BIRTH DEATH PROCESSES
AND
QUEUEING SYSTEMS

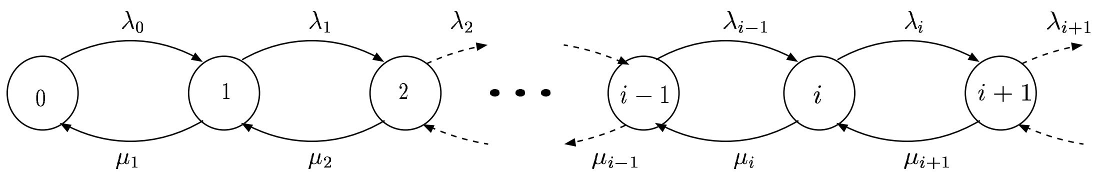
Andrea Bobbio
Anno Accademico 1999-2000

Notation for Queueing Systems

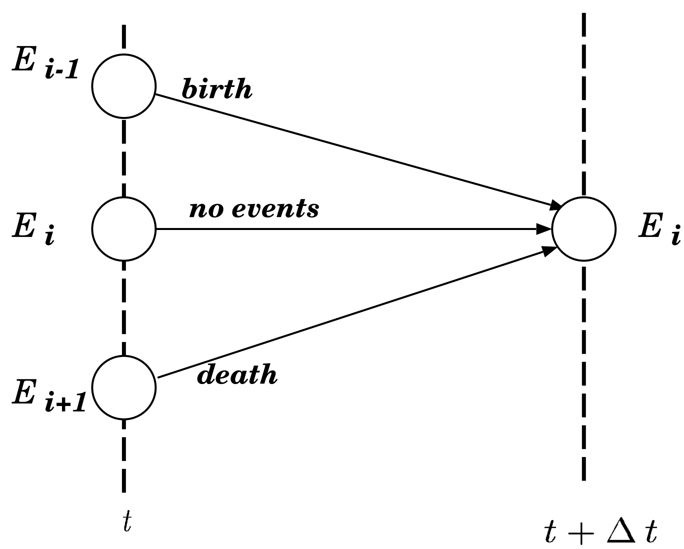
$1/\lambda$	mean time between arrivals
$S = 1/\mu$	mean service time
$\rho = \lambda/\mu$	traffic intensity
N	Number of customers in the queue (including those in service)
N_Q	Number of customers in the queue (excluding those in service)
N_S	Number of customers in service
R	Response time (including the service time)
W	Waiting time ($= R - S$)
U_0	Utilization factor
T	Throughput (Expected number of jobs completed in a time unit)

Birth-Death Processes

Let us identify by state i the condition of the system in which there are i objects. Given the system is in state i, new elements arrive at rate λ_i, and leave at rate μ_i.
The state space transition diagram is:



Let $N(t)$ be the number of elements in the system at time t, and $E_i(t)$ be the event $N(t) = i$.



The figure shows the way in which the event $E_i(t + \Delta t)$ can be generated.

Birth-Death Processes

By the theorem of the total probability, we can write for $i > 0$:

$$Pr\{N(t+\Delta t) = i \,|\, N(t) = i-1\} \;=\; \lambda_{i-1}\,\Delta t \;+\; O(\Delta t)$$

$$Pr\{N(t+\Delta t) = i \,|\, N(t) = i+1\} \;=\; \mu_{i+1}\,\Delta t \;+\; O(\Delta t)$$

$$Pr\{N(t+\Delta t) = i \,|\, N(t) = i\} \;=\; 1 - \lambda_i\,\Delta t - \mu_i\,\Delta t + O(\Delta t)$$

Where:

$$\lim_{\Delta t \to 0} \frac{O(\Delta t)}{\Delta t} \;=\; 0$$

For $i = 0$, we can write:

$$Pr\{N(t+\Delta t) = 0 \,|\, N(t) = 1\} \;=\; \mu_1\,\Delta t \;+\; O(\Delta t)$$

$$Pr\{N(t+\Delta t) = 0 \,|\, N(t) = 0\} \;=\; 1 \;-\; \lambda_0\,\Delta t \;+\; O(\Delta t)$$

