{e}^{-\lambda(t+s)} - 1 + \mathrm{e}^{-\lambda s}}{1 - (1 - \mathrm{e}^{-\lambda s})} = \frac{\mathrm{e}^{-\lambda s} - \mathrm{e}^{-\lambda(t+s)}}{\mathrm{e}^{-\lambda s}} = 1 - \mathrm{e}^{-\lambda t} \\ &= F_\lambda(t) = P(X \leq t)\,. \end{aligned}$$

□

Consequence:
If the service time of a customer is exponentially distributed with parameter $\lambda > 0$ and the customer is already in the service process, then the remaining service time of the customer is again exponentially distributed with the same parameter λ. If the cumulative distribution function of the waiting time of a newly arriving customer is to be determined, no distinction needs to be made between the remaining service times of the customers currently being served and the service times of the other customers waiting before the newly arrived customer.

Theorem 5. *The exponential distribution $F_\lambda(x)$ is the only* continuous *probability distribution with $F(x) = 0$ for $x \leq 0$ that has the property of memorylessness.*

Proof. The proof consists of three steps:

Step 1: Functional equation describing the memorylessness
In order to present the proof as simply as possible, we first define the function $G(t) := 1 - F(t)$, which indicates the probability that the event has *not yet* occurred by time $t \geq 0$. (This definition comes from reliability theory, where $G(t)$ is called the *survival probability*.) From equation (3), using the calculation rules for conditional probabilities, we get:

$$\begin{aligned} & P(X \leq t+s | X \geq s) = P(X \leq t) \iff \frac{F(t+s)-F(s)}{1-F(s)} = F(t) \\ \iff\quad & \frac{1-G(t+s)-(1-G(s))}{G(s)} = 1 - G(t) \\ \iff\quad & -G(t+s)+G(s) = G(s) - G(s)G(t) \iff G(s+t) = G(s)G(t)\,. \end{aligned}$$

Therefore, we have to show that only the survival probability assuming an exponential distribution, i.e., $G_\lambda(t) = \mathrm{e}^{-\lambda t}$, satisfies the equation $G(s+t) = G(s)G(t)$.

Step 2: Explicit representation for $G(t)$
From $G(s+t) = G(s)G(t)$ follows for $a \in \mathbb{N}$ and $b \in \mathbb{N}$:

$$\begin{aligned} G\left(\frac{a}{b}\right) &= G\left(\frac{1}{b}+\frac{a-1}{b}\right) = G\left(\frac{1}{b}\right)\cdot G\left(\frac{a-1}{b}\right) = \ldots = \underbrace{G\left(\frac{1}{b}\right)\cdots G\left(\frac{1}{b}\right)}_{a \text{ factors}} \\ &= \left[G\left(\frac{1}{b}\right)\right]^a . \end{aligned}$$

From $1 = \frac{b}{b}$ follows $G(1) = \left[G\left(\frac{1}{b}\right)\right]^b$ respectively $[G(1)]^{\frac{1}{b}} = G\left(\frac{1}{b}\right)$. This leads to:

$$G\left(\frac{a}{b}\right) = \left[G\left(\frac{1}{b}\right)\right]^a = [G(1)]^{\frac{a}{b}} .$$

This means that the relationship $G(q) = [G(1)]^q$ holds for all positive rational numbers q. Due to the fact that F was assumed to be continuous and therefore G is also continuous, $G(t) = [G(1)]^t$ applies for all real $t \geq 0$. Furthermore, the following applies with the calculation rules for the exponential and logarithms function:

$$G(t) = \mathrm{e}^{\ln\left(G(1)^t\right)} = \mathrm{e}^{t\ln(G(1))}\,.$$

Choosing $\lambda := -\ln(G(1))$ yields $G(t) = \mathrm{e}^{-\lambda t}$ and $F(t) = 1 - \mathrm{e}^{-\lambda t}$. It remains to be shown that $0 < \lambda < \infty$ respectively $0 < G(1) < 1$ applies.

Step 3: Proof that $0 < G(1) < 1$ holds
It was $G(1) = 1 - F(1)$. This immediately implies that $0 \leq G(1) \leq 1$. It remains to show that $0 \neq G(1) \neq 1$.

- If $G(1) = 0$, then $G(x) = [G(1)]^x = 0^x = 0$ would apply for all $x > 0$. This would mean that $F(x) = 1$ for all $x > 0$. However, since F was assumed to be continuous and $F(0) = 0$ according to the assumption, this cannot be the case.

- If $G(1) = 1$, then $G(x) = [G(1)]^x = 1^x = 1$ would hold for all $x \geq 0$. This would mean that $F(x) = 0$ for all $x \geq 0$, which contradicts the fact that $F(x)$ is the cumulative distribution function of a probability distribution and therefore $\lim_{x\to\infty} F(x) = 1$ must hold.

