Reliability and Availability Engineering: Modeling, Analysis, Applications

Chapter 11 - Continuous Time Markov Chain: Queueing Systems

Kishor Trivedi and Andrea Bobbio

Department of Electrical and Computer Engineering
Duke University
Dipartimento di Scienze e Innovazione Tecnologica
Università del Piemonte Orientale

Jan 2017

Outline

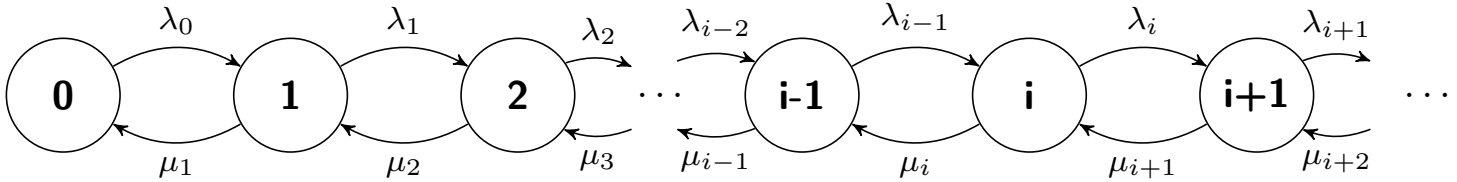
Notation for Queueing Systems

$1/\lambda$	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)

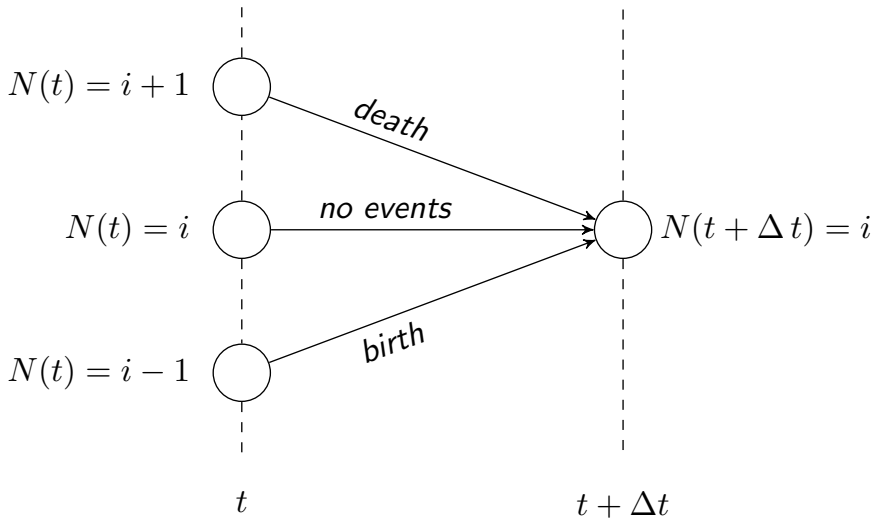
Birth-Death Processes

State i identifies 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 conditioning on the state of the system at time t, we can write for $i > 0$:

$$\begin{aligned} P\{N(t+\Delta t) = i \,|\, N(t) = i-1\} &= \lambda_{i-1}\,\Delta t + o(\Delta t) \\ P\{N(t+\Delta t) = i \,|\, N(t) = i+1\} &= \mu_{i+1}\,\Delta t + o(\Delta t) \\ P\{N(t+\Delta t) = i \,|\, N(t) = i\} &= 1 - \lambda_i\,\Delta t - \mu_i\,\Delta t + o(\Delta t)\,, \end{aligned}$$

and for $i = 0$, we can write:

$$\begin{aligned} P\{N(t+\Delta t) = 0 \,|\, N(t) = 1\} &= \mu_1\,\Delta t + o(\Delta t) \\ P\{N(t+\Delta t) = 0 \,|\, N(t) = 0\} &= 1 - \lambda_0\,\Delta t + o(\Delta t)\,, \end{aligned}$$

where: $\lim_{\Delta t \to 0} \frac{o(\Delta t)}{\Delta t} = 0$

Birth-Death Processes

Let us define: $\pi_i(t) = Pr\{N(t) = i\}$

According to the above relations we can write for $i = 0$ the first line and $i > 0$ the second one:

$$\begin{cases} \pi_0(t+\Delta t) = \mu_1 \Delta t \, \pi_1(t) + (1 - \lambda_0 \Delta t)\, \pi_0(t) + o(\Delta t) \\ \pi_i(t+\Delta t) = \lambda_{i-1}\Delta t \pi_{i-1}(t) + \mu_{i+1}\Delta t \pi_{i+1}(t) + (1 - \lambda_i \Delta t - \mu_i \Delta t)\, \pi_i(t) + o(\Delta t) \end{cases}$$

$$\begin{cases} \dfrac{\pi_0(t+\Delta t) - \pi_0(t)}{\Delta t} = -\lambda_0 \pi_0(t) + \mu_1 \pi_1(t) + \dfrac{o(\Delta t)}{\Delta t} \\ \dfrac{\pi_i(t+\Delta t) - \pi_i(t)}{\Delta t} = -(\lambda_i + \mu_i)\pi_i(t) + \lambda_{i-1}\pi_{i-1}(t) + \mu_{i+1}\pi_{i+1}(t) + \dfrac{o(\Delta t)}{\Delta t}. \end{cases}$$

Birth-Death Processes

Rearranging and taking the limit $\Delta t \to 0$, the following set of linear differential equations is derived:

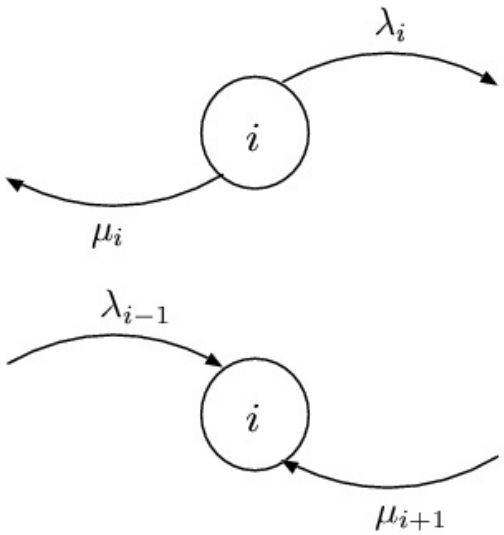
$$\begin{cases} \dfrac{d\,\pi_0(t)}{d\,t} &= -\lambda_0\,\pi_0(t) + \mu_1\,\pi_1(t) & i=0 \\ \dfrac{d\,\pi_i(t)}{d\,t} &= -(\lambda_i+\mu_i)\,\pi_i(t) + \lambda_{i-1}\,\pi_{i-1}(t) + \mu_{i+1}\,\pi_{i+1}(t) & i>0 \end{cases}$$

with initial conditions:

$$\begin{cases} \pi_0(0) &= 1 & i=0 \\ \pi_i(0) &= 0 & i>0 \end{cases}.$$

Transient Balance Equation

Transient continuity (balance) equation in state i.