Let us define: $\quad P_i(t) \;=\; Pr\{N(t) \;=\; i\}$

Birth-Death Processes

According to the above relations we can write:

$$\begin{cases} P_0(t+\Delta t) &= \mu_1 \Delta t\, P_1(t) + (1 - \lambda_0 \Delta t)\, P_0(t) \qquad i = 0 \\ P_i(t+\Delta t) &= \lambda_{i-1} \Delta t P_{i-1}(t) + \mu_{i+1} \Delta t P_{i+1}(t) \\ & \quad + (1 - \lambda_i \Delta t - \mu_i \Delta t)\, P_i(t) \qquad i > 0 \end{cases}$$

$$\begin{cases} \dfrac{P_0(t+\Delta t) - P_0(t)}{\Delta t} &= -\lambda_0 P_0(t) + \mu_1 P_1(t) \qquad i = 0 \\ \dfrac{P_i(t+\Delta t) - P_i(t)}{\Delta t} &= -(\lambda_i + \mu_i) P_i(t) + \lambda_{i-1} P_{i-1}(t) + \mu_{i+1} P_{i+1}(t) \qquad i > 0 \end{cases}$$

Taking the limit $\Delta t \rightarrow 0$, the following set of linear differential equations is derived:

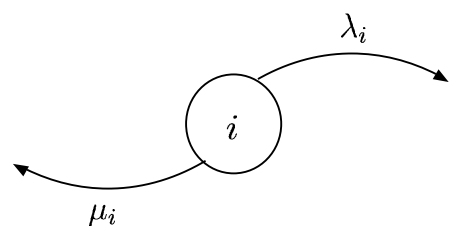
$$\begin{cases} \dfrac{d\,P_0(t)}{d\,t} &= -\lambda_0 P_0(t) + \mu_1 P_1(t) \qquad i = 0 \\ \dfrac{d\,P_i(t)}{d\,t} &= -(\lambda_i + \mu_i)\, P_i(t) + \lambda_{i-1} P_{i-1}(t) + \mu_{i+1} P_{i+1}(t) \qquad i > 0 \end{cases} \tag{1}$$

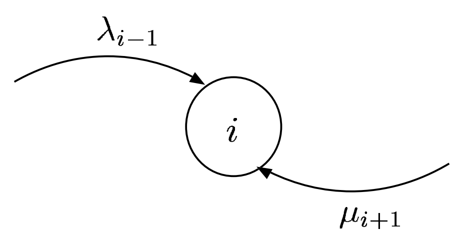
with initial conditions:

$$\begin{cases} P_0(0) &= 1 \qquad i = 0 \\ P_i(0) &= 0 \qquad i > 0 \end{cases}$$

Transient Balance Equation

Transient continuity (balance) equation in state i.
The flow variation in state i equals the difference between the ingoing flow minus the outgoing flow.





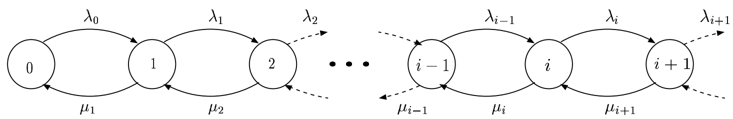
$$\text{variation of flow} = \frac{d\,P_i(t)}{d\,t}$$

$$\text{ingoing flow} = \lambda_{i-1}\,P_{i-1}(t) + \mu_{i+1}\,P_{i+1}(t) \qquad ;\; i \geq 0$$

$$\text{outgoing flow} = (\lambda_i + \mu_i)\,P_i(t)$$

Matrix representation of B/D processes

Given the B/D process of the figure:



we define a transition rate matrix $\mathbf{Q}$ and a state probability row vector $\mathbf{p}(t)$ at time t:

$$\mathbf{Q} = \begin{array}{c|ccccccccc|}
 & 0 & 1 & 2 & 3 & \cdots & i-1 & i & i+1 & \cdots \\
0 & -\lambda_0 & \lambda_0 & & & & & & & \\
1 & \mu_1 & -(\lambda_1+\mu_1) & \lambda_1 & & & & & & \\
2 & 0 & \mu_2 & -(\lambda_2+\mu_2) & \lambda_2 & & & & & \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
i-1 & & & & & & & & & \\
i & & & & & & \mu_i & -(\lambda_i+\mu_i) & \lambda_i & \\
i+1 & & & & & & & & & \\
\vdots & & & & & & & & &
\end{array}$$

$$\mathbf{p}(t) = \{p_0\ p_1\ p_2\ \ldots\ p_i\ \ldots\}$$

The solution equation of (1) can be written in matrix form:

$$\frac{d\,\mathbf{p}(t)}{d\,t} = \mathbf{p}\,\mathbf{Q}$$

Steady-state of B/D processes

For $t \to \infty$, the B/D process may reach a steady-state (equilibrium) condition. Steady state means that the state probabilities do not depend on the time any more.