□

4 Convolution of probability distributions

If two processes that can be described by probability distributions are executed one after the other, we say that the probability density functions of the two probability distributions are convolved with each other. The result of the convolution is the probability density function of the probability distribution that describes the overall process.

Definition 2 (Convolution)**.** If $f(x)$ and $g(x)$ are probability density functions of two probability distributions, the expression

$$f * g(x) := \int_{-\infty}^{\infty} f(t)g(x-t)\,\mathrm{d}t$$

is called the convolution of f and g.

Illustratively, the above formula integrates all conceivable cases in which the total time x can be divided between the two subprocesses f and g.

Since convolution is defined by an integral over the product of two probability density functions, it is the case that for all probability distributions where it is difficult to determine the integral of the probability density functions – which applies to almost all continuous probability distributions except the exponential distribution – it is very difficult or impossible to calculate explicitly. This is the reason why the exponential distribution is always used for service times in the Erlang formulas.

5 Erlang distribution

As we will see next, the Erlang distribution describes a process consisting of the sequential execution of $n \in \mathbb{N}$ subprocesses, each of which has an exponential distribution with the same parameter $\lambda > 0$.

In an M/M/1 system, the waiting time distribution of a newly arriving customer can be represented by the service times of the customers waiting in front of him and the remaining service time of the customer currently being served. However, due to the memorylessness of the exponential distribution (see equation (3)), the remaining service time of the customer currently being served is distributed in the same way as his total service time. This means that if there are $k \geq 0$ customers already waiting in the queue before the newly arriving customer and the service times are exponentially distributed with parameter $\mu \geq 0$, then the waiting time of the new customer is Erlang distributed with parameters μ and $k+1$.

Definition 3 (Probability density function of the Erlang distribution)**.** The function

$$f_{\lambda,n}(x) := \begin{cases} \frac{\lambda^n x^{n-1}}{(n-1)!}\mathrm{e}^{-\lambda x}, & x \geq 0\,, \\ 0, & x < 0 \end{cases}$$

with $\lambda > 0$ and $n \in \mathbb{N}$ is called the probability density function of the Erlang distribution.

Theorem 6 (Connection to the exponential distribution)**.** *The function $f_{\lambda,n}(x)$, $\lambda > 0$, $n \in \mathbb{N}$ is the n-fold convolution of $f_\lambda(x)$.*

Proof. The proof is carried out by induction. Obviously it is true that $f_{\lambda,1}(x) = \lambda\mathrm{e}^{-\lambda x} = f_\lambda(x)$. For the convolution of two exponentially distributed time durations, the following applies:

$$f_{\lambda,1} * f_\lambda(x) = f_\lambda * f_\lambda(x) = \int_0^x \lambda\mathrm{e}^{-\lambda t} \cdot \lambda\mathrm{e}^{-\lambda(x-t)}\,\mathrm{d}t = \left[\lambda^2\mathrm{e}^{-\lambda x}t\right]_{t=0}^{t=x} = \lambda^2 x\mathrm{e}^{-\lambda x} = f_{\lambda,2}(x)$$

for $x \geq 0$ and $f_{\lambda,2}(x) = 0$ else. Now we assume that

$$f_{\lambda,n}(x) = \underbrace{f_\lambda * \cdots * f_\lambda}_{n\ -\text{times}}(x)$$

for some $n \in \mathbb{N}$. Then holds:

$$\begin{aligned}
f_{\lambda,n} * f_\lambda(x) &= \int_0^x \frac{\lambda^n t^{n-1}}{(n-1)!}\mathrm{e}^{-\lambda t} \cdot \lambda \mathrm{e}^{-\lambda(x-t)}\,\mathrm{d}t = \frac{\lambda^{n+1}}{(n-1)!}\mathrm{e}^{-\lambda x}\int_0^x t^{n-1}\,\mathrm{d}t \\
&= \frac{\lambda^{n+1}}{(n-1)!}\mathrm{e}^{-\lambda x}\left[\frac{1}{n}t^n\right]_{t=0}^{t=x} = \frac{\lambda^{n+1}x^n}{n!}\mathrm{e}^{-\lambda x} = f_{\lambda,n+1}(x)
\end{aligned}$$

for $x \geq 0$ and $f_{\lambda,n+1}(x) = 0$ else. □

Theorem 7 (Cumulative distribution function of the Erlang distribution)**.** *The function*

$$F_{\lambda,n}(x) := \begin{cases} 1 - \mathrm{e}^{-\lambda x}\sum_{i=0}^{n-1}\frac{(\lambda x)^i}{i!}, & x \geq 0\ , \\ 0, & x < 0 \end{cases}$$

is the cumulative distribution function of the Erlang distribution with probability density function $f_{\lambda,n}(x)$, $\lambda > 0$, $n \in \mathbb{N}$.

