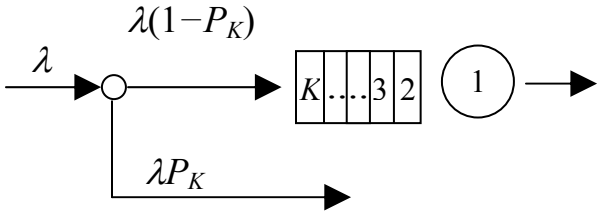
Queueing Theory (4)

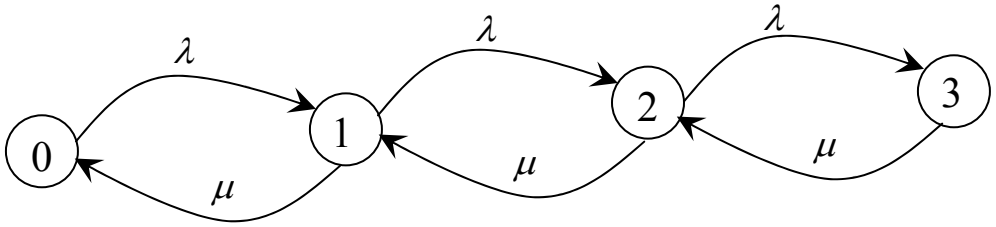
- **The *M*/*M*/1/*K* queue**
  - This is a single-server queue with Poisson arrivals at rate λ, and exponential service times with rate μ.
  - However, the system can accommodate at most K customers (i.e., there are only $K-1$ waiting spaces).
  - K is known as the buffer size.
  - E.g.,
    - A bank that has room for at most K customers.
    - A manufacturing station with a WIP buffer of capacity K.
    - A call center that can handle at most K calls.
  - Arriving customers who find the system "full" leave immediately (these are "lost" customers).



  - The number of customers in the $M/M/1/K$ system $L(t)$ is a birth death process with states 0, 1, 2, …, ,K, $\mu_n = \mu$, and

$$\lambda_n = \begin{cases} \lambda, & \text{if } n < K \\ 0 & \text{if } n \geq K \end{cases}$$

- The transition diagram for $K = 3$ is shown below.



- Define $\rho = \lambda / \mu$.
- Note that ρ needs *not* to be less than 1 here. (why?)
- Applying the general flow balance equation for a birth-death process, the limiting probabilities are given by

$$P_0 = \left(1 + \sum_{n=1}^{K} \frac{\lambda_{n-1}\lambda_{n-2}\ldots\lambda_1\lambda_0}{\mu_n\mu_{n-1}\ldots\mu_2\mu_1} \right)^{-1} = \left(\sum_{n=0}^{K} \left(\frac{\lambda}{\mu} \right)^n \right)^{-1} = \left(\sum_{n=0}^{K} \rho^n \right)^{-1},$$

where $\sum_{n=0}^{K} \rho^n = (1-\rho^{K+1})/(1-\rho)$ if $\rho \neq 1$, and

$$\sum_{n=0}^{K} \rho^n = K+1 \quad \text{if } \rho = 1.$$

- Then,

$$P_0 = \begin{cases} \dfrac{1-\rho}{1-\rho^{K+1}}, & \text{if } \rho \neq 1 \\ \dfrac{1}{K+1}, & \text{if } \rho = 1 \end{cases}$$

- And, for $n = 1,2, \ldots, K$

$$P_n = \begin{cases} \dfrac{\rho^n(1-\rho)}{1-\rho^{K+1}}, & \text{if } \rho \neq 1 \\ \dfrac{1}{K+1}, & \text{if } \rho = 1 \end{cases}$$

- The probability that a customer is lost is P_K .
- The "effective" arrival rate λ_e is the rate of arrivals who join the system. Then,

$$\lambda_e = \lambda(1 - P_K) .$$

- The fraction of time the server is busy is λ_e/μ .
- If $\rho \neq 1$, the mean number in the system is

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

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

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

- If $\rho = 1$,

$$L = \sum_{n=1}^{K} nP_n = \sum_{n=1}^{K} \frac{n}{K+1} = \frac{K(K+1)}{2(K+1)} = \frac{K}{2}.$$

- Little's law implies that the mean waiting time is

$$W = \frac{L}{\lambda_e}$$

- The mean number in queue is

$$L_q = L - \frac{\lambda_e}{\mu}.$$

- The mean waiting time is

$$W_q = \frac{L_q}{\lambda_e}.$$

Fact. $\lambda_e = \lambda(1 - P_K) = \mu(1 - P_0)$.