If a steady-state solution exists, it is characterized by:

$$\lim_{t \to \infty} \frac{d\,P_i(t)}{d\,t} = 0 \qquad (i = 0,\, 1,\, 2,\, \ldots)$$

Let us denote: $P_i = \lim_{t \to \infty} P_i(t)$. The steady state equations become:

$$\begin{cases} 0 = -\lambda_0 P_0 + \mu_1 P_1 & i = 0 \\ 0 = -(\lambda_i + \mu_i) P_i + \lambda_{i-1} P_{i-1} + \mu_{i+1} P_{i+1} & i > 0 \end{cases}$$

that can be rewritten as balance equations (ingoing flow equals outgoing flow) as:

$$\begin{cases} \lambda_0 P_0 = \mu_1 P_1 & i = 0 \\ (\lambda_i + \mu_i) P_i = \lambda_{i-1} P_{i-1} + \mu_{i+1} P_{i+1} & i > 0 \end{cases}$$

Steady-state of B/D processes

The steady-state equation can be written as:

$$\begin{cases} & & \lambda_0 \, P_0 - \mu_1 \, P_1 & = & 0 \\ \\ \lambda_1 \, P_1 - \mu_2 \, P_2 & = & \lambda_0 \, P_0 - \mu_1 \, P_1 & = & 0 \\ \dots & & \dots & & \\ \lambda_i \, P_i - \mu_{i+1} \, P_{i+1} & = & \lambda_{i-1} \, P_{i-1} - \mu_i \, P_i & = & 0 \\ \dots & & \dots & & \end{cases}$$

From the above, the i-th term becomes:

$$\lambda_{i-1} \, P_{i-1} \, = \, \mu_i \, P_i \qquad \Longrightarrow \qquad P_i \, = \, \frac{\lambda_{i-1}}{\mu_i} \, P_{i-1} \qquad\qquad (i \geq 1)$$

$$P_i \, = \, \frac{\lambda_{i-1}}{\mu_i} \, \frac{\lambda_{i-2}}{\mu_{i-1}} \, P_{i-2} \, = \, \frac{\lambda_0 \, \lambda_1 \, \dots \, \lambda_{i-1}}{\mu_1 \, \mu_2 \, \dots \, \mu_i} \, P_0 \, = \, P_0 \, \prod_{j=0}^{i-1} \, \frac{\lambda_j}{\mu_{j+1}}$$

The following normalization condition must hold:

$$\sum_{i \geq 0} \, P_i \, = \, 1$$

Hence:

$$P_0 \, = \, \frac{1}{1 \, + \, \sum_{i \geq 1} \prod_{j=0}^{i-1} \, \frac{\lambda_j}{\mu_{j+1}}}$$

The steady state distribution exists, with $P_i \, > \, 0$, if the series $\sum_{i \geq 1} \prod_{j=0}^{i-1} \, \frac{\lambda_j}{\mu_{j+1}}$ converges.

Standard notation for queueing systems

The standard notation to identify the main elements that define the structure of a queueing system is the following (due to Kendall):

$$A/B/c/d/e$$

where:

- A Is the distribution of the interarrival times;
- B Is the distribution of the service times;
- c Is the number of servers;
- d Is the storage capacity of the system (number of servers plus the storage capacity of the buffer);
- e Is the number of sources that provide clients.

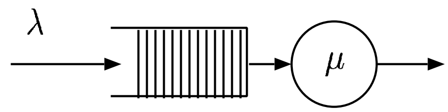
The usual assumption for the interarrival and service time distributions A and B is:

- M Markovian (or exponential);
- G General.

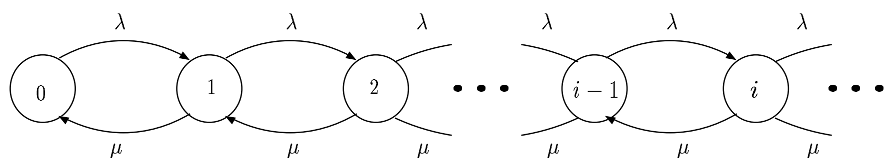
M/M/1

The M/M/1 queueing system is a B/D process characterized by having the arrival rates λ and the service rates μ independent of the state.

