$M/G/1$ queue

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

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 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).

$$\mathrm{E}[W] = \underbrace{\underbrace{\mathrm{E}[N_q]}_{\substack{\text{number of wait-}\\ \text{ing customers}}} \cdot \underbrace{\mathrm{E}[S]}_{\text{mean service time}}}_{\substack{\text{mean time needed to serve the}\\ \text{customers ahead in the queue}}} + \underbrace{\mathrm{E}[R]}_{\substack{\text{unfinished work}\\ \text{in the server}}} \qquad (R = \text{residual service time})$$

- 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 expecation of N_q (number of waiting customers) at the instant of arrival.
- Due to the PASTA property of Poison process, the distributions seen by the arriving cuatomer are the same as those at an arbitrary instant.

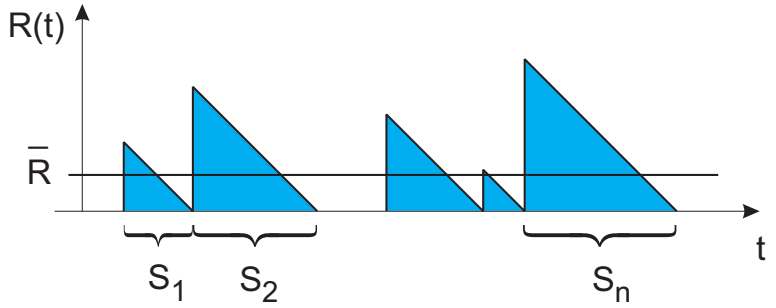
The key observation is that by Little's result the mean queue length $\mathrm{E}[N_q]$ can be expressed in terms of the waiting time (by considering the waiting room as a black box)

$$\mathrm{E}[N_q] = \lambda \mathrm{E}[W] \quad \Rightarrow \quad \boxed{\mathrm{E}[W] = \frac{\mathrm{E}[R]}{1-\rho}}$$

It remains to determine $\mathrm{E}[R]$.
$\rho = \lambda \mathrm{E}[S]$

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 reperesents 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.

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

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

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

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

Mean waiting and sojourn times

$$\begin{cases} \mathrm{E}[W] &= \dfrac{\lambda \mathrm{E}[S^2]}{2(1-\rho)} &= \dfrac{1+C_v^2}{2} \cdot \dfrac{\rho}{1-\rho} \cdot \mathrm{E}[S] \\ \mathrm{E}[T] &= \mathrm{E}[S] + \dfrac{\lambda \mathrm{E}[S^2]}{2(1-\rho)} &= \Big(1 + \dfrac{1+C_v^2}{2} \cdot \dfrac{\rho}{1-\rho}\Big) \cdot \mathrm{E}[S] \end{cases}$$

Squared coefficient of variation C_v^2

$$\begin{aligned} C_v^2 &= \mathrm{V}[S]/\mathrm{E}[S]^2 \\ \mathrm{E}[S^2] &= \mathrm{V}[S] + \mathrm{E}[S]^2 \\ &= (1+C_v^2) \cdot \mathrm{E}[S]^2 \end{aligned}$$

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

Mean number of waiting customers and customers in system

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

Remarks on the PK mean formulae

- Mean values depend only on the expectation $\mathrm{E}[S]$ and variance $\mathrm{V}[S]$ 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 $(1 + C_v^2)/2$.

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

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

$$\begin{cases} \mathrm{E}[N] &= \rho + \dfrac{\rho^2}{1-\rho} &= \dfrac{\rho}{1-\rho} \\ \mathrm{E}[T] &= \Big(1 + \dfrac{\rho}{1-\rho}\Big) \cdot \mathrm{E}[S] &= \dfrac{1}{1-\rho} \cdot \mathrm{E}[S] \end{cases}$$

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

$M/D/1$ queue

In the case of constant service time one has

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

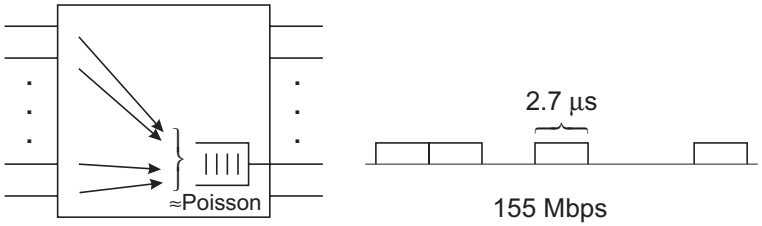
$$\begin{cases} \mathrm{E}[N] &= \rho + \dfrac{1}{2}\dfrac{\rho^2}{1-\rho} \\ \mathrm{E}[T] &= \Big(1 + \dfrac{1}{2}\dfrac{\rho}{1-\rho}\Big) \cdot \mathrm{E}[S] \end{cases}$$

A factor $1/2$ in the "waiting room terms"

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 transmission time is $S = 53 \cdot 8/155\ \mu s = 2.7 \mu s$.

