Network Performance Analysis

Joab R. Winkler

Department of Computer Science

The University of Sheffield

Contents

1. Delay models in networks

One of the most important measures of a data network is the average delay required to deliver a packet from its source to its destination.

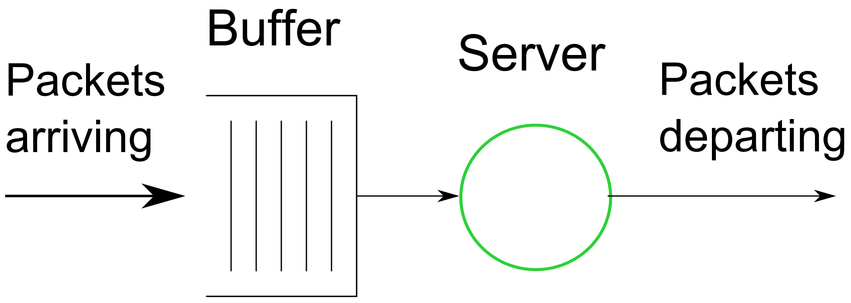
- Delay considerations strongly influence the choice and performance of network algorithms, for example, routing and flow control
- Queuing theory is used to analyse network delay. Assumptions must be made in order to make progress with the theory:
  - This limits the accuracy of the quantitative results, but they do, nevertheless, provide useful results for design and analysis

Why is queuing theory important for the analysis of packet switched networks?

- Packets that arrive at an entry point to a network or an intermediate node on the way to the destination are buffered, processed to determine the link to the next node along the path, and sent over that link when they are to be transmitted.
- The time spent in the buffer waiting for transmission is a significant measure of the performance of such a network. The waiting time depends on:
  - The processing time of each node and the packet length
  - The capacity of the transmission link, measured in packets/second
  - The traffic arriving at the node, also measured in packets/second
  - The service method used to handle the packets

Queuing theory will consider these items, apart from nodal processing time

Consider the simplest model of a queue, shown below



Model of a single server queue

- The queuing theory applied to computer networks can also be applied to customers in a service queue in a shop
- The packets arrive randomly, at an average rate of λ packets/sec
- They queue for service in the buffer and are then served, according to a specified service discipline, at an average rate of μ packets/sec
- More generally, multiple servers may be available, in which case more than one packet may be in service at a time

- As $\lambda \to \mu$, that is, the arrival rate approaches the service rate, a queue will form:
  - For a finite buffer, which is the situation in real life, the queue will saturate as λ exceeds μ, and continue to increase
  - If the buffer is filled, all further packets are blocked on arrival
  - If the buffer has infinite capacity, an assumption that is often made in order to simplify analysis, the queue becomes unstable as $\lambda \to \mu$
- It will be shown that $\lambda < \mu$ ensures stability for a single server queue
- The parameter $\rho = \frac{\lambda}{\mu}$ is called the *utilisation* or *traffic intensity* and plays a critical role in queuing theory
- For a single server queue, as $\rho \to 1$ and then increases, congestion occurs and the packets that arrive are blocked

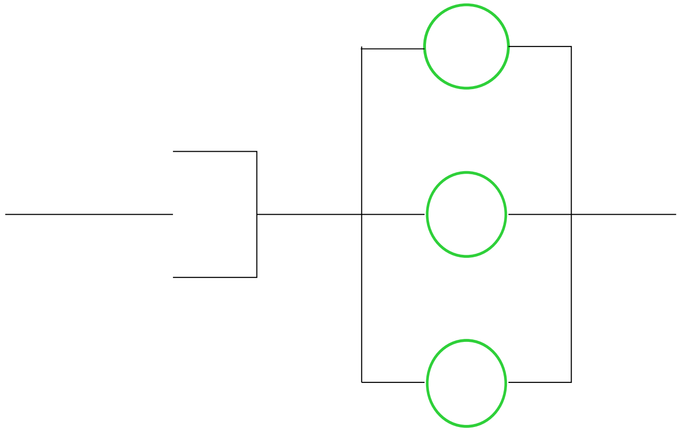
To quantify the analysis of queuing networks, the following is required:

- The arrival rate of the packets, denoted by λ
- The service time distribution
- The service discipline, for example, first come-first served, or last come-first served
- The number of servers, which represents the number of outgoing links

Shorthand notation for queuing systems that contain a single queue:

A queuing system with a *single queue* is denoted by M/M/m/n, where the characters refer to, respectively:

M - the statistics of the arrival process; M - the statistics of the server; m - the number of servers; n - the number of customers the system can hold



An M/M/3 queueing system.

The characters in the first two positions of the M/M/m/n notation indicate:

- M: Markovian statistics
- D: Deterministic timing
- General (arbitrary) statistics
- Geom: Geometric statistics

2. The Poisson process

- The Poisson process is an example of a Markov process:
  - A Markov process is a process whose value at time $t = t_n$ is only a function of its value at time $t = t_{n-1}$ and independent of the function values at all previous times, $t < t_{n-1}$.
  - A Markov process has no memory (history) and it is therefore relatively easy to analyse.
- The Poisson process is the most frequently used arrival process in queuing theory
- It was originally used to model the arrival of calls to a telephone exchange
- The Poisson process may not always be valid because there may, for example, be correlations between individual calls:
  - A caller is more likely to make a second call if he/she was unsuccessful on

the first attempt

- If the exchange is heavily loaded, future calls are blocked, which results in repeated attempts to make calls

Assumptions of the Poisson process:

- The probability P of exactly one customer arriving in the time interval Δt is proportional to Δt, where λ is the proportionality constant
- Only one customer can arrive in the time interval Δt

$$\begin{aligned} P(\text{exactly one arrival in } [t, t+\Delta t]) &= \lambda\Delta t \\ P(\text{exactly no arrivals in } [t, t+\Delta t]) &= 1-\lambda\Delta t \\ P(\text{more than one arrival in } [t, t+\Delta t]) &= 0 \end{aligned}$$

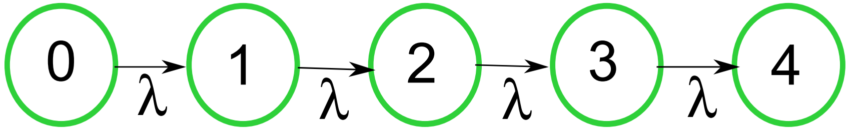
- Let $P_k(t)$ denote the probability of k arrivals in the time interval t
- Let $p_{i,j}(\Delta t)$ be the probability of going from i arrivals to j arrivals in the time interval Δt

$$\begin{aligned} P_k(t+\Delta t) &= P_k(t)p_{k,k}(\Delta t) + P_{k-1}(t)p_{k-1,k}(\Delta t) \\ P_0(t+\Delta t) &= P_0(t)p_{0,0}(\Delta t) \qquad k=0 \text{ (starting condition)} \end{aligned}$$

Notes:

- Only first order terms are used, that is, $\Delta t \gg (\Delta t)^2 \approx 0$
- This equation states that k customers in the time interval $t+\Delta t$ can arise either by having k or $k-1$ customers in the time interval t
- It is assumed that Δt is sufficiently small such that one customer, at most, can arrive in this interval

In the state transition diagram, the circles represent the states of the system (number of arrivals) and the transition rate λ is associated with each transition.



Poisson process state transition diagram.

But

$$p_{k,k}(\Delta t) = (1 - \lambda \Delta t) \quad \text{and} \quad p_{k-1,k}(\Delta t) = \lambda \Delta t$$

and thus

$$\begin{aligned} P_k(t + \Delta t) &= P_k(t)(1 - \lambda \Delta t) + P_{k-1}(t)(\lambda \Delta t) \\ P_0(t + \Delta t) &= P_0(t)(1 - \lambda \Delta t) \qquad k = 0 \text{ (starting condition)} \end{aligned}$$

Rearrange these equations to yield

$$\begin{aligned}\frac{P_k(t+\Delta t)-P_k(t)}{\Delta t} &= -\lambda P_k(t)+\lambda P_{k-1}(t)\\ \frac{P_0(t+\Delta t)-P_0(t)}{\Delta t} &= -\lambda P_0(t)\end{aligned}$$

Let $\Delta t \to 0$, which leads to a set of differential equations for $k \geq 1$:

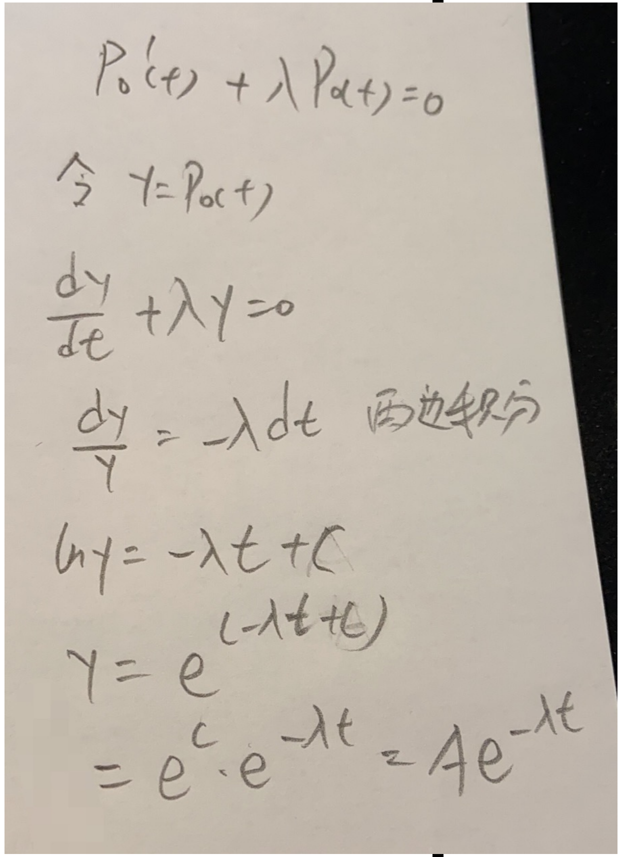
$$\begin{aligned} \frac{dP_k(t)}{dt} &= -\lambda P_k(t) + \lambda P_{k-1}(t) \\ \frac{dP_0(t)}{dt} &= -\lambda P_0(t) \end{aligned}$$

The solution of the second equation is

$$P_0(t) = A \exp(-\lambda t)$$

where A is an arbitrary constant, and thus

$$\frac{dP_1(t)}{dt} = -\lambda P_1(t) + A\lambda \exp(-\lambda t)$$



$P_0'(t) + \lambda P_0(t) = 0$

令 $y = P_0(t)$

$\frac{dy}{dt} + \lambda y = 0$

$\frac{dy}{y} = -\lambda dt$ 两边积分

$\ln y = -\lambda t + C$

$y = e^{(-\lambda t + C)}$

$= e^c \cdot e^{-\lambda t} = Ae^{-\lambda t}$

The solution of this equation is

$$P_1(t) = A\lambda t \exp(-\lambda t)$$

and thus

$$\frac{dP_2(t)}{dt} = -\lambda P_2(t) + A\lambda^2 t \exp(-\lambda t)$$

whose solution is

$$P_2(t) = \frac{A(\lambda t)^2}{2} \exp(-\lambda t)$$

The general solution for arbitrary k is

$$P_k(t) = \frac{A(\lambda t)^k}{k!} \exp(-\lambda t)$$

But what is the value of A?

Recall that $P_k(t)$ denotes the probability of k arrivals in the time interval t. Thus

$$P_0(t) + P_1(t) + P_2(t) + \cdots = \sum_{i=0}^{\infty} P_i(t) = 1$$

and thus

$$A \exp(-\lambda t) \sum_{k=0}^{\infty} \frac{(\lambda t)^k}{k!} = 1$$

which yields $A = 1$.

$$\boxed{P(k|t,\lambda) = \frac{(\lambda t)^k}{k!}\exp(-\lambda t)}$$

defines the Poisson distribution.

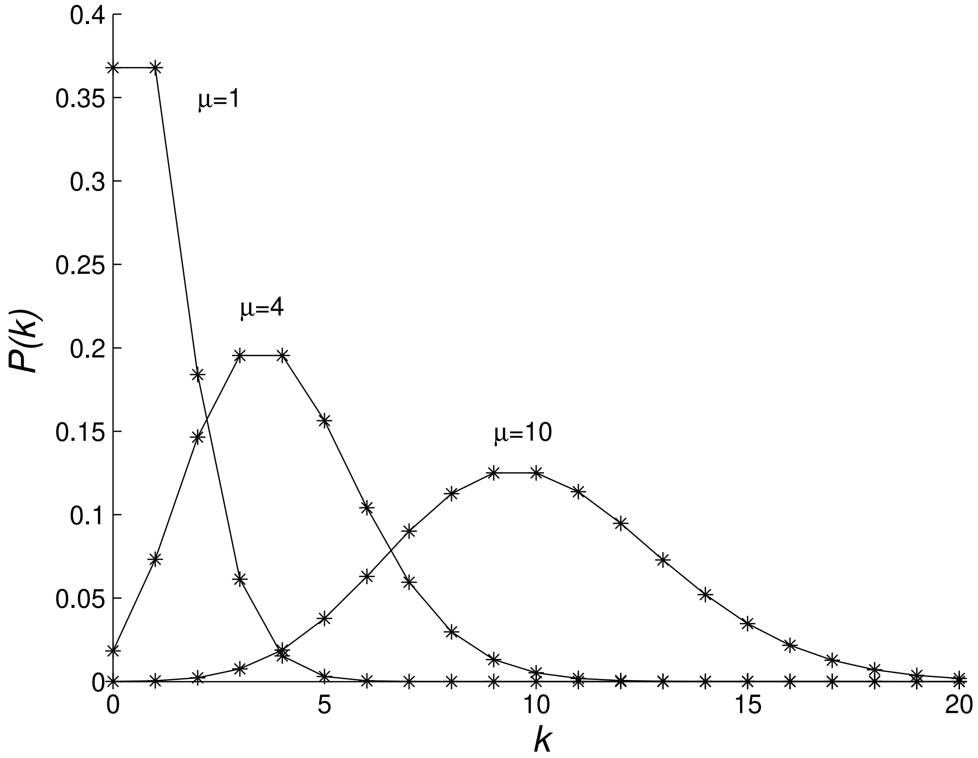
- It defines the probability of k arrivals in the time interval t for a Poisson process of rate λ, which has units $[\text{time}]^{-1}$.
- It is assumed that λ is constant, and thus t is a *time interval*

$$P(k|(t_2 - t_1), \lambda) = P(k|(t_4 - t_3), \lambda), \qquad t_2 - t_1 = t_4 - t_3$$