Proof. The proof is also carried out by induction. For $n = 1$ holds:

$$F_{\lambda,1}(x) = 1 - \mathrm{e}^{-\lambda x}\underbrace{\sum_{i=0}^{1-1}\frac{(\lambda x)^i}{i!}}_{=1} = F_\lambda(x)\ .$$

Since $F_\lambda(x)$ is the cumulative distribution function of the exponential distribution and $f_{\lambda,1}(x) = f_\lambda(x)$ held, the initial step of induction is thus proven. It is now assumed that it is already known that

$$F_{\lambda,n-1}(x) = 1 - \mathrm{e}^{-\lambda x}\sum_{i=0}^{n-2}\frac{(\lambda x)^i}{i!}$$

für some $n \in \mathbb{N}$, $n \geq 2$, for $x \geq 0$. Then holds:

$$\begin{aligned}
F_{\lambda,n}(x) &= \int_0^x f_{\lambda,n}(t)\,\mathrm{d}t = \int_0^x \frac{\lambda^n t^{n-1}}{(n-1)!}\mathrm{e}^{-\lambda t}\,\mathrm{d}t \\
&= \left[\frac{\lambda^n t^{n-1}}{(n-1)!}\left(-\frac{1}{\lambda}\mathrm{e}^{-\lambda t}\right)\right]_{t=0}^{t=x} - \int_0^x \frac{\lambda^n t^{n-2}}{(n-2)!}\left(-\frac{1}{\lambda}\mathrm{e}^{-\lambda t}\right)\,\mathrm{d}t \\
&= -\frac{(\lambda x)^{n-1}}{(n-1)!}\mathrm{e}^{-\lambda x} + \int_0^x f_{\lambda,n-1}(t)\,\mathrm{d}t = -\frac{(\lambda x)^{n-1}}{(n-1)!}\mathrm{e}^{-\lambda x} + F_{\lambda,n-1}(t) \\
&= -\frac{(\lambda x)^{n-1}}{(n-1)!}\mathrm{e}^{-\lambda x} + 1 - \mathrm{e}^{-\lambda x}\sum_{i=0}^{n-2}\frac{(\lambda x)^i}{i!} = 1 - \mathrm{e}^{-\lambda x}\sum_{i=0}^{n-1}\frac{(\lambda x)^i}{i!}
\end{aligned}$$

for $x \geq 0$ and $F_{\lambda,n}(x) = 0$ else. □

Theorem 8. *The function $f_{\lambda,n}(x)$, $\lambda > 0$, $n \in \mathbb{N}$, is the probability density function of a probability distribution, i. e. $f_{\lambda,n}(x) \geq 0$ and*

$$\int_{-\infty}^{\infty} f_{\lambda,n}(x)\,dx = 1\ .$$

Proof. The proposition $f_{\lambda,n}(x) \geq 0$ follows directly from the definition of $f_{\lambda,n}(x)$. Using the series representation of the exponential function, applying L'Hospital's rule n times yields:

$$F_{\lambda,n}(x) = 1 - \mathrm{e}^{-\lambda x}\sum_{i=0}^{n-1}\frac{(\lambda x)^i}{i!} = 1 - \frac{\sum_{i=0}^{n-1}\frac{(\lambda x)^i}{i!}}{\sum_{i=0}^{\infty}\frac{(\lambda x)^i}{i!}} \overset{x\to\infty}{\longrightarrow} 1\ .$$

Therefore, the following applies:

$$\int_{-\infty}^{\infty} f_{\lambda,n}(t)\,\mathsf{d}t = \lim_{x\to\infty} F_{\lambda,n}(x) = 1\,.$$

< □

Theorem 9 (Indicators of the Erlang distribution)**.** *For an Erlang distribution with parameters $\lambda > 0$ and $n \in \mathbb{N}$, the following applies:*

1. $\mathbf{E}[X] = \frac{n}{\lambda}$,
2. $\mathbf{Std}[X] = \frac{\sqrt{n}}{\lambda}$ *and*
3. $\mathbf{CV}[X] = \frac{1}{\sqrt{n}}$ *and* $\mathbf{SCV}[X] = \frac{1}{n}$.

