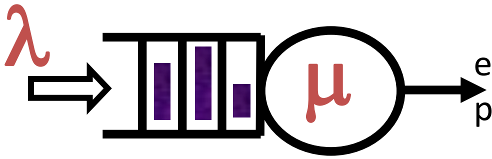
Waiting Systems
Poisson arrival/Exp. Service time

Dr Ahmad Khonsari
ECE Dept.
The University of Tehran
Excerpt from virtamo slides

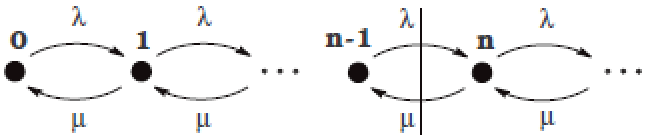
The M/M/1 queue

Number of customers in an M/M/1 queue

Poisson
Arrivals



exponential
period of service



By the method of a cut, one gets the balance condition

$\lambda\pi_{n-1} = \mu\pi_n$ or $\pi_n = \rho\pi_{n-1}$ where $\rho = \lambda/\mu$ (traffic intensity, offered load),

from which we get recursively

$\pi_n = \rho^n\pi_0$ (in order for the queue to be stable, we have to require *< 1)*

The probability of an empty queue 0 is obtained from the normalization condition

$$\pi_0 + \pi_1 + \pi_2 + \ldots = 1$$

$$\pi_0 = 1/\sum_{n=0}^{\infty}\rho^n = 1 - \rho$$

(the probability that the server (ant the queue) is empty $= 1 - \rho \Rightarrow$
probability that the server is busy = ρ)

The queue length distribution of an *M/M/1 queue,* $\pi_n = \mathrm{P}\{\mathrm{N} = \mathrm{n}\}$,

$\underline{\overline{\pi_n = (1-\rho)\,\rho^n}}$ *n = 0, 1, . . .* $\mathrm{Geom}_0(\rho)$ distribution (starts from 0)

The average number of customers in the system

$$\mathrm{E}[N] = \sum_{i=0}^{\infty} i\pi_i \quad = (1-\rho)\sum_{i=0}^{\infty} i\rho^i = (1-\rho)\rho\frac{d}{d\rho}\sum_{i=0}^{\infty}\rho^i \quad = (1-\rho)\rho\frac{d}{d\rho}\left(\frac{1}{1-\rho}\right)$$

$= \dfrac{\rho}{1-\rho}$ the mean of the $\mathrm{Geom}_0(\rho)$ distribution (starts from 0)

$$\mathrm{E}[N] = \frac{\rho}{1-\rho} = \underbrace{\rho}_{\text{customers in the server}} + \frac{\rho^2}{\underbrace{1-\rho}_{\text{waiting customers}}}$$

$$E[N^2] = \textstyle\sum_{i=0}^{\infty} i^2\pi_i = (1-\rho)\sum_{i=0}^{\infty} i^2\rho^i = (1-\rho)\frac{\rho+\rho^2}{(1-\rho)^3} = \frac{\rho+\rho^2}{(1-\rho)^2}$$

==

$\sum_{i=0}^{\infty}\rho^i = 1 + \rho + \rho^2 + \cdots = \frac{1}{1-\rho}$ $\quad$ $\sum_{i=1}^{\infty}\rho^i = 1 + \rho + \rho^2 + \cdots = \rho\,(1 + \rho + \rho^2 + \cdots) = \rho(\frac{1}{1-\rho})$

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

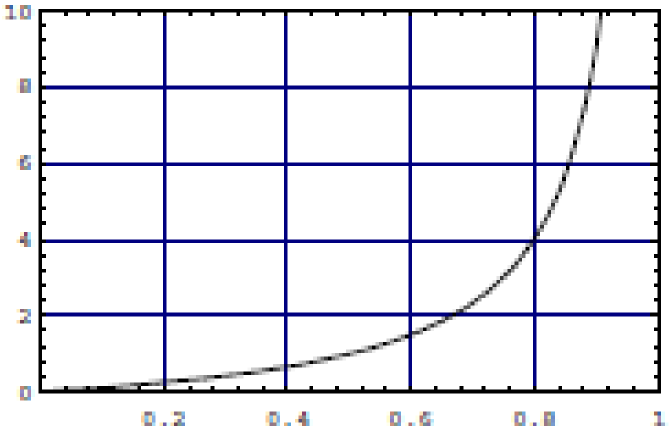
The average number of customers in the system