What is the mean number of cells in the buffer (including the cell being transmitted) ant 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 $\rho = 124/155 = 0.8$.

Then

$$\begin{cases} \mathrm{E}[N] &= 0.8 + \dfrac{1}{2} \cdot \dfrac{0.8^2}{1 - 0.8} = 2.4 \\ \mathrm{E}[T] &= \Big(1 + \dfrac{1}{2} \cdot \dfrac{0.8}{1 - 0.8}\Big) 2.7\,\mu\mathrm{s} = 8.1\,\mu\mathrm{s} \end{cases}$$

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

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

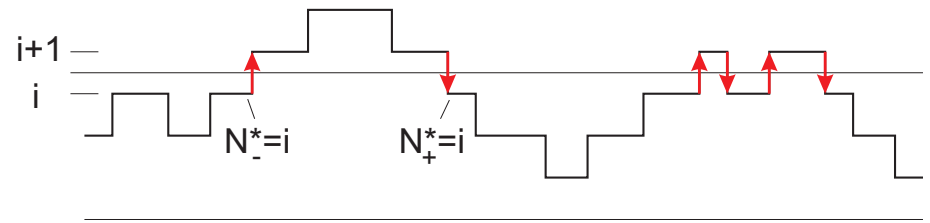
By the PASTA property of Poisson arrivals we have $\boxed{N_-^* \sim N}$

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

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

(so called level crossing property)

Proof:



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

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

Embedded Markov chain (continued)

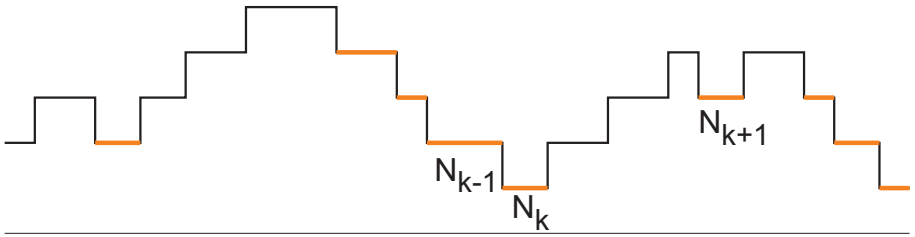
We have shown that $N_+^* \sim N_-^*$ ja $N_-^* \sim N$. $\Rightarrow$ $\boxed{N_+^* \sim N}$

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

$$\begin{cases} N_k &= \text{queue length after the departure od customer } k \\ V_k &= \text{number of new customers arrived during the service time of customer } k. \end{cases}$$



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 and of a random variable V_{k+1} which is independent of N_k and its history:

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

- 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 $\boxed{N_{k+1} = \hat{N}_k + V_{k+1}}$

Upward jumps can be arbitrary large.
Downward one step at a time.

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

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

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

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

$\boxed{\mathcal{G}_N(z) = \mathcal{G}_{\hat{N}}(z) \cdot \mathcal{G}_V(z)}$ The task now is to determine $\mathcal{G}_{\hat{N}}(z)$ and $\mathcal{G}_V(z)$.

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

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

We have obtained the result

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

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) occuring in the interval X, and, in particular, its generating function $\mathcal{G}_K(z)$.

$$\mathcal{G}_K(z) = \mathrm{E}[z^K] = \mathrm{E}[\underbrace{\mathrm{E}\left[z^K \,|\, X\right]}_{K\sim\mathrm{Poisson}(\lambda X)}] = \mathrm{E}[e^{-(1-z)\lambda X}] = X^*((1-z)\lambda) \qquad | \quad X^*(s) = \mathrm{E}[e^{-sX}]$$

Generally, $\boxed{\mathcal{G}_K(z) = X^*((1-z)\lambda)}$ In particular, $\boxed{\mathcal{G}_V(z) = S^*((1-z)\lambda)}$