The usual picture for the M/M/1 is:



The state space of the M/M/1 is:



$$\lambda_i = \lambda \quad \text{for} \quad i \geq 0 \qquad ; \qquad \mu_i = \mu \quad \text{for} \quad i \geq 1$$

By applying the general equilibrium results of a B/D process:

$$P_i = \frac{\lambda}{\mu} P_{i-1} \quad \Longrightarrow \quad P_i = \left(\frac{\lambda}{\mu}\right)^i P_0$$

By applying the normalization condition:

$$\sum_{i=0}^{\infty} P_i = 1 \quad \Longrightarrow \quad P_0 = \frac{1}{1 + \sum_{j=1}^{\infty} \left(\frac{\lambda}{\mu}\right)^i} \tag{2}$$

Let us introduce a new parameter called the *traffic intensity*:

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

Steady state solution of a M/M/1

The denominator of (2) is the geometric series:

$$1 + \rho + \rho^2 + \ldots + \rho^i + \ldots = \sum_{i=0}^{\infty} \rho^i \tag{3}$$

If $\rho < 1$, the series (3) converges to the value

$$\sum_{i=0}^{\infty} \rho^i = \frac{1}{1 - \rho}$$

Hence, if $\rho < 1$ a steady state solution exists, and the M/M/1 is asymptotically stable.

If $\rho < 1$, the state probabilities depend on λ and μ only through the traffic intensity ρ, and are given by:

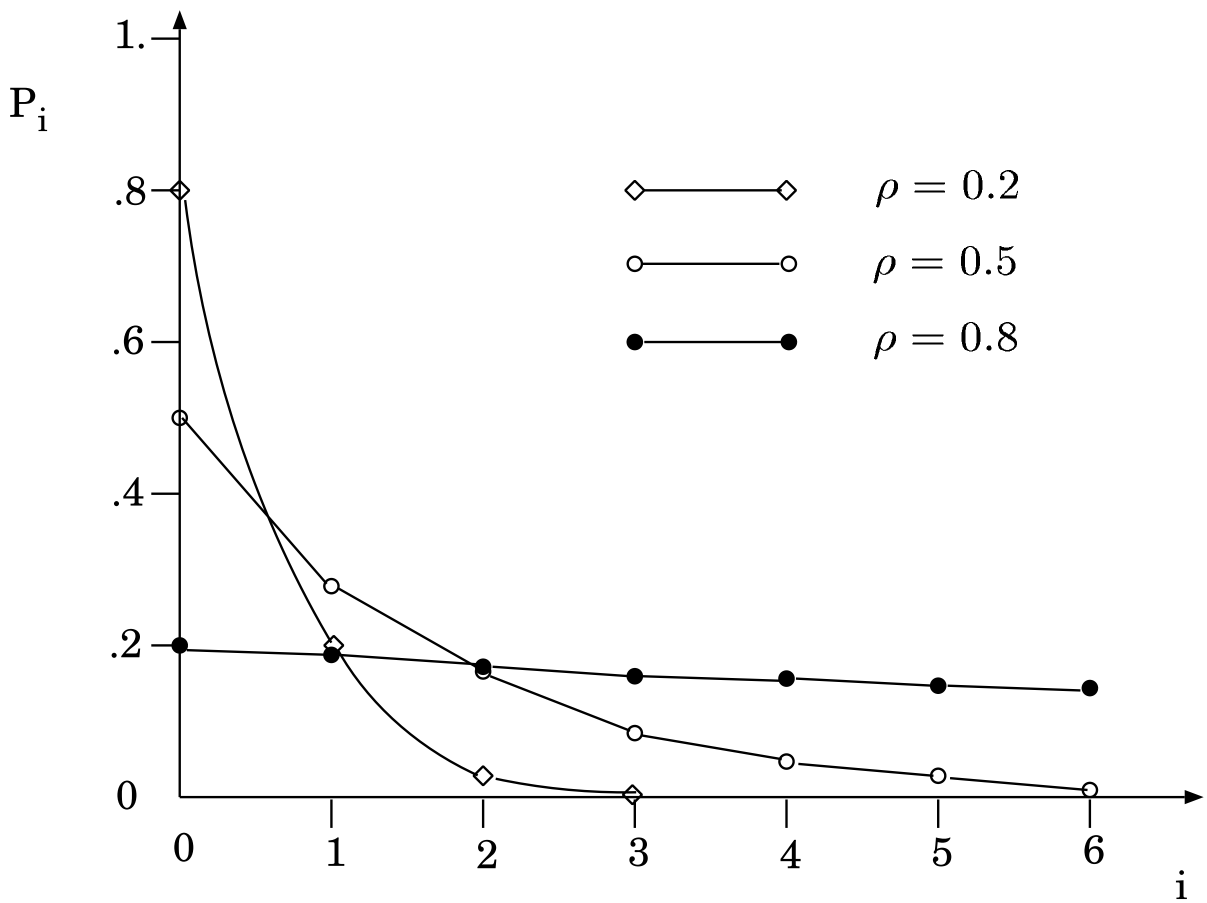
$$\begin{aligned} P_0 &= 1 - \rho \\ P_1 &= (1 - \rho)\,\rho \\ \ldots & \quad \ldots \\ P_i &= (1 - \rho)\,\rho^i \\ \ldots & \quad \ldots \end{aligned}$$