Q: what is VAR[N]

A: $Var[\mathrm{N}] = E[N^2] - E[N]^2 = \frac{\rho+\rho^2}{(1-\rho)^2} - \left(\frac{\rho}{1-\rho}\right)^2 = \frac{\rho+\rho^2}{(1-\rho)^2} - \frac{\rho^2}{(1-\rho)^2} = \frac{\rho}{(1-\rho)^2}$

The tail probability: the probability that there are at least *n customers in the system,*

$P\{N \geq n\} = \sum_{i=n}^{\infty} \pi_i = (1-\rho)\sum_{i=n}^{\infty} \rho^i = (1-\rho)\sum_{k=0}^{\infty} \rho^{k+n}$
$= \rho^n (1-\rho)\sum_{k=0}^{\infty} \rho^k \quad = \rho^n (1-\rho)\frac{1}{(1-\rho)} = \rho^n$



Using Tail Sum formula and

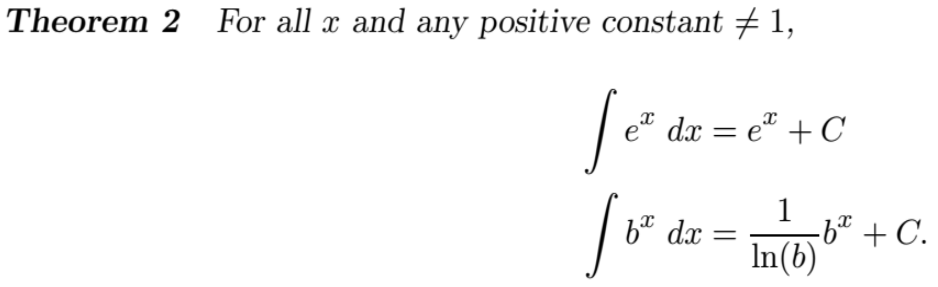
for power series $\sum_{n=0}^{\infty} \rho^n = \frac{1}{(1-\rho)}$ for $|\rho| < 1$:

E[N]= $\sum_{n=0}^{\infty} P\{N > n\} = \rho + \rho^2 + \cdots = \rho\,\frac{1}{(1-\rho)}$ or E[N]= $\sum_{n=1}^{\infty} P\{N \geq n\} = \rho + \rho^2 + \cdots = \rho\,\frac{1}{(1-\rho)}$

##

$\int \rho^n \, d\rho = \frac{\rho^{n+1}}{n+1}$,

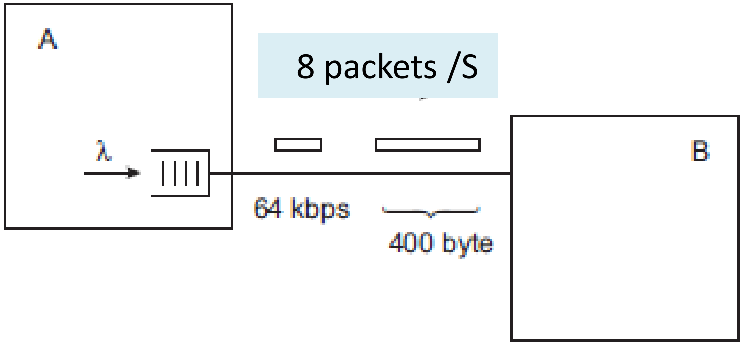
$$\int_0^{\infty} P\{Nn\} \leq dn = \int_0^{\infty} \rho^n \, dn = \left[\frac{\rho^n}{\ln(\rho)}\right]_0^{\infty}$$



Theorem 2 *For all x and any positive constant $\neq 1$,*

$$\int e^x \, dx = e^x + C$$
$$\int b^x \, dx = \frac{1}{\ln(b)} b^x + C.$$

Example.



- Router A sends 8 packets per second, on the average, to router B.
- The mean size of a packet is 400 byte (exponentially distributed).
- The line speed is 64 kbit/s.

How many packets are there on the average in router A waiting for transmission or being transmitted and what is the probability that the number is 10 or more?