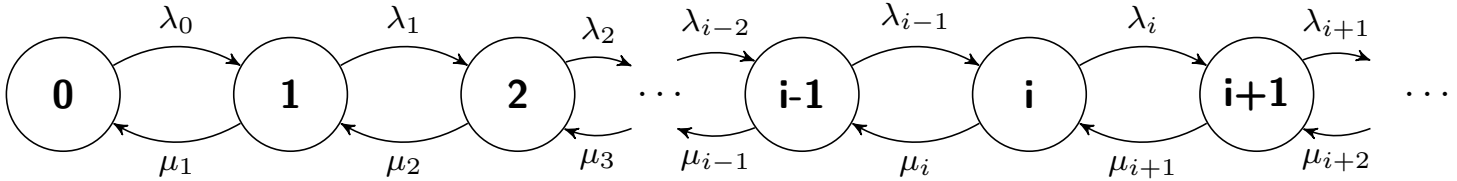


The flow variation in state i equals the difference between the ingoing flow minus the outgoing flow.

$$\begin{aligned} \text{variation of flow} &= \frac{d\,\pi_i(t)}{d\,t} \\ \text{incoming flow} &= \lambda_{i-1}\,\pi_{i-1}(t) + \mu_{i+1}\,\pi_{i+1}(t) \qquad ; \quad i > 0 \\ \text{outgoing flow} &= (\lambda_i + \mu_i)\,\pi_i(t) \end{aligned}$$

Matrix representation of processes

Given the process of the figure:



A BD process with constant birth and death rates is a CTMC with infinitesimal generator $\boldsymbol{Q}$ is given by :

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

Steady-state of B/D processes

The CTMC of a BD process is irreducible assuming that $\lambda_i > 0$ and $\mu_i > 0$ for each i. If a steady state solution exists, it is characterized by:

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

Let us denote: $\pi_i = \lim_{t\to\infty} \pi_i(t)$. The steady state equations become:
Birth-Death Process!steady state

$$\begin{cases} 0 = -\lambda_0\,\pi_0 + \mu_1\,\pi_1 & i = 0 \\ 0 = -(\lambda_i + \mu_i)\,\pi_i + \lambda_{i-1}\,\pi_{i-1} + \mu_{i+1}\,\pi_{i+1} & i > 0 \end{cases}$$

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

$$\begin{cases} \lambda_0\,\pi_0 = \mu_1\,\pi_1 & i = 0 \\ (\lambda_i + \mu_i)\,\pi_i = \lambda_{i-1}\,\pi_{i-1} + \mu_{i+1}\,\pi_{i+1} & i > 0 \end{cases}$$

Steady-state of B/D processes

The steady state equation can be rearranged as:

$$\begin{cases} & \lambda_0\,\pi_0 - \mu_1\,\pi_1 & = & 0 \\ \lambda_1\,\pi_1 - \mu_2\,\pi_2 & = & \lambda_0\,\pi_0 - \mu_1\,\pi_1 & = & 0 \\ \dots & & \dots & & \\ \lambda_i\,\pi_i - \mu_{i+1}\,\pi_{i+1} & = & \lambda_{i-1}\,\pi_{i-1} - \mu_i\,\pi_i & = & 0 \\ \dots & & \dots & & \end{cases}$$

From the above, the i-th term becomes:

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

and finally,

$$\pi_i = \frac{\lambda_{i-1}}{\mu_i}\,\frac{\lambda_{i-2}}{\mu_{i-1}}\,\pi_{i-2} = \frac{\lambda_0\,\lambda_1\,\dots\,\lambda_{i-1}}{\mu_1\,\mu_2\,\dots\,\mu_i}\,\pi_0 = \pi_0 \prod_{j=0}^{i-1} \frac{\lambda_j}{\mu_{j+1}}\,.$$

Steady-state of B/D processes

For the π_i's to form a probability vector, the following normalization condition must hold:

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

hence we obtain:

$$\pi_0 = \frac{1}{1 + \sum_{i \geq 1} \prod_{j=0}^{i-1} \frac{\lambda_j}{\mu_{j+1}}} \,.$$

The steady state probability vector exists, with $\pi_i > 0$, if the series $\sum_{i \geq 1} \prod_{j=0}^{i-1} \frac{\lambda_j}{\mu_{j+1}}$ converges. This is the case when all the states of the CTMC are recurrent non-null.

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):

$A/B/c/K/N/d$ where:

- A Indicates the nature of the distribution of the inter-arrival times;
- B Indicates the nature of the distribution of the service times;
- c Number of servers;
- K Storage capacity (including the servers) – if omitted, it is assumed infinite;
- N Population that can submit jobs – if omitted it is assumed infinite;
- d Scheduling discipline such as FCFS.

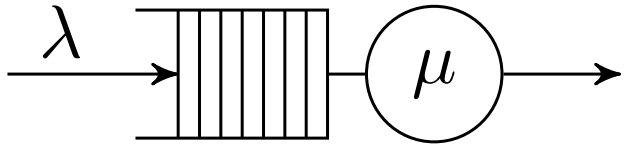
An usual assumption for the inter-arrival and service times A and B is:

- M Memoryless (or exponentially distributed);
- PH Phase Type;
- GI Generally distributed and independent;
- G Generally distributed.

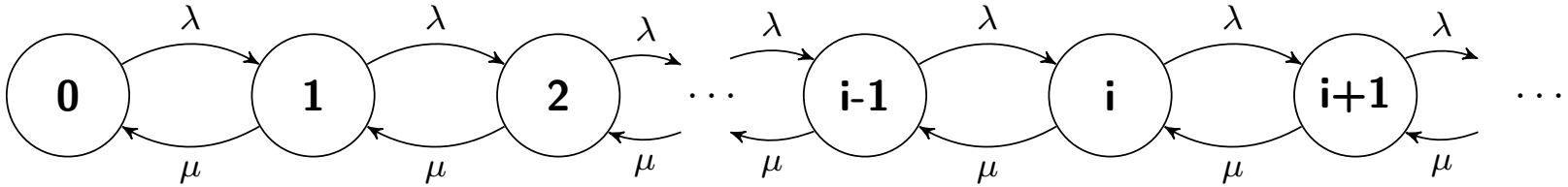
The M/M/1 queue

The M/M/1 is a special case of the general *birth-death* process for which in each state i, $\lambda_i = \lambda$ and $\mu_i = \mu$.

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



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



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

The only constraint on the scheduling discipline is that the server is not left idle if there are requests in the system and that no knowledge of the service times of individual requests is used in scheduling.

This CTMC is irreducible if $\lambda > 0$ and $\mu > 0$.

The M/M/1 queue

Let π_i be the probability of being in state i in the steady state. Then, writing out the balance equations for the CTMC, we get

$$\begin{cases} \lambda\,\pi_0 &= \mu\,\pi_1 \\ (\lambda+\mu)\,\pi_1 &= \lambda\,\pi_0 + \mu\,\pi_2 \\ (\lambda+\mu)\,\pi_2 &= \lambda\,\pi_1 + \mu\,\pi_3 \\ \cdots & \\ (\lambda+\mu)\,\pi_i &= \lambda\,\pi_{i-1} + \mu\,\pi_{i+1} \\ \cdots & \end{cases}$$

Solving the balance equations (or directly from the general B/D process), we get:

$$\pi_i = \left(\frac{\lambda}{\mu}\right)\pi_0 = \rho^i \pi_0\,,$$

where $\rho = \lambda/\mu$, is called the traffic intensity of the system. Imposing the normalization condition $\sum_{i\geq 0}\pi_i = \pi_0 \sum_{i\geq 0}\rho^i = 1$, we get

$$\pi_0 = \frac{1}{\sum_{i\geq 0}\rho^i} = 1-\rho\,. \tag{1}$$

Stability condition for a M/M/1

If $\lambda < \mu$ (i.e., $\rho < 1$) the geometrical series in the denominator

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

converges; hence for the $M/M/1$ queue to be stable, the traffic intensity must be less than unity or the arrival rate should be less than the service rate.
In this case, all the states of the CTMC are recurrent non-null.

If $\lambda = \mu$, then all the states of the CTMC are recurrent null.

If $\lambda \geq \mu$ (i.e., $\rho \geq 1$), the denominator in the expression for π_0 diverges; then all the states of the CTMC are transient.

Thus, if $\lambda \geq \mu$, the system is unstable, and the number of requests in the queue increases without bound.

