

M/G/1 Queue

Dr Ahmad Khonsari

ECE Dept.

The University of Tehran

Excerpt from virtamo slides

M/G/1 queue

{	<i>M (memoryless):</i>	<i>Poisson arrival process, intensity λ</i>
	<i>G (general):</i>	<i>general holding time distribution, mean $\bar{S} = 1/\mu$</i>
	<i>1 :</i>	<i>single server, load $\rho = \lambda\bar{S}$ (in a stable queue one has $\rho < 1$)</i>

The number of customers in the system, $N(t)$, *does not now constitute a Markov process.*

- The probability per time unit for a transition from the state $\{N = n\}$ to the state $\{N = n - 1\}$, *i.e. for a departure of a customer, depends also on the time the customer in service has already spent in the server;*
 - **this information is not contained in the variable $N(t)$**
 - **only in the case of an exponential service time the amount of service already received does not have any bearing (memoryless property)**

In spite of this, the **mean** (queue length, waiting time, and sojourn time) of the *M/G/1 queue* can be found. The results (the Pollaczek-Khinchin formulae) will be derived in the following.

It turns out that even the **distributions** of these quantities can be found. A derivation based on considering an **embedded Markov chain** will be presented after the **mean** formulae.

Pollaczek-Khinchin **mean (expectation)** formula

We start with the derivation of the **expectation** of the **waiting time** W . W is the time the customer has to wait for the service (time in the “waiting room”, i.e. in the actual queue).

$$E[W] = \underbrace{E[N_q] \cdot E[S]}_{\text{mean time needed to serve the customers ahead in the queue}} + \underbrace{E[R]}_{\text{unfinished work in the server}} \quad (R = \text{residual service time})$$

- R is the **remaining service time** of the customer in the server (unfinished work expressed as the time needed to recharge the work).

If the server is idle (i.e. the system is empty), then $R = 0$.

- In order to calculate the **mean waiting time** of an arriving customer one needs the **expectation** of N_q (*number of waiting customers at the instant of arrival*).
- Due to the **PASTA** property of Poisson process, the distributions seen by the arriving customer are the same as those at an arbitrary instant.

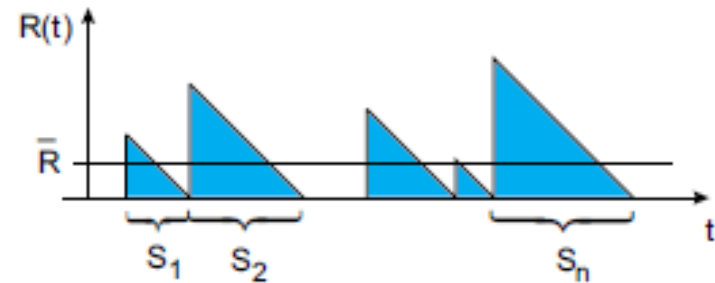
The key observation is that by Little’s result the **mean queue length** $E[N_q]$ can be expressed in terms of the **mean waiting time** (by considering the waiting room as a black box)

$$E[N_q] = \lambda E[W] \quad \Rightarrow \quad \underline{E[W] = \frac{E[R]}{1 - \rho}} \quad \text{It remains to determine } E[R].$$

$$E[W] = \lambda E[W]E[S] + E[R] \quad \rho = \lambda E[S]$$

Pollaczek-Khinchin **mean** formula (continued)

The residual service time can be deduced by using similar graphical argument as was used in explaining the hitchhiker's paradox. The graph represents now the evolution of the **unfinished work in the server, $R(t)$, as a function of time.**



Consider a long interval of **time t** . The **average** value of the sawtooth curve can be calculated by dividing the sum of the areas of the triangles by the length of the interval.

- Now the triangles may be separated by idle periods (queue empty).
- The **number of the triangles, n** , is determined by **the arrival rate λ** ; **mean number is λt** .

$$E[R] = \frac{1}{t} \int_0^t R(t') dt' = \frac{1}{t} \sum_{i=1}^n \frac{1}{2} S_i^2 = \underbrace{\frac{n}{t}}_{\rightarrow \lambda} \cdot \underbrace{\frac{1}{n} \sum_{i=1}^n \frac{1}{2} S_i^2}_{\rightarrow \frac{1}{2} E[S^2]} \quad E[W] = \frac{E[R]}{1 - \rho}$$

$$\underline{\underline{E[W] = \frac{\lambda E[S^2]}{2(1 - \rho)}}}$$

Pollaczek-Khinchin **mean** formula for the **waiting time**

Pollaczek-Khinchin **mean** formula (continued)

From the **mean waiting time** one immediately gets the **mean sojourn time**

$$E[T] = \underbrace{E[S]}_{\text{the customer's own service time}} + E[W]$$

the customer's
own service time

$$\begin{aligned} \rho &= \lambda / \mu, \quad \pi_0 = 1 - \rho \\ E[N] &= \rho + \frac{\rho^2(1 + \sigma^2 \mu^2)}{2(1 - \rho)}, \quad E[N_q] = \frac{\rho^2(1 + \sigma^2 \mu^2)}{2(1 - \rho)} \\ E[T] &= \frac{1}{\mu} + \frac{\lambda(1/\mu^2 + \sigma^2)}{2(1 - \rho)}, \quad E[W] = \frac{\lambda(1/\mu^2 + \sigma^2)}{2(1 - \rho)} \end{aligned}$$

