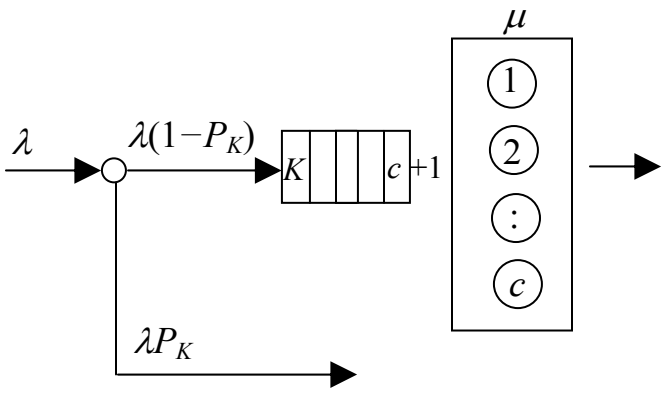
Queueing Theory (3)

- **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,

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

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

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

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

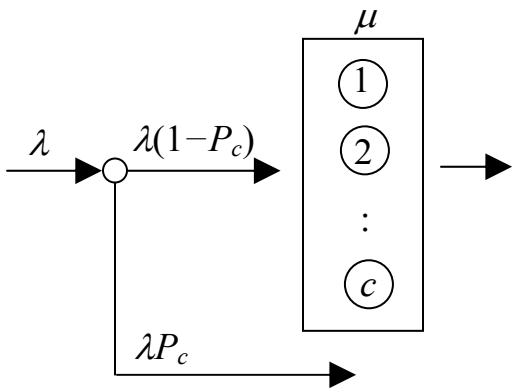
$$\lambda_e = \lambda(1-P_K) = 60$$

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

- 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

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

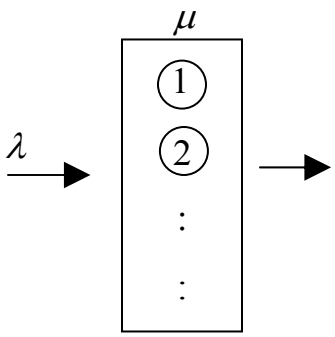
- That is, P_n is *insensitive* to service time variability. It only depends on the mean service time $E[S]$. (More specifically on $a = \lambda E[S]$).

- **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 $\sqrt{a} = \sqrt{30000} = 173.2$.