The utilization of the line (server) is

$$\rho = (8\,\mathrm{s}^{-1} \times 400 \times 8\,\mathrm{bit})/(64 \times 10^3\,\mathrm{bit\ s}^{-1}) = 0.4.$$

This can be also calculated in the form λ/μ, *where*

$\lambda = 8$ packets/s, $\mu = 64$ kbit/s/$(400 \times 8$ bit/packet$) = 20$ packets/s => $\lambda / \mu = 8/20 = 0.4$

Thus E[*N] = 0.4/(1 – 0.4) = 0.67.*

Using tail probability $P\{N \geq n\} = \rho^n$

The probability that the number of packets is 10 or more is $\underline{0.4^{10} = 10^{-4}}$

Sojourn and waiting times in the M/M/1 queue

Little’s result:

The average sojourn time (time in system) $E[T] = E[N]/\lambda$

The average waiting time $E[W] = (E[N] - \rho)/\lambda$

$$E[T] = \frac{\rho}{1-\rho}/\lambda = \frac{1}{1-\rho} \cdot \frac{1}{\mu} = \frac{1}{\mu - \lambda}$$

$$E[W] = (\frac{\rho}{1-\rho} - \rho)/\lambda = (\frac{1}{1-\rho} - 1)\rho/\lambda = \frac{\rho}{1-\rho} \cdot \frac{1}{\mu} = \frac{\lambda}{\mu(\mu - \lambda)}$$

Independence of the scheduling discipline

For the *M/M/1-FIFO queue we have derived the queue length distribution* $\pi_n = (1-\rho)\rho^n$

- This distribution is independent of the scheduling discipline (FIFO, LIFO, PS),

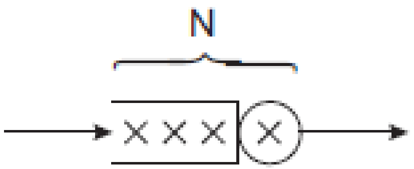
– all these scheduling disciplines lead to exactly the same balance equations (proof is left as an exercise)

- Thus also the mean time in system, $E[T] = 1/(\mu - \lambda)$ *is independent of the discipline* (by Little's result the mean time in system equals the mean queue length divided by λ)
- In contrast, e.g. the distributions of *W and T do depend on the discipline.*

Note. The queue length distribution is not insensitive to the service time distribution in an *M/M/1-FIFO queue. However, in LIFO and PS queues the insensitivity holds.*

The distribution of the sojourn time

Assume that an arriving customer finds *N customers in the* system (including the customer in the server, if any).



By the memoryless property of the exponential distribution also the remaining service time of the customer in service (if any) is distributed as $\sim Exp(\mu)$.

The time *T spent by a customer in the system consists of the time it takes to serve the* customers ahead in the queue and the customer's own service time

$$T = \underbrace{S_1' + S_2 + \ldots + S_N}_{customers} + \underbrace{S_{N+1}}_{own}$$

sum of (*N + 1) rvs with Exp(μ) distribution*

$$\begin{cases} S_i \sim Exp(\mu) \\ N \sim Geom_0(\rho) \end{cases}$$

independent

equilibrium distribution of the queue length (starts from 0), PASTA!

$$f_T(t) = \sum_{n=0}^{\infty} f_{T|N}(t,n) P\{N = n\} = \sum_{n=0}^{\infty} \overbrace{\mu \frac{(\mu t)^n}{n!} e^{-\mu t}}^{Erlang(n+1,\mu)} (1-\rho)\rho^n$$

$$E_n(\mu)$$

$$f(t) = \mu \frac{(\mu t)^{n-1}}{(n-1)!} e^{-\mu t}, \qquad t > 0.$$

$$= \mu(1-\rho)e^{-\mu t} \sum_{n=0}^{\infty} \frac{(\mu\rho t)^n}{n!} = \mu(1-\rho)e^{-\mu(1-\rho)t}$$

$$\underline{\overline{f_T(t) = (\mu - \lambda)e^{-(\mu-\lambda)t}}}$$

exponential distribution $\mathrm{Exp}(\mu - \lambda)$

$$S \equiv S_\infty = \sum_{k=0}^{\infty} r^k = \frac{1}{1-r}$$

The distribution of the sojourn time (continued)

The same result can be derived also by using the result for the Laplace transform of a random Sum.

$$\begin{cases} \mathcal{L}_{N+1}(z) = \dfrac{(1-\rho)z}{1-\rho z} \\ f_S^*(s) = \dfrac{\mu}{\mu+s} \end{cases} \qquad N+1 \sim Geom(1-\rho), \qquad \text{starts from 1}$$

$$f_T^*(s) = \mathcal{L}_{N+1}(f_S^*(s)) = \frac{(1-\rho)\dfrac{\mu}{\mu+s}}{1-\rho\dfrac{\mu}{\mu+s}} = \frac{\mu-\lambda}{(\mu+s)-\lambda} = \frac{(\mu-\lambda)}{(\mu-\lambda)+s}$$

$$\Rightarrow \quad \sim Exp(\mu-\lambda)$$

Distribution of the waiting time

The waiting time *W consists of the service times of the customers in the system upon the* arrival

