Queuing Theory

Little's Theorem: $N = \lambda T$

$$\xrightarrow{\text{arrival rate} = \lambda} \boxed{\text{System}} \xrightarrow{\text{departure rate} = \lambda}$$

- Holds for any (ergodic) system with a steady state
- Def.

  $\alpha(t)$ = the number of arrivals at the system in the interval from time 0 to time *t*.

  = number of arrivals in [0, *t*]

  $\beta(t)$ = the number of customer departures in the interval from time 0 to time *t*.

  = number of departure in [0, *t*]

  $N(t)$ = Number of customers in the system at time *t*

  = $\alpha(t) - \beta(t)$

  N = average (steady-state, long run, expected) number of customers in <u>system</u> (waiting for service or receiving service) in equilibrium

  $= \lim_{t\to\infty} \frac{\int_0^t N(t')dt'}{t}$ [customers

  λ = average (long run) arrival rate of customers

  $= \lim_{t\to\infty} \frac{\text{number of arrivals in } [0,\text{t}]}{t} = \lim_{t\to\infty} \frac{\alpha(t)}{t}$

  T = average time in system of customers in equilibrium

  $= \lim_{\alpha(t)\to\infty} \left\{ \frac{1}{\alpha(t)} \sum_{j=1}^{\alpha(t)} T_j \right\}$

- System = system: $N = \lambda T$

  System = queue: $N_q = \lambda W$

  System = server: $N_s = \lambda EX = \frac{\lambda}{\mu}$

  Because, by definition, $\boxed{T = W + EX}$, we have $\boxed{N = N_q + N_s}$.

- Let τ = interarrival times

  τ_i = the time between the arrival of the *i*-1 and the i^{th} customer

  Assume that all τ_i's are i.i.d., and thus have the same $E[\tau_i] = E\tau$.

  λ = long-term arrival rate at the system = $\frac{1}{E\tau}\left[\frac{\text{customers}}{\text{second}}\right]$

$$\lambda = \lim_{n\to\infty} \frac{n}{\sum_{i=0}^{n} \tau_i} = \frac{1}{\lim_{n\to\infty} \sum_{i=0}^{n} \frac{\tau_i}{n}} = \frac{1}{E\tau}$$

- T = Average time each customer spent in the system

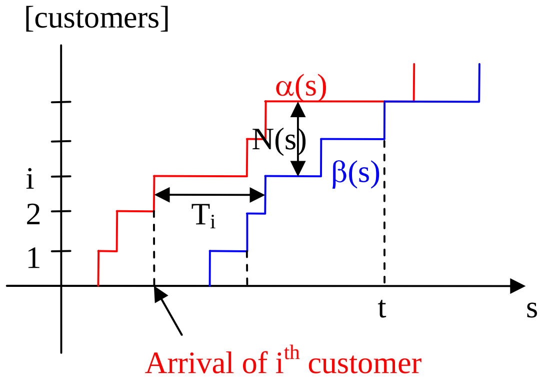
  T_i = Time the i^{th} customer spent in the system

  = the time that elapses between the instant when

  $\alpha(t)$ goes from i-1 to i

  to the instant when

  $\beta(t)$ goes from i-1 to i.



- Let t = a time instant where $\alpha(t) = \beta(t)$, which implies $N(t) = 0$.

  The area between $\alpha(t)$ and $\beta(t)$ from 0 to t:

  1) horizontally, area $= \sum_{j=1}^{\alpha(t)} T_j$

  Note if define d_i = departure time, a_i = arrival time of the i^{th} customer

  Then $T_i = d_i - a_i$.

  Note that $\sum_{j=1}^{\alpha(t)} T_j = \sum_{j=1}^{\alpha(t)} \left(d_j - a_j\right) = \sum_{j=1}^{\alpha(t)} d_j - \sum_{j=1}^{\alpha(t)} a_j$; so, order doesn't matter.

  2) vertically, area $= \int_0^t N(s)ds = \int_0^t \left(\alpha(s) - \beta(s)\right)ds$

  Hence, we have $\int_0^t N(s)ds = \sum_{j=1}^{\alpha(t)} T_j$.

  Thus, $N = \lim_{t\to\infty} \frac{1}{t} \int_0^t N(s)d = \lim_{t\to\infty} \frac{1}{t} \sum_{j=1}^{\alpha(t)} T_j$

  $$= \lim_{t\to\infty} \frac{\alpha(t)}{t} \sum_{j=1}^{\alpha(t)} \frac{T_j}{\alpha(t)} = \left(\lim_{t\to\infty} \frac{\alpha(t)}{t}\right)\left(\lim_{t\to\infty} \sum_{j=1}^{\alpha(t)} \frac{T_j}{\alpha(t)}\right) = \lambda T$$

Queuing Theory

- Standard queuing theory nomenclature

$$\underbrace{\text{Arrival process}}_{\begin{array}{l}\text{Interarrival time } \tau \\ \text{M = exponential} \\ \text{D = deterministic} \\ \text{G = general} \\ \text{Arrival rate: } \lambda = \frac{1}{\mathrm{E}\tau}\end{array}} \;/\; \underbrace{\text{Service time}}_{\begin{array}{l}\text{Service times } X \\ \text{M = exponential} \\ \text{D = deterministic} \\ \text{G = general} \\ \text{Service rate: } \mu = \frac{1}{\mathrm{E}X}\end{array}} \;/\; \underbrace{\text{Servers}}_{\begin{array}{l}\text{1 server} \\ c \text{ servers} \\ \infty\end{array}} \;/\; \underbrace{\text{Max occupancy}}_{\begin{array}{l}K \text{ customers} \\ \text{unspecified if unlimited}\end{array}}$$

  - 1st letter ⇒ nature of the arrival process
    - M = Poisson process (Markov, memoryless) ⇒ exponentially distributed interarrival times.
    - G = general distribution of interarrival times
    - D = deterministic interarrival times
  - 2nd letter ⇒ nature of the probability distribution of the service times.
    - M = exponential
    - G = general
    - D = deterministic
  - 3rd letter ⇒ number of servers
- Successive interarrival times and service times are assumed to be statistically independent of each other.
- Def:

  p_n = Steady state probability of having n customers in the system, n = 0, 1, …

  N = Average number of customers in the system = $\displaystyle\sum_{n=0}^{\infty} np_n$

  T = Average customer time in the system

  N_q = Average number of customers waiting in queue.

  If there are m server, then $N_q = \displaystyle\sum_{n=m+1}^{\infty} (n-m)\,p_n$

  W = Average customer waiting time in queue

  $N_s(t)$ = the number of customers that are being served at time t, and let X denote the service time.

  N_s = the average number of busy servers for a system in steady state

  X = service time, a random variable.

  $h = \mathrm{E}X = \dfrac{1}{\mu}$ = average service time.

- **Utilization factor**:
  - Single server: $\boxed{\rho = \text{proportion of time the server is busy} = p_1 = 1 - p_0 = \frac{\lambda}{\mu} = N_s}$.

Proof. For single-server systems, (1) system has ≥ 1 customers ≡ server is busy; hence, $p_{0,\text{server}} = p_{0,\text{system}} := p_0$. Also, $p_{1,\text{server}} = p_{1,\text{system}} := p_1 = 1 - p_0$. (2) $N_s(t)$ can only be 0 or 1, so N_s represents the proportion of time that the server is busy $\left(p_{1,\text{server}}\right)$. $N_s = 0p_{0,\text{server}} + 1p_{1,\text{server}} = p_{1,\text{server}} = 1 - p_0$. (3) From Little's theorem, $N_s = \lambda EX$. Hence, $1 - p_0 = N_s = \lambda EX$. Note that $1 - p_0$ is the proportion of time that the server is busy. For this reason, the utilization of a single-server system is defined by $\rho = \lambda EX = \frac{\lambda}{\mu}$.

- Similarly, define utilization of a *m*-server system by $\rho = \frac{\lambda EX}{m} = \frac{\lambda}{m\mu}$.