Mean waiting and sojourn times

Squared coefficient of variation C_v^2

$$\begin{cases} E[W] = \frac{\lambda E[S^2]}{2(1 - \rho)} = \frac{1 + C_v^2}{2} \cdot \frac{\rho}{1 - \rho} \cdot E[S] \\ E[T] = E[S] + \frac{\lambda E[S^2]}{2(1 - \rho)} = \left(1 + \frac{1 + C_v^2}{2} \cdot \frac{\rho}{1 - \rho}\right) \cdot E[S] \end{cases} \quad \begin{aligned} C_v^2 &= V[S]/E[S]^2 = \sigma^2 \mu^2 \\ E[S^2] &= V[S] + E[S]^2 \\ &= (1 + C_v^2) \cdot E[S]^2 \end{aligned}$$

By applying Little's result one obtains the corresponding formulae for the **numbers**.

Mean number of waiting customers and customers in system

$$\begin{cases} E[N_q] = \lambda E[W] = \frac{\lambda^2 E[S^2]}{2(1 - \rho)} = \frac{1 + C_v^2}{2} \cdot \frac{\rho^2}{1 - \rho} \\ E[N] = \lambda E[T] = \lambda E[S] + \frac{\lambda^2 E[S^2]}{2(1 - \rho)} = \rho + \frac{1 + C_v^2}{2} \cdot \frac{\rho^2}{1 - \rho} \end{cases}$$

Remarks on the PK mean formulae

- **Mean** values depend only on the **expectation** $E[S]$ and **variance** $V[S]$ of the **service time distribution** but not on higher moments.
- **Mean** values increase linearly with the **variance**.
- Randomness, 'disarray', leads to an increased waiting time and queue length.
- The formulae are similar to those of the *M/M/1 queue*; the only difference is the extra factor $(1 + C_v^2)/2$.

$$\rho = \lambda / \mu, \quad \pi_n = (1 - \rho) \rho^n$$

$$E[N] = \frac{\lambda}{\mu - \lambda} = \frac{\rho}{1 - \rho}, \quad E[N_q] = \frac{\lambda^2}{\mu(\mu - \lambda)} = \frac{\rho^2}{1 - \rho}$$

$$E[T] = \frac{1}{\mu - \lambda} = \frac{1}{\mu(1 - \rho)}, \quad E[W] = \frac{\lambda}{\mu(\mu - \lambda)} = \frac{\rho}{\mu(1 - \rho)}$$