$W = S_1' + S_2 + \ldots + S_N,$ where $\mathrm{S_i} \sim \mathrm{Exp}(\mu)$ and *N* $\sim Geom_0(1-\rho)$ *(starts from 0)*

If *N = 0 there are no terms in the sum and W = 0.*
The tail distribution of *W is derived by conditioning*

$$P\{W > t\} = \underbrace{P\{W > t|N = 0\}P(N = 0)}_{0} + P\{W > t|N > 0\}\underbrace{P\{N > 0\}}_{\rho}$$

$$= \rho . P\{W > t|N > 0\}$$

By the memoryless property of the geometric distribution *N conditioned on N > 0 is distributed as* as $Geom(\rho)$ (starts from 1)

Thus the sum $S_1' + S_2 + \ldots + S_N$ conditioned on *N > 0 is distributed precisely as* $S_1' + S_2 + \ldots + S_{N+1}$ before and obeys the distribution $Exp(\mu - \lambda)\, ; i.e.\ F_T(t) = 1 - e^{-(\mu-\lambda)t}$

$$P\{W > t\} = \rho e^{-(\mu-\lambda)t} \ ; so\ \ P\{W > 0\} = \rho$$

The waiting time is 0 with a finite probability *P{W = 0} = 1 − P{W > 0} =* $1 - \rho$

This, of course, is equal to the empty queue probability P{*N = 0*}

Finite queue: the M/M/1/K system

Let there be K *system places (waiting room + server)*

The equilibrium equations across the cuts are the same as before

$$\pi_n = \rho^n \pi_0 \qquad\qquad n = 0,1,\ldots,K$$

The only difference is in the normalization

$$\sum_{n=0}^{K} \pi_n = 1 \qquad \Rightarrow \qquad \pi_0 = (1+\rho+\ldots+\rho^K)^{-1} = \frac{1-\rho}{1-\rho^{K+1}}$$

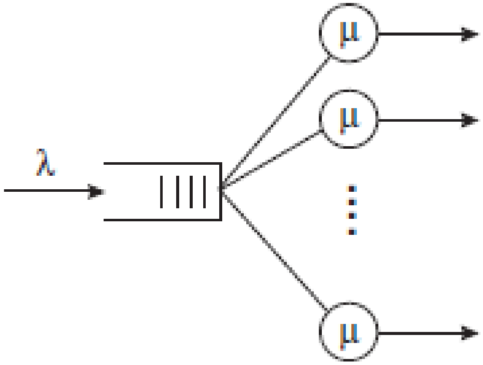
$$\pi_n = \frac{\rho^n}{1+\rho+\ldots+\rho^K} = \frac{1-\rho}{1-\rho^{K+1}}\rho^n \quad n = 0,1,\ldots,K$$ trunc. geom. distribution

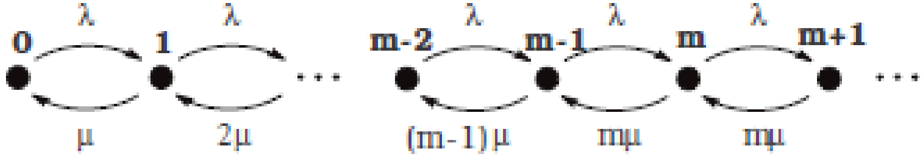
- The probability K *of state K is the probability that an arriving customer finds the* system full (“the buffer overflows”).
- When *K = 1, we have a single server loss system ,*

$$\pi_n = \frac{\rho^n}{1+\rho} \qquad n = 0,1$$

The M/M/m queue (Erlang's waiting system)

- *m parallel servers*
- Poisson arrivals
- Exponential service time distribution





- The state transition diagram is, up to state *m the same as in the loss system.*
- Beyond that state, it is identical with the diagram of an *M/M/1 queue where the capacity* of the server is *mμ.*

The balance equations can again be written by using the method of a cut:

$$\begin{cases} \lambda\pi_{n-1} = n\mu\pi_n, & n \le m \\ \lambda\pi_{n-1} = m\mu\pi_n, & n > m \end{cases}$$

The solution up to a constant factor π_0 is

$$\begin{cases} \pi_n = \pi_0 \dfrac{(m\rho)^n}{n!}, & n \le m \\ \pi_n = \pi_0 \dfrac{m^m\rho^n}{m!}, & n > m \end{cases}$$

$a = \lambda/\mu$ traffic intensity

$\rho = \lambda/m\mu = a/m$ traffic intensity per server.