Proof. The Erlang distribution with parameters $\lambda > 0$ and $n \in \mathbb{N}$ represents the sequential execution of n independent exponential distributions with parameter λ. Due to the independence of the individual subprocesses, not only the expected values can be added, but also the variances. For the expected value of the Erlang distribution, this immediately follows

$$\mathbf{E}[X] = n \cdot \frac{1}{\lambda} = \frac{n}{\lambda}\,.$$

The following applies to the standard deviation:

$$\mathbf{Std}[X] = \sqrt{\mathbf{Var}[X]} = \sqrt{n \cdot \frac{1}{\lambda^2}} = \frac{\sqrt{n}}{\lambda}\,.$$

This means that the following applies directly to the coefficient of variation:

$$\mathbf{CV}[X] = \frac{\mathbf{Std}[X]}{|\mathbf{E}[X]|} = \frac{\frac{\sqrt{n}}{\lambda}}{\frac{n}{\lambda}} = \frac{1}{\sqrt{n}}\,.$$

Furthermore we have $\mathbf{SCV}[X] = (\mathbf{CV}[X])^2 = \frac{1}{n}$. □

6 Birth-and-death processes

Markov processes in which only transitions to the next higher state or the next lower state are possible (e. g., arrivals and departures of individual customers) are called birth-and-death processes. In a birth-and-death processes only transitions to the next higher (a birth respectively an arrival) and to the next lower state (a death respectively a departure) are possible, see figure 2.

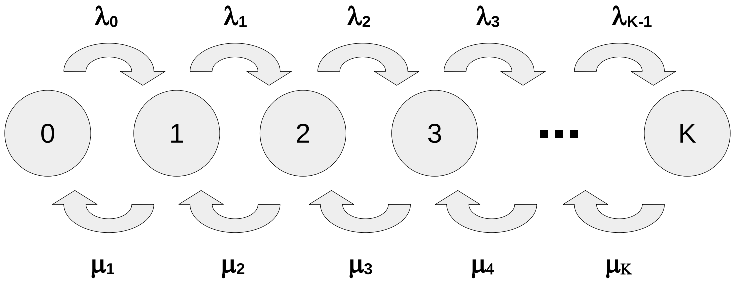

Figure 2: A birth-and-death process as a Markov chain

6.1 Arrival and service rate

For the state-dependent arrival rate at the system in a service process with $K \in \mathbb{N}$ waiting and service places, the following applies to state $n \in \mathbb{N}_0$:

$$\lambda_n^{\mathrm{Sys}} := \begin{cases} \lambda, & n < K \,, \\ 0, & n = K \,, \end{cases} \tag{4}$$

where $\lambda > 0$ is the arrival rate according to the customers. If there are already $n = K$ customers in the system, no further customers can enter the system, i. e., in this case, $\lambda^{\mathrm{Sys}} = 0$.

The following applies to the state-dependent service rate for the entire system:

$$\mu_n^{\mathrm{Sys}} := \begin{cases} n\mu, & n \leq c \,, \\ c\mu, & n > c \,, \end{cases} \tag{5}$$

where $\mu > 0$ is the service rate of an individual operator.

6.2 Transient state probabilities

It is now assumed that the probability $p_n(t)$ that a total of $n \in \mathbb{N}_0$ customers are in the system at a given point in time $t \geq 0$ is known. On this basis, the probability that $n \in \mathbb{N}_0$ customers are in the system at a time $t + \Delta t$ will be determined. Here, $\Delta t > 0$ can be thought of as a small amount of time that has elapsed since time t. In the case of $n = 0$, $p_0(t + \Delta t)$ is composed of the probability that there were already 0 people in the system at time t and no customer arrived, and the probability that there was one customer in the system at time t and that this customer was served within the time Δt. Furthermore, a customer could have arrived and already left the system again during the time period Δt, etc. However, compared to the first two probabilities mentioned, the probabilities for such combined events are negligible – mathematically speaking, they converge to 0 faster than Δt for $\Delta t \to 0$. This is expressed by the Landau symbol $o(\Delta t)$. This results in $n = 0$:

$$p_0(t + \Delta t) = p_0(t)(1 - \Delta t \lambda_0^{\mathrm{Sys}}) + p_1(t)\Delta t \mu_1^{\mathrm{Sys}} + o(\Delta t) \,.$$

For $n = 1, \ldots, K - 1$, there is also the possibility that state n was reached from state $n - 1$:

$$p_n(t + \Delta t) = p_n(t)(1 - \Delta t \lambda_n^{\mathrm{Sys}} - \Delta t \mu_n^{\mathrm{Sys}}) + p_{n-1}(t)\Delta t \lambda_{n-1}^{\mathrm{Sys}} + p_{n+1}(t)\Delta t \mu_{n+1}^{\mathrm{Sys}} + o(\Delta t) \,.$$

Finally, for $n = K$, the state can only be reached from the previous states n and $n - 1$:

$$p_K(t + \Delta t) = p_K(t)(1 - \Delta t \mu_K^{\mathrm{Sys}}) + p_{K-1}(t)\Delta t \lambda_{K-1}^{\mathrm{Sys}} + o(\Delta t) \,.$$