The result can be derived also in a more elementary way

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

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 $\mathcal{G}_K(z)$ has the interpretation of the probability that none of the K arrivals occuring 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:

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

- 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,

$$\mathcal{G}_N(z) = \mathcal{G}_{\hat{N}}(z) \cdot \mathcal{G}_V(z) = \frac{\mathcal{G}_N(z) - (1-\rho)(1-z)}{z} \cdot S^*((1-z)\lambda)$$

From this we can solve $\mathcal{G}_N(z)$

$$\boxed{\mathcal{G}_N(z) = \frac{(1-\rho)(1-z)}{S^*((1-z)\lambda) - z} \cdot S^*((1-z)\lambda) = \frac{(1-\rho)(1-z)}{1 - z/S^*((1-z)\lambda)}}$$

Example. $M/M/1$ queue

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

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

$$\mathcal{G}_N(z) = \frac{(1-\rho)(1-z)}{1 - z((1-z)\rho + 1)} = \frac{(1-\rho)(1-z)}{(1-z)(1-\rho z)} = \frac{1-\rho}{1-\rho z}$$

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

$M/G/1$ 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 occuring in an interval having a given length distribution

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

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

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

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

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

Pollaczek-Khinchin transform formula for the sojourn time

Example. $M/M/1$ queue

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

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

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

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

The following is true generally $\quad T = \underbrace{W}_{\text{wait}} + \underbrace{S}_{\text{service}}$

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 $T^*(s)$,

$$\boxed{W^*(s) = \frac{(1-\rho)s}{s - \lambda + \lambda S^*(s)}}$$ Pollaczek-Khinchin transform formula for the waiting time

The expression can also be rewritten in the form

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

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

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

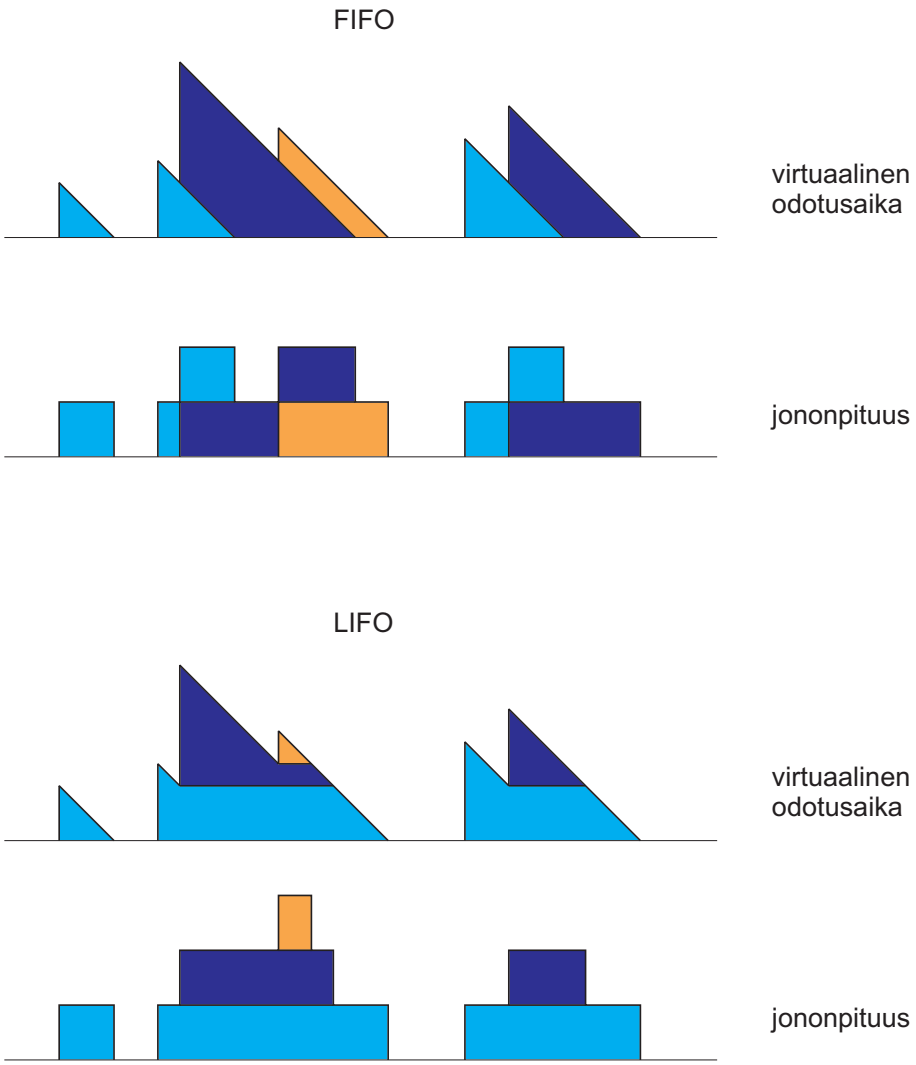
Interpretation of the formula for $W^*(s)$

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

- 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 unfinshed 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 rsedual 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.