$$E[N_q] = \frac{\rho^2(1 + \sigma^2 \mu^2)}{2(1 - \rho)}$$

$$= \left(\frac{\rho^2}{1 - \rho} \right) \left(\frac{1 + (cv)^2}{2} \right)$$

L_Q for M/M/1 queue

Corrects the M/M/1 formula to account for a non-exponential service time dist'n

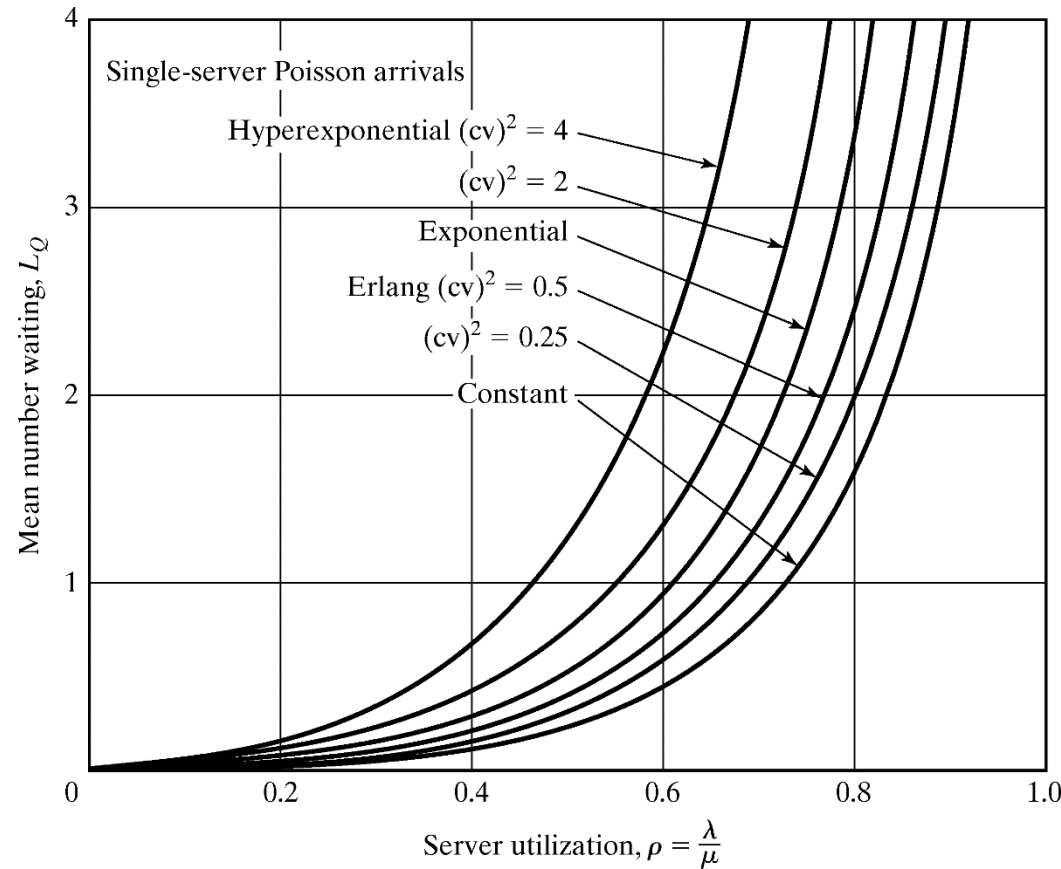
Remarks on the PK mean formulae

$$E[N_q] = \frac{\rho^2(1 + \sigma^2 \mu^2)}{2(1 - \rho)}$$

$$= \left(\frac{\rho^2}{1 - \rho} \right) \left(\frac{1 + (cv)^2}{2} \right)$$

L_Q for $M/M/1$ queue

Corrects the $M/M/1$ formula to account for a non-exponential service time dist'n



Note: $E[N] - E[N_q] = \rho$ is the time-average number of customers being served.

Average length of queue, $E[N_q]$, can be rewritten as:

$$E[N_q] = \frac{\rho^2}{2(1 - \rho)} + \frac{\lambda^2 \sigma^2}{2(1 - \rho)}$$

If λ and μ are held constant, $E[N_q]$ depends on the variability, σ^2 , of the service times.

The PK mean formulae for the M/M/1 and M/D/1 queues

M/M/1 queue

In the case of the exponential distribution one has

$$V[S] = E[S]^2 \quad \Rightarrow \quad C_v^2 = 1$$

$$\begin{cases} E[N] = \rho + \frac{\rho^2}{1-\rho} = \frac{\rho}{1-\rho} \\ E[T] = \left(1 + \frac{\rho}{1-\rho}\right) \cdot E[S] = \frac{1}{1-\rho} \cdot E[S] \end{cases}$$

The familiar formulae for the *M/M/1 queue*

M/D/1 queue

In the case of constant service time one has

$$V[S] = 0 \quad \Rightarrow \quad C_v^2 = 0$$

$$\begin{cases} E[N] = \rho + \frac{1}{2} \frac{\rho^2}{1-\rho} \\ E[T] = \left(1 + \frac{1}{2} \frac{\rho}{1-\rho}\right) \cdot E[S] \end{cases}$$

A factor 1/2 in the “waiting room terms”

$$E[N_q] = \lambda E[W] = \frac{\lambda^2(1/\mu^2 + \sigma^2)}{2(1-\rho)}$$

Example.

Two workers competing for a job, **Able** claims to be faster than **Baker** on **average**, but **Baker** claims to be more consistent, Poisson arrivals at rate $\lambda = 2$ per hour (1/30 per minute).

Able: $1/\mu = 24$ minutes, $\rho = (1/30)/(1/24) = 4/5$ and $\sigma^2 = 20^2 = 400$ minutes²:

$$E[N_q] = \frac{(1/30)^2[24^2 + 400]}{2(1 - 4/5)} = 2.711 \text{ customers}$$

The proportion of arrivals who find Able idle and thus experience no delay is $\pi_0 = 1 - \rho = 1/5 = 20\%$.

Baker: $1/\mu = 25$ minutes, $\rho = (1/30)/(1/25) = 5/6$ and $\sigma^2 = 2^2 = 4$ minutes²:

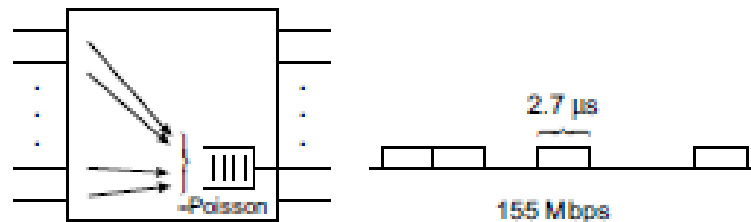
$$E[N_q] = \frac{(1/30)^2[25^2 + 4]}{2(1 - 5/6)} = 2.097 \text{ customers}$$

The proportion of arrivals who find Baker idle and thus experience no delay is $\pi_0 = 1 - \rho = 1/6 = 16.7\%$.

Although working faster on **average**, **Able's** greater service variability results in an **average** queue length about 30% greater than **Baker's**.

Example.

The output buffer of an ATM multiplexer can be modelled as an $M/D/1$ queue.
Constant service time means now that an ATM cell has a fixed size (53 octets) and its transmission time to the link is constant.