Since the state probabilities are known, the system is completely specified, and various measures can be computed.

M/M/1: Probability vs ρ

The state probability P_i as a function of i and for various values of ρ is depicted in the figure:

$$P_i = (1 - \rho)\, \rho^i$$



Expected number of customers in a M/M/1

Server *utilization factor* (probability the server is busy):

$$U_0 = \sum_{i=1}^{\infty} P_i = 1 - P_0 = \rho$$

Expected number of customers $E[N]$
Let N be the number of customers in the queue, including the one in service: the expected number of customers $E[N]$ is given by:

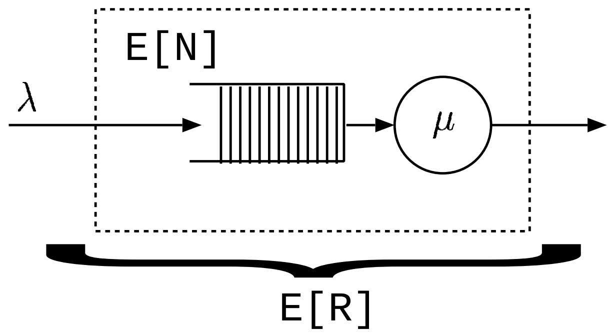
$$\begin{aligned} E[N] &= \sum_{i=0}^{\infty} i \cdot P_i \\ &= 0 \cdot P_0 + 1 \cdot \rho \cdot P_0 + 2 \cdot \rho^2 \cdot P_0 + 3 \cdot \rho^3 \cdot P_0 + \ldots \\ &= P_0 \sum_{i=0}^{\infty} i \cdot \rho^i = (1-\rho) \sum_{i=0}^{\infty} i \cdot \rho^i = \frac{\rho}{1-\rho} \end{aligned}$$

The above proof is based on the sum of the modified geometric series:

$$\begin{aligned} \sum_{i=0}^{\infty} i \cdot \rho^i &= \rho \frac{\partial}{\partial \rho} \sum_{i=0}^{\infty} \rho^i = \rho \frac{\partial}{\partial \rho} \frac{1}{1-\rho} \\ &= \frac{\rho}{(1-\rho)^2} \end{aligned}$$

M/M/1: Little's formula

The Little's formula states that the expected number of customers in the queue $E[N]$ is equal to the arrival rate λ times the expected time spent in the system (the expected *response time*) $E[R]$.



$$E[N] = \lambda \cdot E[R]$$

The expected response time for a M/M/1 queue is obtained by applying Little's formula:

$$E[R] = \lambda^{-1} E[N] = \frac{1}{\lambda} \frac{\rho}{1-\rho} = \frac{1/\mu}{1-\rho}$$

From the above formula, the expected response time $E[R]$ can be interpreted as the ratio between the mean service time $(1/\mu)$ and the probability of the sever to be idle $(1-\rho)$.

M/M/1: Performance measures

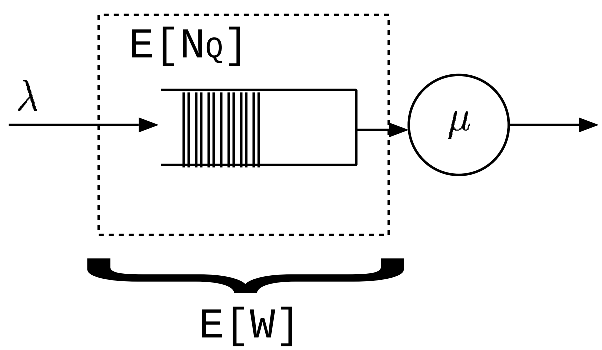
Expected waiting time

Let us define the waiting time $W = R - S$ as the time a customer waits in the queue before service, where R is the response time and S the service time.
The expected waiting time $E[W]$ is given by:

$$E[W] \;=\; E[R] \;-\; E[S] \;=\; \frac{1}{\mu\,(1-\rho)} - \frac{1}{\mu} = \frac{\rho}{\mu\,(1-\rho)}$$