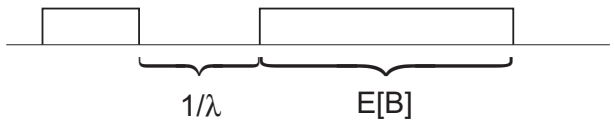
Virtual waiting time (unfinished work) is independent of the scheduling discipline

- 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_t 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 $\begin{cases} B &= \text{length of busy period} \\ I &= \text{length of idle period} \end{cases}$

In a Poisson process the interarrival times are distributed according to $\text{Exp}(\lambda)$. 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, $I \sim \text{Exp}(\lambda)$, thence $\boxed{\text{E}[I] = 1/\lambda}$.

By Little's result, the load of the server $\rho = \lambda \text{E}[S]$ 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:

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

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

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 $\mathrm{E}[N_b]$.

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

$$\Rightarrow \qquad \boxed{\mathrm{E}[N_b] = \frac{1}{1-\rho}}$$

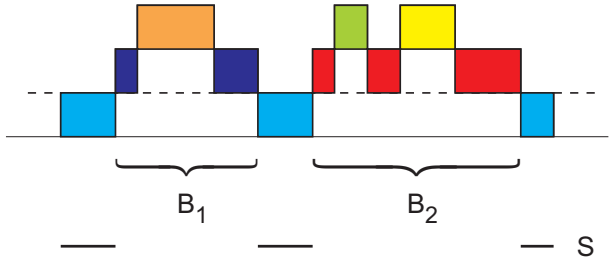
Since the mean service time of the customer is $\mathrm{E}[S]$, it follows further that

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

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

Denote
$$\begin{cases} B &= \text{length of busy period} \\ S &= \text{service time of the customer starting the busy period} \\ V &= \text{number of customers arrving during this service time} \end{cases}$$

The duration of the busy period is independent of the scheduling discipline provided this is work conserving. We can choose schduling 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 $\sim$ Exp(λ) (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 Poisson(λ) process during the interval S.

Distribution of the length of the busy period (continued)

$$B = S + B_1 + B_2 + \cdots + B_V, \qquad V = 0, 1, 2, \ldots, \qquad B \sim B_1 \sim B_2 \sim \cdots \sim B_V$$

$$\begin{aligned} B^*(s) &= \mathrm{E}[e^{-sB}] = \mathrm{E}[\mathrm{E}\left[\mathrm{E}\left[e^{-sB} \,|\, V, S\right] | \, S\right]] = \mathrm{E}[\mathrm{E}\left[\mathrm{E}\left[e^{-s(S+B_1+B_2+\cdots+B_V)} \,|\, V, S\right] | \, S\right]] \\ &= \mathrm{E}[\mathrm{E}\left[e^{-sS}\mathrm{E}\left[e^{-s(B_1+B_2+\cdots+B_V)} \,|\, V, S\right] | \, S\right]] = \mathrm{E}[\mathrm{E}[e^{-sS}\underbrace{\mathrm{E}[e^{-sB}]}_{B^*(s)}{}^V \mid S]] \\ &= \mathrm{E}[e^{-sS}\underbrace{\mathrm{E}\left[B^*(s)^V \,|\, S\right]}_{e^{-(1-B^*(s))\lambda S}}] = \mathrm{E}[e^{-(s+\lambda(1-B^*(s)))S}] \end{aligned}$$

$$\boxed{B^*(s) = S^*(s + \lambda - \lambda B^*(s))}$$ Takács' equation (functional equation) for $B^*(s)$

Example: the first moment of B

$$\mathrm{E}[B] = -B^{*\prime}(0) = \underbrace{S^{*\prime}(0)}_{-\mathrm{E}[S]}(1 - \lambda \underbrace{B^{*\prime}(0)}_{-\mathrm{E}[B]})$$

$$\Rightarrow \quad \mathrm{E}[B] = \frac{\mathrm{E}[S]}{1-\rho}\,, \qquad \text{where } \rho = \lambda \mathrm{E}[S]$$

This is in accordance with our earlier results.
In a similar way, one can derive higher moments of B.