If the link speed is 155 Mbit/s, then the mean **transmission time** is

$$E[S] = \left(\frac{53 \cdot 8}{155} \right) \text{microseconds} = 2.7 \text{microseconds}$$

What is the **mean number** of cells in the buffer (including the cell being transmitted) and the **mean sojourn time** of the cell in the buffer when the **average information rate** on the link is 124 Mbit/s?

The load (utilization) of the link is $\rho = 124/155 = 0.8$

Using

$$E[N] = \rho + \frac{1}{2} \frac{\rho^2}{1 - \rho}$$

$$E[T] = \left(1 + \frac{1}{2} \frac{\rho}{1 - \rho} \right) \cdot E[S]$$

We Have

$$\begin{cases} E[N] = 0.8 + \frac{1}{2} \cdot \frac{0.8^2}{1 - 0.8} = 2.4 \\ E[T] = \left(1 + \frac{1}{2} \cdot \frac{0.8}{1 - 0.8} \right) 2.7 \mu s = 8.1 \text{microseconds} \end{cases}$$

The queue length distribution in an M/G/1 queue

The queue length N_t in an M/G/1 system does not constitute a Markov process.

- The number in system alone does not tell with which probability (per time) a customer in service departs, but this probability depends also on the amount of service already received.

As we saw above, the mean queue length was easy to derive. Also the queue length distribution can be found. There are two different approaches:

1. The first is based on the observation that the unfinished work in the system, X_t (or virtual waiting time V_t), does constitute a Markov process. The Markovian property, is a property of the considered stochastic process, not an intrinsic property of the system.

- The evolution of X_t can be characterized as follows: when there are no arrivals X_t decreases at a constant rate C (when $X_t > 0$). In addition, there is a constant probability per time unit, λ , for a new arrival, bringing to the queue an amount work having a given distribution. No knowledge about the history of X_t is needed.
- A slight technical difficulty is that X_t is a continuous state (real valued) process.

2. The second approach is based on the observation that there is an embedded Markov chain, by means of which the distribution can be solved. In the following we use this method.

Embedded Markov chain

The embedded Markov chain is constituted by the queue left by an departing customer (i.e. number in system at departure epochs). That this indeed is a Markov chain will be justified later.

Denote

$$\left\{ \begin{array}{l} N_-^* = \text{the queue length seen by an arriving customer (queue length just before arrival)} \\ N_+^* = \text{the queue length left by a departing customer} \\ N = \text{queue length at an arbitrary time} \end{array} \right.$$

By the PASTA property of Poisson arrivals we have

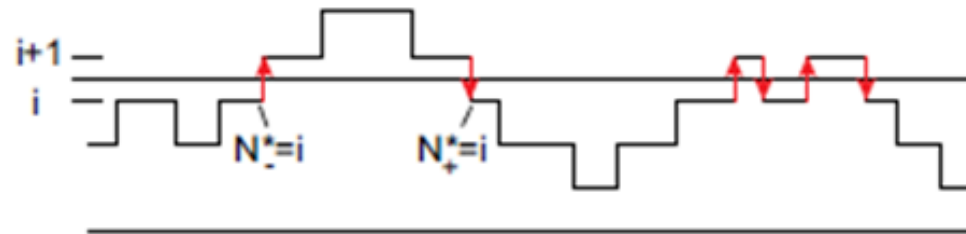
$$\overline{N_-^*} \sim N$$

In addition, for any system with single (in contrast to batch) arrivals and departures, it holds

$$\overline{N_+^*} \sim \overline{N_-^*}$$

(so called level crossing property)

Proof:



The events $\{N_-^* = i\}$ and $\{N_+^* = i\}$ occur pairwise.

$$P\{N_-^* = i\} = P\{N_+^* = i\} \Rightarrow \overline{N_-^*} \sim \overline{N_+^*}$$

Embedded Markov chain (continued)

We have shown that

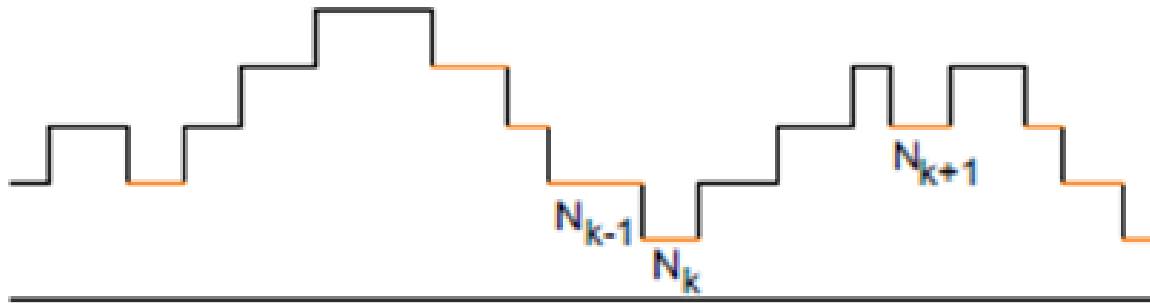
$$N_+^* \sim N_-^* \quad \text{and} \quad N_-^* \sim N \Rightarrow \underline{N_+^* \sim N}$$

Thus to find the distribution of N at an arbitrary time, it is sufficient to find the distribution at instants immediately after departures.