- For finite-capacity systems,

  it is necessary to distinguish between the traffic load offered to a system and the actual load carried by the system

  - The offered load or traffic intensity is a measure of the demand made on the system

    $= \lambda \overline{X}$

  - The carried load is the actual demand met by the system

    $= \lambda\left(1 - P_b\right)\overline{X}$

Occupancy Distribution upon Arrival

- Probabilistic characterization of a queuing system as seen by an arriving customer.
- Unconditional steady-state probabilities

  $p_n = \lim_{t\to\infty} P\left\{N\left(t\right) = n\right\}$

- Steady-state occupancy probabilities upon arrival

  $a_n = \lim_{t\to\infty} P\left\{N\left(t\right) = n \middle| \text{an arrival occured just after time t}\right\}$

- $p_n = a_n$, $n = 0, 1, \ldots$

  for queuing systems

  regardless of the distribution of the service times

  if either

  - the arrival process is Poisson and interarrival times and service times are independent.
  - future arrivals are independent of the current number in the system.

    ⇒

    for every time *t* and increment $\delta > 0$,

    the number of arrivals in the interval $(t, t+\delta)$ is independent of the number in the system at time *t*.

- the arrival process is Poisson and, at any time, the service times of previously arrived customers and the future interarrival times are independent.

Let

$A(t,t+\delta)$ be the event that an arrival occurs in the interval $(t,t+\delta)$

$p_n(t) = \Pr[N(t)=n]$. $\left(\Rightarrow p_n = \lim_{t\to\infty} p_n(t)\right)$.

Then,

$$a_n(t) = P\{N(t)=n \mid \text{an arrival occured just after time t}\}$$
$$= \lim_{\delta\to 0} P\{N(t)=n \mid A(t,t+\delta)\}$$

If the event $A(t,t+\delta)$ is independent of $N(t)$, then

$$a_n(t) = \lim_{\delta\to 0} P\{N(t)=n\} = P\{N(t)=n\} = p_n(t)$$

Taking the limit as $t \to \infty$, from the definition of a_n and p_n, we obtain $a_n = p_n$.

- Ex. non-Poisson arrival process.

  Suppose interarrival times are independent and uniformly distributed between $[a,b]$; $a < b$. Customer service times are all equal to $c < a$ sec.

  - Then, an arriving customer always finds an empty system ($N = 0$) .
  - On the other hand, the average number in the system as seen by an outside observer looking at a system at random time is $N = \lambda T$ where

    $$\lambda = \frac{1}{E\tau} = \frac{1}{\frac{a+b}{2}} = \frac{2}{a+b} \text{ and T = c.}$$

    Thus, $N = \lambda T = \dfrac{2c}{a+b}$.

- Ex. service times and future arrival times are correlated.

  Packet arrival is Poisson process. Transmission time of the nth packet equals one half the interarrival time between packets n and n+1

  - Upon arrival, a packet finds the system empty.
  - On the other hand, the average number in the system as seen by an outside observer looking at a system at random time is

    $$N = \lambda T = \frac{1}{\tau}\left(\frac{\tau}{2}\right) = \frac{1}{2}$$

Occupancy Distribution upon Departure

- The distribution of customers in the system just after a departure has occurred.
- $d_n(t) = P\{N(t)=n \mid \text{a departure occured just before time t}\}$

- steady-state values $d_n = \lim_{t \to \infty} d_n(t)$, $n = 0, 1, \ldots$

- $d_n = a_n$, $n = 0, 1, \ldots$

  if

  - the system reaches a steady-state with all n having positive steady-state probabilities.

  and

  - $N(t)$ changes in unit increments.

For any sample path of the system and for every *n*,

the number in the system will be *n* infinitely often (with probability 1).

$\Rightarrow$

For each time the number in the system increases from *n* to *n+1* due to an arrival, there will be a corresponding future decrease from *n+1* to *n* due to a departure.

$\Rightarrow$

In the long run,

the frequency of transitions from *n* to *n+1* out of transitions from any *k* to *k+1*

equals

the frequency of transitions from *n+1* to *n* out of transitions from any *k+1* to *k*,

which implies that $d_n = a_n$.

M/G/1

- "G" ≡ general (really, GI ≡ general independent)

  Service times are i.i.d.

  Pr[service time ≤ *t*] = $H(t)$ ≡ cdf of the service time; don't have to be continuous

  Mean service time $h = \int_0^\infty t dH(t)$

  $\rho = \lambda h < 1$ which assumes stability.

M/G/1 analysis based on Pollazek-Khinchin theory

- Polla(c)zek-Khinchin theory

- $N_k(z) = Ez^{n_k}$

- $h^*(s) = \int_0^\infty e^{-st} dH(t)$

- $R(z) = h^*(\lambda(1-z))$

$$N(z) = \frac{(1-\rho)(z-1)R(z)}{z-R(z)}$$

- $N = \sum_{m=0}^{\infty} mP(n=m) = N'(1) = \rho + \frac{\lambda^2 \overline{h^2}}{2(1-\rho)}$

$$T = h + \frac{\lambda \overline{h^2}}{2(1-\rho)}$$

$$W_q = T - h = \frac{\lambda \overline{h^2}}{2(1-\rho)}$$

$$N_q = \frac{\lambda^2 \overline{h^2}}{2(1-\rho)}$$

- Distribution of N

$$P(n=m) = \begin{cases} 1-\rho & m=0 \\ \dfrac{\left.\dfrac{d}{dz^m} N(z)\right|_{z=0}}{m!} & m>0 \end{cases}$$

- $\{N(t)\}$ is no longer Markov in non-expo service time case. However, can embed a discrete-time Markov chain at the departure instants
- Define

  n_k = number of customers in system right after (upon) departure of customer k (so, not including customer k itself.)

  s_k = service time of customer k.

  Assume the s_k are i.i.d. with common cdf $H(t) = F_s(t) = P(s_k \le t)$

  Let $h = \int_0^\infty t dH(t)$ be the mean service time.

  r_k = number of new customers arriving during service time of customer k

- $n_k = \begin{cases} n_{k-1} + r_k - 1 & ; \ n_{k-1} > 0 \\ r_k & ; \ n_{k-1} = 0 \end{cases}$

  Proof

  - For $n_{k-1} > 0$, after the (k-1)th customer leave, there are n_{k-1} customer in the system (the kth customer is included here also.) The first one which will be served right away is the kth customer. While the kth customer is served, r_k customers arrive. Thus when the kth customer leave, we have n_{k-1}+r_k-1 customers left in the system. (the -1 comes from the kth customer leaving)
  - For n_{k-1} = 0, after the (k-1)th customer leave, there are no customer in the system. After a while (exponentially distributed random duration), the kth customer

arrives. While the k^{th} customer is served, r_k additional customers arrive. Thus when the k^{th} customer leave, we have 1+r_k-1 = r_k customers left in the system. (The +1 and -1 is from customer k arriving and leaving.)

Another way to think about this: for the first case, n_{k-1} already includes the k^{th} customer so it has to subtract 1 out when the k^{th} customer leave.

- **Generating function**:

$$N_k(z) = Ez^{n_k}$$

$$R_k(z) = Ez^{r_k} = R(z) \text{ ; not a function of } k \text{ because } s_k \text{ are i.i.d.}$$

- **Laplace-Stieltjes transform** of service distribution:

$$h^*(s) = \int_0^\infty e^{-st} dH(t)$$

- Let r be a generic r_k, then

$$R(z) = Ez^{r_k} = Ez^r = \sum_{n=0}^{\infty} P(r=n) z^n = \sum_{n=0}^{\infty} \left(\int_0^\infty P(r=n|s=t) dH(t) \right) z^n$$

$$= \sum_{n=0}^{\infty} \int_0^\infty \frac{(\lambda t)^n e^{-\lambda t}}{n!} dH(t) z^n = \sum_{n=0}^{\infty} \int_0^\infty \frac{(\lambda tz)^n e^{-\lambda t}}{n!} dH(t)$$

Interchange the sum and the integral, then we have

$$R(z) = \int_0^\infty e^{-\lambda t} \sum_{n=0}^{\infty} \frac{(\lambda tz)^n}{n!} dH(t) = \int_0^\infty e^{-\lambda t} e^{\lambda tz} dH(t) = \int_0^\infty e^{-\lambda t(1-z)} dH(t)$$

$$= \int_0^\infty e^{-(\lambda(1-z))t} dH(t) = h^*(\lambda(1-z))$$

So, $\boxed{R(z) = Ez^{r_k} = h^*(\lambda(1-z))}$.