Example A telephone exchange receives, on average, 100 calls a minute, according to a Poisson process. What is the probability that no calls are received in an interval of 5 seconds?

$$P\left(k = 0|t = \frac{1}{12}, \lambda = 100\right) = \exp\left(-\frac{100}{12}\right) = 0.00024$$

□



The Poisson distribution for three values of $\mu = \lambda t$

Recall the assumptions of the Poisson distribution:

$$\begin{aligned} \Delta t \gg (\Delta t)^2 &\approx 0 \\ P(\text{exactly one arrival in } [t, t+\Delta t]) &= \lambda \Delta t \\ P(\text{more than one arrival in } [t, t+\Delta t]) &= 0 \end{aligned}$$

Does the Poisson distribution satisfy these assumptions?

$$P(\text{exactly one arrival in } [t, t+\Delta t]) = P(1|\Delta t, \lambda) = \lambda \Delta t \exp(-\lambda \Delta t)$$

By assumption 1, only consider lowest order terms

$$\begin{aligned} P(\text{exactly one arrival in } [t, t+\Delta t]) &= \lambda \Delta t \exp(-\lambda \Delta t) \approx \lambda \Delta t \\ P(\text{more than one arrival in } [t, t+\Delta t]) &= \sum_{k=2}^{\infty} \frac{(\lambda \Delta t)^k}{k!} \exp(-\lambda \Delta t) \\ &\approx 0 \text{ because } (\Delta t)^k \approx 0 \text{ for } k \geq 2 \end{aligned}$$

Properties of the Poisson distribution

- It is easily checked that

$$\sum_{k=0}^{\infty} P(k|t,\lambda) = \sum_{k=0}^{\infty} \frac{(\lambda t)^k}{k!} \exp(-\lambda t) = 1, \qquad P(k|t,\lambda) \geq 0$$

- The mean of the Poisson distribution. If $\mu = \lambda t$, then the average number of arrivals in a time interval t is

$$\begin{aligned} E\{k\} &= \sum_{k=0}^{\infty} \frac{k\mu^k}{k!} \exp(-\mu) \\ &= \sum_{k=1}^{\infty} \frac{\mu^k}{(k-1)!} \exp(-\mu) \end{aligned}$$

$$
\begin{aligned}
&= \mu \exp(-\mu) \sum_{k=1}^{\infty} \frac{\mu^{k-1}}{(k-1)!} \\
&= \mu \exp(-\mu) \sum_{m=0}^{\infty} \frac{\mu^{m}}{m!} \\
&= \mu
\end{aligned}
$$

Thus $E\left\{k\right\} = \mu = \lambda t$, and thus the average number of arrivals in a time interval t is proportional to t and the arrival rate λ.

- The variance of the Poisson distribution.

$$\text{var}\left\{k\right\} = \sigma_k^2 = E\left\{(k-\mu)^2\right\} = E\left\{k^2\right\} - \mu^2$$

$$\begin{aligned}
E\left\{k^2\right\} &= \sum_{k=0}^{\infty} k^2 \frac{\mu^k}{k!}\exp(-\mu) \\
&= \sum_{k=1}^{\infty} \frac{k\mu^k}{(k-1)!}\exp(-\mu) \\
&= \sum_{k=1}^{\infty} \frac{(k-1)\mu^k}{(k-1)!}\exp(-\mu) + \sum_{k=1}^{\infty} \frac{\mu^k}{(k-1)!}\exp(-\mu) \\
&= \mu^2 \exp(-\mu) \sum_{k=2}^{\infty} \frac{\mu^{(k-2)}}{(k-2)!} + \mu \exp(-\mu) \sum_{k=1}^{\infty} \frac{\mu^{(k-1)}}{(k-1)!} \\
&= \mu^2 + \mu
\end{aligned}$$

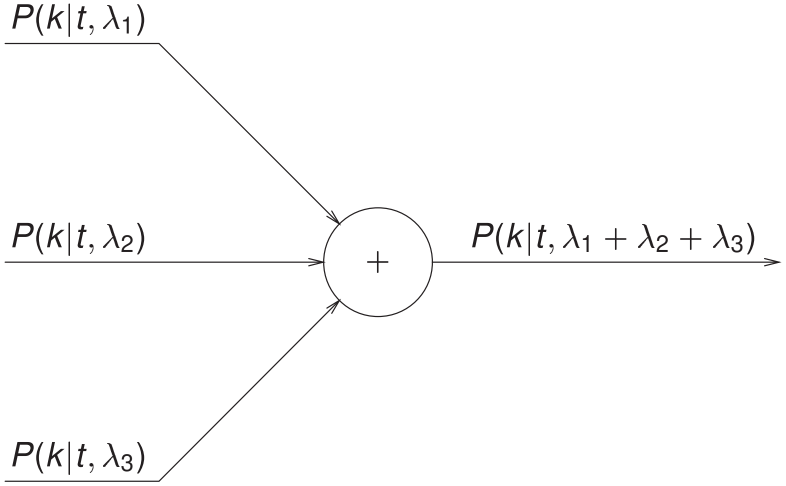
$$\text{var}\{k\} = \sigma_k^2 = E\left\{(k-\mu)^2\right\} = E\left\{k^2\right\} - \mu^2 = \mu$$

$$\boxed{E\{k\} = E\left\{k^2\right\} - \mu^2 = \mu}$$

- Superposition of Poisson processes:

  If N independent Poisson processes with rates $\lambda_1, \lambda_2, \ldots, \lambda_N$ are added together, the resulting process is a Poisson process with rate

$$\lambda = \lambda_1 + \lambda_2 + \cdots + \lambda_N$$



The superposition of Poisson processes

Example There are three queues for registration of MSc students. Students for the first, second and third courses arrive at a rate of 4 students per minute, 2 students per minute, and 5 students per minute.

Assuming that student arrivals are independent and occur as a Poisson process, what is the probability that after 15 minutes there have been **(a)** 90 student arrivals, and **(b)** 180 arrivals?

(a) The sum of the arrival rates is $\lambda = 11$ students per minute, and thus the probability that there are 90 student arrivals after 15 minutes is

$$P(90|15, 11) = \frac{(11 \times 15)^{90}}{90!} \exp(-(11 \times 15)) = 5.5 \times 10^{-11}$$

(b) The probability that there are 180 student arrivals after 15 minutes is

$$P(180|15,11) = \frac{(11 \times 15)^{180}}{180!} \exp(-(11 \times 15)) = ?$$

The numbers in this calculation are too big. Use

$$x \equiv \exp^{\ln x}, \qquad x = \frac{(11 \times 15)^{180}}{180!} \exp(-(11 \times 15))$$

to evaluate the expression:

$$\ln x = 180 \ln(11 \times 15) - (11 \times 15) - \sum_{i=1}^{180} \ln i$$

and thus

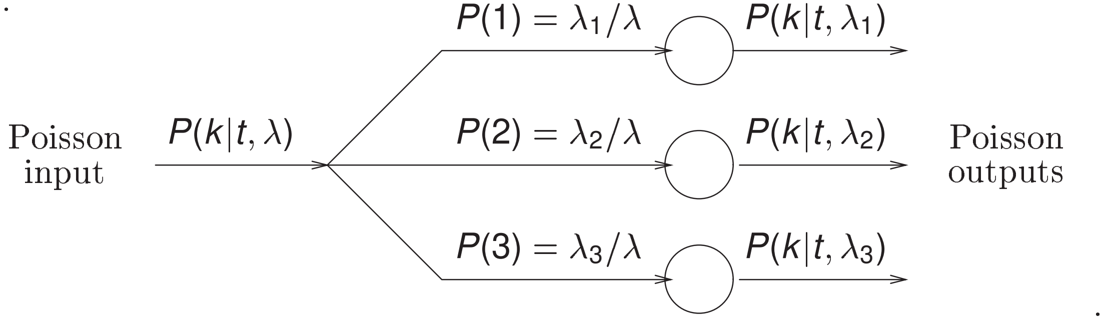
$$x = P(180|15,11) = 0.015$$

□

- Decomposition of Poisson processes:

  Assume there exists a queue in which the arrivals occur with a Poisson process at a rate λ. This channel is split up into N separate output channels, and each input arrival is assigned independently to output channel j with probability $P(j)$. The queue in the jth output channel is a Poisson process with rate

$$\lambda_j = \lambda P(j)$$



The decomposition of Poisson processes

Example There is one registration for all MSc courses. Students arrive in the hall and are assigned to registration desks according to the first letter of their surname. Desk 1 ($d = 1$) takes students with names from A-G, desk 2 ($d = 2$) from H-M, desk 3 ($d = 3$) from N to R and desk 4 ($d = 4$) from S to Z.

Surnames are distributed amongst these four categories with the following probability distribution:

d	1	2	3	4
$P(d)$	0.20	0.27	0.25	0.28

If it is assumed the students arrive independently of their surname as a Poisson process at a rate of 10 students per minute, what is the probability that **(a)** 5 students have arrived at desk 1 after 5 minutes? **(b)** 20 students have arrived at desk 4 after 10 minutes?

(a) Arrivals at desk 1 will occur as a Poisson process with rate $\lambda_1 = 0.2 \times 10 = 2$. The probability that $k = 5$ students will have arrived at desk 1 after $t = 5$ minutes is

$$P\left(k = 5 | t = 5, \lambda_1 = 2\right) = \frac{(2 \times 5)^5}{5!} \exp(-(2 \times 5)) = 0.038$$

(b) Arrivals at desk 4 will occur as a Poisson process with rate $\lambda_4 = 0.28 \times 10 = 2.8$. The probability that $k = 20$ students will have arrived at desk 4 after $t = 10$ minutes is

$$P\left(k = 20 | t = 10, \lambda_4 = 2.8\right) = \frac{(2.8 \times 10)^{20}}{20!} \exp(-(2.8 \times 10)) = 0.025$$

□

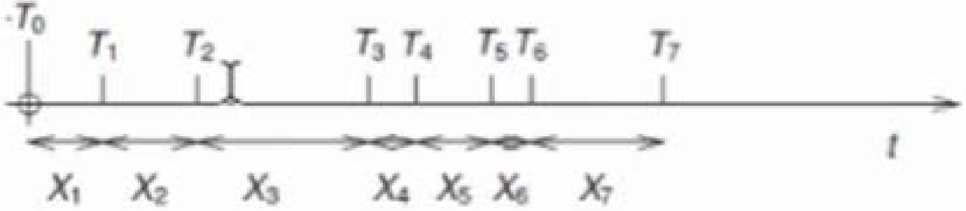
Inter-arrival times

- The time intervals between successive events are called the inter-arrival times

Consider arrival times $T_1, T_2, \ldots, T_n$ of a Poisson process, and let

$$X_i = T_i - T_{i-1}$$

be the times between successive arrivals.



The arrival times T_i and inter-arrival times X_i of a Poisson process

- What is the probability distribution of the random variable X_i?

$$\begin{aligned} P(\text{time interval between arrivals } \leq t) &= 1 - P(\text{time interval between arrivals } > t) \\ &= 1 - P(0|t,\lambda) \\ &= 1 - \exp(-\lambda t) \\ &= \int_0^t f(x)\mathsf{dx} \end{aligned}$$

and thus

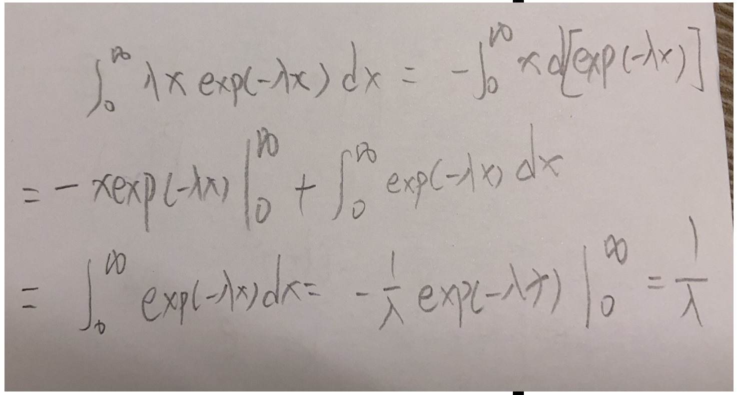
$$f(t) = \frac{\mathsf{d}(1-\exp(-\lambda t))}{\mathsf{d}t} = \lambda \exp(-\lambda t)$$

If the number of arrivals in a time interval T_i have a Poisson distribution with rate λ, then the inter-arrival times have an exponential distribution

$$P(X_i) = \lambda \exp(-\lambda X_i)$$

- The mean of the exponential distribution is

$$E\left\{x\right\} = \int_0^\infty \lambda x \exp(-\lambda x)\,\textsf{dx} = \frac{1}{\lambda}$$



$$\int_0^\infty \lambda x \exp(-\lambda x)\,dx = -\int_0^\infty x\,d\left[\exp(-\lambda x)\right]$$