We focus on the Markov chain N_+^* , which in the following will for brevity be denoted just N .

In particular, denote

$$\left\{ \begin{array}{l} N_k = \text{queue length after the departure of customer } k \\ V_k = \text{number of new customers arrived during the service time of customer } k. \end{array} \right.$$



Embedded Markov chain (continued)

Claim: The discrete time process N_k constitutes a Markov chain (however, not of a birth-death type process).

Proof: Given N_k ; N_{k+1} can be expressed in terms of it (N_k) and of a random variable V_{k+1} which is independent of N_k and its history:

$$N_{k+1} = \begin{cases} N_k - 1 + V_{k+1}, & N_k \geq 1 \\ V_{k+1}, & N_k = 0 \end{cases} \quad (= N_k + V_{k+1})$$

- If $N_k \geq 1$, then upon the departure of customer k , customer $k + 1$ is in the queue and enters the server.

When ultimately customer $k + 1$ departs, the queue length is decremented by one. Meanwhile (during the service of customer $k + 1$), there have been V_{k+1} arrivals.

- If $N_k = 0$, customer k leaves an empty queue. Upon the arrival of customer $k + 1$ the queue length is first incremented and then decremented by one when customer $k + 1$ departs.

The queue consists of those customers who arrived during the service of customer $k + 1$.

- As the service times are independent and the arrivals are Poissonian, the V_k are independent of each other. Moreover, V_{k+1} is independent of the queue length process before the departure of customer k , i.e. of N_k and its previous values.

The stochastic characterization of N_{k+1} depends on N_k but not on the earlier history. QED.

Embedded Markov chain (continued)

Denote $\hat{N}_k = (N_k - 1)^+ = \begin{cases} N_k - 1, & N_k \geq 1 \\ N_k (= 0), & N_k = 0 \end{cases}$

Then $\overline{N_{k+1}} = \overline{\hat{N}_k + V_{k+1}}$ Upward jumps can be arbitrary large.
Downward one step at a time.

- In equilibrium (when the initial information has been washed out) the random variables N_k, N_{k+1}, \dots have the same distribution.
- So are the distributions of the random variables $\hat{N}_k, \hat{N}_{k+1}, \dots$ the same (mutually).
- Random variables V_k, V_{k+1}, \dots have from the outset the same distributions (mutually).

Denote the random variables obeying the equilibrium distributions without indices, so that

$$\overline{N} = \overline{\hat{N} + V}$$

Since V and \hat{N} are independent, we have for the generating functions

$$\overline{G_N(z)} = \overline{G_{\hat{N}}(z)} \cdot \overline{G_V(z)}$$

The task now is to determine $G_N(z)$ and $G_V(z)$

Expressing the generating function of \hat{N} in terms of the generating function of N

$$\begin{aligned}
 \mathcal{G}_{\hat{N}}(z) &= E[z^{\hat{N}}] \\
 &= z^0 \cdot \frac{P\{\hat{N} = 0\}}{P\{N=0\} + P\{N=1\}} + \sum_{i=1}^{\infty} z^i \frac{P\{\hat{N} = i\}}{P\{N=i+1\}} \\
 &= P\{N = 0\} + \frac{1}{z} \sum_{i=1}^{\infty} z^i P\{N = i\} \\
 &= \underbrace{P\{N = 0\}}_{1-\rho} \left(1 - \frac{1}{z}\right) + \frac{1}{z} \underbrace{\sum_{i=0}^{\infty} z^i P\{N = i\}}_{\mathcal{G}_N(z)}
 \end{aligned}$$

We have obtained the result

$$\underline{\underline{\mathcal{G}_{\hat{N}}(z) = \frac{\mathcal{G}_N(z) - (1 - \rho)(1 - z)}{z}}} \quad \text{where} \quad \rho = \lambda E[S]$$

The number of arrivals from a Poisson process during a service time

Let X be an arbitrary random variable (representing an interval of time).

We wish to determine the distribution of the number of arrivals, K , from a Poisson process (intensity λ) occurring in the interval X , and, in particular, its generating function $G_K(z)$.

$$G_K(z) = E[z^K] = E\left[\underbrace{E[z^K | X]}_{K \sim \text{Poisson}(\lambda X)}\right] = E[e^{-(1-z)\lambda X}] = X * ((1-z)\lambda) \quad X^*(s) = E[e^{-sX}]$$

Generally, $\underline{G_K(z) = X^*((1-z)\lambda)}$ In particular, $\underline{G_V(z) = S^*((1-z)\lambda)}$

The result can be derived also in a more elementary way

$$\begin{aligned} G_K(z) &= \sum_{i=0}^{\infty} z^i P\{K = i\} = \sum_{i=0}^{\infty} z^i \int_0^{\infty} \overbrace{P\{K = i | X = x\}}^{\frac{(\lambda x)^i}{i!} e^{-\lambda x}} f_X(x) dx \\ &= \int_0^{\infty} f_X(x) e^{-\lambda x} \sum_{i=0}^{\infty} \frac{(\lambda x z)^i}{i!} dx = \int_0^{\infty} f_X(x) e^{-\lambda x} e^{\lambda x z} dx = \int_0^{\infty} f_X(x) e^{-(1-z)\lambda x} dx \\ &= X^*((1-z)\lambda) \end{aligned}$$

The number of arrivals from a Poisson process during a service time (continued)

The result can be interpreted by the method of collective marks:

- In the method of collective marks $G_K(z)$ has the interpretation of the probability that none of the K arrivals occurring in the interval X is marked, when each arrival is independently marked with the probability $(1 - z)$.