Note: require Poisson to prove $R(z) = h^*(\lambda(1-z))$

- Quantity of principal interest is $\lim_{k\to\infty} N_k(z) = N(z)$ (Will show later that $N'(1) = N$)

Recall that $n_k = \begin{cases} n_{k-1} + r_k - 1 & ; \ n_{k-1} > 0 \\ r_k & ; \ n_{k-1} = 0 \end{cases}$; hence,

$$N_k(z) = Ez^{n_k} = P(n_{k-1} = 0) E\left(z^{n_k} \middle| n_{k-1} = 0\right) + P(n_{k-1} > 0) E\left(z^{n_k} \middle| n_{k-1} > 0\right)$$

$$= P(n_{k-1} = 0) Ez^{r_k} + P(n_{k-1} > 0) E\left(z^{n_{k-1}+r_k-1} \middle| n_{k-1} > 0\right)$$

We already have $Ez^{r_k} = R(z) = h^*(\lambda(1-z))$. So, consider $E\left(z^{n_{k-1}+r_k-1} \middle| n_{k-1} > 0\right)$ = $\frac{1}{z} E\left(z^{n_{k-1}+r_k} \middle| n_{k-1} > 0\right)$. Now, because n_{k-1} and r_k are independent,

$$\begin{aligned}
E\left(z^{n_{k-1}+r_k-1}\middle|n_{k-1}>0\right) &= \frac{1}{z}E\left(z^{n_{k-1}}\middle|n_{k-1}>0\right)E\left(z^{r_k}\middle|n_{k-1}>0\right)\\
&= \frac{1}{z}E\left(z^{n_{k-1}}\middle|n_{k-1}>0\right)E\left(z^{r_k}\right)\\
&= \frac{1}{z}E\left(z^{n_{k-1}}\middle|n_{k-1}>0\right)R(z)
\end{aligned}$$

$$\begin{aligned}
P\left(n_{k-1}>0\right)E\left(z^{n_{k-1}+r_k-1}\middle|n_{k-1}>0\right) &= P\left(n_{k-1}>0\right)E\left(z^{n_{k-1}+r_k-1}\middle|n_{k-1}>0\right)\\
&= \cancel{P\left(n_{k-1}>0\right)}\,\frac{1}{z}\left(\sum_{n=1}^{\infty}\frac{P\left(n_{k-1}=n\right)}{\cancel{P\left(n_{k-1}>0\right)}}z^n\right)R(z)\\
&= \frac{1}{z}\left(\sum_{n=1}^{\infty}P\left(n_{k-1}=n\right)z^n\right)R(z)\\
&= \frac{1}{z}\left(\sum_{n=0}^{\infty}P\left(n_{k-1}=n\right)z^n - P\left(n_{k-1}=0\right)z^0\right)R(z)\\
&= \frac{1}{z}\left(N_{k-1}(z)-P\left(n_{k-1}=0\right)\right)R(z)
\end{aligned}$$

Thus $N_k(z)=P(n_{k-1}=0)R(z)+\frac{1}{z}(N_{k-1}(z)-P(n_{k-1}=0))R(z)$

- As $k\to\infty$ with $\rho<1$, we get
  - $N_k(z)$ and $N_{k-1}(z)\to N(z)$ and
  - $P\left(n_{k-1}=0\right)\to p_0$.

  We already know that, for a single server, $\rho=1-p_0$.

  Thus, we have

$$\begin{aligned}
N(z) &= (1-\rho)R(z)+\frac{1}{z}\left(N(z)-(1-\rho)\right)R(z)\\
zN(z) &= z(1-\rho)R(z)+N(z)R(z)-(1-\rho)R(z)\\
N(z) &= \frac{(1-\rho)(z-1)R(z)}{z-R(z)}
\end{aligned}$$

- To find N,

$$N=\sum_{m=0}^{\infty}mP(n=m)$$

$$N(z)=Ez^n=\sum_{m=0}^{\infty}P(n=m)z^m$$

$$\frac{d}{dz}N(z)=N'(z)=\sum_{m=0}^{\infty}mP(n=m)z^{m-1}$$

$$N'(1)=\sum_{m=0}^{\infty}mP(n=m)=N$$

  - May be easier to use a Taylor series approach and expand around *z* = 1. Introduce *u* = *z*-1, so we can expand around *u* = 0.

- Let

  $b(u)=R(z)\big|_{z=u+1}=h^*(\lambda(1-z))=h^*(-\lambda u)=\int_0^{\infty}e^{\lambda ut}dH(t)$, and

  $G(u)\equiv N(z)\big|_{z=u+1}=\dfrac{(1-\rho)ub(u)}{u+1-b(u)}$.

- By Taylor's Theorem:

  $b(u)=b(0)+b'(0)u+\dfrac{b''(0)}{2}u^2+o(u^2)$ as $u\to 0$.

  Note that $b_0=b(0)=\int_0^{\infty}e^{\lambda 0t}dH(t)=\int_0^{\infty}dH(t)=1$. Also, $b'(u)=\lambda\int_0^{\infty}te^{\lambda ut}dH(t)$. Hence, $b_1=b'(0)=\lambda\int_0^{\infty}te^{\lambda 0t}dH(t)=\lambda\int_0^{\infty}tdH(t)=\lambda h=\rho$. The second derivative $b''(u)=\lambda^2\int_0^{\infty}t^2e^{\lambda ut}dH(t)$. Therefore, $b''(0)=\lambda^2\int_0^{\infty}t^2e^{\lambda 0t}dH(t)=\lambda^2\int_0^{\infty}t^2dH(t)=\lambda^2\overline{h^2}$, and $b_2=\dfrac{b''(0)}{2}=\dfrac{\lambda^2\overline{h^2}}{2}$.

- $N=\rho+\dfrac{\lambda^2\overline{h^2}}{2(1-\rho)}$

Proof. $G(u) = \frac{(1-\rho)ub(u)}{u+1-b(u)} = \frac{(1-\rho)ub(u)}{u+1-b_0-b_1u-b_2u^2+o(u^2)}$

$$= \frac{(1-\rho)\cancel{u}b(u)}{\underset{1}{\cancel{u}}+\cancel{1-}1-b_1\cancel{u}-b_2u^{\cancel{2}}-o\left(u^{\cancel{2}}\right)}$$

$$= \frac{(1-\rho)b(u)}{(1-b_1)-b_2u-o(u)}$$

$$= \frac{1-\rho}{1-b_1}b(u)\frac{1}{1-\frac{b_2}{1-b_1}u+o(u)}$$

Now, note that as $x \to 0$, $\frac{1}{1-x+o(x)} = 1+x+o(x)$.

Pf. First, note that $\frac{1}{1-x} = 1+x+o(x)$. We will show that if

$\frac{1}{f(x)} = g(x)+o(x)$, then $\frac{1}{f(x)+o(x)} = g(x)+o(x)$.

Start with $\frac{1}{f(x)} = g(x)+o(x)$, we have $\lim_{x\to 0}\left(\frac{1}{xf(x)} - \frac{g(x)}{x}\right) = 0$.

Now, $\lim_{x\to 0}\frac{1}{x}\left(\frac{1}{f(x)+o(x)} - g(x)\right) = \lim_{x\to 0}\left(\frac{1}{xf(x)+\underset{0}{\cancel{o(x^2)}}} - \frac{g(x)}{x}\right)$.

$$= 0$$

Hence, $G(u) = \frac{1-\rho}{1-b_1}\left(1+b_1u+o(u)\right)\left(1+\frac{b_2}{1-b_1}u+o(u)\right)$

$$= \frac{1-\rho}{1-b_1}\left(1+\left(b_1+\frac{b_2}{1-b_1}\right)u+o(u)\right)$$

From $r'(0)=0$ for continuous $r(x)=o(x)$ as $x \to 0$, we then have

$$G'(0) = \frac{d}{du}G(u)\Big|_{u=0} = \frac{1-\rho}{1-b_1}\left(b_1+\frac{b_2}{1-b_1}\right)$$

Thus, $N'(1) = G'(0) = \frac{1-\rho}{1-b_1}\left(b_1+\frac{b_2}{1-b_1}\right)$.