6.3 Stationary state probabilities

Term transformation in the three equations above initially yields (for $n = 1, \ldots, K - 1$ in the middle equation):

$$\begin{aligned} \frac{p_0(t + \Delta t) - p_0(t)}{\Delta t} &= -p_0(t)\lambda_0^{\mathrm{Sys}} + p_1(t)\mu_1^{\mathrm{Sys}} + \frac{o(\Delta t)}{\Delta t} \,, \\ \frac{p_n(t + \Delta t) - p_n(t)}{\Delta t} &= -p_n(t)(\lambda_n^{\mathrm{Sys}} + \mu_n^{\mathrm{Sys}}) + p_{n+1}(t)\mu_{n+1}^{\mathrm{Sys}} + p_{n-1}(t)\lambda_{n-1}^{\mathrm{Sys}} + \frac{o(\Delta t)}{\Delta t} \,, \\ \frac{p_K(t + \Delta t) - p_K(t)}{\Delta t} &= -p_K(t)\mu_K^{\mathrm{Sys}} + p_{K-1}(t)\lambda_{K-1}^{\mathrm{Sys}} + \frac{o(\Delta t)}{\Delta t} \,. \end{aligned} \tag{6}$$

If we now choose increasingly shorter time intervals for Δt, i. e., if we perform the limit transition $\Delta t \to 0$, we obtain the following for $n = 0, \ldots, K$ on the left-hand sides of the equations (6):

$$\lim_{\Delta t \to 0} \frac{p_n(t + \Delta t) - p_n(t)}{\Delta t} = p_n'(t) \,.$$

Furthermore, $\lim_{\Delta t \to 0} \frac{o(\Delta t)}{\Delta t} = 0$ applies according to the definition of $o(\Delta t)$. Thus, for $\Delta t \to 0$, from (6):

$$\begin{aligned} p_0'(t) &= -p_0(t)\lambda_0^{\text{Sys}} + p_1(t)\mu_1^{\text{Sys}} \, , \\ p_n'(t) &= -p_n(t)(\lambda_n^{\text{Sys}} + \mu_n^{\text{Sys}}) + p_{n+1}(t)\mu_{n+1}^{\text{Sys}} + p_{n-1}(t)\lambda_{n-1}^{\text{Sys}} \, , \\ p_K'(t) &= -p_K(t)\mu_K^{\text{Sys}} + p_{K-1}(t)\lambda_{K-1}^{\text{Sys}} \, . \end{aligned} \tag{7}$$

Looking at a queueing system in steady state means looking at the system after a very long time, like for $t \to \infty$. The steady state is characterized in a system that reaches it, i. e., in which $\rho < 1$, by the fact that the state probabilities stabilize, i. e., $p_n'(t) \to 0$ applies for all $n = 0, \ldots, K$ for $t \to \infty$. If we set $p_n'(t) = 0$ in (7), we obtain the following linear system of equations for the stationary, i. e., time-independent state probabilities $p_n := \lim_{t\to\infty} p_n(t)$:

$$\begin{aligned} p_1 &= \frac{\lambda_0^{\text{Sys}}}{\mu_1^{\text{Sys}}} p_0 \, , \\ p_{n+1} &= \frac{\lambda_n^{\text{Sys}}}{\mu_{n+1}^{\text{Sys}}} p_n + \frac{\mu_n^{\text{Sys}}}{\mu_{n+1}^{\text{Sys}}} p_n - \frac{\lambda_{n-1}^{\text{Sys}}}{\mu_{n+1}^{\text{Sys}}} p_{n-1} \, , \\ p_K &= \frac{\lambda_{K-1}^{\text{Sys}}}{\mu_K^{\text{Sys}}} p_{K-1} \, . \end{aligned} \tag{8}$$

Theorem 10. *With the above notations, the following applies:*

$$p_n = \prod_{i=1}^{n} \frac{\lambda_{i-1}^{\text{Sys}}}{\mu_i^{\text{Sys}}} p_0 \quad \textit{für } n = 1, \ldots, K \quad \textit{and} \quad p_0 = \left[\sum_{n=1}^{K} \prod_{i=1}^{n} \frac{\lambda_{i-1}^{\text{Sys}}}{\mu_i^{\text{Sys}}} + 1 \right]^{-1} \, . \tag{9}$$