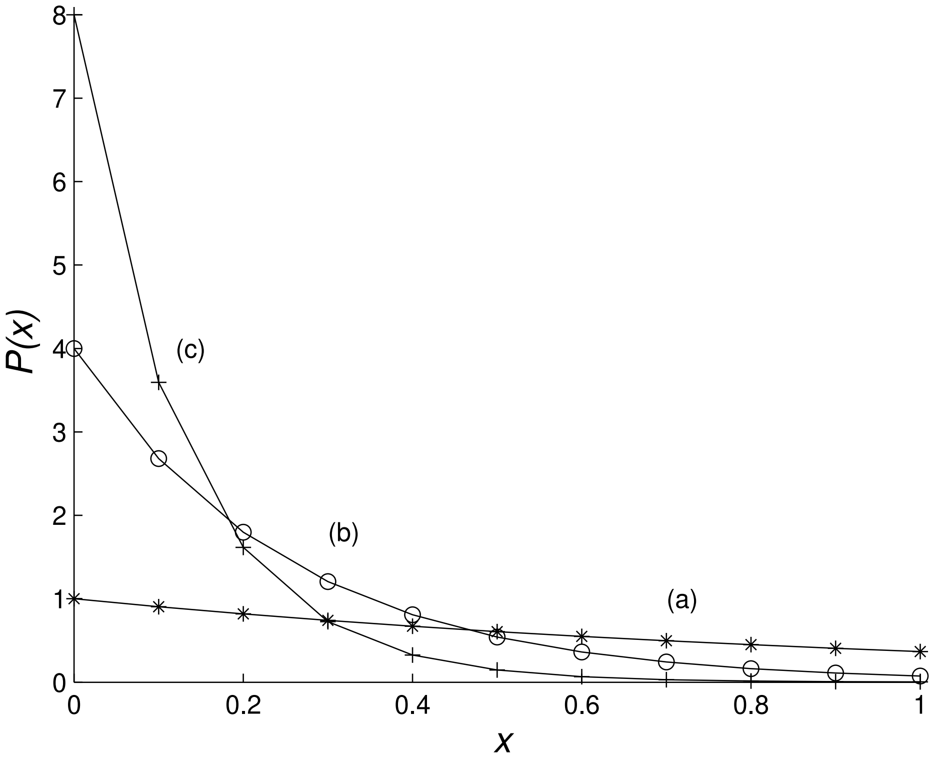
$$= -x\exp(-\lambda x)\Big|_0^\infty + \int_0^\infty \exp(-\lambda x)\,dx$$

$$= \int_0^\infty \exp(-\lambda x)\,dx = -\frac{1}{\lambda}\exp(-\lambda x)\Big|_0^\infty = \frac{1}{\lambda}$$

- The variance of the exponential distribution is

$$E\left\{x^2\right\} - \Big(E\left\{x\right\}\Big)^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

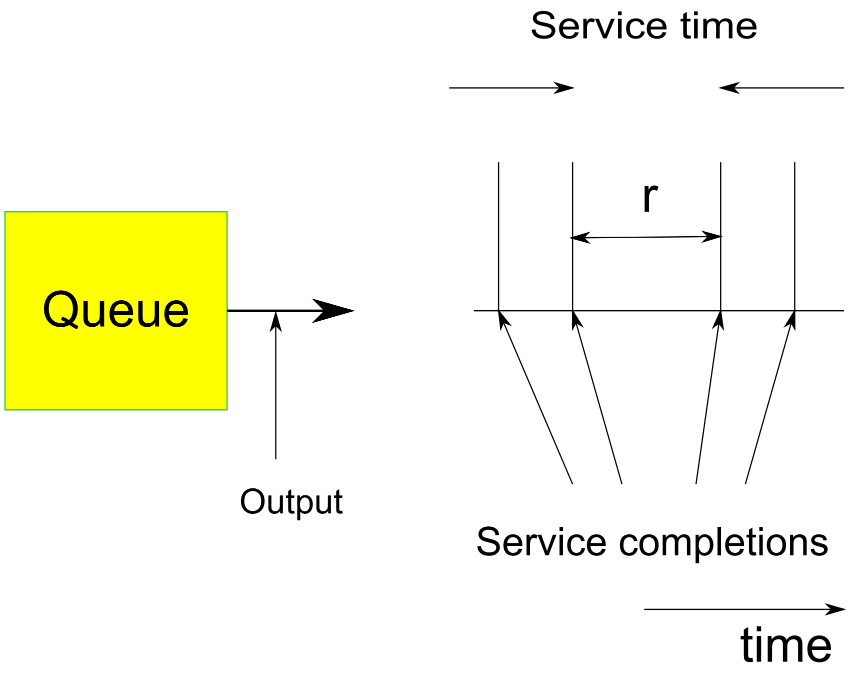
The average time between arrivals is expected because if the rate of arrivals is λ, then the average time between arrivals is $\frac{1}{\lambda}$.



The exponential distribution for (a) $\lambda = 1$, (b) $\lambda = 4$ and (c) $\lambda = 8$.

What is the connection between the Poisson arrival process and the exponential service time distribution?

Consider a queue in which customers (packets, calls, etc.) wait for service, and mark the times at which a customer completes service.



Service completions at the output of a queue

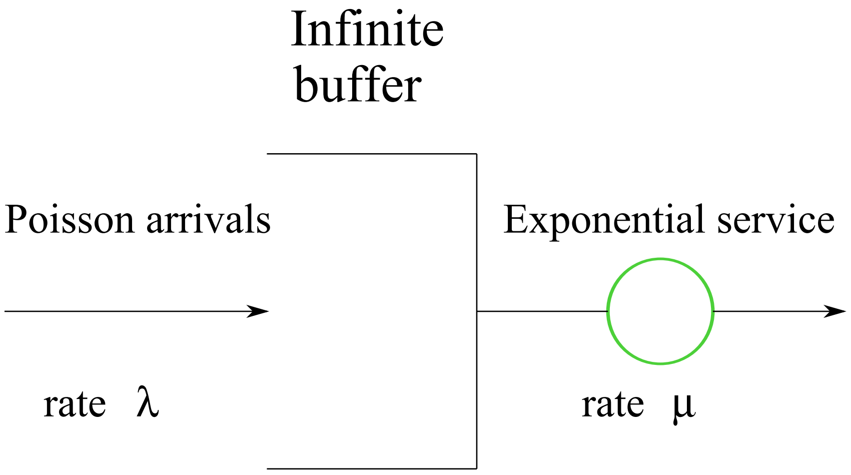
- Let the random variable that represents the time between completions be r. This must be the service time because a customer is served as soon as the service point is empty.

- Consider the situation in which r has an exponential distribution with mean $1/\mu$

$$P(r) = \mu \exp(-\mu r)$$

- The exponential distribution has been found to be useful for modelling the duration of telephone calls and the time to transmit packets of fixed length over a computer network.

- When the arrival distribution is not Poisson or the service time is not exponential, the history of the process must be considered and the process is therefore not memoryless. This makes the analysis significantly harder.



The M/M/1 queue

Arrivals: Poisson distribution

$$P(k|t,\lambda) = \frac{(\lambda t)^k}{k!} \exp(-\lambda t), \qquad \text{mean} = \lambda t, \qquad \text{rate} = \lambda$$

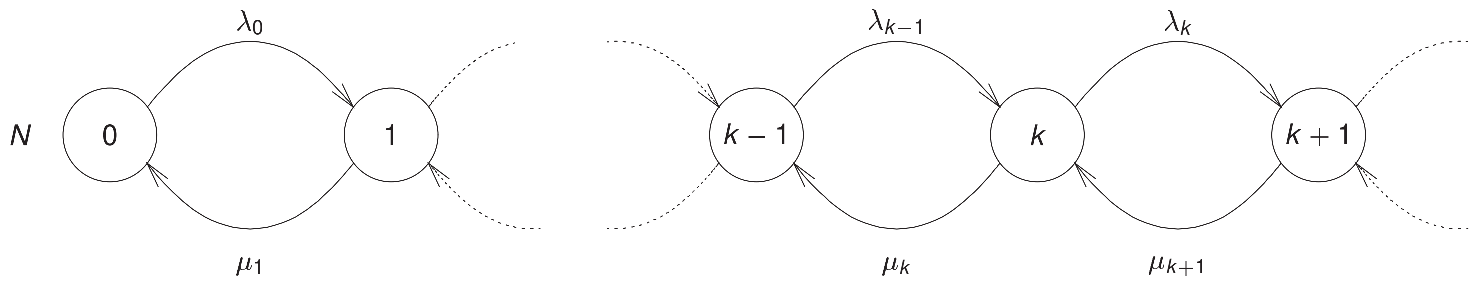
Service: Exponential distribution

$$P(r) = \mu \exp(-\mu r), \qquad \text{mean} = \frac{1}{\mu}, \qquad \text{rate} = \mu$$

- An M/M/1 queue is characterised by one queue and one server
- Customers, calls, packets, etc. arrive according to a Poisson distribution, and they are served according to an exponential distribution
- The M/M/1 queue is also called a *birth death process* because customers, calls, packets etc. arrive at the queue (a ‘birth’) and they are served (a ‘death’)
- A system is in **state** n if there are n people in the system, which includes people in the queue and people being served

If the probabilities at state n are known, then all other information on the queue can deduced.

- It is assumed that the system is operating at steady state, which implies the probabilities do not change with time:
  - Given an initial condition, for example, the number of people in the queue, the probabilities should approach steady-state conditions, and thus the arrival and service time distributions are invariant with respect to time



State diagram for an M/M/1 queue

- λ_k is the arrival rate of customers when the system is in state k
- μ_k is the service rate of customers when the system is in state k

Recall the assumptions of the Poisson distribution:

$$\begin{aligned} P(\text{exactly one arrival in } [t, t+\Delta t]) &= \lambda \Delta t \\ P(\text{exactly no arrivals in } [t, t+\Delta t]) &= 1 - \lambda \Delta t \\ P(\text{more than one arrival in } [t, t+\Delta t]) &= 0 \end{aligned}$$

Use the same model for the service rate:

$$\begin{aligned} P(\text{exactly one service completion in } [t, t+\Delta t]) &= \mu \Delta t \\ P(\text{exactly no service completions in } [t, t+\Delta t]) &= 1 - \mu \Delta t \\ P(\text{more than one service completion in } [t, t+\Delta t]) &= 0 \end{aligned}$$

Let

$$\begin{aligned} P_k(t) &= P(k \text{ arrivals in the time interval } t) \\ p_{i,j}(\Delta t) &= P(\text{going from } i \text{ arrivals to } j \text{ arrivals in a time interval of } \Delta t) \end{aligned}$$

$$\begin{aligned} P_k(t+\Delta t) &= P_k(t)p_{k,k}(\Delta t) + P_{k-1}(t)p_{k-1,k}(\Delta t) \\ &\quad +P_{k+1}(t)p_{k+1,k}(\Delta t) \\ P_0(t+\Delta t) &= P_0(t)p_{0,0}(\Delta t) + P_1(t)p_{1,0}(\Delta t) \quad k=0 \text{ (starting condition)} \end{aligned}$$

How does this equation satisfy the memoryless property?

- If the system is in state k at time $t+\Delta t$, then it must have been in state k, or state $k-1$, or state $k+1$ at time t

Notes:

- Only first order terms are used
- This equation states that k customers in the system at time $t+\Delta t$ can occur by having $k-1$ or k or $k+1$ customers in the system at time t
- This follows from the definition of a birth-death process

Use the equations on the previous slides:

$$\begin{aligned} P_k(t+\Delta t) &= P_k(t)(1-\lambda\Delta t)(1-\mu\Delta t) + P_{k-1}(t)(\lambda\Delta t)(1-\mu\Delta t) \\ &\quad + P_{k+1}(t)(\mu\Delta t)(1-\lambda\Delta t) \\ P_0(t+\Delta t) &= P_0(t)(1-\lambda\Delta t) + P_1(t)(\mu\Delta t)(1-\lambda\Delta t), \quad k=0 \end{aligned}$$

Multiply out these expressions and let $\Delta t \to 0$:

$$\begin{aligned} \frac{P_k(t+\Delta t)-P_k(t)}{\Delta t} \quad \frac{dP_k(t)}{dt} &= -(\lambda+\mu)P_k(t) + \lambda P_{k-1}(t) + \mu P_{k+1}(t), \quad k \geq 1 \\ \frac{P_0(t+\Delta t)-P_0(t)}{\Delta t} \quad \frac{dP_0(t)}{dt} &= -\lambda P_0(t) + \mu P_1(t) \end{aligned}$$

- These equations describe the evolution in time of the state probabilities of the M/M/1 queuing system
- But what do they represent physically?

Flows and balancing

The rates λ and μ are average values across all states, but the values of λ and μ depend on the state of the system:

$$\lambda = \lambda(P_0 + P_1 + P_2 + \cdots) \qquad \text{and} \qquad \mu = \mu(P_0 + P_1 + P_2 + \cdots)$$

In state k, the average arrival and service rates are

$$\lambda_k = \lambda P_k \qquad \text{and} \qquad \mu_k = \mu P_k$$

λ_k and μ_k are the arrival probability flux and service probability flux at state k.

- The probability flux at state k is defined as

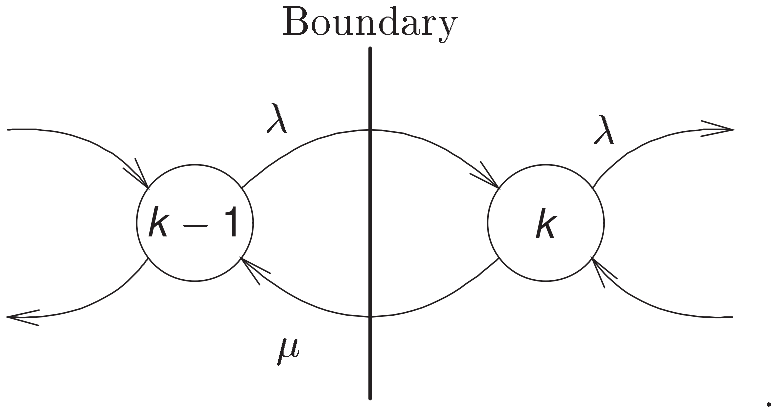
  probability flux = probability of system in state k × transition rate

  - The probability flux along a transition is the average number of times that the event that corresponds to the transition occurs

Example If a transition has a rate of 10 sec^{-1}, and the probability of the transition at which the state occurs is 0.1, then the transition is traversed by the system, on average, once every second. □

The equations in circuit theory are obtained from the principle that charge cannot build up at a point.