We finally have

$$N = N'(1) = \frac{1-\rho}{1-b_1}\left(b_1 + \frac{b_2}{1-b_1}\right) = \frac{1-\cancel{\rho}}{\cancel{1}-\rho}\left(\rho + \frac{\frac{\lambda^2 \overline{h^2}}{2}}{1-\rho}\right) = \rho + \frac{\lambda^2 \overline{h^2}}{2(1-\rho)}.$$

- $N_q = N - \rho = \dfrac{\lambda^2 \overline{h^2}}{2(1-\rho)}$
- $T = \dfrac{N}{\lambda} = \dfrac{\rho + \frac{\lambda^2 \overline{h^2}}{2(1-\rho)}}{\lambda} = h + \dfrac{\lambda \overline{h^2}}{2(1-\rho)}$
- $W_q = T - h = \dfrac{\lambda \overline{h^2}}{2(1-\rho)}$
- $P(n=m) = \begin{cases} 1-\rho & m=0 \\ \dfrac{\left.\frac{d^m}{dz^m} N(z)\right|_{z=0}}{m!} & m>0 \end{cases}$

Proof. $N(z) = \sum_{m=0}^{\infty} P(n=m) z^m = P(n=0) + P(n=1) z^1 + P(n=2) z^2 + \dots$

$$P(n=0) = N(0) = \frac{(1-\rho)(0-1)R(0)}{0-R(0)} = 1-\rho$$

$$N'(z) = \sum_{m=1}^{\infty} m P(n=m) z^{m-1} = P(n=1) + P(n=2) z + \dots$$

$$N'(0) = P(n=1).$$

Hence, $P(n=m) = \begin{cases} 1-\rho & m=0 \\ \dfrac{\left.\frac{d^m}{dz^m} N(z)\right|_{z=0}}{m!} & m>0 \end{cases}$

- Distribution of waiting time.

$$w^*\left(\lambda(1-z)\right) = \frac{(1-\rho)(z-1)}{z - h^*\left(\lambda(1-z)\right)}$$

$$w^*(\tilde{s}) = \frac{1-\rho}{1 - \frac{\lambda}{\tilde{s}}\left(1 - h^*(\tilde{s})\right)}$$

Observe that, in the steady-state, the random variable *n* that represents the system population at the point of departure of a customer may also be thought of as the arrivals during the total system time (sojourn time) of that customer.

Said sojourn time is the sum of the waiting random variable, w, and the service random variable, s.

The same sort of reasoning that gave us $R(z)=h^*(\lambda(1-z))$ can be applied to give us the moment generating function of the number of arrivals during *w* + *s* as $N(z)=f^*(\lambda(1-z))$

where f^* is the Laplace transform of the distribution of *w* + *s*

during *s* →	*r*, *R*
during *w* + *s* →	*n*, *N*

Since *w* and *s* are independent, the pdf of *w* + *s* is the convolution of the pdf of *w* and pdf of *s*. This implies that the Laplace transform is the multiplication: $f^*(s)=w^*(s)h^*(s)$

$$N(z)=f^*(\lambda(1-z))=w^*(\lambda(1-z))h^*(\lambda(1-z)).$$

where w^* is the L-S transform of the distribution of *w*.

$$w^*(\lambda(1-z))=\frac{N(z)}{h^*(\lambda(1-z))}=\frac{\dfrac{(1-\rho)(z-1)R(z)}{z-R(z)}}{h^*(\lambda(1-z))}=\frac{\dfrac{(1-\rho)(z-1)\,\cancel{h^*(\lambda(1-z))}}{z-h^*(\lambda(1-z))}}{\cancel{h^*(\lambda(1-z))}}$$

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

Let $\tilde{s}=\lambda(1-z)$

$$w^*(\tilde{s})=\frac{(1-\rho)\left(-\dfrac{\tilde{s}}{\lambda}\right)}{\left(1-\dfrac{\tilde{s}}{\lambda}\right)-h^*(\tilde{s})}=\frac{1-\rho}{-\dfrac{\lambda}{\tilde{s}}+1+\dfrac{\lambda}{\tilde{s}}h^*(\tilde{s})}=\frac{1-\rho}{1-\dfrac{\lambda}{\tilde{s}}\left(1-h^*(\tilde{s})\right)}$$

- We did explicitly use the fact that the number of arrivals during a service of length *s* is Poisson with parameter λs.

  Our justification for equating the statistics just after a departure instant in equilibrium to those at a randomly chosen instant in equilibrium also depended on the Poisson nature of the arrivals. (need $d_n = a_n = p_n$).

- Average length of an idle period = $\dfrac{1}{\lambda}$

  Proof. Since an idle period occurs when the system is waiting for a customer to arrive after the queue becomes empty. At the moment that server becomes empty, by memoryless property, have to wait $\mathcal{E}(\lambda)$ with average $\frac{1}{\lambda}$ for the next customer to arrive, independent of how long it has already been from the moment when the last customer arrived.

- Average length of busy period = $\frac{1}{\mu-\lambda}$.

  Proof. Let B = average length of buy period. We have shown that average length of an idle period is $\frac{1}{\lambda}$. Note that the busy period and idle period are alternating sequence. Hence,

$$\rho=\lim_{n\to\infty}\frac{\sum_{i=1}^{n}\tau_{busy,i}}{\sum_{i=1}^{n}\left(\tau_{idle,i}+\tau_{busy,i}\right)}=\frac{\lim_{n\to\infty}\frac{\sum_{i=1}^{n}\tau_{busy,i}}{n}}{\lim_{n\to\infty}\frac{\sum_{i=1}^{n}\tau_{idle,i}}{n}+\lim_{n\to\infty}\frac{\sum_{i=1}^{n}\tau_{busy,i}}{n}}=\frac{B}{B+\frac{1}{\lambda}}.$$

  Solving for B, we get $B=\frac{\rho}{\lambda(1-\rho)}=\frac{h}{1-\rho}=\frac{1}{\mu-\lambda}$.

- Avergae number of customers served in a busy period = $\frac{1}{1-\rho}$

  Idea. $=\frac{B}{h}$.

M/G/1 analysis based on the concept of the mean residual service time

- R_i = Residual service time seen by the i^{th} customer.
  By this we mean that if customer j is already being served when i **<u>arrive</u>**s,
  R_i is the remaining time until customer j's service time is complete.
  If no customer is in service (i.e., the system is empty when i arrives), the $R_i = 0$.
- R = mean residual time = $\lim_{i\to\infty} ER_i$
- $\overline{R}$ = mean residual service time given that one is arrived when the server is busy

  By renewal theory: $\overline{R}=\frac{1}{2}\frac{\overline{X^2}}{\overline{X}}$

  - Note: If M/M/1, service time is exponentially distributed, and thus memoryless. Therefore, given that the service time does not end there, what's left is also exponentially distributed with the same mean. So, $\overline{R}=\frac{1}{\mu}$.

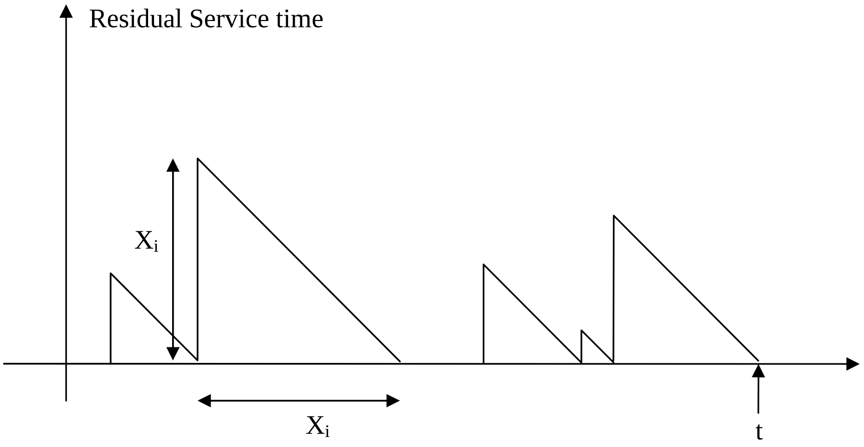
Using the above equation gives the same result: $\overline{R} = \frac{1}{2}\frac{\frac{1}{\mu^2}+\left(\frac{1}{\mu}\right)^2}{\frac{1}{\mu}} = \frac{1}{\mu}$.