Stability condition for a M/M/1

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

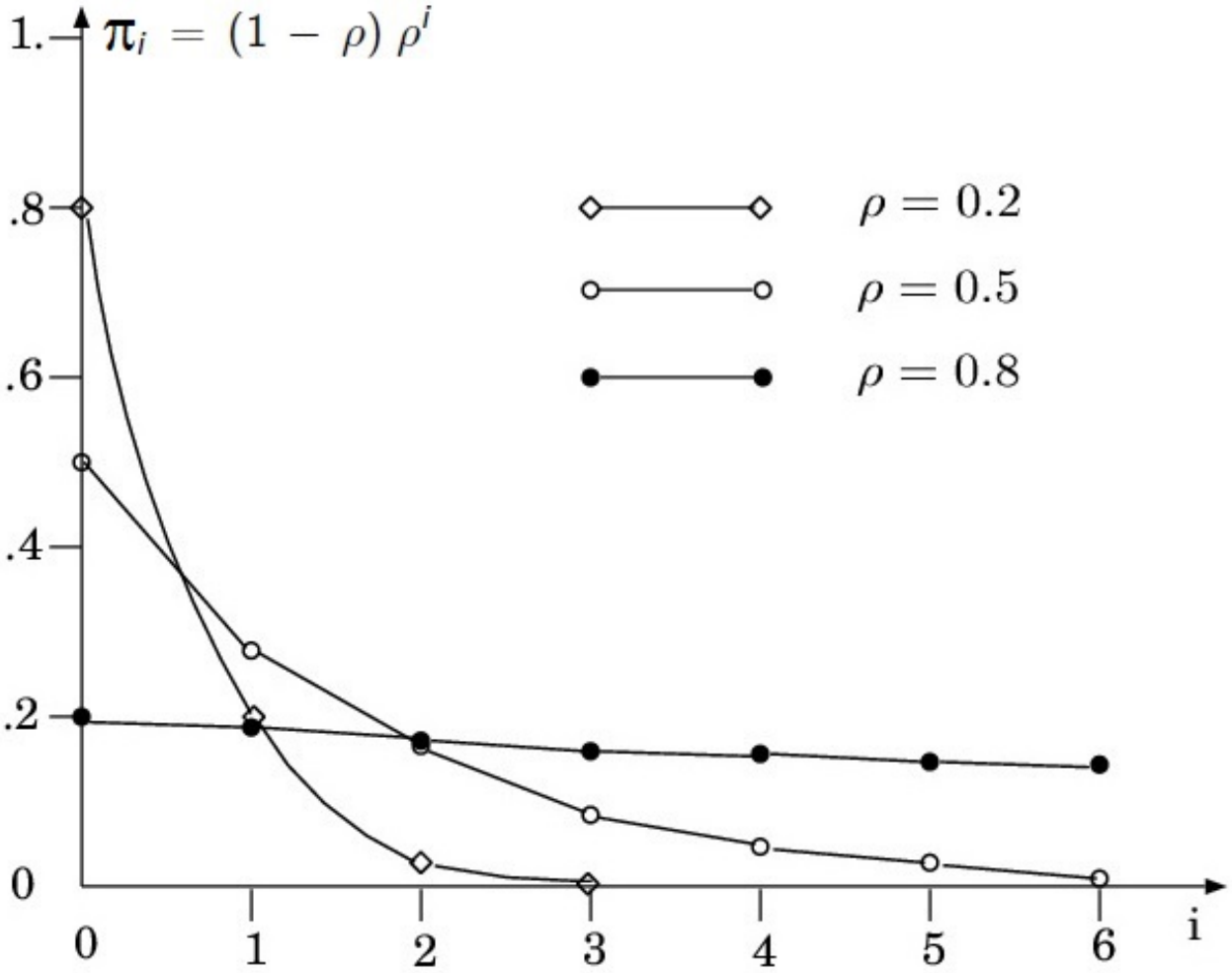
If $\rho < 1$, the state probabilities depend on λ and μ only through the traffic intensity ρ, and are given by:

$$\begin{aligned} \pi_0 &= 1-\rho \\ \pi_1 &= (1-\rho)\,\rho \\ \dots &\quad \dots \\ \pi_i &= (1-\rho)\,\rho^i \\ \dots &\quad \dots \end{aligned}$$

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 π_i as a function of i and for various values of ρ is depicted in the figure:



Metric of interest for the $M/M/1$ queue

Server Utilization

Note that the server is busy as long as the system is not in state 0.

Thus the *Utilization* of the server, meaning the fraction of the time that the server is busy in steady state, is given by:

$$U_0 = \sum_{j=1}^{\infty} \pi_j = 1 - \pi_0 = \rho\,.$$

This is the first metric of interest for the $M/M/1$ queue.

Expected number of customers in a M/M/1

We denote by N the total number of customers in the system (including the customer under service) in steady state.

In State i we have $N = i$.

The steady state expected number of customers in the system $E[N]$ is derived as:

$$E[N] = \sum_{i=0}^{\infty} i \cdot \pi_i = \pi_0 \sum_{i=0}^{\infty} i \cdot \rho^i = (1-\rho) \sum_{i=0}^{\infty} i \cdot \rho^i = \frac{\rho}{1-\rho}.$$

To derive $E[N]$ in the above Equation we have made use of a property of the sum of the geometric series:

$$\sum_{i=0}^{\infty} i \cdot \rho^i = \rho \frac{\partial}{\partial \rho} \sum_{i=0}^{\infty} \rho^i = \rho \frac{\partial}{\partial \rho} \frac{1}{1-\rho} = \frac{\rho}{(1-\rho)^2}.$$

Variance of the number of customers in a M/M/1

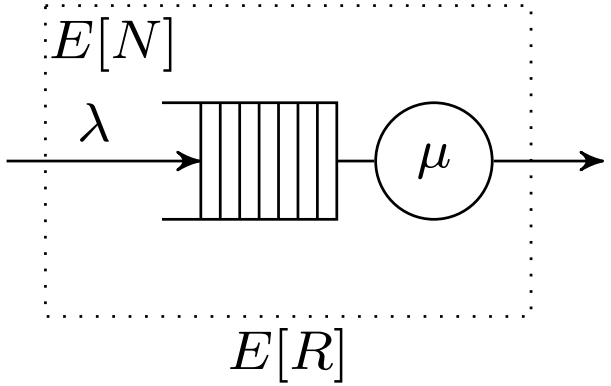
The variance of the number of customers in the queue, $Var[N]$ is,

$$Var[N] = \sum_{i=0}^{\infty} i^2 \pi_i - (E[N])^2 = \frac{\rho}{(1-\rho)^2}.$$

Expected response time $E[R]$ in a M/M/1

Let the random variable R represent the *response time*, defined as the time elapsed from the instant of the request arrival until the instant of its completion and departure from the system, in steady state.