Proof. The proof for p_n, $n = 1, \ldots, K$, is carried out by induction. The initial condition for induction follows directly from the first equation in (8). Let us now assume that equation (9) holds for some $n \in \mathbb{N}$. Then, from the second equation of (8), we obtain for $n + 1$:

$$\begin{aligned} p_{n+1} &= \frac{\lambda_n^{\text{Sys}}}{\mu_{n+1}^{\text{Sys}}} p_n + \frac{\mu_n^{\text{Sys}}}{\mu_{n+1}^{\text{Sys}}} p_n - \frac{\lambda_{n-1}^{\text{Sys}}}{\mu_{n+1}^{\text{Sys}}} p_{n-1} \\ &= \frac{\lambda_n^{\text{Sys}}}{\mu_{n+1}^{\text{Sys}}} \prod_{i=1}^{n} \frac{\lambda_{i-1}^{\text{Sys}}}{\mu_i^{\text{Sys}}} p_0 + \frac{\mu_n^{\text{Sys}}}{\mu_{n+1}^{\text{Sys}}} \prod_{i=1}^{n} \frac{\lambda_{i-1}^{\text{Sys}}}{\mu_i^{\text{Sys}}} p_0 - \frac{\lambda_{n-1}^{\text{Sys}}}{\mu_{n+1}^{\text{Sys}}} \prod_{i=1}^{n-1} \frac{\lambda_{i-1}^{\text{Sys}}}{\mu_i^{\text{Sys}}} p_0 \\ &= \prod_{i=1}^{n+1} \frac{\lambda_{i-1}^{\text{Sys}}}{\mu_i^{\text{Sys}}} p_0 + \underbrace{\left[\frac{\mu_n^{\text{Sys}} \lambda_{n-1}^{\text{Sys}}}{\mu_{n+1}^{\text{Sys}} \mu_n^{\text{Sys}}} - \frac{\lambda_{n-1}^{\text{Sys}}}{\mu_{n+1}^{\text{Sys}}} \right]}_{=0} \prod_{i=1}^{n-1} \frac{\lambda_{i-1}^{\text{Sys}}}{\mu_i^{\text{Sys}}} p_0 \, . \end{aligned}$$

The queueing system is always be in one of the states $0, \ldots, K$, i. e., the sum of all probabilities is 1 ($\sum_{n=0}^{K} p_n = 1$). Using the formula p_n, we obtain:

$$p_0 + \sum_{n=1}^{K} \prod_{i=1}^{n} \frac{\lambda_{i-1}^{\text{Sys}}}{\mu_i^{\text{Sys}}} p_0 = 1 \, .$$

Solving this equation for p_0 immediately yields the statement for p_0. □

6.4 Stationary state probabilities for the actual arrival and service rates

Previously, the formulas for the state probabilities used the arrival rate relative to the entire system λ^{Sys} and the service rate of the entire system μ^{Sys}. In the following, we will now use the arrival rates of the individual customers λ and the service rates of the individual operators μ. The variables λ^{Sys} and μ^{Sys} are defined differently for three ranges: for $n = 0$, for $n = 1, \ldots, c$ and for $n = c+1, \ldots, K$. For this reason, p_n is also calculated separately for these ranges in the following:

Calculation of p_0

If we substitute the state probabilities from (4) and (5) for p_0 in (9), we

$$\begin{aligned} p_0 &= \left[\sum_{n=1}^{K}\prod_{i=1}^{n}\frac{\lambda_{i-1}^{\text{Sys}}}{\mu_i^{\text{Sys}}}+1\right]^{-1} = \left[\sum_{n=1}^{K}\left(\lambda^n\cdot\frac{1}{\prod_{i=1}^{\min(n,c)} i\mu\cdot\prod_{i=\min(n,c)+1}^{n} c\mu}\right)+1\right]^{-1} \\ &= \left[\sum_{n=1}^{c}\frac{\lambda^n}{\mu^n n!}+\sum_{n=c+1}^{K}\frac{\lambda^n}{\mu^c c!\cdot(c\mu^{n-c})}+1\right]^{-1} = \left[\sum_{n=1}^{c}\frac{\lambda^n}{\mu^n n!}+\sum_{n=c+1}^{K}\frac{\lambda^n}{\mu^n c! c^{n-c}}+1\right]^{-1} .\end{aligned}$$

With $a := \frac{\lambda}{\mu}$ the following applies:

$$p_0 = \left[\sum_{n=1}^{c}\frac{a^n}{n!}+\sum_{n=c+1}^{K}\frac{a^n}{c!c^{n-c}}+1\right]^{-1} .$$

By choosing

$$C_n := \begin{cases} \dfrac{a^n}{n!} & \text{für } n \le c\,, \\ \dfrac{a^n}{c!c^{n-c}} & \text{für } c < n \le K \end{cases} \tag{10}$$

finally follows

$$p_0 = \left[\sum_{n=0}^{K} C_n\right]^{-1} .$$

Calculation of p_n for $n = 1, \ldots, c$

If we substitute the state probabilities from (4) and (5) for p_n in (9), we obtain the following with the definition of C_n in (10) for $n = 1,\ \ldots c$:

$$p_n = \prod_{i=1}^{n}\frac{\lambda_{i-1}^{\text{Sys}}}{\mu_i^{\text{Sys}}}\cdot p_0 = \prod_{i=1}^{n}\frac{\lambda}{n\mu}\cdot p_0 = \frac{a^n}{n!}\cdot p_0 = C_n p_0 .$$