- The process of marked arrivals is obtained by a random selection of a Poisson process, and is thus a Poisson process with intensity $(1 - z)\lambda$.

- The interpretation of the Laplace transform in terms of collective marks: $X^*(s)$ is the probability that there are no arrivals in the interval X from a Poisson process with intensity s :

$$X^*(s) = E[e^{-sX}] = E[P\{\text{no arrivals in } X|X\}] = P\{\text{no arrivals in } X\}$$

- When the intensity of the marking process is $(1 - z)\lambda$, the probability of no marks is $X^*((1 - z)\lambda)$.

Pollaczek-Khinchin transform formula for the queue length

By collecting the results together,

$$G_N(z) = G_N^{\wedge}(z) \cdot G_V(z) = \frac{G_N(z) - (1 - \rho)(1 - z)}{z} \cdot S^*((1 - z)\lambda)$$

From this we can solve $G_N(z)$

$$G_N(z) = \frac{(1 - \rho)(1 - z)}{S^*((1 - z)\lambda) - z} \cdot S^*((1 - z)\lambda) = \frac{(1 - \rho)(1 - z)}{1 - z/S^*((1 - z)\lambda)}$$

Example. M/M/1 queue

$$S \sim \text{Exp}(\mu) \Rightarrow S^*(s) = \frac{\mu}{s + \mu}$$

$$S^*((1 - z)\lambda) = \frac{\mu}{(1 - z)\lambda + \mu} = \frac{1}{(1 - z)\rho + 1} \quad | \quad \rho = \lambda/\mu$$

$$G_N(z) = \frac{(1 - \rho)(1 - z)}{1 - z((1 - z)\rho + 1)} = \frac{(1 - \rho)(1 - z)}{(1 - z)(1 - \rho z)} = \frac{1 - \rho}{1 - \rho z}$$

$$= (1 - \rho)(1 + (\rho z) + (\rho z)^2 + \dots) \quad (\text{generates the distribution in an } M/M/1 \text{ queue})$$

M/G/1 queue: distribution of the sojourn time T

Above we have derived a formula for the distribution of the queue length N (*fully*, N_+^*) left by a customer, which was noted to be the same as the distribution at an arbitrary instant.

From this result we can infer more, viz. the distribution of the total time, T , *spent in the system* (sojourn time). For the expectation, we have already obtained the Pollaczek-Khinchin mean formula.

The key observation is that the queue, N , *left by a customer consists of those customers who have arrived during the time is system of the departing customer.*

Again we can apply the general result concerning the generating function of the number of arrivals from a Poisson process occurring in an interval having a given length distribution

$$\overline{\mathcal{G}_N(z)} = \overline{T^*((1-z)\lambda)} \quad \text{where } T^*(\cdot) \text{ is the Laplace transform of the sojourn time.}$$

Note: By evaluating the derivative with respect to z at $z = 1$ one gets

$$E[N] = \mathcal{G}'_N(1) = -\lambda \underbrace{T^{*'}(0)}_{-E[T]} = \lambda E[T]$$

In view of the Little's result, this is as it should be.

M/G/1 queue: distribution of the sojourn time (continued)

We have obtained

$$T^*((1-z)\lambda) = \frac{(1-\rho)(1-z)}{S^*((1-z)\lambda) - z} S^*((1-z)\lambda)$$

Here z is a free variable. Denote $s = (1-z)\lambda$, i.e. $z = 1 - s/\lambda$, whence

$$\underline{T^*(s) = \frac{(1-\rho)s}{s - \lambda + \lambda S^*(s)} S^*(s)}$$

Pollaczek-Khinchin transform formula
for the sojourn time

Example. M/M/1 queue

$$S \sim \text{Exp}(\mu) \Rightarrow S^*(s) = \frac{\mu}{s + \mu}$$

$$T^*(s) = \frac{(1-\rho)s}{s - \lambda + \lambda \frac{\mu}{s + \mu}} \frac{\mu}{s + \mu} = \frac{(1-\rho)s}{s - \lambda \frac{s}{s + \mu}} \frac{\mu}{s + \mu} = \frac{\mu - \lambda}{s + (\mu - \lambda)}$$

$\Rightarrow T \sim \text{Exp}(\mu - \lambda)$, in accordance with earlier result.