- $R = \frac{1}{2}\lambda\overline{X^2}$

Proof. We know that the probability of server being busy for single server is $p_1 = \lambda EX = \rho$. Hence,

$$R = \overline{R}p_0 + 0\,p_1 = \overline{R}\rho = \frac{1}{2}\rho\frac{\overline{X^2}}{\overline{X}} = \frac{1}{2}\lambda\overline{X^2}$$

Proof. (Graphical argument)



$r(\tau)$ = the remaining time for completion of the customer in service at time τ

When a new service of duration X begins, $r(\tau)$ starts at X and decays linearly for X time units.

$$\int_0^t r(\tau)d\tau = \sum_{i=1}^{M(t)}\frac{1}{2}X_i^2$$

M(t) = number of triangles in [0,t] = number of service completions in [0,t]

$$R = \lim_{t\to\infty}\frac{1}{t}\int_0^t r(\tau)d\tau = \lim_{t\to\infty}\frac{1}{2}\frac{M(t)}{t}\sum_{i=1}^{M(t)}\frac{X_i^2}{M(t)} = \frac{1}{2}\left(\lim_{t\to\infty}\frac{M(t)}{t}\right)\left(\lim_{t\to\infty}\sum_{i=1}^{M(t)}\frac{X_i^2}{M(t)}\right)$$

$$= \frac{1}{2}\lambda\overline{X^2}$$

- $W = \frac{\lambda\overline{X^2}}{2(1-\rho)}$

Proof. Note that the time waiting in the queue of the i^{th} customer = residual service time seen by the i^{th} customer + time used to service all customers already in the queue.

$$W = R + \frac{1}{\mu} N_q = R + \frac{1}{\mu}\lambda W = R + \rho W .$$

$$W = \frac{R}{1-\rho} = \frac{\frac{1}{2}\lambda \overline{X^2}}{1-\rho} = \frac{\lambda \overline{X^2}}{2(1-\rho)} .$$

- The average customer in queue N_q and the mean residual time *R* as seen by an arriving customer are also equal to the average number in queue and mean residual time seen by an outside observer at a random time.

  This is due to the Poisson character of the arrival process, which implies that the occupancy distribution upon arrival is typical.
- M/G/1 is a renewal process when busy

  M/G/1 has occasional (with probability 1-ρ of occurrence) $\mathcal{E}(\lambda)$ random variable inserted into service time renewal process.
- M/G/1 queue can have $\rho < 1$ but infinite *W* if the second moment $\overline{X^2} \to \infty$
- The formula is valid for any order of servicing customers as long as the order is determined independently of the required service time

  To see this, suppose the *i*th and *j*th customers are both in the queue and that they exchange places.

  The expected queuing time of customer i will then be exchanged with that for customer j, but the average, over all customers, is unchanged.

  Since any service order can be considered as a sequence of reversals in queue position, the P-K formula remains valid.

M/G/1 with vacations

- At the end of each busy period, the server goes on "vacation" for some random interval of time.

  A new arrival to an idle system, rather than going into service immediately, waits for the end of the vacation period.

  If the system is still idle at the completion of a vacation, a new vacation starts immediately.
- Let V_i's be the durations of the successive vacations taken by the server.

  Assume V_i's are i.i.d. random variables and independent of the customer interarrival times and service times.
- $\overline{R} = \frac{1}{2}\frac{\overline{X^2}}{\overline{X}}$ = mean residual time given that arrive when the server is serving someone

  $\frac{1}{2}\frac{\overline{V^2}}{\overline{V}}$ = mean residual time given that arrive when the server is on vacation (idle)

$$R = \frac{1}{2}\frac{\overline{X^2}}{\overline{X}}P\{\text{server busy}\} + \frac{1}{2}\frac{\overline{V^2}}{\overline{V}}P\{\text{server idle}\}$$
$$= \frac{1}{2}\frac{\overline{X^2}}{\overline{X}}\rho + \frac{1}{2}\frac{\overline{V^2}}{\overline{V}}(1-\rho) = \frac{1}{2}\overline{X^2}\lambda + \frac{1}{2}\frac{\overline{V^2}}{\overline{V}}(1-\rho)$$

- $W = \dfrac{R}{1-\rho} = \dfrac{\lambda\overline{X^2}}{2(1-\rho)} + \dfrac{1}{2}\dfrac{\overline{V^2}}{\overline{V}}$
- $\boxed{W_{\text{w/ vacation}} = \dfrac{\lambda\overline{X^2}}{2(1-\rho)} + \dfrac{1}{2}\dfrac{\overline{V^2}}{\overline{V}} = W_{\text{w/o vacation}} + \dfrac{1}{2}\dfrac{\overline{V^2}}{\overline{V}}}$

M/M/1

- $\rho := \dfrac{\lambda}{\mu} = 1 - p_0 = N_s = p_{1_{server}}$

  $p_n = \rho^n(1-\rho)\ ;\ n = 0,1,\ldots$

  $N = \dfrac{\rho}{1-\rho} = \dfrac{\lambda}{\mu-\lambda}$

  $T = \dfrac{1}{\mu-\lambda}$

  $W = \dfrac{\rho}{\mu-\lambda}$

  $N_q = \dfrac{\rho^2}{1-\rho}$

- Transient if $\rho > 1$; Null recurrence if $\rho = 1$; Ergodic if $\rho < 1$
- $\rho = 1 - p_0$ = utilization factor = the long-term proportion of time the server is busy

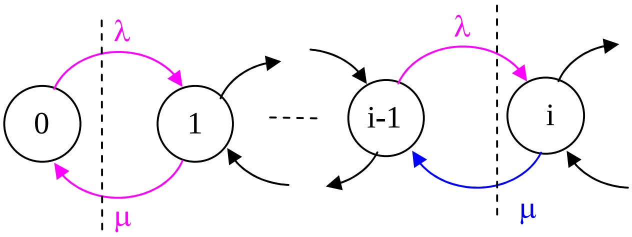
  Proof.

  (1) If the system has ≥ 1 customers, the server is busy (serving surely 1 customer). This occur with probability $1 - p_0$. Note also that if the server is busy, then the system has ≥ 1 customers (at least one in the server). Hence, $p_{0_{system}} = p_{0_{server}} = p_0$. If the system has 0 customer, the server is idle (serving 0 customer). This occur with probability p_0. So, the long-term proportion of time the server is busy $= 1 - p_0$. Average number of customer in the server = $N_s = p_0 \times 0 + p_1 \times 1 = p_1 = 1 - p_0$

  (2) Now, Apply Little's theorem to the server. Then $N_s = \lambda EX = \lambda\dfrac{1}{\mu} := \rho$.

From (1) and (2), $\rho := \frac{\lambda}{\mu} = 1 - p_0 = N_s = p_{1_{server}}$.

- State diagram



- $p_n = \rho^n (1-\rho)$; $n = 0, 1, \ldots$

  Proof. This is a birth-and-death process.

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

  Proof. $N = \sum_{n=0}^{\infty} n p_n = \sum_{n=0}^{\infty} n \rho^n (1-\rho) = (1-\rho)\frac{\rho}{(1-\rho)^2}$.