- **Example 8**
  - Cars arrive to the drive-in at Hot Dog King according to a Poisson process with rate 40 per hour. If a total of more than four cars are in line (including the car at the window) a car will not enter the line. The average service time at the window is exponentially distributed with mean 4 minutes.
  - What is the mean number of cars waiting?
  - This can be modeled as an $M/M/1/K$ with $\lambda = 40$, $\mu = 15$, and $K = 4$. Then, $\rho = \lambda/\mu = 40/15 = 2.666 \neq 1$, and

$$L = \frac{\rho}{1-\rho} - \frac{\rho^{K+1}(K+1)}{1-\rho^{K+1}} = \frac{2.667}{1-2.667} - \frac{2.667^5 \times 5}{1-2.667^5} = 3.44.$$

  - In addition,

$$\lambda_e = \mu(1 - P_0) = \mu\left(1 - \frac{1-\rho}{1-\rho^{K+1}}\right) = 15\left(1 - \frac{1-2.667}{1-2.667^5}\right) = 14.81$$

  - Finally, $L_q = L - \lambda_e/\mu = 3.44 - 14.81/15 = 2.45$.

- On average, how many cars are served per hour?
- $\lambda_e = 14.81$.
- On average, how long will a car stay at the drive-in window before receiving food?
- $W = L/\lambda_e = 3.44/\ 14.81 = 0.23$ hours $\approx$ 14 minutes.

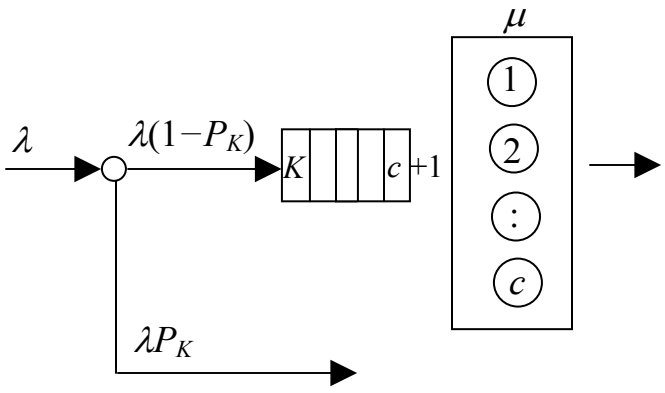
- **Example 9**
  - Seas Beginnings, a small mail order firm, has one phone operator. Calls arrive to Seas Beginnings at a Poisson rate of 60 per hour, and it takes an exponentially distributed time with mean 1 minute to handle a call. When the operator is busy, an incoming call is put on hold (with "nice" music) in one of K phone lines Seas Beginnings. If all K lines are busy (meaning that one call is being handled and $K - 1$ are on hold), a caller gets a busy signal and calls a competitor (Air End). Seas Beginnings wants at most 1% of caller to get a busy signal. How many phone lines should be provided?
  - This can be modeled as an $M/M/1/K$ with $\lambda = \mu = 60$ / hour. Then, $\rho = 1$, and $P_K = 1/\ (K+1)$. The desired service level requires $P_K \leq 0.01 \Rightarrow K+1 > 100 \Rightarrow K = 100$.

- What about mean caller delay?
- $W_q = (L - \lambda_e/\mu)/\ \lambda_e = [K/2 - 1/(K+1)]/\{\lambda[1 - 1/(K+1)]\}$

  $= 0.84$ hours ≈ 50 minutes. (too long).
- How can this system be improved?
- Add more operators to bring mean delay down while maintaining a "rejection" probability of 1%. For this, we need to generalize $M/M/1/K$ to multi-server. This is done next.

- **The $M/M/c/K$ queue**
  - This is a generalization of $M/M/1/K$ to many servers. Specifically, this is a Markovian queue with c servers and $K - c$ waiting spaces (where $K > c$).
  - The number of customers in the $M/M/c/K$ system, $L(t)$, is a birth death process with states 0, 1, 2, …, ,K, and

$$\lambda_n = \begin{cases} \lambda, & \text{if } n < K \\ 0 & \text{if } n \geq K \end{cases} \qquad \mu_n = \begin{cases} n\mu, & \text{if } n < c \\ c\mu & \text{if } c \leq n \leq K \end{cases}$$