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



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

Knowing the expression for $E[N]$, we can now get an expression for $E[R]$:

$$E[R] = E[N]/\lambda = \frac{1}{\lambda}\frac{\rho}{1-\rho} = \frac{1/\mu}{1-\rho} = \frac{\text{average service time}}{\text{prob. that the server is idle}}$$

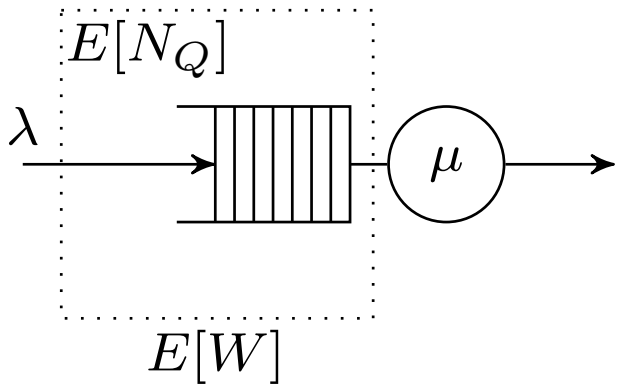
M/M/1: Performance measures

Expected waiting time $E[W]$ - 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 $E[N_Q]$

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



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

M/M/1: Performance measures

Expected number of customers in service $E[N_S]$ - 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:

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

Expected throughput $E[T]$ - In all the state $i > 0$, the throughput is μ, but in state 0 the throughput is 0. Hence the expected throughput, $E[T]$ is (see (**??**)):

$$E[T] = \sum_{i=1}^{\infty} \mu\,\pi_i = \mu \sum_{i=1}^{\infty} \pi_i = \mu\,(1-\pi_0) = \mu\,\rho = \lambda\,.$$

M/M/1: Performance measures

All the above metrics can be expressed in the framework of Markov Reward Models by associating a suitable reward rate r_i with state i of the CTMC, as per the following Table.

$M/M/1$ metrics as expected steady state reward rates $E[X] = \sum_{i=0}^{\infty} r_i \pi_i$

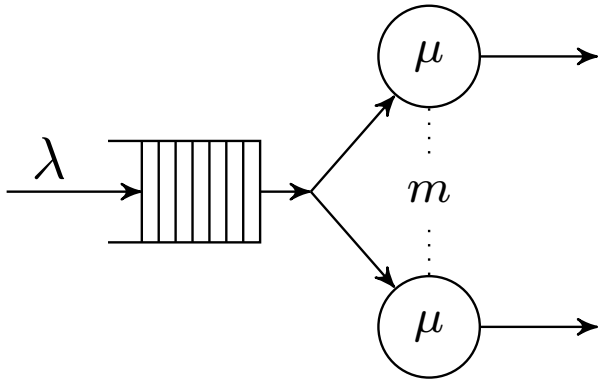
Mean No. in the system $E[N]$	$r_i = i$	$\frac{\rho}{1-\rho}$
Mean Response time $E[R]$	$r_i = i/\lambda$	$\frac{1/\mu}{1-\rho}$
Throughput $E[T]$	$\begin{cases} r_0 = 0 \\ r_i = \mu \quad (i > 0) \end{cases}$	λ

Summary of results for the M/M/1

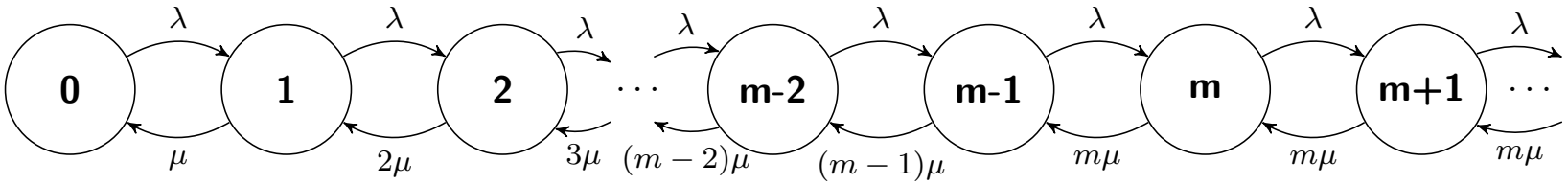
λ , μ			arrival rate and service rate
ρ	$=$	λ/μ	traffic intensity
$E[N]$	$=$	$\frac{\rho}{1-\rho}$	Expected number of customers in the queue (including those in service)
$E[R]$	$=$	$\frac{1/\mu}{1-\rho}$	Expected response time
$E[T]$	$=$	$\mu\,\rho = \lambda$	Expected throughput
$E[W]$	$=$	$E[R] - E[S]$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

This queueing system has a Poisson arrival process of rate λ, a single shared queue and m identical servers, each with service rate μ.



In the state diagram the service rate grows linearly from μ in state 1 up to $m\,\mu$ in state m where all the m servers are busy, and then retains the value $m\,\mu$ for states $i > m$.



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

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

M/M/m - Queueing system with m servers

The state probabilities thus satisfy:

$$\begin{cases} \pi_i = \pi_0 \prod_{j=0}^{i-1} \frac{\lambda}{(j+1)\,\mu} = \pi_0 \cdot \left(\frac{\lambda}{\mu}\right)^i \frac{1}{i!} & i < m \\ \pi_i = \pi_0 \prod_{j=0}^{m-1} \frac{\lambda}{(j+1)\,\mu} \cdot \prod_{k=m}^{i-1} \frac{\lambda}{m\,\mu} = \pi_0 \left(\frac{\lambda}{\mu}\right)^i \frac{1}{m!\, m^{i-m}} & i \geq m \end{cases}$$

Define $\rho = \frac{\lambda}{m\,\mu}$; the stability condition requires $\rho < 1$. Rewriting the state probabilities in terms of ρ, we obtain:

$$\pi_i = \begin{cases} \pi_0 \frac{(m\,\rho)^i}{i\,!} & i < m \\ \pi_0 \frac{\rho^i\, m^m}{m!} & i \geq m \end{cases}$$

M/M/m - Queueing system with m servers

Applying the normalization condition, we obtain:

$$\pi_0 = \left\{ \sum_{i=0}^{m-1} \frac{(m\,\rho)^i}{i\,!} + \sum_{i=m}^{\infty} \frac{\rho^i\, m^m}{m!} \right\}^{-1} . \tag{2}$$

The second summand in Equation (2) can be rewritten 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 π_0 becomes:

$$\pi_0 = \left\{ \sum_{i=0}^{m-1} \frac{(m\,\rho)^i}{i\,!} + \frac{(m\,\rho)^m}{m\,!} \frac{1}{1-\rho} \right\}^{-1} .$$

M/M/m - Queueing system with m servers

The expected number of customers in the system $E[N]$ is:

$$E[N] = \sum_{i=0}^{\infty} i\,\pi_i = m\,\rho + \rho\,\frac{(m\,\rho)^m}{m!}\,\frac{\pi_0}{(1-\rho)^2}\,.$$

The expected number of busy servers $E[M]$ is:

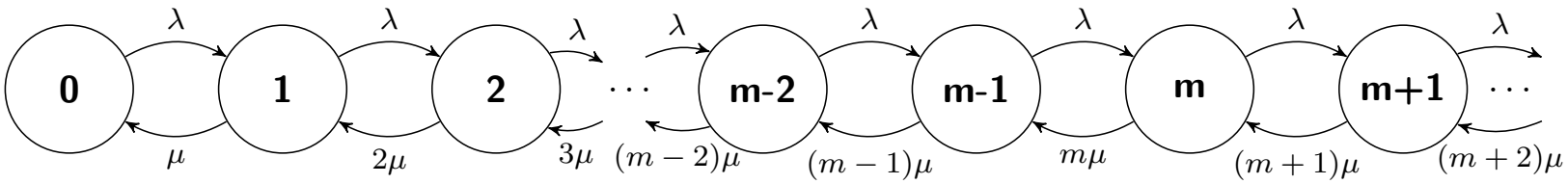
$$E[M] = \sum_{i=0}^{m-1} i\,\pi_i + m\sum_{i=m}^{\infty}\pi_i = m\,\rho = \frac{\lambda}{\mu}\,.$$

The probability that an arriving customer finds all the servers busy and joins the queue is given by:

$$\pi_{[queue]} = \sum_{i=m}^{\infty}\pi_i = \frac{\pi_m}{1-\rho} = \frac{(m\,\rho)^m}{m!}\,\frac{\pi_0}{1-\rho}\,.$$