- Use the same principle for queuing theory by considering a boundary between states $k-1$ and k
- Compare the average number of times per second one enters a state with the average number of times per second one leaves a state



Flow rates at a boundary for an M/M/1 queue.

Consider the average number of times one enters a state, and the average number of times one leaves a state.

- The probability flux that leaves state k at time t is

$$(\lambda + \mu)P_k(t)$$

- The probability flux that enters state k at time t is

$$\lambda P_{k-1}(t) + \mu P_{k+1}(t)$$

The difference between these two probabilities is the change in the probability that the system is in state k:

$$\frac{dP_k(t)}{dt} = -(\lambda + \mu)P_k(t) + \lambda P_{k-1}(t) + \mu P_{k+1}(t), \quad k \geq 1$$

This is the same equation that was derived earlier

The equation for $\frac{dP_0(t)}{dt}$ can be derived in a similar manner.

- The M/M/1 system yields a set of differential equations that describe the evolution of the state probabilities as a function of time
- This is useful if a study of the transients of a queuing system is required. For example, if interest is restricted to the first 10 minutes after service commences from an empty queue, the initial conditions are

$$P_0(0) = 1 \qquad \text{and} \qquad P_k(0) = 0, \quad k \geq 1$$

What happens in the steady state, when all transients have decayed to zero? In this circumstance

$$\frac{dP_k(t)}{dt} = 0, \qquad k \geq 0$$

This yields

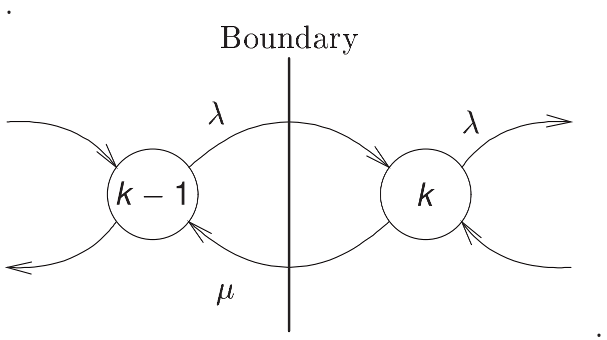
$$(\lambda + \mu)P_k - \lambda P_{k-1} - \mu P_{k+1} = 0, \quad k \geq 1$$

What happens at $k = 0$?

$$\lambda P_0 - \mu P_1 = 0$$

- This is a set of linear equations
- Steady state conditions are assumed and thus the argument t has been deleted

What is the solution of this set of equations?



Flow rates at a boundary for an M/M/1 queue.

Consider the steady state flux (flow rates) at the boundary between two states:

- Flux from state $k-1$ to state k is λP_{k-1}
- Flux from state k to state $k-1$ is μP_k

For flux to be balanced:

$$\lambda P_{k-1} = \mu P_k$$

It is checked that this equation satisfies

$$\begin{aligned}(\lambda+\mu)P_k-\lambda P_{k-1}-\mu P_{k+1} &= 0, \quad k\geq 1\\ \lambda P_0-\mu P_1 &= 0\end{aligned}$$

The solution

$$\lambda P_{k-1}=\mu P_k$$

leads to

$$P_k=\frac{\lambda}{\mu}P_{k-1}$$

and thus

$$P_k=\left(\frac{\lambda}{\mu}\right)^k P_0=\rho^k P_0$$

where $\rho=\frac{\lambda}{\mu}$ is called the *utilisation factor* for the $M/M/1$ queue.

- Use the normalisation condition to calculate P_0:

$$\sum_{k=0}^{\infty} P_k = 1$$

and it follows from $P_k = \rho^k P_0$ that

$$P_0 \sum_{k=0}^{\infty} \rho^k = 1$$

which yields

$$P_0 = \frac{1}{\sum_{k=0}^{\infty} \rho^k}$$

- If $\rho < 1$, that is, $\lambda < \mu$,

$$\sum_{k=0}^{\infty} \rho^k = \frac{1}{1-\rho} \quad \text{for } \rho < 1$$

Thus

$$P_0 = (1 - \rho)$$

and the distribution for the *steady state* probabilities is

$$P_k = \rho^k (1 - \rho), \qquad \rho < 1$$

Faster arrivals than service

Notice that if the arrival rate (birth rate) is greater than the service rate (death rate), then $\rho > 1$ and the series

$$\sum_{k=0}^{\infty} \rho^k$$

does not converge. This is indicative of the fact that if the birth rate is greater than the death rate, a steady state distribution does not exist and the queue continues to grow in size.

The probability distribution

$$P_k = \rho^k (1 - \rho), \qquad \rho < 1$$

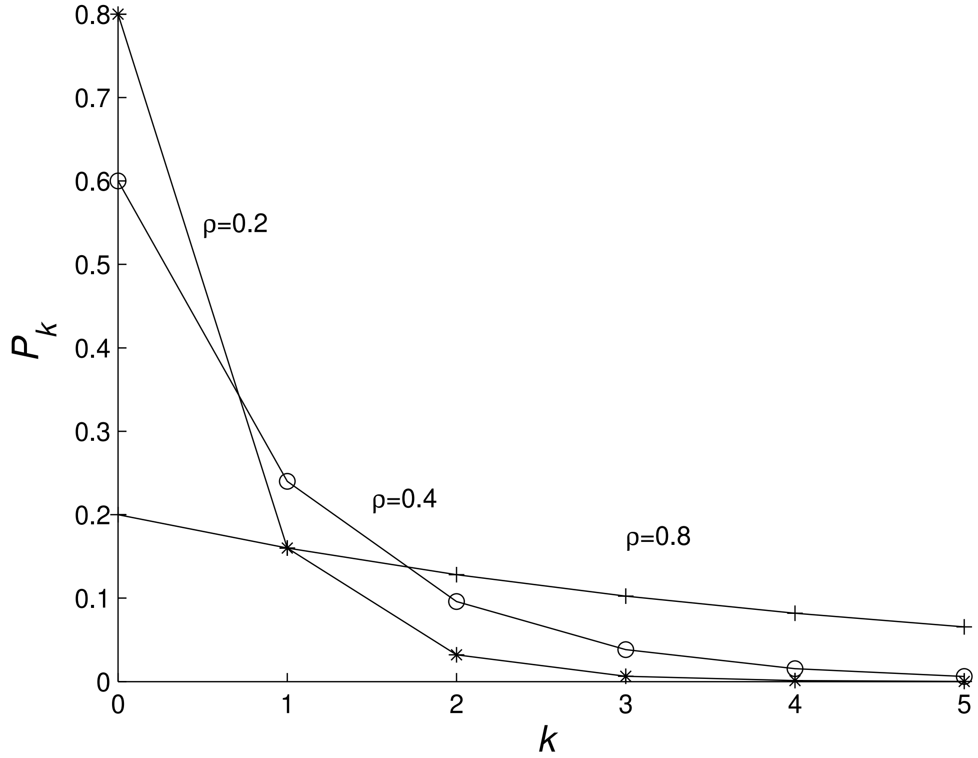
is called a geometric distribution.

- The probability that the system is empty is $P_0 = 1 - \rho$, and thus the probability that the system is not empty is ρ, which explains why this parameter is called the utilisation

Typical questions that are of interest:

- What is the average number of people in the system?
- What is the average delay per customer?
- What is the average waiting time in the queue?

The second and third points require Little's theorem.



The state probabilities for an M/M/1 queue for three values of ρ.

The average number of people in the system

This is the expected (average) state of the system:

$$E\{k\} = \sum_{k=0}^{\infty} kP_k = (1-\rho)\sum_{k=0}^{\infty} k\rho^k = \rho(1-\rho)\sum_{k=0}^{\infty} k\rho^{k-1}$$

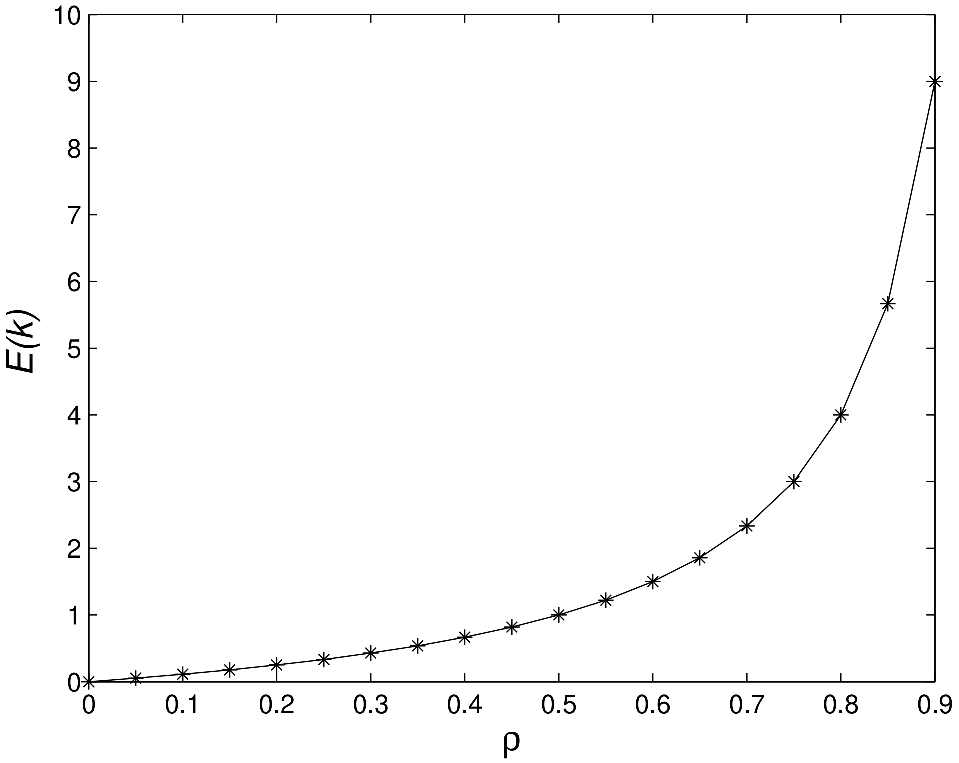
The right hand side can be written as

$$\rho(1-\rho)\frac{\partial \sum_{k=0}^{\infty}\rho^k}{\partial\rho} = \rho(1-\rho)\frac{\partial}{\partial\rho}\left[\frac{1}{1-\rho}\right] = \rho(1-\rho)\frac{1}{(1-\rho)^2}$$

and thus

$$E\{k\} = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu-\lambda}$$

- $E\{k\} \to \infty$ as $\rho \to 1$
- The server cannot keep up with the arrival rate if $\rho > 1$, and the length of the queue increases without bound



The average state of an M/M/1 queue.

The summation on the previous slide can be calculated using another method.

$$S = \sum_{k=0}^{\infty} k\rho^k = \rho + 2\rho^2 + 3\rho^3 + 4\rho^4 \cdots$$

and thus

$$\rho S = \rho^2 + 2\rho^3 + 3\rho^4 + 4\rho^5 + \cdots$$

Subtract the second equation from the first equation

$$S(1-\rho) = \rho + \rho^2 + \rho^3 + \rho^4 + \cdots = \frac{\rho}{1-\rho}$$

and thus

$$S = \frac{\rho}{(1-\rho)^2}$$

Little's theorem

Little's formula applies under steady state conditions.

$$\bar{N} = \lambda \bar{T}$$

$\bar{N}$: The average number of people in the system

λ : The steady state arrival rate

$\bar{T}$: The average total time spent by a customer in the system

Example The queue at a nightclub on a Thursday has, on average, 20 people and people arrive at a rate of 40 per hour. According to Little's formula, how long would one expect to wait to enter?

The waiting time is given by Little's formula:

$\bar{N} = 20$, $\lambda = 40$ and $\bar{T} = \frac{\bar{N}}{\lambda} = \frac{20}{40} = \frac{1}{2}$hour. $\square$

Use Little's law to calculate the average delay in the system (the waiting time in the queue and the service time).

$$\bar{T} = \frac{\bar{N}}{\lambda} = \frac{E\{k\}}{\lambda} = \frac{\rho}{\lambda(1-\rho)} = \frac{1}{\mu - \lambda}$$

The service rate is μ, and thus the average service time (the average time taken for a customer to be served) is $\frac{1}{\mu}$. The average time spent in the queue is therefore

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

Use Little's formula again to calculate the average number of customers in the queue

$$\bar{N}_Q = \lambda W = \frac{\rho^2}{1-\rho}$$

Note the difference in the two applications of Little's law:

- The equation $\bar{N} = \lambda \bar{T}$ is applied to the entire system (the queuing system and the serving system), where λ is the arrival rate in the queuing system
- The equation

$$\bar{N}_Q = \lambda W = \frac{\rho^2}{1 - \rho}$$

is applied to the queuing system only

Thus

- Little's theorem applies to the total system (the queuing and service processes) and the queuing process only, and λ is the arrival rate, which is defined by the queuing system, for both applications

But

- Does Little's theorem apply to service process only?

Consider the service process.