The probability π_0 of state 0 is determined by the normalization condition $\sum_n \pi_n = 1$

$$\pi_0 = (\underbrace{\sum_{m=0}^{m-1} \frac{(m\rho)^n}{n!}}_{u} + \underbrace{\frac{(m\rho)^m}{m!\,(1-\rho)}}_{v})^{-1} \Rightarrow \pi_0 == (u+v)^{-1}$$

The probability P_q *that upon an arrival all servers are busy and the customer has to wait is*

$$P_q = C(m,a) = \sum_{n=m}^{\infty} \pi_n = \sum_{n=m}^{\infty} \frac{\pi_0 m^m \rho^n}{m!} = \frac{\pi_0 (m\rho)^m}{m!\,(1-\rho)} = \frac{v}{u+v}$$

Erlang's *C formula*

$a = m\rho, \rho = a/m$

The mean number of waiting customers $\overline{N}_q$

$$\overline{N}_q = \sum_{n=0}^{\infty} n\pi_{m+n} = \sum_{n=0}^{\infty} n\pi_0 \frac{m^m \rho^{m+n}}{m!} = P_q \sum_{n=0}^{\infty} n(1-\rho)\rho^n \qquad \rho = \lambda/m\mu$$

The sum is of the same form as the mean queue length in an *M/M/1 queue. Thus*

$$\overline{\underline{\overline{N}_q = P_q \frac{\rho}{1-\rho}}} \qquad \overline{N} = m\rho + \overline{N}_q \qquad \Longrightarrow \overline{\underline{\overline{N} = m\rho + P_q \frac{\rho}{1-\rho}}}$$

By Little’s result we obtain the mean waiting and sojourn times:

$$\begin{cases} W = \dfrac{\overline{N}_q}{\lambda} = P_q \dfrac{\rho}{1-\rho}/\lambda = P_q \cdot \dfrac{1}{m\mu - \lambda} \\ T = \dfrac{\overline{N}}{\lambda} = \dfrac{1}{\mu} + W = \dfrac{1}{\mu} + P_q \cdot \dfrac{1}{m\mu - \lambda} \end{cases}$$

The distribution of the waiting time

$$P\{W > t\} = \underbrace{P\{W > t|N < m\}}_{0}P\{N < m\} + P\{W > t|N \geq m\}\underbrace{P\{N \geq m\}}_{P_q}$$

$$\overline{\underline{P\{W > t\} = P_q e^{-(m\mu-\lambda)t}}}$$

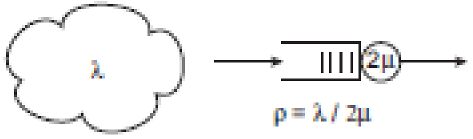
When *N m the system behaves as an M/M/1 queue* with capacity $m\mu$

Example 1

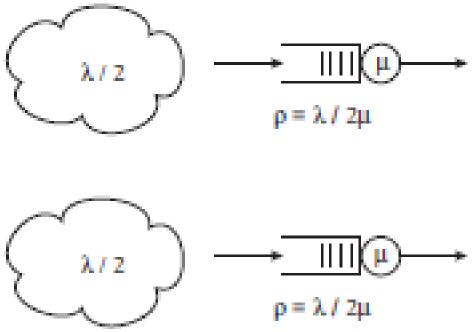
A printer is attached to the LAN of the department. The printing jobs are assumed to with arrive a Poissonian intensity λ and the actual printing times are assumed to obey the distribution Exp(μ).

The capacity of the printer has become insufficient with regard to the increased load. In order to improve the printing service, there are three alternatives:

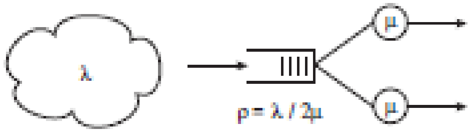
1. Replace the old printer by a new one twice as fast, i.e. with service rate 2μ.



2. Add another similar printer (service rate μ) *and divide* the users in two groups of equal size directing the works in each group to their own printer. The arrival rate of jobs to each printer is $\lambda/2$.



3. The same as alternative 2, but now there is a common printer queue where all jobs are taken and the job at the head of the queue is sent to whichever printer becomes free first.