Calculation of p_n for $n = c+1, \ldots, K$

If we substitute the state probabilities from (4) and (5) for p_n in (9), we obtain the following with the definition of C_n in (10) für $n = c+1, \ldots, K$:

$$p_n = \prod_{i=1}^{n}\frac{\lambda_{i-1}^{\text{Sys}}}{\mu_i^{\text{Sys}}}\cdot p_0 = \prod_{i=1}^{c}\frac{\lambda}{n\mu}\cdot\prod_{i=c+1}^{n}\frac{\lambda}{c\mu}\cdot p_0 = \frac{a^n}{c!c^{n-c}}\cdot p_0 = C_n p_0 .$$

The above considerations for p_n can be summarized in the following theorem:

Theorem 11. *With the choice of C_n according to (10), the following applies to the state probabilities p_n in a birth and death process with arrival and service rates according to (4) and (5):*

$$p_n = \begin{cases} \left[\sum_{i=0}^{K} C_i\right]^{-1} & \textit{für } n = 0\,, \\ C_n p_0 & \textit{für } n > 0\,. \end{cases} \tag{11}$$

7 The Erlang C formula

With the choice of $K := \infty$, the following initially applies to C_n (see formula (10)):

$$C_n := \begin{cases} \dfrac{a^n}{n!} & \text{für } n < c\,, \\ \dfrac{a^n}{c!c^{n-c}} & \text{für } n \geq c\,. \end{cases}$$

The probability that there are no customers in the system, p_0, is calculated in this case by

$$1 = p_0 + \sum_{n=1}^{\infty} C_n p_0 = p_0 \left(1 + \sum_{n=1}^{\infty} C_n\right)$$

to:

$$\begin{aligned} p_0 &= \left[1 + \sum_{n=1}^{\infty} C_n\right]^{-1} = \left[\sum_{n=0}^{c-1} \frac{a^n}{n!} + \sum_{n=c}^{\infty} \frac{a^n}{c!c^{n-c}}\right]^{-1} = \left[\sum_{n=0}^{c-1} \frac{a^n}{n!} + \frac{a^c}{c!} \sum_{n=0}^{\infty} \frac{a^n}{c^n}\right]^{-1} \\ &= \left[\sum_{n=0}^{c-1} \frac{a^n}{n!} + \frac{a^c}{c!} \cdot \frac{1}{1-\frac{a}{c}}\right]^{-1} = \left[\sum_{n=0}^{c-1} \frac{a^n}{n!} + \frac{a^c \cdot c}{c!(c-a)}\right]^{-1} . \end{aligned}$$

In the calculation

$$\sum_{n=0}^{\infty} \left(\frac{a}{c}\right)^n = \frac{1}{1-\frac{a}{c}}$$

In the penultimate calculation step, the geometric series was used and the assumption was made that in the case of a queueing model without cancelations, $\lambda < c\mu$ must apply, i. e., $\frac{a}{c} < 1$ must be true, which is necessary for the convergence of the series.

On this basis, the waiting time distribution $P(W \leq t)$ can now be determined for an Erlang C model:

$$\begin{aligned} P(W \leq t) &= \sum_{n=0}^{c-1} C_n p_0 + \sum_{n=c}^{\infty} F_{c\mu,n-c+1}(t) C_n p_0 \\ &= \sum_{n=0}^{c-1} C_n p_0 + \sum_{n=c}^{\infty} \int_0^t \frac{(c\mu)^{n-c+1} x^{n-c}}{(n-c)!} \mathrm{e}^{-c\mu x} \,\mathrm{d}x C_n p_0 \\ &= \sum_{n=0}^{c-1} C_n p_0 + \int_0^t \sum_{n=c}^{\infty} \frac{(c\mu)^{n-c+1} x^{n-c}}{(n-c)!} \mathrm{e}^{-c\mu x} \,\mathrm{d}x C_n p_0 \\ &= \sum_{n=0}^{c-1} C_n p_0 + p_0 \frac{a^c c\mu}{c!} \int_0^t \mathrm{e}^{-c\mu x} \sum_{n=0}^{\infty} \frac{(\mu x a)^n}{n!} \,\mathrm{d}x \\ &= \sum_{n=0}^{c-1} C_n p_0 + p_0 \frac{a^c c\mu}{c!} \int_0^t \mathrm{e}^{-c\mu x} \cdot \mathrm{e}^{a\mu x} \,\mathrm{d}x \\ &= \sum_{n=0}^{c-1} C_n p_0 + p_0 \frac{a^c c\mu}{c!} \left[-\frac{1}{(c-a)\mu} \mathrm{e}^{-(c-a)\mu x}\right]_{x=0}^{x=t} \\ &= \sum_{n=0}^{c-1} C_n p_0 - p_0 \frac{a^c c}{c!(c-a)} \left(\mathrm{e}^{-(c-a)\mu t} - 1\right) \\ &= 1 - p_0 \frac{a^c c}{c!(c-a)} \mathrm{e}^{-(c-a)\mu t} . \end{aligned}$$