- Let $a = \lambda/\mu$, and $\rho = a/c$. Applying birth-death flow balance equation gives

$$P_0 = \begin{cases} \left(\sum_{n=0}^{c-1} \frac{a^n}{n!} + \frac{a^c(1-\rho^{K-c+1})}{c!(1-\rho)} \right)^{-1}, & \text{if } \rho \neq 1 \\ \left(\sum_{n=0}^{c-1} \frac{a^n}{n!} + \frac{a^c(K-c+1)}{c!} \right)^{-1}, & \text{if } \rho = 1 \end{cases}$$

- Then,

$$P_n = \begin{cases} \frac{a^n}{n!} P_0, & \text{if } n < c \\ \frac{a^n}{c!c^{n-c}} P_0 & \text{if } c \leq n \leq K \end{cases}$$

- And,

$$L_q = \begin{cases} \frac{a^c \rho}{c!(1-\rho)^2} \left[1 - \rho^{K-c+1} - (1-\rho)(K-c+1)\rho^{K-c} \right] P_0, & \text{if } \rho \neq 1 \\ \frac{c^c}{c!} \left[\frac{(K-c)(K-c+1)}{2} \right] P_0, & \text{if } \rho = 1 \end{cases}$$

- The effective arrival rate is $\lambda_e = \lambda(1 - P_K)$. Other measures of performance are found as follows.

$$W_q = \frac{L_q}{\lambda_e},$$

$$W = W_q + \frac{1}{\mu},$$

$$L = \lambda_e W.$$

- **Example 10**
  - How many more operators should Sea Beginnings needs mean delay down while maintaining a "rejection" probability of 1%.
  - Consider adding two servers. The resulting *M/M*/2/100 system has $\lambda = \mu = 60$, $a = 1$, and $\rho = 0.5$.
  - Then,

$$P_0 = \left(\sum_{n=0}^{c-1} \frac{a^n}{n!} + \frac{a^c (1-\rho^{K-c+1})}{c!(1-\rho)} \right)^{-1} = \left(1+1+\frac{1-0.5^{99}}{2\times 0.5} \right)^{-1} = 0.333$$

$$P_K = \frac{a^K}{c!c^{K-c}} P_0 = \frac{0.333}{2\times 2^{98}} = 0$$

$$L_q = \frac{a^c \rho}{c!(1-\rho)^2}\left[1-\rho^{K-c+1}-(1-\rho)(K-c+1)\rho^{K-c}\right]P_0$$

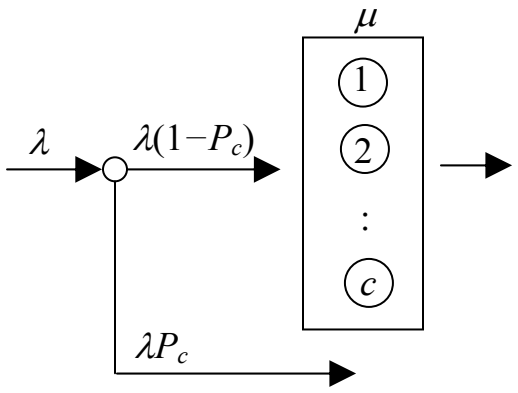
$$= \frac{0.5}{2(0.5)^2}\left[1-0.5^{99}-0.5\times 99\times 0.5^{98}\right](0.333) = 0.333$$

$$\lambda_e = \lambda(1-P_K) = 60$$

$$W_q = \frac{L_q}{\lambda_e} = \frac{0.333}{60} \text{ hours} = 1/3 \text{ min}$$

  - But obviously here, there are more lines than needed. In your HW, you will determine the minimum number of operators and lines that achieve the desired service level.

- **The *M*/*M*/*c*/*c* Erlang loss model**
  - This a special case of *M*/*M*/*c*/*K* with $K = c$.
  - That is, there is no waiting. Incoming customers that find all servers busy leave the system.



  - Applying the formulas for *M/M/c/K* with $K = c$,

$$P_n = \frac{a^n / n!}{\sum_{n=0}^{c} \frac{a^n}{n!}}, \quad n = 0, 1, 2, \ldots, c$$

  - In particular, *Erlang's loss formula* is

$$B(c, a) \equiv P_c = \frac{a^c / c!}{\sum_{n=0}^{c} \frac{a^n}{n!}}.$$

  - Note that $B(c,a) = P\{\text{all servers are busy}\}$
    $= P\{\text{an arrival will be rejected}\}$.
  - Erlang, a Swedish engineer, developed this model for a simple telephone network.
  - This is considered the first application of queueing theory.