- The steady state number of people being served is

$$N_S = \bar{N} - N_Q = \frac{\lambda}{\mu - \lambda} - \frac{\rho^2}{1-\rho} = \rho$$

  because $\rho = \lambda/\mu$

- Consider the product λT_S in Little's formula, where T_S is the average time taken for a customer to be served.

$$T_S = \frac{1}{\mu} \qquad \text{and thus} \qquad \lambda T_S = \frac{\lambda}{\mu} = \rho = N_S \quad \text{from above}$$

Summary:

- Little's theorem applies to the total system (the queuing and service processes), the queuing process only, and the serving system only

4. The birth-death process

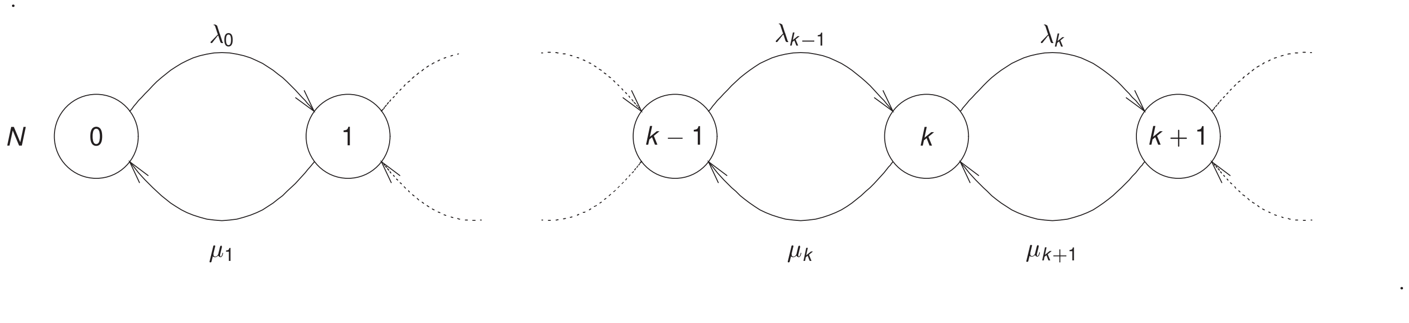
The analysis of the M/M/1 queuing system is easily extended to single queue and single server systems in which the arrival and departure service rates are dependent upon the state of the system.

- These processes are called birth-death processes

Examples of birth-death processes:

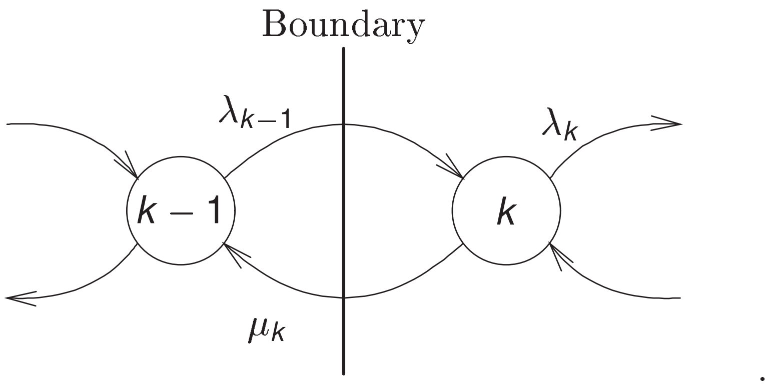
- A queuing system of the form M/M/m, that is, the queue and serving processes are defined by Markovian statistics, there is one queue and there are m servers
- A system in which the arrival queue cannot grow indefinitely, but is limited to a finite number of people (packets, calls, etc.)

- Extend the analysis of the M/M/1 system to the situation that occurs when $\lambda = \lambda_k$ and $\mu = \mu_k$:
  - The probability of exactly one arrival in the time interval Δt is $\lambda_k \Delta t$
  - The probability of no arrivals in the time interval Δt is $(1 - \lambda_k \Delta t)$
  - The probability of exactly one departure in the time interval Δt is $\mu_k \Delta t$
  - The probability of no departures in the time interval Δt is $(1 - \mu_k \Delta t)$



Flow rates at a boundary for a birth-death process

Obtain the steady-state conditions in the same way as for an M/M/1 process.



Flow rates at a boundary for a birth-death process

$$(\lambda_k + \mu_k)P_k = \lambda_{k-1}P_{k-1} + \mu_{k+1}P_{k+1} \qquad k \geq 1$$

λ_k : The rate of arrival of the customers, given that the system is in state k

P_k : The probability, at steady state, that the system is in state k

μ_k : The rate of departure of the customers, given that the system is in state k

The solution of the equation is

$$\lambda_k P_k = \mu_{k+1} P_{k+1}$$

and thus

$$\begin{aligned}
P_1 &= \left(\frac{\lambda_0}{\mu_1}\right) P_0 \\
P_2 &= \frac{\lambda_1}{\mu_2} P_1 = \left(\frac{\lambda_0 \lambda_1}{\mu_1 \mu_2}\right) P_0 \\
P_3 &= \frac{\lambda_2}{\mu_3} P_2 = \left(\frac{\lambda_0 \lambda_1 \lambda_2}{\mu_1 \mu_2 \mu_3}\right) P_0
\end{aligned}$$

and the general solution is

$$P_k = \left(\frac{\prod_{i=0}^{k-1} \lambda_i}{\prod_{i=1}^{k} \mu_i}\right) P_0, \qquad k \geq 1$$

The constant P_0 is calculated from the normalisation condition

$$\sum_{k=0}^{N} P_k = P_0 + P_0 \sum_{k=1}^{N} \left(\frac{\prod_{i=0}^{k-1} \lambda_i}{\prod_{i=1}^{k} \mu_i} \right) = 1$$

- The existence of a steady state solution depends only on the convergence of the sum:
  - The sum always exists if N is finite
  - If the sum converges as $N \to \infty$, then P_0 exists and thus a steady state solution exists
  - If the sum diverges as $N \to \infty$, then a steady state solution does not exist

Question Why are the the equations

$$P_k = \left(\frac{\prod_{i=0}^{k-1} \lambda_i}{\prod_{i=1}^{k} \mu_i} \right) P_0 \quad \text{and} \quad \sum_{k=0}^{N} P_k = P_0 \left[1 + \sum_{k=1}^{N} \left(\frac{\prod_{i=0}^{k-1} \lambda_i}{\prod_{i=1}^{k} \mu_i} \right) \right] = 1$$

important?

Answer Many performance measures are functions of these state probabilities.

- The average rate at which people leave the system is

$$\bar{Y} = \sum_{k=1}^{\infty} \mu_k P_k$$

  Note the summation starts at $k = 1$ because the system is empty at $k = 0$ and thus the throughput is zero. The throughput is a weighted average of the service rates, where the state probabilities serve as the weights.

- The average number of customers in the system is

$$\bar{N} = \sum_{k=1}^{\infty} k P_k$$

Both these measures are weighted averages.

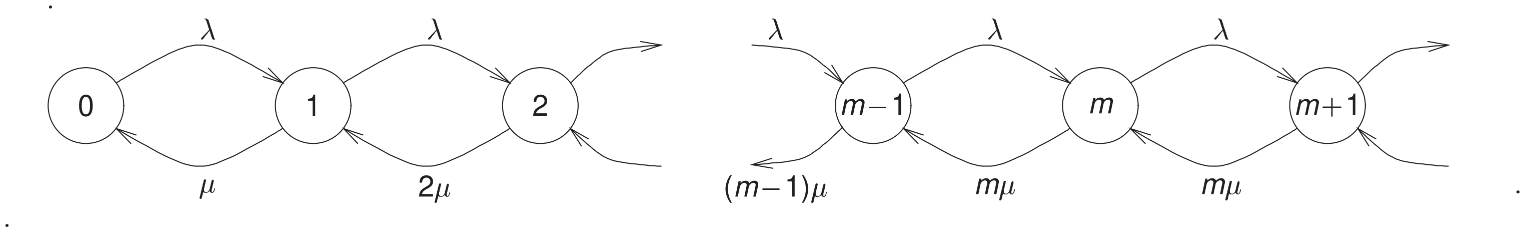
5. The M/M/m queuing system

We have considered the following:

- The M/M/1 queue - one queue, without limit on the number of customers in the queue, and one server, such that the steady state arrival and service rates are constant and independent of the state
- The birth-death process, which is an extension of the M/M/1 process in which the steady state arrival and service rates are dependent upon the state of the system

Many practical situations involve several servers, for example, a call centre, and a switch that has one incoming port and two or more outgoing ports, such that the outgoing port chosen by a data packet is random.

- The M/M/m system has one queue and m servers. This system is also the Erlang delay system.



The M/M/m state transition diagram.

Let λ and μ be the arrival and service rates for an M/M/1 system.

- Since there is only one queue, the steady state arrival rate is λ for all states k
- If the number of people in the system is less than or equal to m, then the queue is empty and the service rate for state k is $\mu_k = k\mu$ because all the people in the system are being served
- If there are more than m people in the system, then there must be at least one person in the queue

The state dependent arrival and service rates are

$$\begin{aligned}
\lambda_k &= \lambda, \qquad k = 0, 1, 2, \ldots \\
\mu_k &= \begin{cases} k\mu & k = 0, 1, 2, \ldots, m-1, m \\ m\mu & k \geq m \end{cases}
\end{aligned}$$

The steady state balance equations are obtained by considering the flux entering state k from state $k-1$, and the flux entering state $k-1$ from state k:

$$\begin{aligned}
\lambda P_{k-1} &= k\mu P_k, \qquad k \leq m \\
\lambda P_{k-1} &= m\mu P_k, \qquad k \geq m
\end{aligned}$$

To solve these equations, consider initially the condition $k \leq m$. Then

$$P_k = \frac{1}{k}(\rho m) P_{k-1}, \qquad \frac{\lambda}{\mu} = \rho m$$

yields

$$\begin{aligned} P_1 &= (\rho m)\, P_0 \\ P_2 &= \frac{(\rho m)}{2} P_1 = \frac{(\rho m)^2}{2!} P_0 \\ P_3 &= \frac{(\rho m)}{3} P_2 = \frac{(\rho m)^3}{3!} P_0 \end{aligned}$$

and in general

$$P_k = \left(\frac{(m\rho)^k}{k!} \right) P_0, \qquad k < m$$

Consider now the condition $k \geq m$.

$$P_k = \frac{\lambda}{m\mu} P_{k-1} = \rho P_{k-1}$$

yields

$$\begin{aligned} P_m &= \rho P_{m-1} \\ P_{m+1} &= \rho P_m = \rho^2 P_{m-1} \\ P_{m+2} &= \rho P_{m+1} = \rho^3 P_{m-1} \end{aligned}$$

and in general

$$P_k = \rho^{k-m+1} P_{m-1} \qquad k = m, m+1, \ldots$$

But from the previous slide

$$P_{m-1} = \left(\frac{(m\rho)^{m-1}}{(m-1)!} \right) P_0$$

and thus

$$P_k = \left(\frac{m^m \rho^k}{m!} \right) P_0, \qquad k = m, m+1, \ldots$$

Thus the steady state probabilities of the M/M/m queue are

$$P_k = \begin{cases} P_0 \left(\frac{(m\rho)^k}{k!} \right) & k \leq m \\ P_0 \left(\frac{m^m \rho^k}{m!} \right) & k \geq m \end{cases}$$

where

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

How is the initial state probability calculated? Use

$$\sum_{k=0}^{\infty} P_k = 1$$

Thus

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

Since

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

for $\rho < 1$, it follows that the steady state probabilities for an M/M/m queue are

$$P_k = \begin{cases} P_0 \left(\frac{(m\rho)^k}{k!} \right) & k \leq m \\ P_0 \left(\frac{m^m \rho^k}{m!} \right) & k \geq m \end{cases} , \qquad \rho = \frac{\lambda}{m\mu} < 1$$

where

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

An important parameter in M/M/m systems is the probability of queuing.

Consider a telephone exchange in which there are m outgoing trunk lines. The probability that a customer will wait is equal to the probability that all the servers are busy

$$P_{k\geq m} = \sum_{k=m}^{\infty} P_k = P_0 \sum_{k=m}^{\infty} \frac{m^m \rho^k}{m!} = P_0 \left(\frac{m^m \rho^m}{m!(1-\rho)} \right)$$

- This equation is important in telephone trunk design for queued systems in which calls are held rather than blocked
- It has been tabulated and it is called *Erlang C formula* or Erlang's formula of the second kind
- The probability that one or more calls are in the queue is

$$P_{k\geq m+1} = \sum_{k=m+1}^{\infty} P_k = \sum_{k=m}^{\infty} P_k - P_m = P_0 \left(\frac{m^m \rho^{m+1}}{m!(1-\rho)} \right)$$

Define the offered load a and the function $C(m, a)$ as

$$a = \frac{\lambda}{\mu} = m\rho \qquad \text{and} \qquad C(m, a) = P_{k \geq m} = P_0 \left(\frac{a^m}{m! \left(1 - \frac{a}{m}\right)} \right)$$

- Care must be exercised when $C(m, a)$ is computed for large values of m because overflow may occur for the terms a^m and $m!$
- Evaluate $C(m, a)$ by computing its logarithm

  $$\ln C(m, a) = \ln P_0 + m \ln a - \ln m! - \ln \left(1 - \frac{a}{m}\right)$$

  but care must still be exercised with the function $\ln P_0$
- The Erlang C curve is a graph of $C(m, a)$, the probability of waiting, against a, the offered load, for different values of m

- The units of $a = \frac{\lambda}{\mu}$ are *Erlangs*, and

$$a = \frac{\text{arrival rate}}{\text{service rate for each customer}}$$

- The parameter ρ is given by

$$\rho = \frac{\lambda}{m\mu} = \frac{a}{m} = \frac{\lambda}{m\mu} = \frac{1}{\text{no. servers}} \times \frac{\text{arrival rate}}{\text{service rate for each customer}}$$

and it is the utilisation per server, trunk line etc.

Example A system that handles $\lambda = 2$ messages/second, where each message lasts, on average, 3 seconds has an intensity of 6 Erlangs because $\mu = \frac{1}{3}$. □

Example A telephone exchange A is to serve $10,000$ subscribers. Assume the subscribers generate calls that follow a Poisson distribution with a rate of 10 calls per minute, and that the calls last, on average, 3 minutes. How many trunk lines are required in order that the probability of waiting is 0.01?

Since $\lambda = 10$ and $\mu = \frac{1}{3}$, it follows that $a = \lambda/\mu = 30$ Erlangs. The Erlang C curve shows that $m = 45$. □

Example The average length of a call to a customer service centre is 8 minutes. Calls arrive at the centre at a rate of one every 2 minutes during the busy period. If the inter-arrival and service times have an exponential distribution, how many personnel are required in order that the probability of waiting be 0.1?