$M/G/1$ queue: algorithmic approach to the queue length distribution

Previously we have derived a result, the Pollaczek-Khinchin transform formula, for the generating function of the queue length distribution.

- The result is theoretically important.
- It can be used, e.g. to derive moments of the distribution.
- However, the formula is not very practical for computing the distribution itself (probabilities of different queue lengths) because the dependence on z is quite complicated.

The queue length distribution, however, can be determined algorithmically (not as a closed form formula) quite straight forwardly. We will derive the algorithm in the following.

To this end, consider again the embedded Markov chain N_k, i.e. the queue lengths immediately after the departure epochs. We have found that this chain evolves as follows:

$$N_{k+1} = \begin{cases} N_k - 1 + V & \text{when} \quad N_k \geq 1 \\ V & \text{when} \quad N_k = 0 \end{cases}$$

where V is the number of arrivals of new customers during the service time of customer $k+1$.

Distribution of the number of arrivals, V, during the service time

Denote

$$\begin{cases} k_i & = \mathrm{P}\{V = i\} \\ f_S(x) & = \text{the pdf of the service time } S \end{cases}$$

According to the law of the total probability, it holds

$$k_i = \mathrm{P}\{V = i\} = \int_0^\infty \mathrm{P}\{V = i \,|\, S = x\} f_S(x)dx = \int_0^\infty \frac{(\lambda x)^i}{i!} e^{-\lambda x} f_S(x)dx$$

$$\boxed{k_i = \int_0^\infty \frac{(\lambda x)^i}{i!} e^{-\lambda x} f_S(x)dx} \qquad i = 0, 1, \ldots$$

When the arrival intensity λ and the pdf of the service time $f_S(x)$ are given, then the probabilities k_i can be calculated, at least numerically. In the case of some simple distributions, the integration can be done analytically.

Example. Exponential service time distribution ($M/M/1$ queue)

$$k_i = \int_0^\infty \frac{(\lambda x)^i}{i!} e^{-\lambda x} \mu e^{-\mu x} dx = \Big(\frac{\lambda}{\lambda+\mu}\Big)^i \frac{\mu}{\lambda+\mu} \frac{1}{i!} \underbrace{\int_0^\infty y^i e^{-y} dy}_{i!} = \Big(\frac{\lambda}{\lambda+\mu}\Big)^i \frac{\mu}{\lambda+\mu} \quad \Big| \quad \begin{array}{l}\text{change of variable} \\ y = (\lambda+\mu)x\end{array}$$

Geometrical distribution: "λ and μ compete"; "λ wins" i times, until it "loses", i.e. "μ wins".

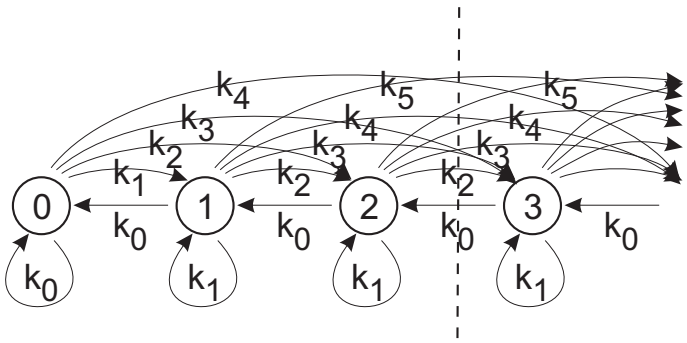
Transition probability matrix of the embedded chain

From the evolution equation of the chain N_k one can identify the transition probabilities

$$p_{i,j} = \mathrm{P}\{N_{k+1} = j \,|\, N_k = i\} = \begin{cases} \mathrm{P}\{V = j-i+1\} &= k_{j-i+1} & \text{if } i \geq 1 \\ \mathrm{P}\{V = j-i\} &= k_{j-i} & \text{if } i = 0 \end{cases}$$