- An interesting feature of the Erlang loss model is that the system size distribution formula, holds for any service time distribution.
- That is, for an $M/G/c/c$ system

$$P_n = \frac{a^n / n!}{\sum_{n=0}^{c} \frac{a^n}{n!}}, \quad n = 0, 1, 2, \ldots, c$$

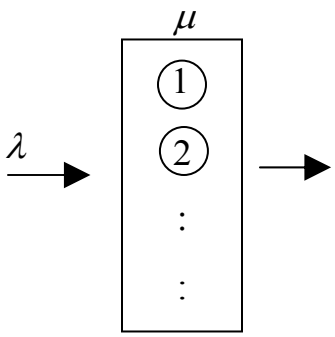
- That is, P_n is *insensitive* to service time variability. It only depends on the mean service time $E[S]$. (More specifically on $a = \lambda E[S]$).

- **Example 11**
  - What is the minimal number of servers needed, in an $M/M/c/c$ Erlang loss system, to handle an offered load $a = \lambda/\mu = 2$ erlangs, with a loss no higher than 2%?
  - Starting with $c = 1$, increase c until $B(c, a) < 0.02$.

c	***B*(*c*, 2)**
1	2/3
2	2/5
3	4/19
4	$2/21 \approx 0.095$
5	$4/109 \approx 0.095$
6	$4/381 \approx 0.01$

  - Therefore, 6 servers are needed to achieve the desired service level.

- **The $M/M/\infty$ unlimited service model**
  - This is an *M/M/c* queue with an infinite number of servers.



  - It applies for example to a self-service situation.
  - The number of customers in the $M/M/\infty$ system $L(t)$ is a birth-death process with $\lambda_n = \lambda$, and $\mu_n = n\mu$, $n = 0,1,2, \ldots$
  - Applying the birth-death flaw balance equations gives

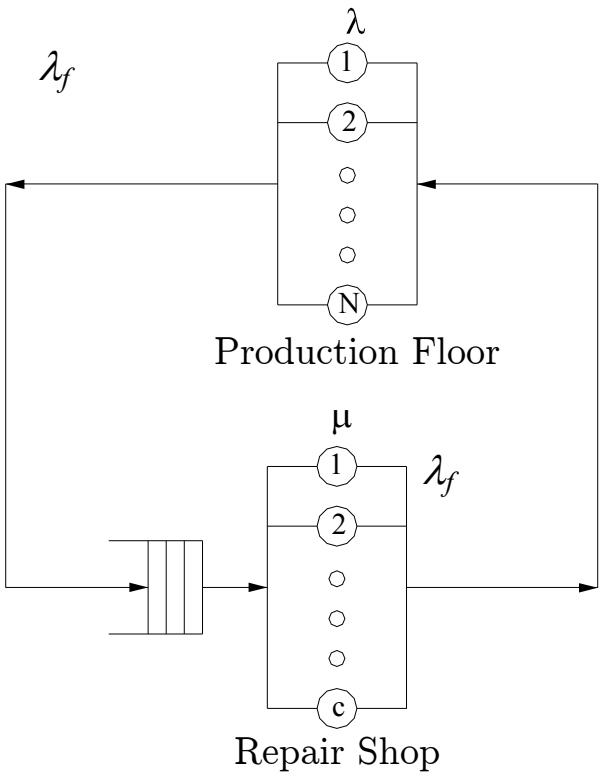
$$P_n = \frac{a^n}{n!} e^{-a}, \quad n = 0,1,2,\ldots,$$

  - That is, the number of busy servers is a Poisson random variable with mean $a = \lambda/\mu$.
  - It can be shown that this Poisson distribution is *insensitive* to service times variability. That is, it holds for $M/G/\infty$ queue.
  - Note that the mean number of busy servers is a.

- **Example 12**
  - Television station KCAD in a large metropolitan area wishes to know the average number of viewers it can expect on a Saturday evening prime-time program. It has found from past surveys that people turning on their television sets on Saturday evening during prime time can be described rather well by a Poisson distribution with a mean of 100,000/hour. There are five major TV stations in the area, and it is believed that a given person chooses among these essentially at random. Surveys have also showed that a person tunes in for an average time of 90 minutes.
  - This can be modeled as an $M/G/\infty$ with λ = 100,000 /5 = 20,000 persons/hour and μ = 1/(3/2) = 2/3. Then, the mean number of viewers is $a = \lambda/\mu = 30{,}000$.
  - What is the standard deviation of the number of viewers?
  - The standard deviation is $\sqrt{a} = \sqrt{30000} = 173.2$.