The average call rate is $\lambda = 0.5$ calls/minute, and $\mu = 1/8$ customers per minute. It follows that $a = \lambda/\mu = 4$, and the Erlang C curve shows that 8 personnel are required. □

- The probability of delay $C(m, a)$ is important because many parameters of interest can be expressed in terms of it

Example Consider the average number of people waiting for service, which is the difference between the average number of people in the system and the average number of people in service.

The average number of customers in the queue (not in service) is

$$\begin{aligned}\bar{N}_Q &= \sum_{k=m+1}^{\infty}(k-m)P_k = P_0\sum_{k=m+1}^{\infty}(k-m)\frac{m^m\rho^k}{m!}\\ &= P_0\left(\frac{m^m\rho^m}{m!}\right)\sum_{k=1}^{\infty}k\rho^k\end{aligned}$$

Since

$$\sum_{k=1}^{\infty} k\rho^k = \frac{\rho}{(1-\rho)^2}$$

it follows that the total number of people in the queue is

$$\bar{N}_Q = C(m,a)\frac{\rho}{1-\rho}$$

where $C(m,a)$ is the probability of queueing.

When $m = 1$, this formula reduces to the formula for the M/M/1 queue because $C(m,a) = \rho$ for this queue.

Use Little's formula to calculate the average time spent by a customer in the queue:

$$\bar{W}_Q = \frac{\bar{N}_Q}{\lambda} = \frac{\rho C(m,a)}{\lambda(1-\rho)} = \frac{C(m,a)}{m\mu(1-\rho)}$$

Since the service rate for each of the m servers is μ, it follows that the total average delay for each customer is

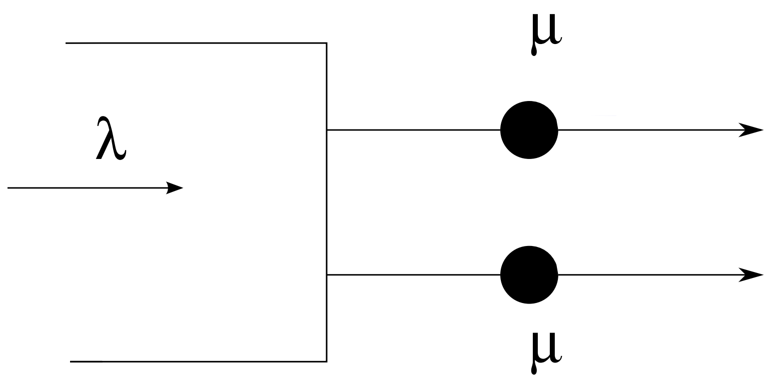
$$\bar{T} = \frac{1}{\mu} + \bar{W}_Q = \frac{1}{\mu} + \frac{\rho C(m,a)}{\lambda(1-\rho)} = \frac{1}{\mu} + \frac{C(m,a)}{m\mu - \lambda}$$

The average number of customers in the system is therefore

$$\bar{N} = \lambda \bar{T} = \frac{\lambda}{\mu} + \frac{\rho C(m,a)}{1-\rho}$$

Example Consider two outgoing transmission links that connect a packet switch in a packet-switched network. Each data packet uses one of the outgoing links, chosen at random. What effect does the addition of the second link have on the operation of the system?

- Assume that the packets arrive at the input of the double transmission link with a Poisson distribution with an average rate of λ packets/sec.
- Assume that the service rate for each transmission link is μ packets/second.



This is an M/M/2 system, and thus

$$P_0 = \left[\sum_{k=0}^{1} \frac{(2\rho)^k}{k!} + \frac{(2\rho)^2}{2(1-\rho)} \right]^{-1} = \frac{1-\rho}{1+\rho}, \qquad \rho = \frac{\lambda}{2\mu}$$

The probability that there are $k \geq 1$ packets in the system is

$$P_k = P_0 \left(2\rho^k\right) = \frac{2(1-\rho)\rho^k}{1+\rho}, \qquad k \geq 1$$

The average number of packets in the system is therefore

$$\sum_{k=0}^{\infty} kP_k = \frac{2(1-\rho)}{1+\rho} \sum_{k=0}^{\infty} k\rho^k = \frac{2\rho}{1-\rho^2}$$

Recall

$$\sum_{k=0}^{\infty} k\rho^k = \frac{\rho}{(1-\rho)^2}$$

If the system were M/M/1 with $(\lambda, 2\mu)$, the average number of packets in the system would be

$$\frac{\rho}{1-\rho}, \qquad \rho = \frac{\lambda}{2\mu}$$

The average number of packets in an M/M/2 system is greater than the average number of packets in an M/M/1 system:

$$\frac{2\rho}{1-\rho^2} > \frac{\rho}{1-\rho}$$

Since

$$(\rho-1)^2 > 0 \Rightarrow \rho^2 - 2\rho + 1 > 0 \Rightarrow 2 - 2\rho > 1 - \rho^2$$

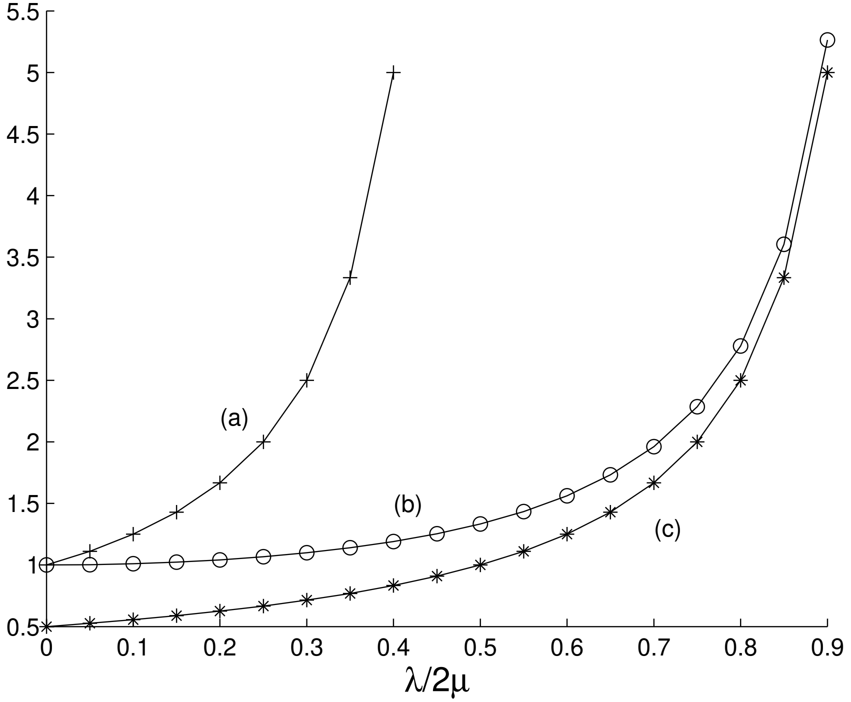
it follows that

$$\frac{2\rho}{1-\rho^2} > \frac{\rho}{1-\rho}$$

From Little's theorem, the average total delay in the system (queuing time and service time) is

$$\left(\frac{1}{\lambda}\right)\frac{2\rho}{1-\rho^2}=\frac{1}{\mu(1-\rho^2)}$$

- The additional transmission link reduces the total time delay through the system.



The product ($\mu \times$ average total delay) for (a) an M/M/1 queue, (b) an M/M/2 queue and (c) an M/M/1 queue with a service rate of 2μ, against $\lambda/2\mu$.

□

Example The average length of a call to a customer service centre is 8 minutes. Calls arrive at the centre at a rate of one every 2 minutes during the busy period. If the inter-arrival and service times have an exponential distribution, how many personnel are required in order that the probability of waiting be 0.1?

The service rate is $\mu = 1/8$. It was shown in a previous example that 8 personnel are required, and thus $\rho = \lambda/m\mu = 0.5$. The average time spent by a customer in the queue is

$$\bar{W}_Q = \frac{C(m,a)}{m\mu(1-\rho)} = \frac{0.1}{8 \times \frac{1}{8}(1-\frac{1}{2})} = 0.2 \text{ minutes } = 12 \text{ seconds}$$

The average total time $\bar{W}_Q$ spent in the queue by *all* customers (people who need to queue and people who do not need to queue) can be written as

$$\bar{W}_Q = \bar{W}_1 P_Q + (1 - P_Q)\bar{\bar{W}}_1$$

which is a weighted average, where $P_Q = C(m, a)$ is the probability of queuing. $\bar{W}_1$ is the average waiting for those people who need to queue, and $\bar{\bar{W}}_1$ is the average waiting time for people who do not need to queue. But $\bar{\bar{W}}_1 = 0$, by definition.

The average queueing time for the people who need to queue is therefore

$$\frac{\bar{W}_Q}{C(m, a)} = \frac{0.2}{0.1} = 2 \text{ minutes}$$

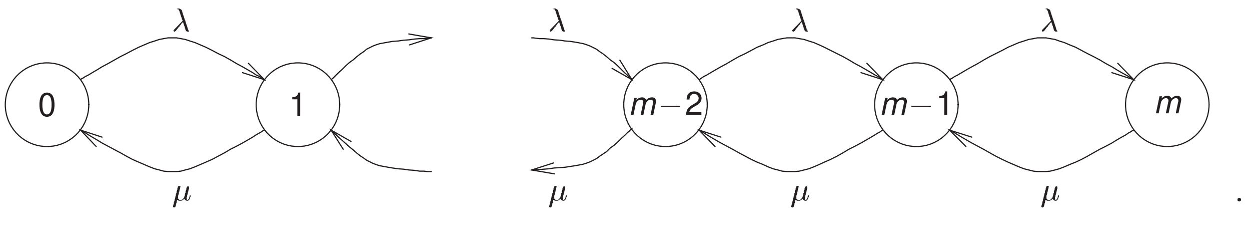
□

6. The M/M/1/m queuing system

It has been assumed thus far that the queue can be unbounded, but this may not always occur in practice because a buffer may be of restricted size.

- If there are m customers in the system, a new customer is turned away or blocked

Assume the system can contain a maximum of m customers, that is, the queue contains a maximum of $m-1$ customers.

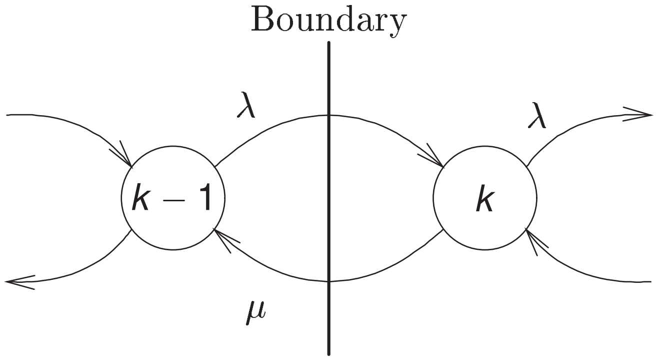


State diagram for the $M/M/1/m$ queue. Note that there are no transitions to states above the mth state.

The state dependent arrival and service rates are

$$\begin{aligned}\lambda_k &= \lambda, \qquad k = 0, 1, \ldots, m-1 \\ \mu_k &= \mu, \qquad k = 1, 2, \ldots, m\end{aligned}$$

.



.

Flow rates at a boundary for the M/M/1/m queue

The steady state balance equations are obtained by considering the flux entering state k from state $k-1$, and the flux entering state $k-1$ from state k:

$$\lambda_{k-1}P_{k-1} + \mu_{k+1}P_{k+1} = \mu_k P_k + \lambda_k P_k, \qquad k = 1, \ldots, m-1$$

The solution of this equation is

$$\lambda_k P_k = \mu_{k+1} P_{k+1}, \qquad k = 0, \ldots, m-1$$

This is the same equation as for the M/M/1 queue, but for finite k. Its solution is

$$\begin{aligned}
P_1 &= \left(\frac{\lambda_0}{\mu_1}\right) P_0 \\
P_2 &= \left(\frac{\lambda_1}{\mu_2}\right) P_1 = \left(\frac{\lambda_0\lambda_1}{\mu_1\mu_2}\right) P_0 \\
P_3 &= \left(\frac{\lambda_2}{\mu_3}\right) P_2 = \left(\frac{\lambda_0\lambda_1\lambda_2}{\mu_1\mu_2\mu_3}\right) P_0
\end{aligned}$$

and in general

$$P_k = \left(\frac{\prod_{i=0}^{k-1} \lambda_i}{\prod_{i=1}^{k} \mu_i}\right) P_0 = \left(\frac{\lambda}{\mu}\right)^k P_0, \qquad k = 0, \ldots, m$$

The normalisation condition

$$\sum_{i=0}^{m} P_i = \left(1 + \rho + \rho^2 + \cdots + \rho^m\right) P_0 = 1, \qquad \rho = \frac{\lambda}{\mu}$$

yields

$$P_0 = \frac{1-\rho}{1-\rho^{m+1}}$$

Compare with the expression for an M/M/1 queue

$$P_0 = \frac{1-\rho}{1-\rho^{m+1}} \approx 1-\rho \qquad \text{if} \quad \rho \ll 1$$

- This is expected because if $\mu \gg \lambda$, the queue cannot grow and will remain small, thus minimising the probability of blocking