Example 1 (continued)

Lut us compare the performance of the alternatives at different loads. As measure of performance we use the mean sojourn time of a job $\overline{T}$ (time in system, from the arrival of the printing job to the full completion of the job).

1. In this case we have an *M/M/1 queue with parameters* λ *and 2μ.* $\mathrm{E[T]} = \dfrac{1}{\mu - \lambda}$

$$\rho = \frac{\lambda}{2\mu} \qquad \overline{T}_1 = \frac{1}{2\mu - \lambda} = \frac{1}{1-\rho} \cdot \frac{1}{2\mu}$$

2. Now we have two separate *M/M/1 queues with parameters* λ */2 and μ.*

$$\rho = \frac{\lambda/2}{\mu} = \frac{\lambda}{2\mu} \qquad \overline{T}_2 = \frac{1}{\mu - \lambda/2} = \frac{1}{1-\rho} \cdot \frac{1}{\mu}$$

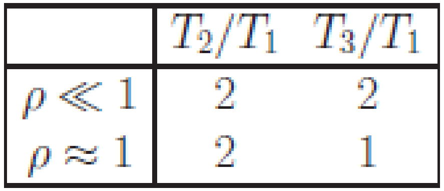
The load per server is the same as before. Now just everything happens two times slower (both arrivals and the service).

3. In the case of a common printing queue, an appropriate model is the *M/M/2 queue with* parameters λ *and μ.*

$$\rho = \frac{\lambda}{2\mu} \qquad T_3 = \frac{1}{\mu} + P_q \frac{1}{2\mu - \lambda} \approx \begin{cases} \dfrac{1}{\mu} & \rho << 1 \\ \dfrac{1}{1-\rho} \cdot \dfrac{1}{2\mu} & \rho \approx 1 \end{cases}$$

Example 1: Summary of the comparison

Take case 1 as the reference: calculate the sojourn times in cases 2 and 3 in relation to that in case 1



	T_2/T_1	T_3/T_1
$\rho \ll 1$	2	2
$\rho \approx 1$	2	1

- Alternative 1, i.e. one fast printer is the best one.
- In alternative 2, the sojourn time is twice as long as in case 1.
- In case 3, the second printer does not help at all at low loads: each job is taken directly into the service (without waiting) but the actual printing takes twice the time as with the fast printer.
- At heavy loads, the mean sojourn time of case 3 is the same as in case 1 (in both cases it consists mainly of the waiting). Two slow printers fed by a common queue discharge the work in the queue as efficiently as one fast printer.
- This is not the case for the alternative 2. When the queues are separate, it is possible that one printer stays idle while there are jobs waiting in the queue for the other printer. This deteriorates the overall performance in such a way that also at high loads alternative 2 is on the average two times slower than alternative 1.

Example 2

- A telephone switch is modelled as an *M/M/m system (when all lines are busy, the callers* are let to wait by signaling them the ring tone)
- How many lines (*m) are needed that the probability that a caller has to wait longer than* time t_{max} *is less than 1 % ?*