$M/M/\infty$: infinite number of servers

A special case of the $M/M/m$ queueing system is the $M/M/\infty$ queueing system where there is an infinite number of servers and hence each arriving customer goes immediately into service without waiting in the queue. The state diagram of the queueing system is shown in Figure:



The general BD 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:

$$\pi_i = \pi_0 \prod_{j=0}^{i-1} \frac{\lambda}{(j+1)\,\mu} = \pi_0 \frac{1}{i!} \left(\frac{\lambda}{\mu}\right)^i .$$

M/M/∞: infinite number of servers

TThe normalization condition provides:

$$\pi_0 = \frac{1}{1 + \sum_{i=1}^{\infty} \frac{1}{i!} \left(\frac{\lambda}{\mu}\right)^i} = e^{-\lambda/\mu}.$$

Hence, the state probabilities assume the following form and are clearly seen to follow a Poisson pmf with parameter λ/μ:

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

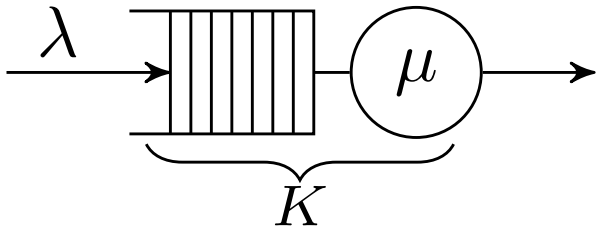
$$E[N] = \lambda/\mu \qquad ; \qquad E[R] = \frac{E[N]}{\lambda} = \frac{1}{\mu}.$$

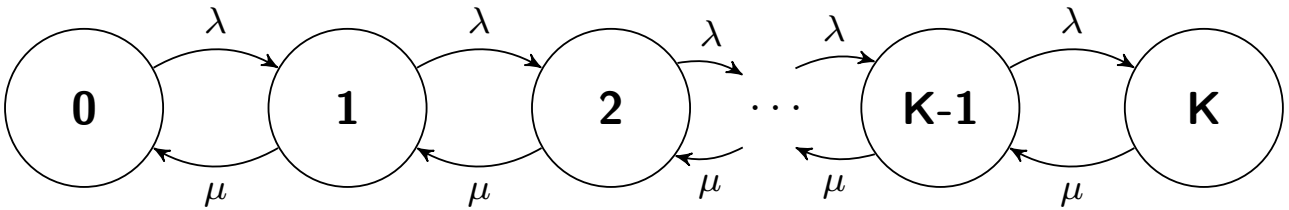
Since each arriving customer finds an available server, no waiting time is involved and the response time distribution is the same as the service time distribution

$$F_R(t) = 1 - e^{-\mu t}.$$

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 BD process can be particularized as follows:

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

M/M/1/K: finite storage

The state probabilities satisfy

$$\begin{cases} \pi_i = \pi_0 \prod_{j=0}^{i-1} \frac{\lambda}{\mu} = \pi_0 \cdot \rho^i & i \leq K \\ \pi_i = 0 & i > K \end{cases}$$

From the normalization condition:

$$\pi_0 = \frac{1}{1 + \sum_{j=1}^{K} \rho^j} = \frac{1}{1 + \frac{\rho(1-\rho^K)}{1-\rho}} = \frac{1-\rho}{1-\rho^{K+1}}.$$

Since the $M/M/1/K$ queue is a finite-state CTMC, it is stable for any value of the *traffic intensity* ρ.

M/M/1/K: finite storage

The state probabilities are:

$$\begin{cases} \pi_i = \dfrac{(1-\rho)\,\rho^i}{1-\rho^{K+1}} & i \le K \\ \pi_i = 0 & i > K \end{cases} \tag{3}$$

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

$$\lim_{\rho\to 1} \pi_i = \lim_{\rho\to 1} \frac{(1-\rho)\,\rho^i}{1-\rho^{K+1}} = \lim_{\rho\to 1} \frac{-\rho^i + i\,(1-\rho)\,\rho^{i-1}}{-(K+1)\,\rho^K} = \frac{1}{K+1}\,.$$

The *rejection probability* is the probability that an arriving customer will find the queue full and is rejected (or blocked). Since the queue is full when in state K, the *rejection probability* is:

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

The rate at which jobs are rejected is $\lambda\pi_K$ and the rate at which jobs are accepted is $\lambda_{acc} = \lambda(1-\pi_K)$.

M/M/1/K: finite storage

The expected number of customers $E[N]$ is:

$$E[N] = \sum_{i=0}^{K} i \cdot \pi_i = \sum_{i=0}^{K} i \cdot \frac{(1-\rho)\,\rho^i}{1-\rho^{K+1}} = \frac{1-\rho}{1-\rho^{K+1}} \sum_{i=0}^{K} i \cdot \rho^i$$

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

Above formula is based on the following finite series:

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

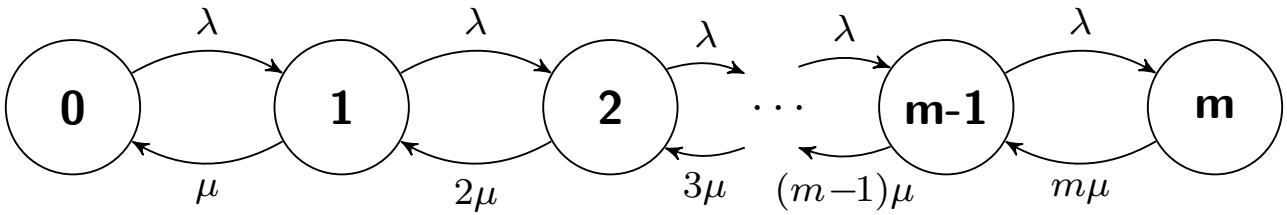
From Equation of $E[N]$, it follows:

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

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

M/M/m/m: no waiting line

The $M/M/m/m$ queue does not have a waiting line and an arriving customer enters service if and only if one of the m servers is idle. The state diagram of the system is shown in Figure **??**.



The general BD can be expressed mathematically as follows:

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

$$\begin{cases} \pi_i = \pi_0 \displaystyle\prod_{j=0}^{i-1} \frac{\lambda}{\mu_{j+1}} = \pi_0 \cdot \frac{\lambda^i}{i\,!\,\mu^i} & i \leq m \\ \pi_i = 0 & i > m \end{cases}$$

M/M/m/m: no waiting line

From the normalization condition:

$$\pi_0 = \frac{1}{1 + \sum_{j=1}^{m} \frac{\lambda^i}{i!\,\mu^i}} = \frac{1}{1 + \sum_{j=1}^{m} \frac{\rho^i}{i!}},$$

where the traffic intensity ρ is defined as $\rho = \lambda/\mu$. The state probabilities are:

$$\begin{cases} \pi_i = \dfrac{\rho^i}{i!} \left(\sum_{j=0}^{m} \dfrac{\rho^j}{j!} \right)^{-1} & i \leq m \\ \pi_i = 0 & i > m \end{cases} \tag{4}$$

and the rejection probability is obtained by substituting $i = m$ in (4),

$$\pi_m = \frac{\rho^m}{m!} \left(\sum_{j=0}^{m} \frac{\rho^j}{j!} \right)^{-1}. \tag{5}$$

M/M/m/m: no waiting line

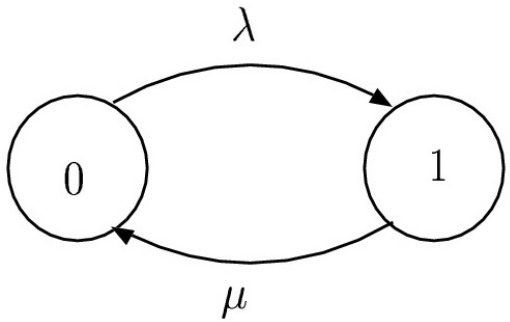
The expected number of customers in the system is:

$$\begin{aligned} E[N] &= \sum_{j=0}^{m} j\,\pi_j = \pi_0 \sum_{j=0}^{m} j\,\frac{\rho^i}{i\,!} \\ &= \rho - \rho\,\frac{\rho^m}{m\,!}\,\pi_0 = \rho\,(1 - \pi_m)\,. \end{aligned}$$

M/M/1/1: no waiting line

With $m = 1$, the system reduces to a $M/M/1/1$ queue that corresponds to a 2-state CTMC with the state diagram shown in Figure.