M/G/1 queue: distribution of the waiting time W

The following is true generally

$$T = \underbrace{W}_{\text{wait}} + \underbrace{S}_{\text{service}}$$

Since W and S are independent, we have for the Laplace transforms

$$T^*(s) = W^*(s) \cdot S^*(s)$$

and identify from the formula for $T^*(s)$.

$$\underline{W^*(s) = \frac{(1 - \rho)s}{s - \lambda + \lambda S^*(s)}}$$

Pollaczek-Khinchin transform formula for the waiting time

The expression can also be rewritten in the form

$$W^*(s) = \frac{1 - \rho}{1 - \rho \frac{1 - S^*(s)}{sE[S]}} \quad | \quad \rho = \lambda E[S]$$

Denote now by R the residual service time in the server conditioned on that there is a customer in the server. One can show (left as an exercise) that the density function of R is

$$f_R(t) = \frac{1 - F_S(t)}{E[S]} \Rightarrow R^*(s) = \frac{1 - S^*(s)}{sE[S]} \Rightarrow \underline{W^*(s) = \frac{1 - \rho}{1 - \rho R^*(s)}}$$

Interpretation of the formula for $W^*(s)$

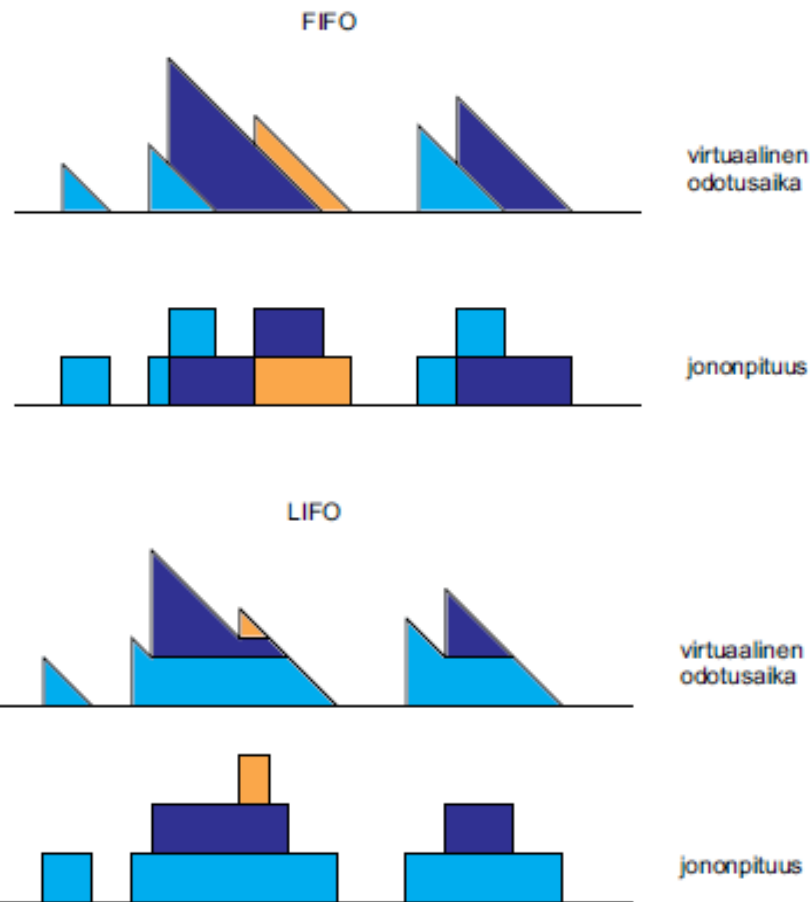
$$W^*(s) = \frac{1 - \rho}{1 - \rho R^*(s)} = \sum_{n=0}^{\infty} (1 - \rho) \rho^n R^*(s)^n$$

- The real waiting time W of the customers is, by the PASTA property of Poisson arrivals, distributed as the virtual waiting time (unfinished work expressed as the time it takes to recharge the work) at an arbitrary instant.
- Virtual waiting time is independent of the scheduling discipline (justified later) and is in the ordinary FIFO queue the same as e.g. in a PS queue (Processor Sharing).
- The queue length distribution of an $M/M/1$ -PS queue, $\pi_n = (1 - \rho) \rho^n$, is independent of the service time distribution and applies also for the $M/G/1$ queue (this does not hold for the FIFO discipline).
- The unfinished work in a PS queue at an arbitrary instant is composed of the residual service times of the customers in the system. One can show that, conditioned on the number of customers in the queue, n , the residual service times are independent and distribute as R . The total residual service time of the customer thus has the Laplace transform $R^*(s)^n$.
- By the law of total probability, the above formula gives the Laplace transform of the virtual waiting time in a PS queue, and thus also that in the FIFO queue.

Virtual waiting time (unfinished work) is independent of the scheduling discipline

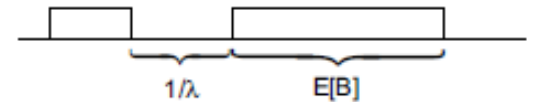
- If the discipline is work conserving, i.e. the server is busy always when there are customers in the system, the busy periods are the same no matter in which order the service is given to different customers in the system; the total work is “anonymous work”.

- The scheduling affects N_t but not X or V_t



Busy period of an M/G/1 queue: waiting time

The server is alternatingly busy and idle. The busy period is a continuous period where the server is uninterruptedly busy. Two busy periods are separated by an idle period.