The state transition matrix and diagram are thus

$$\mathbf{P} = \begin{pmatrix} k_0 & k_1 & k_2 & k_3 & \cdots \\ k_0 & k_1 & k_2 & k_3 & \cdots \\ 0 & k_0 & k_1 & k_2 & \cdots \\ 0 & 0 & k_0 & k_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$



The method of a cut yields the recursion equations

$$\begin{cases} k_0\pi_1 &= (k_1 + k_2 + \ldots)\pi_0 \\ k_0\pi_2 &= (k_2 + k_3 + \ldots)\pi_0 + (k_2 + k_3 + \ldots)\pi_1 \\ k_0\pi_3 &= (k_3 + k_4 + \ldots)\pi_0 + (k_3 + k_4 + \ldots)\pi_1 + (k_2 + k_3 + \ldots)\pi_2 \\ \vdots & \\ k_0\pi_i &= (k_i + k_{i+1} + \ldots)\pi_0 + (k_i + k_{i+1} + \ldots)\pi_1 + \ldots + (k_2 + k_3 + \ldots)\pi_{i-1} \end{cases}$$

from which the π_i can be solved consecutively, starting from the known value $\pi_0 = 1 - \rho$.

Recursion (continued)

Denote: $a_i = k_{i+1} + k_{i+2} + k_{i+3} + \ldots = \mathrm{P}\{V > i\}$

i.e. a_i is the probability that during the service time S there are at least $i+1$ arrivals.

In terms of this, the general recursion step can be written as

$$\boxed{\pi_i = \frac{1}{k_0}\left(a_{i-1}\pi_0 + \sum_{j=1}^{i-1} a_{i-j}\pi_j\right)}$$

Since $k_0 + k_1 + k_2 + \ldots = 1$, we have
$k_0 = 1 - (k_1 + k_2 + \ldots) = 1 - a_0$.

The recursion begins from $\pi_0 = 1 - \rho$.

Recursion (continued)

One need not evaluate the sums of the series in the definition of a_i, but we can derive a single integral expression for these coefficients. Denote by X_j a general interarrival time in a Poisson process with intensity λ, $X_j \sim \text{Exp}(\lambda)$. Then, by the definition of a_i we have

$$a_i = \text{P}\{S > X_1 + X_2 + \cdots + X_{i+1}\}$$

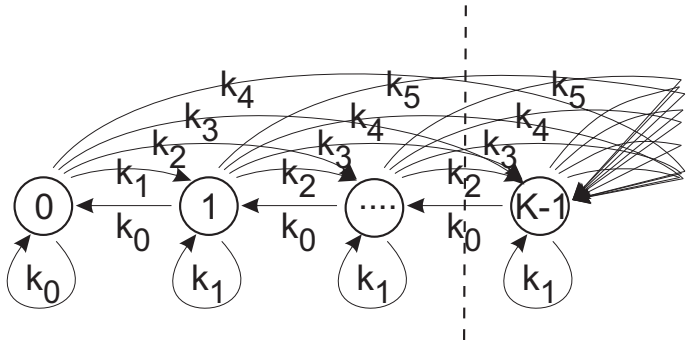
The sum $X_1 + X_2 + \cdots + X_{i+1}$ is distributed as Erlang$(i+1, \lambda)$ with the pdf $\lambda e^{-\lambda x}(\lambda x)^i/i!$.

Given that the sum has the value x, the probability is given by the tail distribution of S, $G_S(x) = 1 - F_S(x)$. By the law of the total probability, it then follows that

$$\boxed{a_i = \int_0^\infty G_S(x) \frac{(\lambda x)^i}{i!} \lambda e^{-\lambda x} dx = \int_0^\infty G_S(y/\lambda) \frac{y^i}{i!} e^{-y} dy}$$

From this the a_i can be computed numerically.

Finite $M/G/1/K$ queue



The queue has K system places (server + waiting places).

The queue length distribution can be derived from that of the infinite system. To this end, we need a few deduction steps.

Consider again the embedded Markov chain constituted by the queue length, N, left by a departing customer.

The first observation is that N cannot be grater than $K-1$. Before the departure the queue may have been full, $N = K$, but after the departure there is one less customer.

The second observation is that up to the state $K-1$ the state transition diagram is precisely the same as in the case of an infinite system. The only difference is that, all the transitions, which in the infinite system had taken the system to a state $N > K-1$, will now imply that during the service time of the departed customer the queue has become full and overflown. In all the cases, after the departure, the system will be in state $K-1$ as shown in the figure.