- $T = \frac{N}{\lambda} = \frac{\frac{\rho}{1-\rho}}{\lambda} = \frac{1}{\mu - \lambda}$

- $W = T - \overline{X} = \frac{1}{\mu - \lambda} - \frac{1}{\mu} = \frac{\lambda}{\mu(\mu-\lambda)} = \frac{\rho}{\mu - \lambda}$

- $N_q = \lambda W = \frac{\lambda\rho}{\mu - \lambda} = \frac{\frac{\lambda}{\mu}\rho}{1 - \frac{\lambda}{\mu}} = \frac{\rho^2}{1-\rho}$

  or

  $$N_q = 0 p_0 + 0 p_1 + 1 p_2 + 2 p_3 + \ldots + (i-1) p_i + \ldots$$
  $$= \sum_{i=2}^{\infty} (i-1) p_i = \sum_{i=2}^{\infty} (i-1) \rho^i (1-\rho)$$
  $$= (1-\rho) \sum_{m=1}^{\infty} m \rho^{m+1} = (1-\rho)\rho \sum_{m=1}^{\infty} m \rho^{m+1} = (1-\rho)\rho \sum_{m=1}^{\infty} m \rho^m$$
  $$= (1-\rho)\rho \sum_{m=0}^{\infty} m \rho^m = (1-\rho)\rho \frac{\rho}{(1-\rho)^2} = \frac{\rho^2}{1-\rho}$$

- **Effect of scale on performance**
  - *m* separate M/M/1 systems each: λ, μ

$$\mathrm{E}T = \frac{1/\mu}{1-\rho}$$

- One consolidated system: $m\lambda$, $m\mu$

  $$\rho' = \frac{\lambda'}{\mu'} = \frac{m\lambda}{m\mu} = \frac{\lambda}{\mu} = \rho$$

  $$\mathrm{E}T' = \frac{1/\mu'}{1-\rho} = \frac{1/m\mu}{1-\rho} = \frac{1}{m}ET \text{ (less delay)}$$

- The improved performance of the combined system arises from improved global usage of the processors.
  - In the separate systems,

    some of the queues may be empty while others are not.

    Consequently, some processors can be idle, even though there is work to be done in the system.
  - In the combined system,

    the processor will stay busy as long as customers are waiting to be served

Applying M/G/1 analysis to M/M/1

- Mean residual service time

  As noted above, since service time for M/M/1 is exponentially distributed, and thus memoryless. Therefore, given that the service time does not end when the packet arrive, what's left of service time for the currently serviced packet is also exponentially distributed with the same mean. So, $\overline{R} = \frac{1}{\mu}$.

  Thus, $W = \left(\rho\frac{1}{\mu} + (1-\rho)0\right) + \frac{1}{\mu}N_Q$. The average waiting time in the queue is the summation of 1) the residual time of the currently serviced packet which is 0 if server is idle and $\frac{1}{\mu}$ if server is busy and 2) The time required to service the customers already waiting in the queue which is N_Q times the average service time.

  and $W = \rho\frac{1}{\mu} + \frac{1}{\mu}\lambda W = \frac{\rho}{\mu-\lambda}$ as before.

- $dH(t) = \mu e^{-\mu t}dt$

  $h = \frac{1}{\mu}$, $\overline{h^2} = \frac{2}{\mu^2}$

  $$W = \frac{\lambda\overline{h^2}}{2(1-\rho)} = \frac{\lambda\frac{2}{\mu^2}}{2(1-\rho)} = \frac{\rho}{\mu(1-\rho)} = \frac{\rho}{\mu-\lambda} \text{ as expected.}$$

- pmf for *N* for M/M/1 system:

$$h(t)=\mu e^{-\mu t}\xrightarrow{L}h^*(s)=\mu\frac{1}{s+\mu}=\frac{\mu}{s+\mu}$$

$$N(z)=\frac{(1-\rho)(z-1)h^*(\lambda(1-z))}{z-h^*(\lambda(1-z))}=\frac{(1-\rho)(z-1)\frac{\mu}{\lambda(1-z)+\mu}}{z-\frac{\mu}{\lambda(1-z)+\mu}}$$

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

$$=\frac{(1-\rho)\mu}{\mu-\lambda z}=\frac{1-\rho}{1-\rho z}$$

By expanding N(z) in a power series, we have $N(z)=\sum_{i=0}^{\infty}(1-\rho)(\rho z)^i$.

Since $N(z)=\sum_{i=0}^{\infty}P(n=i)z^i$, $P(n=i)=(1-\rho)\rho^i$ for k = 0, 1, 2, …

- pdf of *W* for M/M/1 system: $f_W(t)=(1-\rho)\delta(t)+(1-\rho)\lambda e^{-\mu(1-\rho)t}$

$$w^*(s)=\frac{1-\rho}{1-\frac{\lambda}{s}\left(1-h^*(s)\right)}=\frac{1-\rho}{1-\frac{\lambda}{s}\left(1-\frac{\mu}{s+\mu}\right)}=\frac{1-\rho}{1-\frac{\lambda}{s}\left(\frac{s}{s+\mu}\right)}=\frac{1-\rho}{1-\frac{\lambda}{s+\mu}}$$

$$=(1-\rho)\frac{s+\mu}{s+\mu-\lambda}=(1-\rho)\left(\frac{(s+\mu-\lambda)+\lambda}{s+\mu-\lambda}\right)=(1-\rho)\left(1+\frac{\lambda}{s+(\mu-\lambda)}\right)$$

$$f_W(t)=(1-\rho)\left(\delta(t)+\lambda e^{-(\mu-\lambda)t}\right)=(1-\rho)\delta(t)+(1-\rho)\lambda e^{-\mu(1-\rho)t}\ ;\ t>0$$

- pdf of *T* for M/M/1 system: $f_T(t)=\mu(1-\rho)e^{-\mu(1-\rho)t}$

$$T^*(s)=w^*(s)h^*(s)$$

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

$$f_T(t)=(1-\rho)\mu e^{-(\mu-\lambda)t}=\mu(1-\rho)e^{-\mu(1-\rho)t}$$

M/D/1

- “D” ⇒ deterministic
  service time = h for every customer
- $\overline{h}=h$, $\overline{h^2}=h^2$

$$W = \frac{\lambda \overline{h^2}}{2(1-\rho)} = \frac{\lambda h^2}{2(1-\lambda h)}$$

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

$$N_q = \lambda W = \frac{\lambda\rho}{2\mu(1-\rho)} = \frac{\rho^2}{2(1-\rho)}$$

$$T = W + h = \frac{\rho}{2\mu(1-\rho)} + \frac{1}{\mu} = \frac{1}{\mu}\left(\frac{\rho}{2(1-\rho)} + 1\right) = \frac{1}{\mu}\left(\frac{\rho + 2 - 2\rho}{2(1-\rho)}\right) = \frac{2-\rho}{2\mu(1-\rho)}$$

$$N = N_q + \rho = \frac{\rho^2}{2(1-\rho)} + \rho = \frac{\rho^2 + 2\rho - 2\rho^2}{2(1-\rho)} = \frac{2\rho - \rho^2}{2(1-\rho)}$$

M/M/1/K

- Equilibrium (stable):
  - $p_n\lambda = p_{n+1}\mu$

    $p_{n+1} = \frac{\lambda}{\mu} p_n = \left(\frac{\lambda}{\mu}\right)^{n+1} p_0 = \rho^{n+1} p_0$; n = 0, 1, …, K-1
  - $\sum_{n=0}^{K} p_n = 1 \Rightarrow \sum_{n=0}^{K} \rho^n p_0 = p_0 \frac{1-\rho^{K+1}}{1-\rho} = 1 \Rightarrow p_0 = \frac{1-\rho}{1-\rho^{K+1}}$
- $p_n = \frac{1-\rho}{1-\rho^{K+1}} \rho^n$; n = 0, 1, …, K
- P(blocking or loss) = p_K = proportion of time that the system is full = $\frac{1-\rho}{1-\rho^{K+1}} \rho^K$

  Proof. This is a truncated birth-and-death process.
- For $\rho < 1 \Rightarrow \lambda < \mu$
  - the probabilities decrease exponentially as n increases
  - N tends to cluster around n = 0
  - adding more buffers (K) is beneficial since the result is a reduction in loss probability
- For $\rho = 1$
  - all state are equally probable
  - $p_n = \frac{1}{K+1}$
- For $\rho > 1 \Rightarrow \lambda > \mu$
  - p_n increase with n

- p_n tend to cluster toward $n = K \Rightarrow$ the system tends to be full
- adding buffers is counterproductive since the system will fill up the additional buffers.