$$P\{W > t\} = P_q e^{-(m\mu-\lambda)t}$$

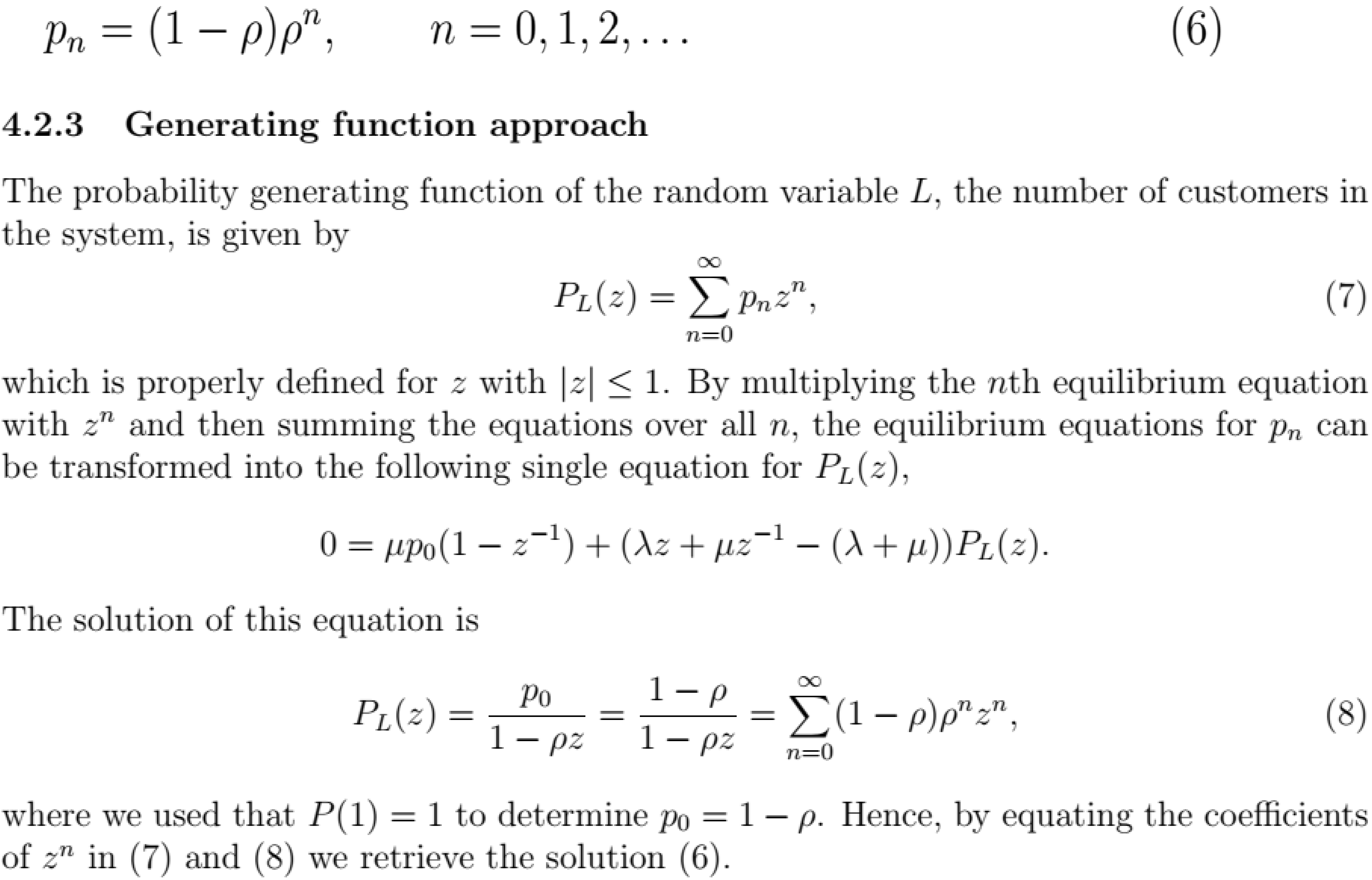
$$P\{W > t_{max}\} < 0.01 \Rightarrow \; P_q e^{-(m\mu-\lambda)t_{max}} < 0.01 \Rightarrow$$

$$100P_q < e^{(m\mu-\lambda)t_{max}} \Rightarrow \; \log(100P_q) < (m\mu-\lambda)t_{max} \;\; \Rightarrow \;\; \mathrm{m} > \frac{\log(100P_q) + \lambda t_{max}}{\mu t_{max}}$$

P_q *is a function of m (monotonically decreasing); thus the inequality is still an implicit one.*

It can be solved by trying sequentially values *m = 1, 2, 3, . . . until the inequality is satisfied.*

By letting the callers to wait for a free line for a while before blocking them, the number of blocked calls can be reduced or, conversely, the load of the system ρ *can be increased in* comparison with a loss system with the same blocking probability.



$$p_n = (1-\rho)\rho^n, \qquad n = 0, 1, 2, \ldots \tag{6}$$

4.2.3 Generating function approach

The probability generating function of the random variable L, the number of customers in the system, is given by

$$P_L(z) = \sum_{n=0}^{\infty} p_n z^n, \tag{7}$$

which is properly defined for z with $|z| \leq 1$. By multiplying the nth equilibrium equation with z^n and then summing the equations over all n, the equilibrium equations for p_n can be transformed into the following single equation for $P_L(z)$,

$$0 = \mu p_0(1 - z^{-1}) + (\lambda z + \mu z^{-1} - (\lambda + \mu))P_L(z).$$

The solution of this equation is

$$P_L(z) = \frac{p_0}{1-\rho z} = \frac{1-\rho}{1-\rho z} = \sum_{n=0}^{\infty} (1-\rho)\rho^n z^n, \tag{8}$$

where we used that $P(1) = 1$ to determine $p_0 = 1 - \rho$. Hence, by equating the coefficients of z^n in (7) and (8) we retrieve the solution (6).