Expected number of customers in the line

The expected number of customers in the line (awaiting for service) is obtained by applying Little's rule to the queue only:



$$E[N_Q] \;=\; \lambda \cdot E[W] \;=\; \frac{\rho^2}{1-\rho}$$

Number of customers in service

The expected number of customers in service is:

$$E[N_S] \;=\; E[N] \;-\; E[N_Q] \;=\; \rho$$

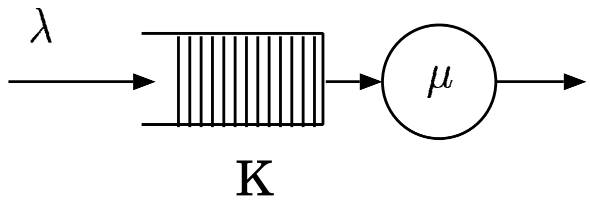
From the Little's rule applied to the server, only:

$$E[N_S] \;=\; \lambda \cdot E[S] \;=\; \frac{\lambda}{\mu} \;=\; \rho$$

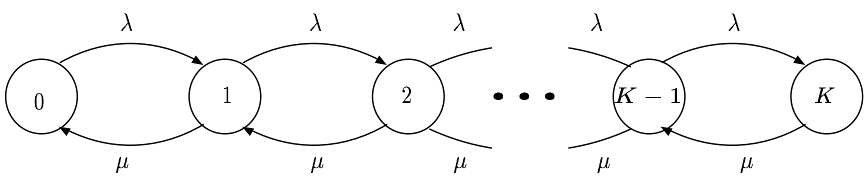
Summary of results for the M/M/1

λ		arrival rate
μ		service rate
ρ	$= \lambda/\mu$	traffic intensity
$E[N]$	$= \dfrac{\rho}{1-\rho}$	Expected number of customers in the queue (including those in service)
$E[R]$	$= \dfrac{1/\mu}{1-\rho}$	Expected response time
$E[S]$	$= \dfrac{1}{\mu}$	Expected service time
$E[W]$	$= E[R] - E[S] = \dfrac{\rho}{\mu(1-\rho)}$	Expected waiting time
$E[N_Q]$	$= \lambda \cdot E[W] = \dfrac{\rho^2}{1-\rho}$	Expected number of waiting customers
$E[N_S]$	$= E[N] - E[N_Q] = \lambda\, E[S] = \rho$	Expected number of customers in service

M/M/1/K: finite storage



The storage capacity of the system is K (one customer in service and $K-1$ customers in the waiting line) and the exceeding customers are refused.



The general B/D process can be particularized as follows:

$$\lambda_i = \begin{cases} \lambda & i < K \\ \\ 0 & i \geq K \end{cases} \qquad ; \qquad \mu_i = \mu$$

The state probabilities satisfy

$$\begin{cases} P_i = P_0 \prod_{j=0}^{i-1} \dfrac{\lambda}{\mu} = P_0 \cdot \rho^i & i \leq K \\ \\ P_i = 0 & i > K \end{cases}$$

From the normalization condition:

$$P_0 = \frac{1}{1 + \sum_{j=1}^{K} \rho^j} = \frac{1}{1 + \dfrac{\rho(1-\rho^K)}{1-\rho}} = \frac{1-\rho}{1-\rho^{K+1}}$$

M/M/1/K: finite storage

The M/M/1/K queue is stable for any positive value of the *traffic intensity* ρ.

The state probabilities are:

$$\begin{cases} P_i = \dfrac{(1-\rho)\,\rho^i}{1-\rho^{K+1}} & i \leq K \\ \\ P_i = 0 & i > K \end{cases}$$

For $\rho \rightarrow 1$ the above formula is undefined. We find the limit resorting to De l'Hospital rule:

$$\begin{aligned} \lim_{\rho \rightarrow 1} P_i &= \lim_{\rho \rightarrow 1} \frac{(1-\rho)\,\rho^i}{1-\rho^{K+1}} \\ &= \lim_{\rho \rightarrow 1} \frac{-\rho^i + i\,(1-\rho)\,\rho^{i-1}}{-(K+1)\,\rho^K} = \frac{1}{K+1} \end{aligned}$$

Let us define the *rejection probability* as the probability of an arriving customer to find the queue full and to be rejected.