- $N = \begin{cases} \dfrac{\rho}{1-\rho} - \dfrac{(K+1)\rho^{K+1}}{1-\rho^{K+1}} & \text{for } \rho \neq 1 \\ \dfrac{K}{2} & \text{for } \rho = 1 \end{cases}$

  - For $\rho < 1$,

$$\sum_{n=0}^{K} np_n = \sum_{n=0}^{K} n \frac{1-\rho}{1-\rho^{K+1}} \rho^n = \frac{1-\rho}{1-\rho^{K+1}} \sum_{n=0}^{K} n\rho^n$$

$$\begin{aligned} 1\cdot\rho + \quad 2\cdot\rho^2 + 3\cdot\rho^3 + \ldots + K\rho^K \quad &= \quad s \\ 1\rho^2 + 2\cdot\rho^3 + 3\cdot\rho^3 + \ldots + \quad K\rho^{K+1} \quad &= \quad \rho s \end{aligned}$$

$$s(1-\rho) = \rho + \rho^2 + \rho^3 + \ldots + \rho^K - K\rho^{K+1}$$

$$s = \frac{1}{1-\rho}\left(\frac{\rho-\rho^{K+1}}{1-\rho}\right) - \frac{K\rho^{K+1}}{1-\rho}$$

$$\frac{1-\rho}{1-\rho^{K+1}} \sum_{n=0}^{K} n\rho^n = \frac{1-\rho}{1-\rho^{K+1}} \frac{\rho-\rho^{K+1}}{(1-\rho)^2} - \frac{1-\rho}{1-\rho^{K+1}} \frac{K\rho^{K+1}}{1-\rho}$$

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

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

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

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

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

  - For $\rho = 1$,

$$\sum_{n=0}^{K} np_n = \sum_{n=0}^{K} n\frac{1}{K+1} = \frac{1}{K+1}\frac{K(K+1)}{2} = \frac{K}{2}$$

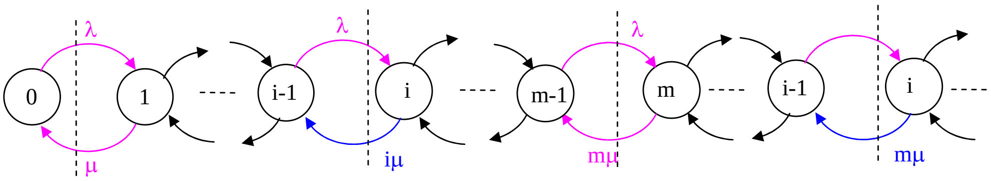
- $T = \dfrac{N}{\lambda(1-p_K)}$
- For $\rho \to 0$,

- $N = \frac{\rho}{1-\rho} - \frac{(K+1)\rho^{K+1}}{1-\rho^{K+1}} \to 0$
- $W \to 0$
- $p_K = P_{\text{loss}} = \frac{1-\rho}{1-\rho^{K+1}}\rho^K = 0$
- $T = W+X \to T = X$

- For $\rho \to \infty$,
  - $N \to K$
  - $p_K \to 1$
  - $T \to \frac{K}{\mu} = KEX$

M/M/m

- m server.



- $\rho = \frac{1}{m}\frac{\lambda}{\mu} < 1$
- $p_0 = \frac{1}{\sum_{i=0}^{m-1}\frac{(m\rho)^i}{i!} + \frac{(m\rho)^m}{m!(1-\rho)}}$
- $p_i = \begin{cases} \frac{(m\rho)^i}{i!}p_0 & 0 \le i \le m \\ \frac{m^m\rho^i}{m!}p_0 & i \ge m \end{cases}$
- Erlang C formula $P_Q = \frac{p_0(m\rho)^m}{m!(1-\rho)} = \frac{p_m}{1-\rho} = \mathrm{P}\{\mathrm{W} > 0\}$
- $N_Q = \frac{\rho P_Q}{1-\rho}$
- $W = \frac{\rho P_Q}{\lambda(1-\rho)}$

- $T = \frac{1}{\mu} + W$
- $N = N_Q + m\rho = m\rho + \frac{\rho P_Q}{1-\rho}$

- For 1 ≤ i ≤ m

  $$i\mu p_i = \lambda p_{i-1} \Rightarrow p_i = \frac{1}{i}\frac{\lambda}{\mu} p_{i-1} = \frac{1}{i}(m\rho)\, p_{i-1}$$

  $$p_i = \frac{(m\rho)^i}{i!} p_0, 1 \le i \le m$$

  For i > m

  $$m\mu p_i = \lambda p_{i-1} \Rightarrow p_i = \frac{1}{m}\frac{\lambda}{\mu} p_{i-1} = \rho p_{i-1}$$

  $$p_i = \rho^{i-m} p_m = \rho^{i-m}\frac{(m\rho)^m}{m!} p_0 = \frac{m^m \rho^i}{m!} p_0$$

  $$p_i = \frac{m^m \rho^i}{m!} p_0, i \ge m+1$$

  $$p_i = \begin{cases} \frac{(m\rho)^i}{i!} p_0 & 1 \le i \le m \\ \frac{m^m \rho^i}{m!} p_0 & i \ge m+1 \end{cases} = \begin{cases} \frac{(m\rho)^i}{i!} p_0 & 0 \le i \le m \\ \frac{m^m \rho^i}{m!} p_0 & i \ge m+1 \end{cases} = \begin{cases} \frac{(m\rho)^i}{i!} p_0 & 0 \le i \le m-1 \\ \frac{m^m \rho^i}{m!} p_0 & i \ge m \end{cases}$$

  Note: for i = m, can use any equation.

- **P_Q** = P{Queuing} = probability that an arrival will find all servers busy and will be forced to wait in queue

- **Erlang C formula**: $P_Q = \frac{p_0 (m\rho)^m}{m!(1-\rho)} = \frac{p_m}{1-\rho}$

  - probability that an arrival will find all servers busy and will be forced to wait in queue = P_{block}
  - Since an arriving customer finds the system in "typical" state

    $$P_Q = P\{N \ge m\}$$

    $$= \sum_{i=m}^{\infty} p_i = \sum_{i=m}^{\infty} \frac{m^m \rho^i}{m!} p_0 = p_0 \sum_{i=m}^{\infty} \frac{m^m \rho^i}{m!} = p_0 \sum_{k=0}^{\infty} \frac{m^m \rho^{k+m}}{m!} \quad ; k = i - m$$

    $$= p_0 \sum_{k=0}^{\infty} \frac{m^m \rho^k \rho^m}{m!} = p_0 \frac{m^m \rho^m}{m!} \sum_{k=0}^{\infty} \rho^k = p_0 \frac{(m\rho)^m}{m!(1-\rho)}$$

- The probability of a call request finding all of the *m* circuits of a transmission line busy, assumed that such a call request "remains in queue," that is, continuously attempts to find a free circuit.

- $p_0 = \dfrac{1}{\sum_{i=0}^{m-1} \dfrac{(m\rho)^i}{i!} + \dfrac{(m\rho)^m}{m!(1-\rho)}}$

$$\sum_{i=0}^{\infty} p_i = \sum_{i=0}^{m-1} p_i + \sum_{i=m}^{\infty} p_i = \sum_{i=0}^{m-1} \frac{(m\rho)^i}{i!} p_0 + P_Q = \sum_{i=0}^{m-1} \frac{(m\rho)^i}{i!} p_0 + p_0 \frac{(m\rho)^m}{m!(1-\rho)}$$

$$= p_0 \left(\sum_{i=0}^{m-1} \frac{(m\rho)^i}{i!} + \frac{(m\rho)^m}{m!(1-\rho)} \right)$$

Set $\sum_{i=0}^{\infty} p_i = 1$ gives $p_0 = \dfrac{1}{\sum_{i=0}^{m-1} \dfrac{(m\rho)^i}{i!} + \dfrac{(m\rho)^m}{m!(1-\rho)}}$

- $N_Q = \sum_{i=0}^{\infty} i p_{i+m} = \sum_{i=0}^{\infty} i \dfrac{m^m \rho^{i+m}}{m!} p_0 = \dfrac{(\rho m)^m}{m!} p_0 \sum_{i=0}^{\infty} i \rho^i = \dfrac{(\rho m)^m}{m!} p_0 \dfrac{\rho}{(1-\rho)^2} = P_Q \dfrac{\rho}{1-\rho}$