4.4 Distribution of the sojourn time and the waiting time

It is also possible to derive the distribution of the sojourn time. Denote by L^a the number of customers in the system just before the arrival of a customer and let B_k be the service time of the kth customer. Of course, the customer in service has a residual service time instead of an ordinary service time. But these are the same, since the exponential service time distribution is memoryless. So the random variables B_k are independent and exponentially distributed with mean $1/\mu$. Then we have

$$S = \sum_{k=1}^{L^a+1} B_k. \tag{10}$$

By conditioning on L^a and using that L^a and B_k are independent it follows that

$$P(S > t) = P(\sum_{k=1}^{L^a+1} B_k > t) = \sum_{n=0}^{\infty} P(\sum_{k=1}^{n+1} B_k > t) P(L^a = n). \tag{11}$$

The problem is to find the probability that an arriving customer finds n customers in the system. PASTA states that the fraction of customers finding on arrival n customers in the system is equal to the fraction of time there are n customers in the system, so

$$P(L^a = n) = p_n = (1 - \rho)\rho^n. \tag{12}$$

Substituting (12) in (11) and using that $\sum_{k=1}^{n+1} B_k$ is Erlang-$(n + 1)$ distributed, yields

$$\begin{aligned} P(S > t) &= \sum_{n=0}^{\infty} \sum_{k=0}^{n} \frac{(\mu t)^k}{k!} e^{-\mu t} (1 - \rho)\rho^n \\ &= \sum_{k=0}^{\infty} \sum_{n=k}^{\infty} \frac{(\mu t)^k}{k!} e^{-\mu t} (1 - \rho)\rho^n \\ &= \sum_{k=0}^{\infty} \frac{(\mu \rho t)^k}{k!} e^{-\mu t} \\ &= e^{-\mu(1-\rho)t}, \qquad t \geq 0. \end{aligned} \tag{13}$$

Hence, S is exponentially distributed with parameter $\mu(1-\rho)$. This result can also be obtained via the use of transforms. From (10) it follows, by conditioning on L^a, that

$$\begin{aligned}
\widetilde{S}(s) &= E(e^{-sS}) \\
&= \sum_{n=0}^{\infty} P(L^a = n) E(e^{-s(B_1+\ldots+B_{n+1})}) \\
&= \sum_{n=0}^{\infty} (1-\rho)\rho^n E(e^{-sB_1}) \cdots E(e^{-sB_{n+1}}).
\end{aligned}$$

Since B_k is exponentially distributed with parameter μ, we have

$$E(e^{-sB_k}) = \frac{\mu}{\mu+s},$$

so

$$\widetilde{S}(s) = \sum_{n=0}^{\infty} (1-\rho)\rho^n \left(\frac{\mu}{\mu+s}\right)^{n+1} = \frac{\mu(1-\rho)}{\mu(1-\rho)+s},$$

from which we can conclude that S is an exponential random variable with parameter $\mu(1-\rho)$. So, for this system, the probability that the actual sojourn time of a customer is larger than a times the mean sojourn time is given by

$$P(S > aE(S)) = e^{-a}.$$

Hence, sojourn times of 2, 3 and even 4 times the mean sojourn time are not uncommon.

To find the distribution of the waiting time W, note that $S = W + B$, where the random variable B is the service time. Since W and B are independent, it follows that

$$\widetilde{S}(s) = \widetilde{W}(s) \cdot \widetilde{B}(s) = \widetilde{W}(s) \cdot \frac{\mu}{\mu + s} .$$

and thus,

$$\widetilde{W}(s) = \frac{(1-\rho)(\mu + s)}{\mu(1-\rho) + s} = (1-\rho) \cdot 1 + \rho \cdot \frac{\mu(1-\rho)}{\mu(1-\rho) + s} .$$

From the transform of W we conclude that W is with probability $(1-\rho)$ equal to zero, and with probability ρ equal to an exponential random variable with parameter $\mu(1-\rho)$. Hence

$$P(W > t) = \rho e^{-\mu(1-\rho)t}, \qquad t \geq 0. \tag{14}$$

The distribution of W can, of course, also be obtained along the same lines as (13). Note that

$$P(W > t | W > 0) = \frac{P(W > t)}{P(W > 0)} = e^{-\mu(1-\rho)t},$$

so the *conditional waiting time* $W | W > 0$ is exponentially distributed with parameter $\mu(1-\rho)$.