Since the queue is full when in state K, the *rejection probability* is:

$$P_K = \frac{(1-\rho)\,\rho^K}{1-\rho^{K+1}}$$

M/M/1/K: finite storage

Expected number of customers $E[N]$

$$\begin{aligned} E[N] &= \sum_{i=0}^{K} i \cdot P_i = \sum_{i=0}^{K} i \cdot \frac{(1-\rho)\,\rho^i}{1-\rho^{K+1}} \\ &= \frac{1-\rho}{1-\rho^{K+1}} \sum_{i=0}^{K} i \cdot \rho^i \qquad (4) \\ &= \frac{\rho}{1-\rho^{K+1}} \, \frac{1-(K+1)\,\rho^K + K\,\rho^{K+1}}{(1-\rho)} \end{aligned}$$

The above formula (4) is based on the following finite series sum:

$$\begin{aligned} \sum_{i=0}^{K} i \cdot \rho^i &= \rho \, \frac{\partial}{\partial \rho} \sum_{i=1}^{K} \rho^i = \rho \, \frac{\partial}{\partial \rho} \, \frac{\rho\,(1-\rho^K)}{1-\rho} \\ &= \rho \, \frac{1-(K+1)\,\rho^K + K\,\rho^{K+1}}{(1-\rho)^2} \end{aligned}$$

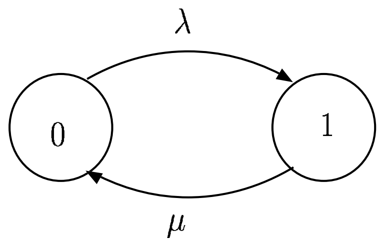
From formula (4), it follows:

$$\lim_{\rho \to 0} E[N] = 0 \quad ; \quad \lim_{\rho \to \infty} E[N] = K \quad ; \quad \lim_{\rho \to 1} E[N] = \frac{K}{2}$$

where the last limit ($\rho \to 1$) is obtained by applying twice the De l'Hospital rule.

M/M/1/1: no waiting line

The queue does not have a waiting line and the arriving customer enters service only if the server is idle.

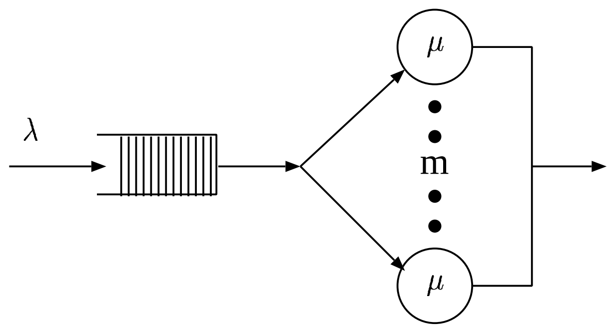


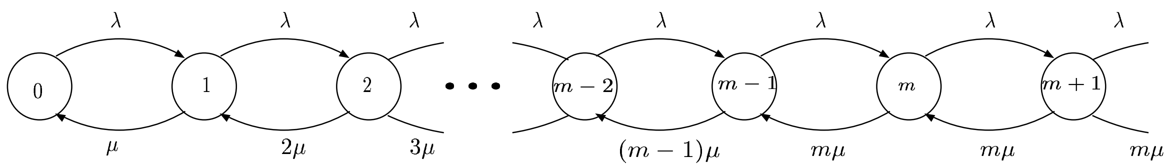
From the M/M/1/K case, we get:

$$\begin{cases} P_0 = \dfrac{1}{1+\rho} = \dfrac{\mu}{\lambda+\mu} \\ \\ P_1 = \dfrac{\rho}{1+\rho} = \dfrac{\lambda}{\lambda+\mu} \end{cases}$$

M/M/m - Queueing system with m servers

The queue has one arrival line and m identical servers with service rate μ. The structure of the queue and its state space are represented in the figures:





The general B/D process can be particularized as follows:

$$\lambda_i = \lambda \qquad i \geq 0 \qquad ; \qquad \mu_i = \begin{cases} i\,\mu & 0 < i < m \\ m\,\mu & i \geq m \end{cases}$$

The state probabilities satisfy:

$$\begin{cases} P_i = P_0 \displaystyle\prod_{j=0}^{i-1} \frac{\lambda}{(j+1)\,\mu} = P_0 \cdot \left(\frac{\lambda}{\mu}\right)^i \frac{1}{i!} & i < m \\ P_i = P_0 \displaystyle\prod_{j=0}^{m-1} \frac{\lambda}{(j+1)\,\mu} \cdot \prod_{k=m}^{i-1} \frac{\lambda}{m\,\mu} = P_0 \left(\frac{\lambda}{\mu}\right)^i \frac{1}{m!\,m^{i-m}} & i \geq m \end{cases}$$

M/M/m - Queueing system with m servers

Let us define the traffic intensity as $\rho = \dfrac{\lambda}{m\,\mu}$.