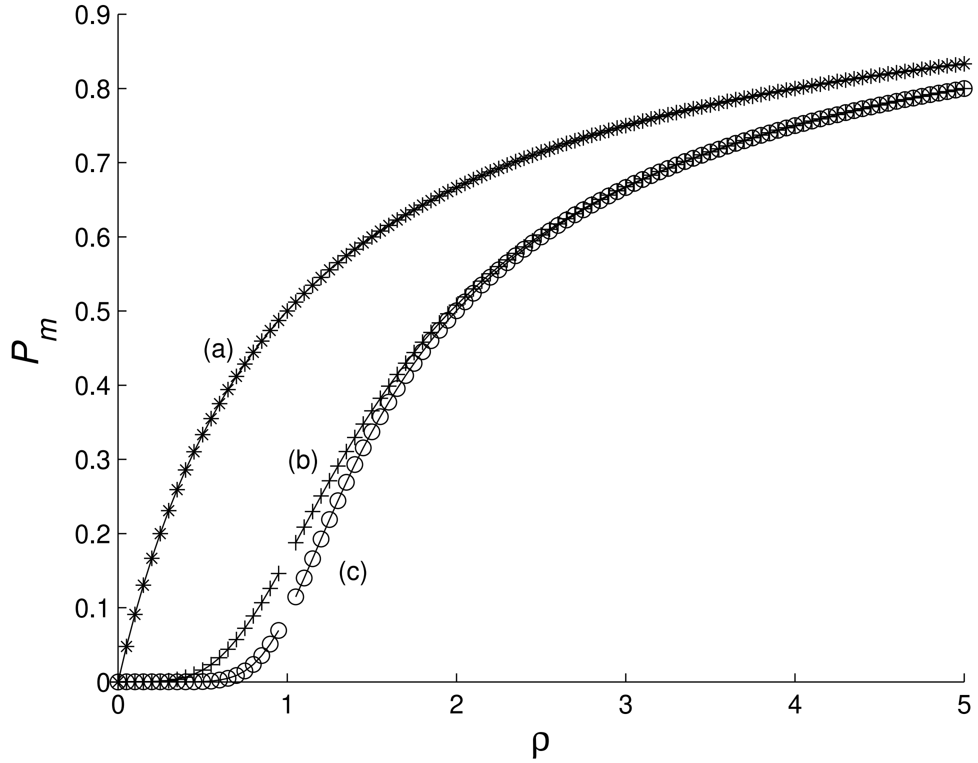
Thus

$$P_k = \frac{(1-\rho)\rho^k}{1-\rho^{m+1}} \qquad k = 0, \ldots, m$$

Example The blocking probability P_m is the probability that the system is full. The tables below show the values of P_m for (a) $m = 5$ as a function of ρ and (b) $\rho = 0.9$ for various values of m.

$\rho = \frac{\lambda}{\mu}$	$P_m = P_5$
0.10	9×10^{-6}
0.50	0.016
0.75	0.072
1.00	0.166
2.00	0.508
5.00	0.800

m	P_m for $\rho = 0.9$
1	0.474
2	0.299
3	0.212
5	0.126
10	0.051
20	0.014
100	2.7×10^{-6}



The blocking probability P_m for (a) $m = 1$, (b) $m = 5$ and (c) $m = 10$.

- The blocking probability increases as ρ increases
- What is the blocking probability as $\rho \to 1$? Let $\rho = 1 - \epsilon$ and consider $\epsilon \to 0$.

$$\begin{aligned}
\lim_{\rho \to 1} P_m &= \lim_{\rho \to 1} \frac{(1-\rho)\rho^m}{1-\rho^{m+1}} \\
&= \lim_{\epsilon \to 0} \frac{\epsilon(1-\epsilon)^m}{1-(1-\epsilon)^{m+1}} \\
&= \lim_{\epsilon \to 0} \frac{\epsilon(1-\epsilon)^m}{(m+1)\epsilon} \\
&= \frac{1}{m+1}
\end{aligned}$$

- $\rho \leq 1$ for an M/M/1 queue, but ρ may be greater than one for an M/M/1/m queue because excess customers are turned away

□

The example shows that the analysis of a simple queue with blocking is tractable.

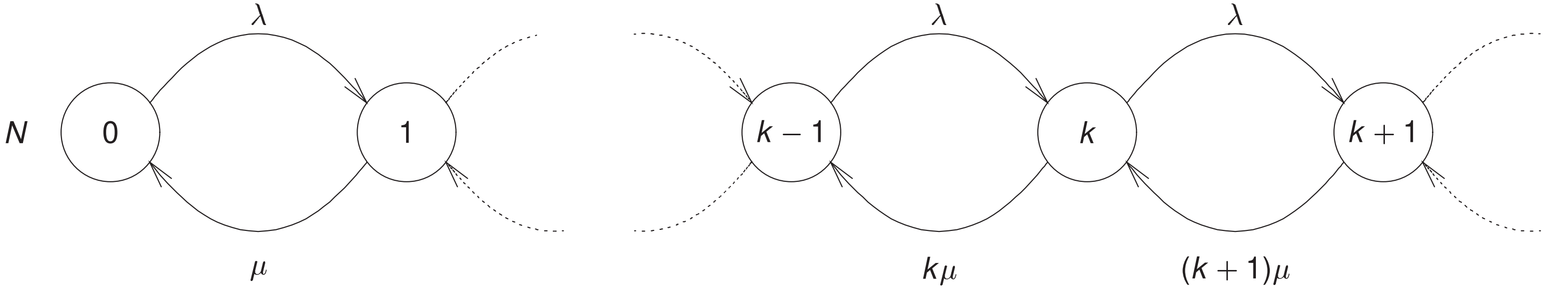
- Is this also true for networks in which queues form?

Consider the following:

- The presence of blocking introduces dependencies between the queuing systems that are not easily considered:
  - When one queuing system is full, departures from the systems that feed it are blocked
- Analytic solutions for networks of queuing systems that are defined by Markov chains exist if there is ample buffer space, such that the probability of blocking is small and can be neglected

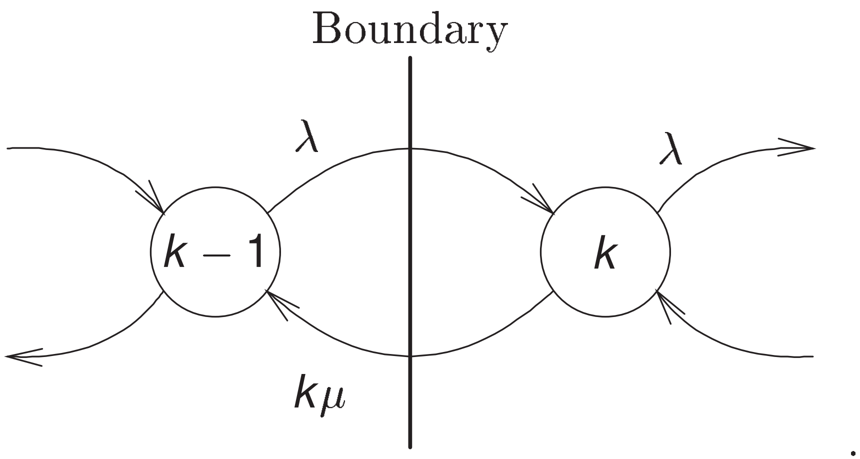
7. The M/M/∞ queuing system

- The M/M/1/m queue covers the case in which the queue length is bounded
- The case when the number of servers is *unbounded* is important because:
  - It gives bounds on the performance as the number of servers increases
  - It can be used to model the effect of delay in systems
- There is one server for every customer in the system



State diagram for the $M/M/\infty$ queue.

Make use of the standard detailed balance equations for this queue



Flow rates at a boundary for the M/M/∞ queue

- There is infinite service capacity in the queue and thus the service rate at state k is $\mu_k = k\mu, k = 1, 2, \ldots$
- The rate of arrivals stays constant, $\lambda_k = \lambda, k = 0, 1, \ldots$

The flow rates at the boundary yield

$$\lambda_{k-1} P_{k-1} = \mu_k P_k$$

and since $\lambda_k = \lambda$ and $\mu_k = k\mu$, this equation simplifies to

$$\lambda P_{k-1} = k\mu P_k$$

Thus

$$P_k = \frac{\lambda}{k\mu} P_{k-1}$$

and therefore

$$P_k = \frac{1}{k!} \left(\frac{\lambda}{\mu} \right)^k P_0$$

The normalisation condition for this queue is

$$\sum_{k=0}^{\infty} P_k = 1$$

and thus

$$P_0 = \left[\sum_{k=0}^{\infty} \left(\frac{\lambda}{\mu} \right)^k \frac{1}{k!} \right]^{-1} = \exp\left(-\frac{\lambda}{\mu} \right)$$

It follows that

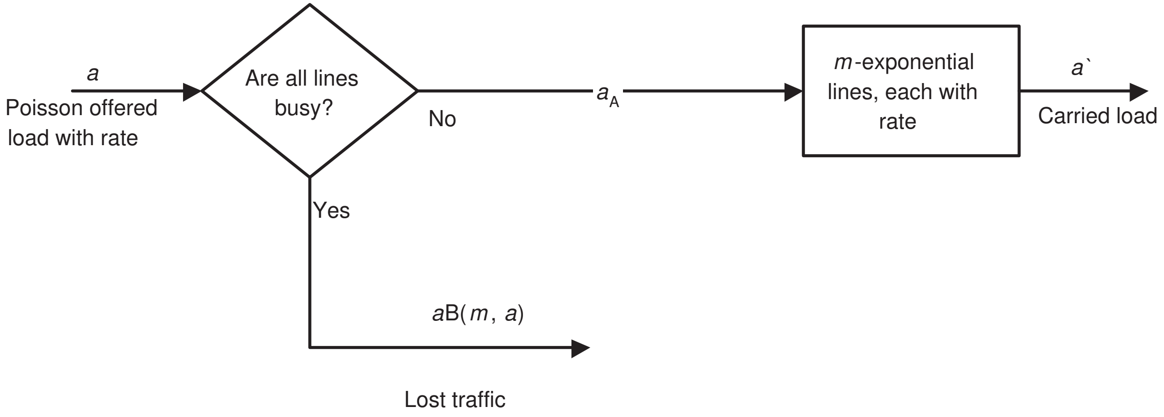
$$P_k = \frac{1}{k!} \left(\frac{\lambda}{\mu} \right)^k \exp\left(-\frac{\lambda}{\mu} \right)$$

- This is a Poisson distribution with parameter λ/μ, where $0 < \lambda/\mu < \infty$
- The infinite number of servers allows λ/μ to be greater than one
- The capacity of an M/M/∞ queue is infinite because a queue is never allowed to form
- The probability that the system is not empty is

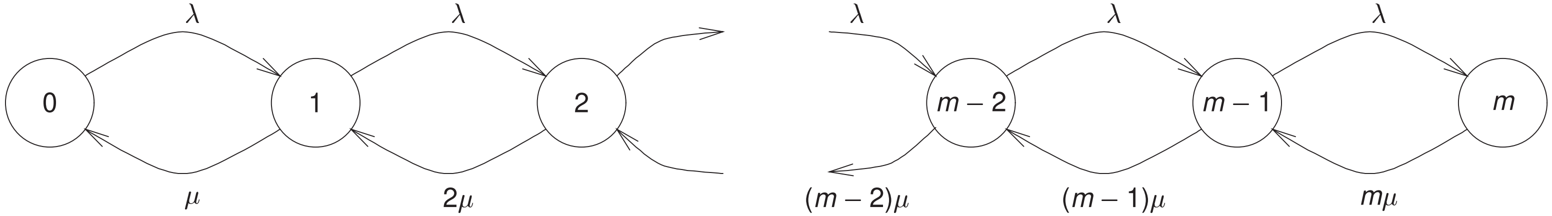
$$1 - P_0 = 1 - \exp\left(-\frac{\lambda}{\mu}\right)$$

8. The M/M/m/m queuing system

- The Erlang delay system contains m servers, without restriction on the number of people in the system
- There is a maximum number of customers in the $M/M/m/m$ system, such that if a customer arrives and all the servers are busy, then he leaves the system. There are m servers, and a maximum of m customers in the system.



Flow chart of the M/M/m/m system



State Diagram for the Erlang loss system, M/M/m/m

Recall the difference between a and ρ:

- If there is only one server, then

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

- If there are k servers, each with rate μ, then a is the offered load, or the load normalised with respect to the number of servers, and measured in Erlangs

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

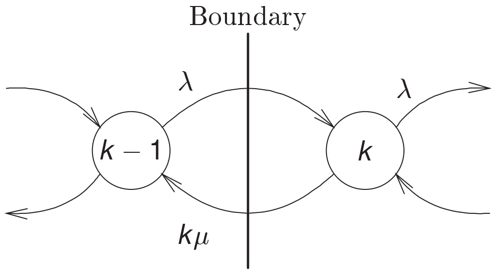
- The arrival rate is

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

- The service rate is

$$\mu_k = \begin{cases} k\mu & k = 0, 1, \ldots, m \\ 0 & k > m \end{cases}$$

When the system is in state $k \leq m$, we have the boundary flow shown below



Flow rates at a boundary

- This leads to the following balance equation

$$\lambda P_{k-1} = k\mu P_k, \qquad k = 1, \ldots, m$$

and thus

$$P_k = \frac{\lambda}{k\mu} P_{k-1} = \frac{a}{k} P_{k-1}, \qquad k = 1, \ldots, m$$

- The solution of this equation is

$$P_k = P_0 \left(\prod_{n=1}^{k} \frac{a}{n} \right) = \frac{a^k}{k!} P_0, \qquad k = 0, \ldots, m$$

- The normalisation condition yields

$$P_0 \sum_{k=0}^{m} \frac{a^k}{k!} = 1$$

that is,

$$P_0 = \frac{1}{\sum_{k=0}^{m} \frac{a^k}{k!}}$$

- Thus

$$P_k = \frac{\frac{a^k}{k!}}{\sum_{n=0}^{m} \frac{a^n}{n!}}, \qquad a = \frac{\lambda}{\mu}, \qquad k = 0, \ldots, m$$

Note: If $m \to \infty$ the $M/M/\infty$ queue is recovered, as expected.