- **The *M*/*M*/*c* // *N* machine repair model**
  - This is a model with a *N* machines.
  - The machines can be in one of two states: Operational or In-Repair.
  - Operational machines are on the production floor.

- Each machine fails at an exponential rate λ independent of other machines.
- A failed machine is moved to the repair shop staffed by $c \leq N$ repairmen. If all repairmen are busy the machine waits in queue.
- Repair times are exponentially distributed with rate μ.



- This model is different from other "open" models in that there is a limited number of entities N that move around the system. This is called a *closed queueing network*.
- The number of machines in the repair shop, $L(t)$, is a birth-death process with

$$\lambda_n = (N-n)\lambda,\ n \leq N$$

$$\mu_n = \begin{cases} n\mu, & \text{if } n < c \\ c\mu & \text{if } c \le n \le N \end{cases}$$

- Applying the birth-death flaw balance equations gives

$$P_0 = \left(\sum_{n=0}^{c-1} \frac{N!}{(N-n)!n!} a^n + \sum_{n=c}^{N} \frac{N!}{(N-n)!c!c^{n-c}} a^n \right)^{-1}$$

$$P_n = \begin{cases} \dfrac{N!a^n}{(N-n)!n!} P_0, & \text{if } n < c \\ \dfrac{N!a^n}{(N-n)!c!c^{n-c}} P_0, & \text{if } c \le n \le N \end{cases}$$

- The mean number in the repair system (mean number of down machines) is

$$L = \sum_{n=0}^{N} np_n .$$

- The mean number of up machines is $N - L$.
- The most important measure of performance of such system is the effective arrival rate λ_f, also known as the *aggregate failure rate* or the *throughput*. By conditioning on $L(t)$,

$$\lambda_f = \sum_{n=0}^{N} (N-n)\lambda p_n = N\lambda - L\lambda = (N-L)\lambda .$$

- Other measure of performance are
  - The mean number in repair queue, $L_q = L - \lambda_f / \mu$,
  - The waiting time for repair, $W_q = L_q / \lambda_f$,

- The total time in the repair system, $W = W_q + 1/\mu$.

- **Example 13**
  - The Train SemiConductor Company uses five robots in the manufacturing of its circuit boards. The robots break down periodically, and the company has two repair people to do service when robots fail. When one is fixed, the time until the next breakdown is thought to be exponentially distributed with mean 30 hours. The shop always has enough of a work backlog to ensure that all robots in operating condition will be working. The repair time for each service is thought to be exponentially distributed with mean 3 hours.
  - The shop manager wished to know the average number of robots operational at any given time, the expected down time of a robot, and the expected fraction of time the repairmen are idle.
  - This can be modeled as an $M/M/2$ // 5 queue with $\lambda = 1/30$, $\mu = 1/3$. Then, $a = \lambda/\mu = 0.1$.
  - Then,

$$P_0 = \left(\sum_{n=0}^{c-1} \frac{N!}{(N-n)!n!} a^n + \sum_{n=c}^{N} \frac{N!}{(N-n)!c!c^{n-c}} a^n \right)^{-1}$$

$$= \left(1 + 5(0.1) + 10(0.1)^2 + 15(0.1)^3 + 15(0.1)^4 + 7.5(0.1)^5 \right)^{-1}$$

$$= 0.619$$

$$L = \sum_{n=0}^{5} nP_n = \left[5(0.1)(1) + 10(0.1)^2(2) + 15(0.1)^3(3) + 15(0.1)^4(4) + 7.5(0.1)^5(5) \right] P_0$$

$$= 0.465$$

$$\lambda_f = (N - L)\lambda = (5 - 0.465)(1/30) = 0.151$$

- Therefore, the average number of operational robots is

$$N - L = 4.535.$$

- The expected fraction of time a repairman is idle is

$$P_0 + (1/2)P_1 = 0.619\ [1 + (0.5)(5)(0.1)] = 0.773.$$

- The expected down time of a robot is

$W = L / \lambda_f = 3.075$ hours .

Remark.

- The following algorithm facilitates the computations.

Step 0. Set $P_0 = 1$, $S = 1$, $L = 0$.

Step 2. For $n = 1$ to $c - 1$

Set $P_n = [(N - n)/n]\ P_{n-1}$, $S = S + P_n$, $L = L + nP_n$

Step 2. For $n = c$ to N

Set $P_n = [(N - n)/c]\ P_{n-1}$, $S = S + P_n$, $L = L + nP_n$

Step 3. Set $L = L/S$.

Step 4. For $n = 0$ to N

Set $P_n = P_n/S$.