Denote $\left\{ \begin{array}{l} B = \text{length of busy period} \\ I = \text{length of idle period} \end{array} \right.$

In a Poisson process the interarrival times are distributed according to $Exp(\lambda)$. Because of the memoryless property, the idle periods (time from the end of an busy period to the start of the next one) obey the same distribution, $I \sim Exp(\lambda)$, thence

$$\underline{\underline{E[I] = 1/\lambda}}$$

By Little's result, the load of the server $\rho = \lambda E[S]$ is the same as the expected number of customers in the server. As there can be at most one customer at a time in the server, the expected number equals the probability that there is a customer in the server, and further, this equals the proportion of time the server is busy:

$$\lambda E[S] = \frac{E[B]}{E[B] + E[I]} = \frac{E[B]}{E[B] + 1/\lambda} \Rightarrow \underline{\underline{E[B] = \frac{\rho}{1 - \rho} \cdot \frac{1}{\lambda} = \frac{E[S]}{1 - \rho}}}$$

In the case of an M/M/1 queue this the same as $E[T]$, i.e. mean sojourn time!

Appendix

Mean number of customers served during a busy period

A busy period consists of full service times of a set of customers.

Let the number of customers served during the busy period be N_b . We deduce the expectation $E[N_b]$.

- The first customer of a busy period finds the system empty, the others find it non-empty.
- Thus an arriving customer finds the system empty with probability $1/E[N_b]$.
- The probability that the system is empty at an arbitrary instant is $1 - \rho$.
- By the PASTA property, these probabilities are equal.

$$\Rightarrow \quad \overline{\overline{E[N_b] = \frac{1}{1 - \rho}}}$$

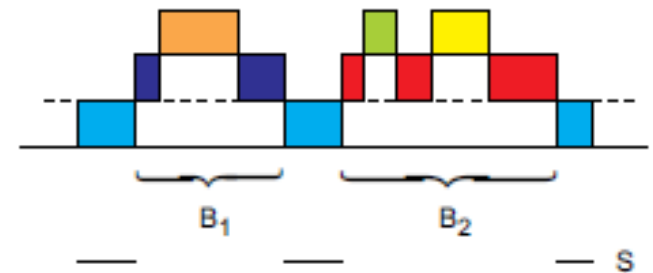
Since the mean service time of the customer is $E[S]$, it follows further that

$$E[B] = \frac{E[S]}{1 - \rho}$$

Distribution of the length of the busy period in an M/G/1 queue

Denote $\left\{ \begin{array}{l} B = \text{length of busy period} \\ S = \text{service time of the customer starting the busy period} \\ V = \text{number of customers arriving during this service time} \end{array} \right.$

The duration of the busy period is independent of the scheduling discipline provided this is work conserving. We can choose scheduling as we wish. It is easiest to consider a stack, i.e. a LIFO queue.



- The first arrival within the busy period interrupts the service of the customer who started the busy period.
- By considering the period starting from this instant of interruption to the point when the service of the first customer is resumed, we notice that this period itself forms a busy period which is distributed in the same way as B , we call it a “mini busy period”.
- **It may be paradoxical that inside a busy period there are subperiods with the same distribution as the busy period itself.** However, their expected number is < 1 .
- The number of mini busy periods is V : *the service of the first customer is executed in pieces, each of them having a duration $\sim \text{Exp}(\lambda)$ (except for the last one); thus the number of mini busy periods is the same as the number of borderlines of the pieces, which equals the number of arrivals from a $\text{Poisson}(\lambda)$ process during the interval S .*

Distribution of the length of the busy period (continued)

$$B = S + B_1 + B_2 + \dots + B_\nu, \quad \nu = 0, 1, 2, \dots, \quad B \sim B_1 \sim B_2 \sim \dots \sim B_\nu$$

$$\begin{aligned} B^*(s) &= E[e^{-sB}] = E[E[E[e^{-sB} | \nu, S] | S]] = E[E[E[e^{-s(S+B_1+B_2+\dots+B_\nu)} | \nu, S] | S]] \\ &= E[E[e^{-sS} E[e^{-s(B_1+B_2+\dots+B_\nu)} | \nu, S] | S]] = E[E[e^{-sS} \underbrace{E[e^{-sB}]^\nu}_{B^*(s)} | S]] \\ &= E[e^{-sS} \underbrace{E[B^*(s)^\nu | S]}_{e^{-(1-B^*(s))\lambda S}}] = E[e^{-(s+\lambda(1-B^*(s)))S}] \end{aligned}$$

$$\underline{B^*(s) = S^*(s + \lambda - \lambda B^*(s))}$$

Tak'acs' equation (functional equation) for $B^*(s)$

Example: the first moment of B

$$\begin{aligned} E[B] &= -B^*(0)' = \underbrace{S^*(0)'}_{-E[S]} (1 - \lambda \underbrace{B^*(0)'}_{-E[B]}) \\ \Rightarrow E[B] &= \frac{E[S]}{1 - \rho}, \quad \text{where} \quad \rho = \lambda E[S] \end{aligned}$$

This is in accordance with our earlier results.

In a similar way, one can derive higher moments of B .