Probability of blocking

- For the M/M/m queue, we were interested in the probability that a customer waits in the queue for service
- In the M/M/m/m queue, there is no queue because the maximum state of the system is the same as the number of servers, and thus the probability of waiting is zero
- The probability of blocking is of interest in an M/M/m/m queue:
  - The probability that a customer arrives at the system when it is full and is therefore denied service
- The blocking probability is given by P_m

$$B(m,a) = P_m = \frac{\frac{a^m}{m!}}{\sum_{n=0}^{m} \frac{a^n}{n!}}$$

- Use Erlang B curves to compute it

What is the average number of people in the system?

$$\bar{N} = \sum_{k=0}^{m} kP_k = \frac{1}{\sum_{n=0}^{m} \frac{a^n}{n!}} \sum_{k=0}^{m} \frac{ka^k}{k!} = \frac{1}{\sum_{n=0}^{m} \frac{a^n}{n!}} \sum_{k=1}^{m} \frac{ka^k}{k!}$$

The last term simplifies to

$$\frac{a}{\sum_{n=0}^{m} \frac{a^n}{n!}} \sum_{k=1}^{m} \frac{a^{k-1}}{(k-1)!} = \frac{a}{\sum_{n=0}^{m} \frac{a^n}{n!}} \left[\sum_{k=0}^{m} \frac{a^k}{k!} - \frac{a^m}{m!} \right]$$

and thus

$$\bar{N} = a(1 - P_m) = a\left(1 - B(m, a)\right)$$

- The average number of people in the system is equal to a multiplied by the probability that customers are not rejected from the system

What happens as a increases?

$$\bar{N} = a(1 - P_m) = a\left(1 - \frac{\frac{a^m}{m!}}{\sum_{n=0}^{m} \frac{a^n}{n!}}\right) = \frac{a\left(\sum_{n=0}^{m} \frac{a^n}{n!} - \frac{a^m}{m!}\right)}{\sum_{n=0}^{m} \frac{a^n}{n!}}$$

If a is large, only consider the largest terms in the numerator and denominator:

$$\lim_{a \to \infty} \bar{N} = \frac{a\left(\frac{a^{m-1}}{(m-1)!}\right)}{\frac{a^m}{m!}} = m$$

- The average rate of the number of calls accepted by the system is :

$$\gamma = \lambda(1 - P_m)$$

- The average service rate is

$$\bar{\mu} = \sum_{k=0}^{m} kP_k\mu = \mu\sum_{k=0}^{m} kP_k = \bar{N}\mu$$

because if the system is in state k, then k service tills are occupied and the total service rate for these k tills is $k\mu$

Example

1. For a blocking probability of 0.01, what is the number of trunks needed for 6 Erlangs of offered load?
2. For a blocking probability of 0.01, what is the number of trunks needed for 18 Erlangs of offered load?

Answers:

1. From the Erlang B curve with $a = 6$ and $P_m = 0.01, 13$ trunks are required.
2. From the Erlang B curve with $a = 18$ and $P_m = 0.01, 28$ trunks are required.

□

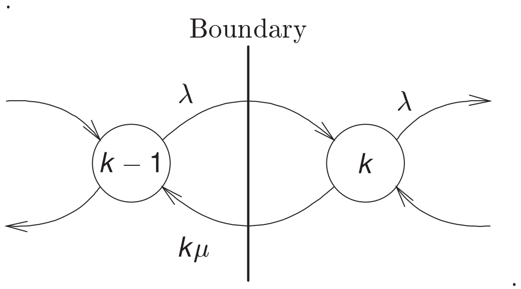
Example A call centre employs thirty operators selling tickets for a major sporting event. The call centre has sixty lines in total. When all operators are active, new calls are placed in a queue for the first available operator. The call centre is to be modelled as an $M/M/m/n$ queue, *i.e.* a queue in which there are $m = 30$ servers and 60 input lines. The maximum state of the system is $n = 60 + 30 = 90$.

1. Write down the balance equations for the two cases $k < m$ and $m \leq k < n$.
2. For each of the two cases outlined in the previous part write down the steady state distribution as a function of P_0, the probability that the system is empty.
3. Use the fact that the steady state distribution must be normalised to show that

$$P_0 = \left[\sum_{k=0}^{m} \frac{a^k}{k!} + \sum_{k=m+1}^{n} \frac{a^k}{m! m^{k-m}} \right]^{-1}, \qquad a = \frac{\lambda}{\mu}$$

where λ is the arrival rate and μ is the rate of the service time distribution.

1. **Case 1:** $k \leq m$. When the system is in state k that is less than or equal to the number of servers, the boundary flow is as shown below:



This leads to the following equations

$$\begin{aligned}\lambda_k &= \lambda, \qquad k = 0, 1, \ldots, m-1 \\ \mu_k &= k\mu, \qquad k = 0, 1, \ldots, m\end{aligned}$$

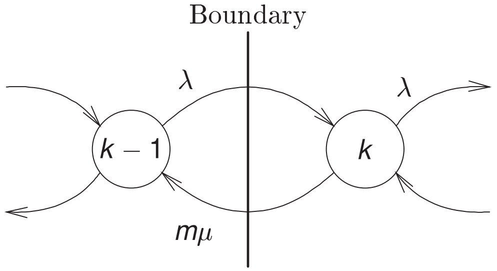
and thus

$$\lambda P_{k-1} = k\mu P_k, \qquad k = 1, \ldots, m$$

1 cont'd. **Case 2:** $m \leq k \leq n$. All the servers are busy.

$$\begin{aligned} \lambda_k &= \lambda, \qquad k = m, \ldots, n \\ \mu_k &= m\mu, \qquad k = m, \ldots, n \end{aligned}$$

The flow at the boundary is as shown below:



This leads to

$$\lambda P_{k-1} = m\mu P_k$$

2. **Case 1:** $k \leq m$. From the detailed balance equation

$$P_k = \frac{\lambda}{k\mu} P_{k-1}$$

If $a = \frac{\lambda}{\mu}$, then

$$\frac{\lambda}{k\mu} = \frac{a}{k}$$

and thus

$$P_k = \frac{a}{k} P_{k-1}$$

from which it follows that

$$P_k = \frac{a^k}{k!} P_0 \quad \text{for} \quad k \leq m$$

2. cont'd **Case 2:** $m \leq k \leq n$. The detailed balance equation for this case is

$$P_k = \frac{\lambda}{m\mu} P_{k-1} = \frac{a}{m} P_{k-1}$$

and thus

$$P_k = \left(\frac{a}{m}\right)^{k-m} P_m \qquad \text{where} \qquad P_m = \frac{a^m}{m!} P_0$$

Hence

$$P_k = \frac{a^k}{m^{k-m} m!} P_0 \quad \text{for} \quad k > m$$

and therefore

$$P_k = \begin{cases} \frac{a^k}{k!} P_0 & k \leq m \\ \frac{a^k}{m^{k-m} m!} P_0 & k \geq m \end{cases}$$

3. The quantity P_0 is calculated by normalisation. Since

$$\sum_{k=0}^{n} P_k = 1$$

it follows that

$$P_0 \left[\sum_{k=0}^{m} \frac{a^k}{k!} + \sum_{k=m+1}^{n} \frac{a^k}{m! m^{k-m}} \right] = 1$$

and thus

$$P_0 = \left[\sum_{k=0}^{m} \frac{a^k}{k!} + \sum_{k=m+1}^{n} \frac{a^k}{m! m^{k-m}} \right]^{-1}$$

This expression enables the formulae for P_k to be computed. □

9. The M/G/1 queuing system

Markovian systems are simple to analyse, but many systems exhibit more complex behaviour. For example:

- In a packet switching system, the packets are not distributed exponentially in length, but they are either of constant length, or of random length, between bounds imposed by the protocol of the system
- Measurements of packet lengths in real systems show that the distribution of packet lengths is bi-modal:
  - Most of the packets are either short, corresponding to interactive terminal traffic, or long, corresponding to file transfers and other high volume transactions

The simplest of these more general models is the M/G/1 queue, which is characterised by:

- Arrivals that form a Poisson process [M]
- The service times have an arbitrary (general) probability distribution [G]
- There is one queue (arrival stream) and one server (output channel along which the packets are sent) [1]
- The buffer is assumed to be infinite in size, and thus there are no limits on the arrival stream

Assumptions of the M/G/1 queue

- No assumptions are made about the length of each message
- No assumptions are made about the distribution of the lengths of each message, except that its first and second moments are finite
- The message lengths are independent of the arrival times, and the lengths of other messages
- Transmissions occur at a fixed rate, and thus message length and transmission times are related

What is the average waiting time for an M/G/1 queue?

- A general service time distribution is now assumed, and thus the memoryless property cannot be used
- A simple balance equation for the state of the queues cannot be used.

Use the **Pollaczek-Khinchin** formula, which states the average waiting time in the queue:

$$\bar{W} = \frac{\lambda E\left\{X^2\right\}}{2(1-\rho)}$$

- $\bar{W}$ is expected waiting time
- $E\left\{X^2\right\}$ is the second moment of the service time distribution
- The utilisation factor ρ is given by $\rho = \frac{\lambda}{\mu} = \lambda\bar{X}$
- $\bar{X} = E\left\{X\right\}$ is the mean of the service time distribution, $\bar{X} = 1/\mu$

The total waiting time is therefore

$$T = \bar{X} + \bar{W} = \bar{X} + \frac{\lambda E\left\{X^2\right\}}{2(1-\rho)}$$

Recall that the Pollaczek-Khinchin formula makes no assumption about the service time distribution. But:

- The formula is a function of the average arrival rate λ, the mean service rate μ, and the variance of the service time distribution.
- The formula is independent of higher moments of the service time distribution, even though the service time distribution is general.

Example Compute the average waiting time in the queue associated with an exponentially distributed service time.

The exponentially distribution is given by

$$p\left(X\right)=\mu\exp\left(-\mu X\right)$$

and its mean and variance are

$$\bar{X}=E\left\{X\right\}=\frac{1}{\mu},\qquad \text{var}\left(X\right)=E\left\{X^2\right\}-\left(E\left\{X\right\}\right)^2=\frac{1}{\mu^2}$$

and thus

$$E\left\{X^2\right\}=\frac{2}{\mu^2}$$

Substituting into the Pollaczek-Khinchin formula yields

$$\bar{W}=\frac{\rho}{\mu\left(1-\rho\right)}=\frac{\rho}{\mu-\lambda}$$

which is identical to those we derived for the $M/M/1$ queue. □

Example of an M/G/1 system - the ARQ system

There exist various methods to perform error-detection and correction in a communication link.

In an ARQ system:

- Each correctly received packet is acknowledged by an acknowledgement packet (ACK) sent from the receiver to the sender
- This acknowledgement packet includes the packet number that was sent and a timeout mechanism
- If the sender receives the ACK, it knows the packet was transmitted successfully
- Otherwise, if a certain length of time elapses (the timeout) and the sender receives no ACK, the packet is sent again

This is the *stop and wait algorithm*, the simplest ARQ system

Consider a system in which the length of each packet is one time unit, and that the maximum wait for an acknowledgement is $n - 1$ time units before a packet is retransmitted, where n is a known integer. The timer for the acknowledgement starts when the last bit of the packet leaves the sender. Thus:

- The total delay for each retransmission is $(n - 1) + 1 = n$ time units
- If k retransmissions are required, the total delay for the retransmissions is kn
- Since it is assumed that an acknowledgement is received after a delay of kn, and the transmission time of each packet is one time unit, the total delay for a packet to be received correctly is $kn + 1$

Assumption of the analysis of the ARQ system:

- The packets arrive at the transmitter according to a Poisson distribution with rate λ
- A packet is rejected at the receiver with probability p, independent of other packets. This is not a realistic assumption because if one packet is rejected, it is likely that several consecutive packets will be rejected, and thus the assumption of independence is violated

Since k retransmissions are required, it follows that:

- The start of the first transmission of a given packet and the end of the last transmission of a given packet is $kn + 1$ time units with probability $(1 - p)p^k$, which is a geometric distribution

$$P(X = kn + 1) = (1 - p)p^k$$

This is the probability that a packet is rejected at the receiver k times and successful transmission occurs on the $(k+1)$th attempt, that is, the receiver takes $X = kn + 1$ time units to process the packet.

The expected service time of the packet is calculated from the first moment:

$$\begin{aligned} E\{X\} &= \sum_{k=0}^{\infty}(kn+1)(1-p)p^k \\ &= (1-p)\left(\sum_{k=0}^{\infty}p^k + n\sum_{k=0}^{\infty}kp^k\right) \\ &= 1+\frac{np}{1-p} \end{aligned}$$

since

$$\sum_{k=0}^{\infty}p^k = \frac{1}{1-p} \quad \text{and} \quad \sum_{k=0}^{\infty}kp^k = \frac{p}{(1-p)^2}$$

The second moment of the service time is

$$\begin{aligned}
E\left\{X^2\right\} &= \sum_{k=0}^{\infty}(1+kn)^2(1-p)p^k \\
&= (1-p)\left(\sum_{k=0}^{\infty}p^k + 2n\sum_{k=0}^{\infty}kp^k + n^2\sum_{k=0}^{\infty}k^2p^k\right) \\
&= 1+\frac{2np}{1-p}+\frac{n^2\left(p+p^2\right)}{(1-p)^2}
\end{aligned}$$

since

$$\sum_{k=0}^{\infty}k^2p^k = \frac{p+p^2}{(1-p)^3}$$

These moments can be used with the Pollaczek-Khinchin formula to obtain statistics for the queue.

The utilisation factor is

$$1 - P(k = 0) = 1 - P(X = 1) = 1 - p$$

Also, from the Pollaczek-Khinchin

$$\rho = \lambda \bar{X} = \lambda E\left\{X\right\} = \lambda + \frac{\lambda n p}{1 - p}$$

The average waiting time in the queue is

$$\bar{W} = \frac{\lambda E\left\{X^2\right\}}{2(1 - \lambda \bar{X})}$$

and thus the total waiting time is

$$T = \bar{X} + \bar{W} = \bar{X} + \frac{\lambda E\left\{X^2\right\}}{2(1 - \lambda \bar{X})}$$