The stability condition requires $\rho < 1$.

By rewriting the state probabilities in terms of the traffic intensity, we obtain:

$$P_i \; = \begin{cases} P_0 \, \dfrac{(m\,\rho)^i}{i\,!} & i < m \\[2ex] P_0 \, \dfrac{\rho^i\, m^m}{m!} & i \geq m \end{cases}$$

From the normalization condition, we obtain:

$$P_0 \; = \left\{ \sum_{i=0}^{m-1} \frac{(m\,\rho)^i}{i\,!} + \sum_{i=m}^{\infty} \frac{\rho^i\, m^m}{m!} \right\}^{-1} \qquad (5)$$

The second sum in (5) can be written as:

$$\sum_{i=m}^{\infty} \frac{\rho^i\, m^m}{m!} = \frac{\rho^m\, m^m}{m\,!} \sum_{k=0}^{\infty} \rho^k = \frac{(m\,\rho)^m}{m\,!} \, \frac{1}{1-\rho}$$

So that (5) becomes:

$$P_0 \; = \left\{ \sum_{i=0}^{m-1} \frac{(m\,\rho)^i}{i\,!} + \frac{(m\,\rho)^m}{m\,!} \, \frac{1}{1-\rho} \right\}^{-1}$$

M/M/m - Queueing system with m servers

Expected number of customers in the queue:

$$E[N] \;=\; \sum_{i=0}^{\infty} \; i\,P_i \;=\; m\,\rho \,+\, \rho\,\frac{(m\,\rho)^m}{m\,!}\;\frac{P_0}{(1-\rho)^2}$$

Expected number of busy servers:

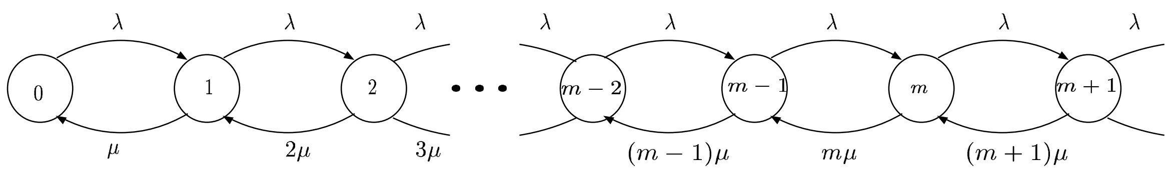
$$E[M] \;=\; \sum_{i=0}^{m-1} \; i\,P_i \,+\, m \; \sum_{i=m}^{\infty} \; P_i \;=\; m\,\rho \;=\; \frac{\lambda}{\mu}$$

Probability that an arriving customer should join the queue (equal to the probability that an arriving customer finds all the servers busy):

$$P_{[queue]} \;=\; \sum_{i=m}^{\infty} \; P_i \;=\; \frac{P_m}{1-\rho} \;=\; \frac{(m\,\rho)^m}{m\,!}\;\frac{P_0}{1-\rho}$$

M/M/∞: infinite number of servers

The state space of the queue is represented in the figure:



The general B/D process can be particularized as follows:

$$\begin{cases} \lambda_i = \lambda & i \geq 0 \\ \mu_i = i\,\mu & i \geq 0 \end{cases}$$

The state probabilities become:

$$P_i = P_0 \prod_{j=1}^{i-1} \frac{\lambda}{(j+1)\,\mu} = P_0 \frac{1}{i\,!} \left(\frac{\lambda}{\mu}\right)^i$$

The normalization condition provides:

$$P_0 = \frac{1}{1 + \sum_{i=1}^{\infty} \frac{1}{i\,!} \left(\frac{\lambda}{\mu}\right)^i} = e^{-\lambda/\mu}$$

Hence, the state probabilities assume the following form and are Poisson distributed:

$$P_i = e^{-\lambda/\mu} \frac{(\lambda/\mu)^i}{i\,!}$$

$$E[N] = \lambda/\mu \qquad ; \qquad E[R] = \frac{E[N]}{\lambda} = \frac{1}{\mu}$$