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, the steady state probabilities are then given by:

$$\begin{cases} \pi_0 = \dfrac{1}{1+\rho} = \dfrac{\mu}{\lambda+\mu} \\ \pi_1 = \dfrac{\rho}{1+\rho} = \dfrac{\lambda}{\lambda+\mu} \end{cases}$$

Telecommunication Switching System: Performance

The system consists of n trunks (or channels) with an infinite caller population. A call will be lost (referred to as blocking) when it finds all n trunks are busy upon its arrival.

The call arrival process is assumed to be Poisson with rate λ and the call holding times are exponentially distributed with rate μ.

Without considering link failure, the pure performance model is an $M/M/n/n$ queue, and the transient probability of being in State j can be computed by solving the following set of differential equations:

$$\begin{aligned}
\frac{d\pi_0(t)}{dt} &= -\lambda\pi_0(t) + \mu\pi_1(t) \\
\frac{d\pi_k(t)}{dt} &= -(\lambda + k\mu)\pi_k(t) + \lambda\pi_{k-1}(t) + (k+1)\mu\pi_{k+1}(t), \quad k = 1, 2, ..., n-1 \\
\frac{d\pi_n(t)}{dt} &= -n\mu\pi_n(t) + \lambda\pi_{n-1}(t).
\end{aligned}$$

with the initial condition $\pi_k(t = 0) = \pi_k(0)$.

Telecommunication Switching System: Performance

The transient blocking probability is

$$P_{bk}(t) = \pi_n(t)$$

The expected number of lost calls in the interval $(0, t]$ is given as

$$N_{loss}(t) = \int_0^t \lambda P_{bk}(x)\, dx$$

.

n-identical component failure/repair model

Consider a system with n identical components, each with a failure rate λ and repair rate μ.

This is known as the machine repairmen model.

The components are repaired with two possible repair policies:

i) - independent repair;

ii) - shared repair with a single repair person.

n-identical component failure/repair model

The failure/repair process of the system can be modeled as a BD process with $n+1$ states. We assume as a state index i the number of failed components.

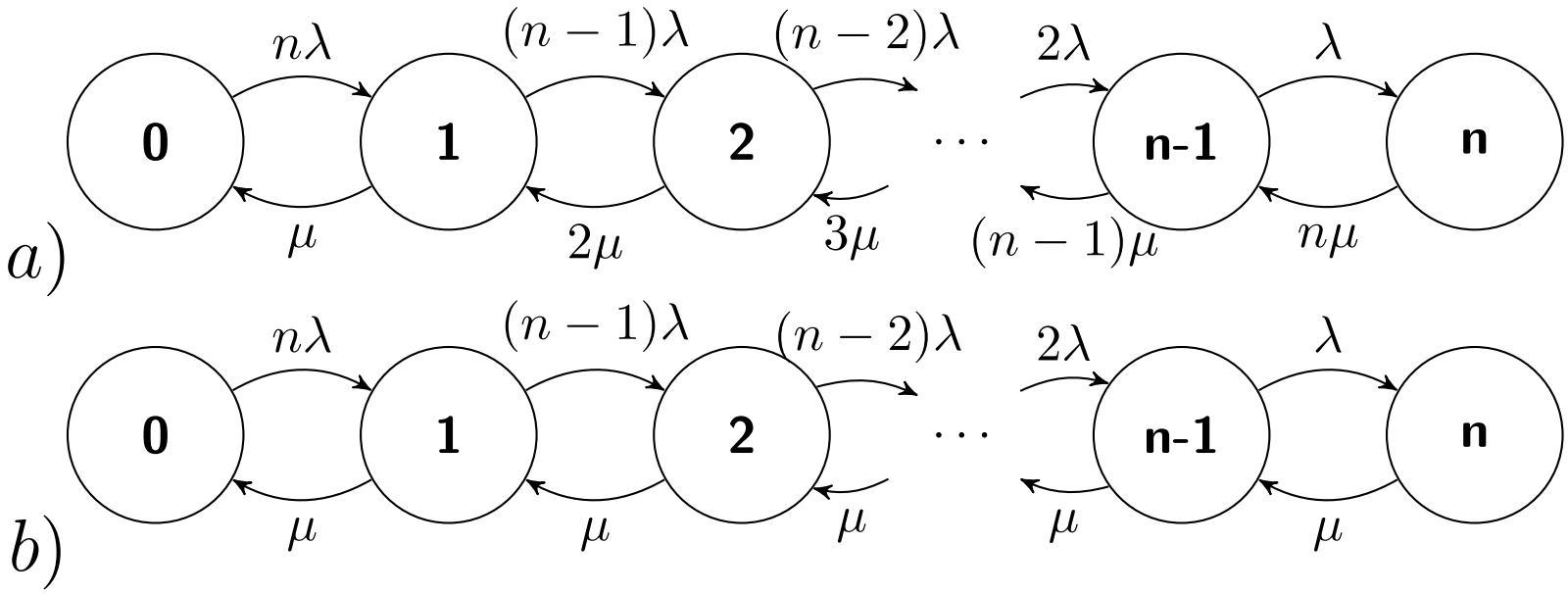
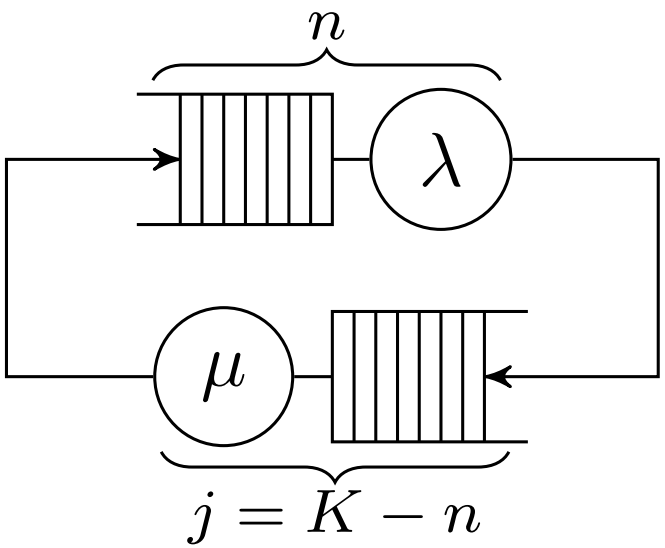


Figure a) shows the case of independent repair (as many repair-persons as failed components);

Figure b) the case of shared repair with a single repair person.

Closed $M/M/1$ queue

A closed $M/M/1$ queueing system is a system where the total number of jobs in the system is constant, as represented in the Figure.



K is the total number of jobs in the system; inactive jobs are queued in the upper queue, and are sent to the server with rate λ.

The active jobs, that are queued for service in the lower queue, are served at the rate μ.