Using the definition

$$P_1 := p_0 \frac{a^c c}{c!(c-a)} = \frac{\frac{a^c c}{c!(c-a)}}{\sum_{n=0}^{c-1} \frac{a^n}{n!} + \frac{a^c \cdot c}{c!(c-a)}} \tag{12}$$

this ultimately yields the usual Erlang C formula:

$$P(W \leq t) = 1 - P_1 \mathrm{e}^{-(c-a)\mu t} \, .$$

7.1 Characteristics

In the case of an M/M/c system, the other characteristics can be derived directly from the already determined values p_n and C_n:

- *Average queue length:*
  Using the definition of P_1 from (12), the following initially applies for $n > c$:

  $$p_n = C_n p_0 = \frac{a^n}{c! c^{n-c}} p_0 = P_1 \rho^{n-c}(1-\rho) \, .$$

  This immediately gives us the average queue length with the definition of the expected value:

  $$\begin{aligned} \mathbf{E}[N_Q] &= \sum_{n=c+1}^{\infty} (n-c) p_n = P_1(1-\rho) \sum_{n=c+1}^{\infty} (n-c)\rho^{n-c} \\ &= P_1(1-\rho) \sum_{n=1}^{\infty} n\rho^n = P_1(1-\rho) \sum_{n=0}^{\infty} n\rho^n = P_1(1-\rho)\rho \sum_{n=1}^{\infty} n\rho^{n-1} \, . \end{aligned}$$

  Since $f'(\rho) = n\rho^{n-1}$ holds for $f(\rho) := \rho^n$, the following also applies:

  $$\mathbf{E}[N_Q] = P_1(1-\rho)\rho \sum_{n=1}^{\infty} [\rho^n]' \, .$$

  Due to the absolute convergence of the series caused by the fact that $\rho < 1$, limit process and differentiation can be interchanged. Applying the geometric series ($\sum_{n=0}^{\infty} a^n = \frac{1}{1-a}$ for $|a| < 1$), the following applies:

  $$\begin{aligned} \mathbf{E}[N_Q] &= P_1(1-\rho)\rho \left[\sum_{n=1}^{\infty} \rho^n\right]' = P_1(1-\rho)\rho \left[\frac{1}{1-\rho}\right]' = P_1(1-\rho)\rho \frac{0-(-1)}{(1-\rho)^2} \\ &= P_1 \frac{\rho}{1-\rho} = P_1 \frac{a}{c-a} \, . \end{aligned}$$

- *Average number of customers in the system:*
  The average number of customers in the system is calculated as the sum of the average queue length $\mathbf{E}[N_Q]$ and the average number of customers being served a:

  $$\mathbf{E}[N] = P_1 \frac{a}{c-a} + a \, .$$

  The fact that the workload a corresponds precisely to the average number of customers being served is a consequence of Little's law.

- *Average waiting time:*
  Using Little's law, the mean waiting time can be directly calculated from $\mathbf{E}[N_Q]$:

  $$\mathbf{E}[W] = \frac{1}{\lambda} \mathbf{E}[N_Q] = P_1 \frac{1}{\lambda} \frac{a}{c-a} = P_1 \frac{\frac{1}{\mu}}{c-a} = P_1 \frac{1}{c\mu - \lambda} \, .$$

- *Average residence time:*
  The average residence time is composed of the average waiting time $\mathbf{E}[W]$ and the average service time $\mathbf{E}[S] = \frac{1}{\mu}$:
  $$\mathbf{E}[V] = P_1 \frac{1}{c\mu - \lambda} + \frac{1}{\mu} \ .$$

In summary:

$$\begin{aligned}
\mathbf{E}[N_Q] &= P_1 \frac{a}{c-a} \ , \\
\mathbf{E}[N] &= P_1 \frac{a}{c-a} + a \ , \\
\mathbf{E}[W] &= P_1 \frac{1}{c\mu - \lambda} \ , \\
\mathbf{E}[V] &= P_1 \frac{1}{c\mu - \lambda} + \frac{1}{\mu} \ .
\end{aligned}$$