The key thing is that the transitions across all the cuts between states lower than $K-1$ are precisely the same as in the infinite system. The ratios of the state probabilities are as before. The distribution is obtained by normalization.

Finite $M/G/1/K$ queue (continued)

Denote

$$\begin{cases} \pi_i^{(\infty)} &= \text{probability that a departing customer leaves queue } i \text{ in an infinite system} \\ \pi_i^{(K)} &= \text{probability that a departing customer leaves queue } i \text{ in a finite system} \end{cases}$$

As said, the finite system probabilities are obtained by simply normalizing

$$\boxed{\pi_i^{(K)} = \frac{\pi_i^{(\infty)}}{\sum_{j=0}^{K-1} \pi_j^{(\infty)}}} \qquad i = 0, 1, \ldots, K-1$$

The probabilities $\pi_i^{(\infty)}$ are computed with the algorithm given before.

By the level crossing argument, one deduces that $\pi_i^{(K)}$ is also the probability that an arriving customer accepted into the queue finds the queue length i upon arrival:

- the frequency of transitions from state i to state $i+1$ equals the frequency of transitions from state $i+1$ to state i
- now we just have to observe that not all arriving customers are accepted to the queue, and thus $\pi_i^{(K)}$ is not the state probability as seen by an arriving customer

Finite $M/G/1/K$ queue (continued)

Denote the, thus far unknown, state probabilities as seen by an arriving customer by p_i, $i = 0, 1, \ldots, K$ (notice that upon arrival the queue can be full, i.e. in the state K). Due to the PASTA property p_i is also the equilibrium probability at an arbitrary instant.

Denote

$$\begin{cases} X^* &= \text{queue length seen by arriving customer} \\ A &= \text{customer is accepted in the queue} = \{X^* < K\} \\ \bar{A} &= \text{customer is rejected} = \{X^* = K\} \end{cases}$$

For $i < K$ we have

$$p_i = \mathrm{P}\{X^* = i\} = \underbrace{\mathrm{P}\{X^* = i \,|\, A\}}_{\pi_i^{(K)}} \underbrace{\mathrm{P}\{A\}}_{1-p_K} + \underbrace{\mathrm{P}\{X^* = i \,|\, \bar{A}\}}_{0} \underbrace{\mathrm{P}\{\bar{A}\}}_{p_K} = (1 - p_K)\pi_i^{(K)}, \qquad i < K$$

In order to determine the unknown p_K, we make use of the condition which says that, on average, the frequency at which customers are admitted to the queue equals the frequency with which customers depart from the system ($\rho = \lambda \mathrm{E}[S]$):

$$\lambda(1 - p_K) = (1 - p_0)/\mathrm{E}[S] = (1 - (1 - p_K)\pi_0^{(K)})/\mathrm{E}[S] \quad \Rightarrow \quad 1 - p_K = \frac{1}{\rho + \pi_0^{(K)}}$$

$$\boxed{p_i = \frac{\pi_i^{(K)}}{\rho + \pi_0^{(K)}}} \qquad i = 0, 1, \ldots, K-1 \qquad \boxed{p_K = 1 - \frac{1}{\rho + \pi_0^{(K)}}}$$

Finite $M/G/1/K$ queue (summary)

We gather all the results together, skipping the intermediate steps, and modifying the result:

First calculate recursively the queue length distribution in an infinite system:

$$\begin{cases} \pi_0^{(\infty)} = 1 - \rho \\ \pi_i^{(\infty)} = \dfrac{1}{1 - a_0} \left(a_{i-1}\pi_0^{(\infty)} + \sum\limits_{j=1}^{i-1} a_{i-j}\pi_j^{(\infty)} \right) \end{cases} \qquad a_i = \int_0^\infty G_S(x) \frac{(\lambda x)^i}{i!} \lambda e^{-\lambda x} dx$$

where $G_S(x)$ is the tail distribution of the service time S.

Then compute the tail probability (usually small) of the infinite queue

$$q_K = \sum_{i=K}^{\infty} \pi_i^{(\infty)} = 1 - \sum_{i=0}^{K-1} \pi_i^{(\infty)}$$

In terms of these, the final result reads

$$\boxed{p_i = \frac{\pi_i^{(\infty)}}{1 - q_K\rho}} \quad i = 0, 1, \ldots, K-1 \qquad \boxed{p_K = \frac{(1-\rho)q_K}{1 - q_K\rho}}$$