Closed $M/M/1$ queue

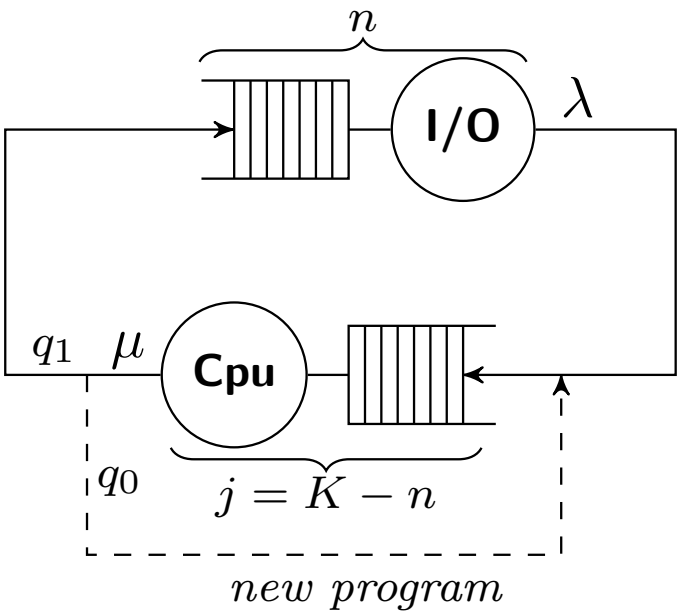
We use the number of jobs in the server subsystem, $j = K - n$, as the index of a state in the state space, where n is the number of jobs in the waiting queue.

With this assignment, the state diagram of the closed $M/M/1$ queue is similar to the state diagram of the $M/M/1/K$ queue, and a similar solution approach can be adopted.

However, while in the $M/M/1/K$ the appropriate model should display a self-loop on state K since jobs keep on arriving even when the system is full (though rejected in this state), in the closed $M/M/1$ queue case, no jobs can arrive in state K since the total number of jobs is fixed.

Cyclic queueing system

Consider the cyclic queueing model for a multiprogramming system depicted in Figure.



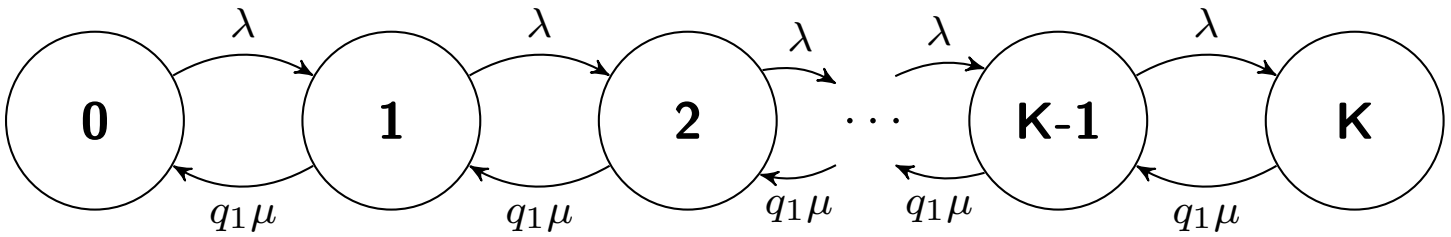
The successive CPU bursts are exponentially distributed with rate μ and the successive I/O bursts are exponentially distributed with rate λ.

At the end of a CPU burst the job requires an I/O operation with probability q_1 and leaves the system with probability q_0 ($q_1 + q_0 = 1$). At the end of a job completion a statistically identical job enters the system leaving the number of jobs constant (level of multiprogramming)

Cyclic queueing system

At the end of a job completion a statistically identical job enters the system leaving the number of jobs constant (level of multiprogramming).

The state diagram is shown in Figure:



Comparing the state diagram of the Figure with the state diagram of the $M/M/1/K$ queue, we can see that the solution of this case can be derived from the solution of the $M/M/1/K$ by setting

$$\rho = \frac{\lambda}{q_1\,\mu}$$

.

Queues with Breakdown

A number of situations occur in real systems where the single service station is incapacitated from time to time to render service to the incoming customers (server breakdown).

When a server failure occurs, the job (if any) in service will be interrupted.

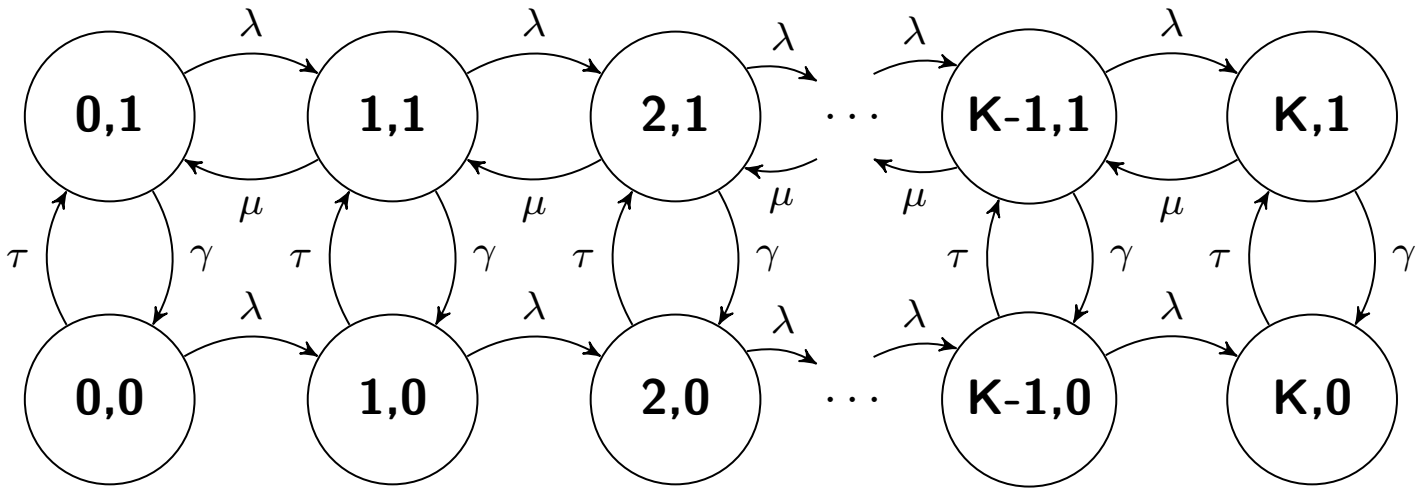
The job may be dropped or may simply be preempted. In the former case, the job may be later retried.

In the case of preemption, the job will be executed once the server is repaired. There are several possibilities to consider as the preemptive resume (*prs*) policy, and preemptive repeat (*prt*) policy.

Further, in the *prt* case, there are possible sub categories depending whether the interrupted job is restarted with a different time request (but distributed with the same Cdf) *preemptive repeat different (prd)* or with an identical time request *preemptive repeat identical (pri)*

$M/M/1/K$ queue with breakdown

In the $M/M/1/K$ queue, we now allow the server to fail at the rate γ and get repaired at the rate τ, obtaining, under the *prs* policy, the state diagram in Figure.



In the Figure, the state label (i, b) indicates that there are i jobs in the system and b indicates the state of the server (with the usual assignment $b = 1$ sever up, $b = 0$ sever down).

$M/M/1/K$ queue with breakdown

Obtaining a closed-form solution for the steady state probabilities is difficult.

The solution for the CTMC can be obtained numerically given the values of the various parameters.

Then we can obtain several measures of interest by an appropriate assignment of reward rates to the states of the CTMC.

For a customer, some interesting measures to compute could be the blocking probability, the rate of blocking and the mean response time (for completed jobs).

For the whole system, an important measure is the expected number of jobs in the system.