- $N = N_Q + N_S$

$N_S = m\rho = \dfrac{\lambda}{\mu}$

$N = N_Q + m\rho$

- Another way to find N_S

$$N_S = \sum_{i=0}^{m-1} i p_i + \sum_{i=m}^{\infty} m p_i = \sum_{i=0}^{m-1} i \frac{(m\rho)^i}{i!} p_0 + mP_Q$$

$$= \sum_{i=1}^{m-1} \frac{(m\rho)^i}{(i-1)!} p_0 + mp_0 \frac{(m\rho)^m}{m!(1-\rho)}$$

$$= p_0 \left(\left(\sum_{i=1}^{m} \frac{(m\rho)^i}{(i-1)!} - \frac{(m\rho)^m}{(m-1)!} \right) + m \frac{(m\rho)^m}{m!(1-\rho)} \right)$$

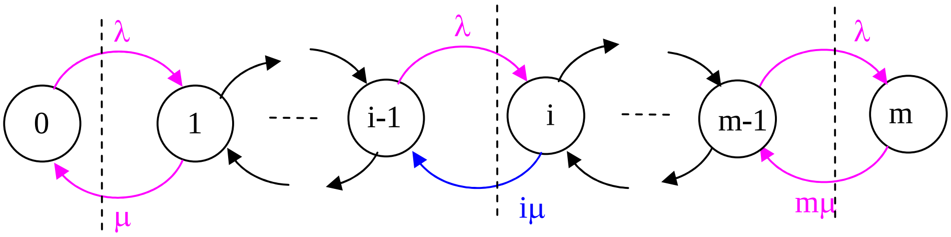
$$= p_0 \left(\sum_{k=0}^{m-1} \frac{(m\rho)^{k+1}}{k!} + \left(-\frac{m(1-\rho)(m\rho)^m}{m!(1-\rho)} + m \frac{(m\rho)^m}{m!(1-\rho)} \right) \right)$$

$$= p_0 \left(m\rho \sum_{k=0}^{m-1} \frac{(m\rho)^k}{k!} + m\rho \frac{(m\rho)^m}{m!(1-\rho)} \right)$$

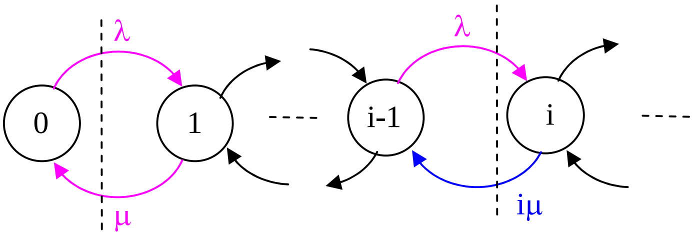
$$= \frac{m\rho \left(\cancel{\sum_{k=0}^{m-1} \frac{(m\rho)^k}{k!} + \frac{(m\rho)^m}{m!(1-\rho)}} \right)}{\cancel{\sum_{i=0}^{m-1} \frac{(m\rho)^i}{i!} + \frac{(m\rho)^m}{m!(1-\rho)}}} = m\rho$$

M/M/m/m ⇒ Erlang model

- $\rho = \dfrac{\lambda}{\mu}$
- $p_0 = \dfrac{1}{\sum_{i=0}^{m} \dfrac{\rho^i}{i!}}$
- $p_n = p_0 \dfrac{\rho^n}{n!} = \dfrac{\dfrac{\rho^n}{n!}}{\sum_{i=0}^{m} \dfrac{\rho^i}{i!}}$
- **Erlang B formula**: $p_{lost} = p_m = \dfrac{\dfrac{\rho^m}{m!}}{\sum_{i=0}^{m} \dfrac{\rho^i}{i!}}$



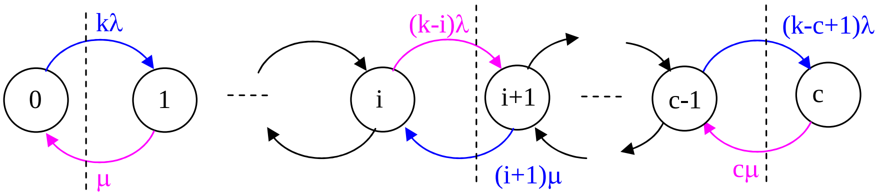
M/M/∞



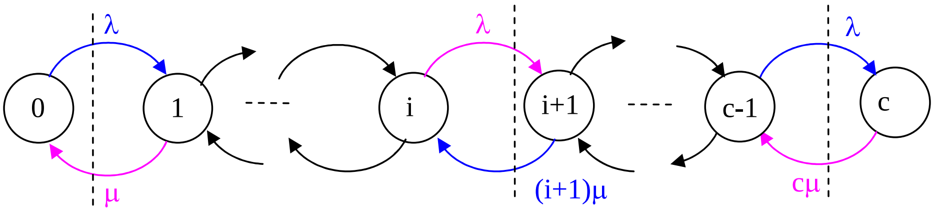
- $\rho = \dfrac{\lambda}{\mu}$
- $p_0 = \dfrac{1}{\sum_{i=0}^{\infty} \dfrac{\rho^i}{i!}} = \dfrac{1}{e^{\rho}} = e^{-\rho}$
- $p_i = p_0 \dfrac{\rho^i}{i!} = e^{-\rho} \dfrac{\rho^i}{i!} \Rightarrow$ Poisson

Comparison

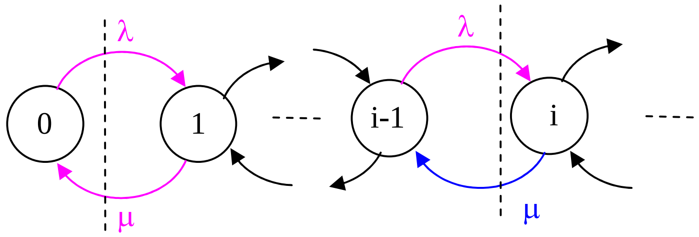
Engsett:



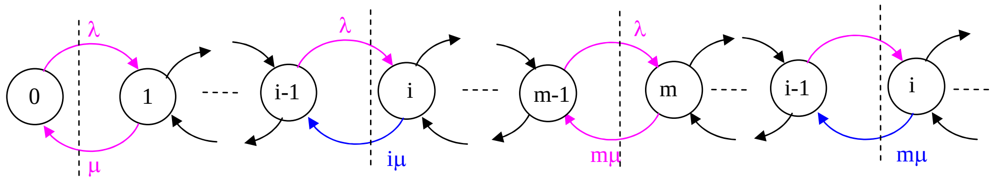
Erlang:



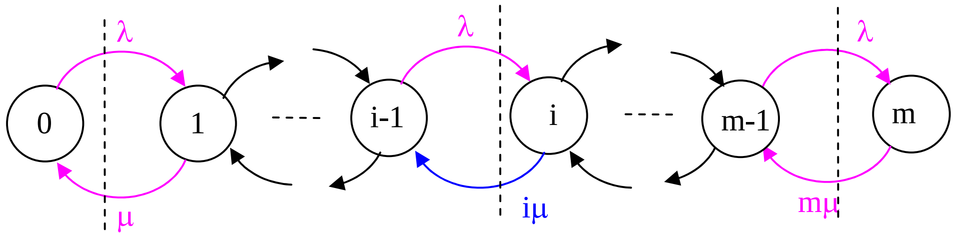
M/M/1



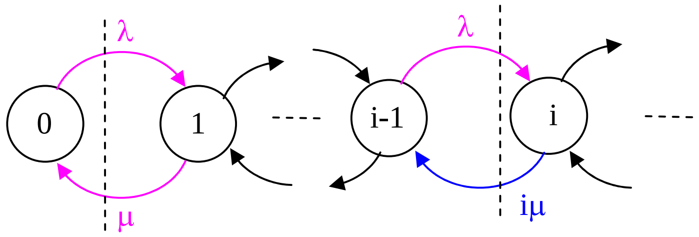
M/M/m



M/M/m/m



M/M/∞



Etc

- Burke's theorem:

  For an M/M/1. M/M/c, or M/M/∞ queuing system at steady state with arrival rate λ, then

  - The departure process is Poisson with rate λ
  - At each time t, the number of customers in the system n(t) is independent of the sequence of departure times prior to t.