$M/M/1/K$ queue with breakdown

Expected number of jobs in the system at time t (an example of the expected reward rate at time t):

$$E[N(t)] = \sum_{i=0}^{K} i \cdot (\pi_{i,1}(t) + \pi_{i,0}(t)) .$$

A cumulative transient analysis will yield the expected number of jobs completed in the interval $(0, t]$:

$$E[C(t)] = \sum_{i=0}^{K} \mu \cdot \int_0^t \pi_{i,1}(x)dx ,$$

and the expected number of jobs blocked (rejected) in the same interval:

$$E[B(t)] = \lambda \cdot \int_0^t (\pi_{K,1}(x) + \pi_{K,0}(x))dx ,$$

which are both examples of expected accumulated reward in $(0, t]$.

Stiff CTMC: $M/M/1/K$ queue with breakdown

The $M/M/1/K$ queue with breakdown is an excellent example of a *stiff* Markov chain due to the presence of transitions with rates of vastly differing orders of magnitude.

The arrival and departure transitions are fast transitions, while the failure and repair transitions are slow transitions since their rates are much smaller ($\lambda, \mu >> \gamma, \tau$).

One way to deal with stiff CTMC is to use stiffly stable numerical methods, but two other options can be envisaged based on

- aggregation
- decomposition.

$M/M/1/K$ queue with breakdown: Aggregation

The idea behind state aggregation is to classify each state into a fast state (one with at least one fast transition rate out of it) or a slow state, and to consider that the fast states have reached their asymptotic value in the time scale of the slow transitions.

The algorithm then proceeds to group the set of all fast states into one or more *fast recurrent* subsets and one *fast transient* subset of states.

The fast subsets can then be analyzed in isolation to find their steady state condition.

The algorithm then replaces each fast recurrent subset with a slow state, and the fast transient subset with a probabilistic switch to construct a smaller non-stiff Markov chain that can be analyzed using conventional techniques.

$M/M/1/K$ queue with breakdown: Aggregation

Details of the application of the aggregation technique to solve the $M/M/1/K$ queue with breakdown are in [*]:

[*] A. Bobbio and K. S. Trivedi, "An aggregation technique for the transient analysis of stiff Markov chains," *IEEE Transactions on Computers*, vol. C-35, pp. 803–814, 1986.

$M/M/1/K$ queue with breakdown: Decomposition

The decomposition approach has been adopted in a pioneering work by Meyer [*] to model and solve the system *performability*.

The *performability* model separates the overall system representation into an upper level dependability model that typically encompasses all the slow events, and a lower level performance model that captures all the fast events in the system.

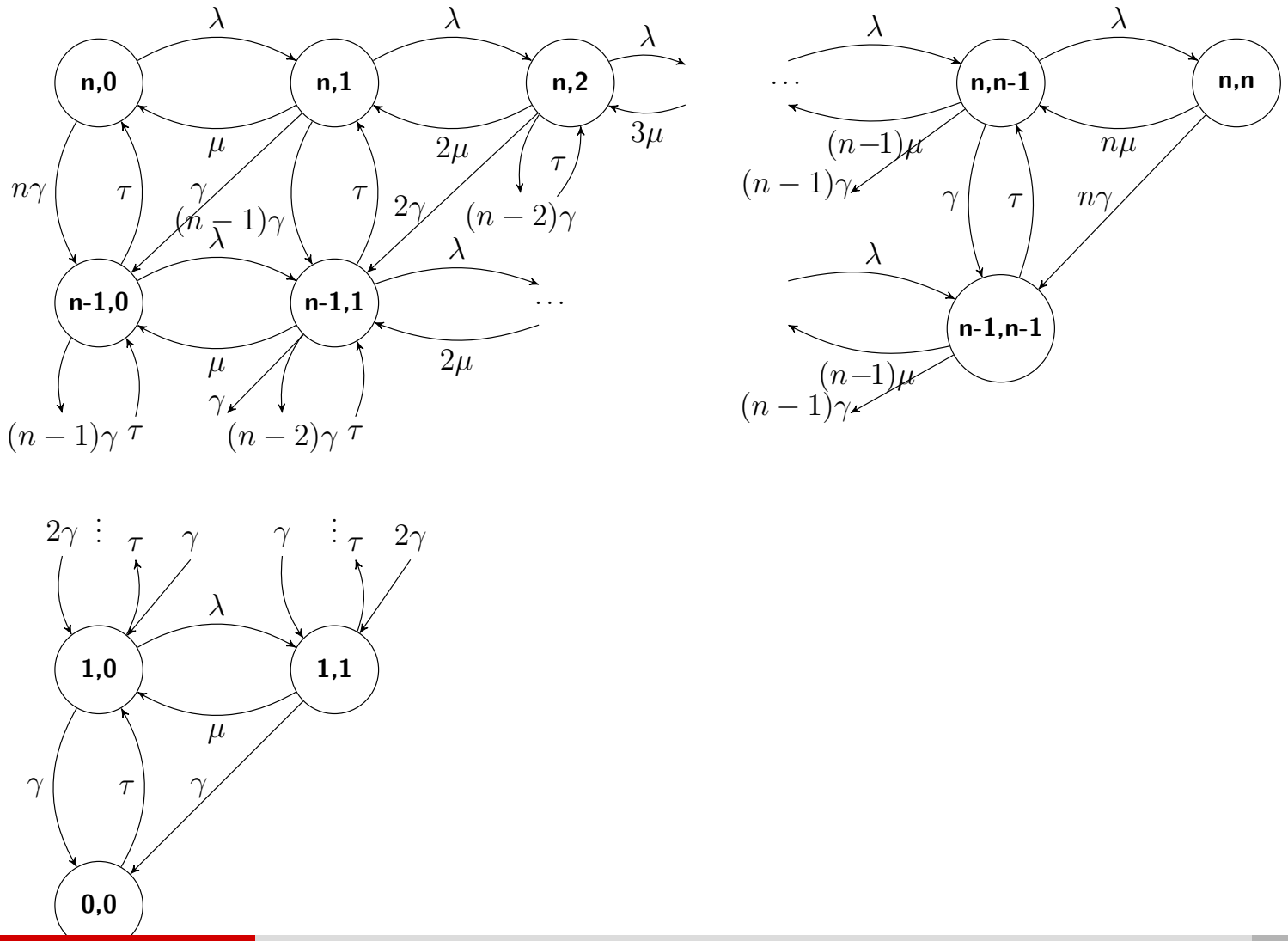
The performance metrics obtained from the performance models corresponding to each dependability state, are then incorporated into the higher level dependability model in the form of reward rates.

Details of the method are in:

[*] J. Meyer, "On evaluating the performability of degradable systems," *IEEE Transactions on Computers*, vol. C-29, pp. 720–731, 1980.

Telecommunication Switching System Model

The composite performance–availability model is shown in the Figure.



Telecommunication Switching System Model

State (i, j) indicates that there are i non-failed trunks in the system and j of them are carrying ongoing calls.

λ and μ are the time-independent arrival and service rates, while γ and τ are the time-independent failure and repair rates.

We assume a single repair-person. Transitions due to arrival of new calls and due to completion of existing calls are self-explanatory.

From each State (i, j) two failure transitions emanate, since we distinguish whether the failure involves a busy trunk (with rate $j\,\gamma$) with loss of the current call, or an idle trunk with rate $(i - j)\,\gamma$ with no loss of calls.

Telecommunication Switching System Model

The steady-state probability of being in State (i, j) is denoted as $\pi_{i,j}$. The steady-state and transient blocking probability P'_{bk} and $P'_{bk}(t)$ can be computed as

$$P'_{bk} = \sum_{k=0}^{n} \pi_{k,k} \quad , \qquad P'_{bk}(t) = \sum_{k=0}^{n} \pi_{k,k}(t)$$

where State (k, k) in this model represents the situation that all the functioning trunks are busy or all trunks are down.

The expected number of lost calls till time t is given as

$$E[N'_{loss}(t)] = \int_0^t \lambda P'_{bk}(x)\, dx.$$