Queueing Theory (2)

- **Distribution of waiting time in *M*/*M*/1**
  - Let T_q be the waiting time in queue of a customer.
  - Then it can be shown that,

$$P\{T_q > t\} = \rho e^{-(\mu-\lambda)t} .$$

  - Let T be the total time of a customer in the system (in queue plus in service). Then, it can be shown in a similar way that

$$P\{T > t\} = e^{-(\mu-\lambda)t} .$$

- **Example 2**
  - In the bank of Example 1, what is the probability that the waiting time of a customer exceeds 2 hours?
  - $\mathrm{P}\{T_q > t\} = \rho e^{-(\mu-\lambda)t} = 0.9e^{-2} = 0.122.$
  - What is the probability that a customer spends more than 3 hours in the bank?
  - $\mathrm{P}\{T > t\} = e^{-(\mu-\lambda)t} = e^{-3} = 0.05.$

- **Example 3**
  - Consider an Interface Message Processor (IMP). The packet size in bits is exponentially distributed with mean $1/\mu$ bits/packet. The capacity of the communication channel is C bits/sec. Packets arrive at random (i.e. exponential inter

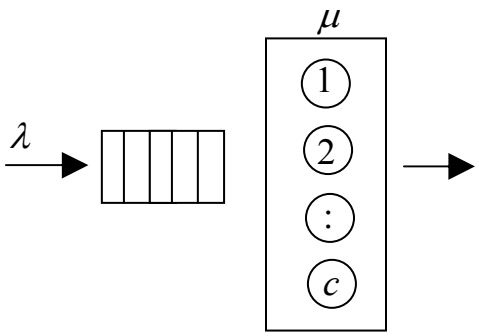
arrival times) with arrival rate λ packets/sec. Find the queueing delay for packets at the IMP.

- Model this as a $M/M/1$ with arrival rate λ packets/sec and service rate $C/(1/\mu) = \mu C$ packets/sec.
- Then, the mean delay per packet is $W_q = 1/(\mu C - \lambda)$.

- **Example 4 (Dedicated vs. shared channels).**
  - Two computers are connected by a 64 kbps line. There are eight parallel sessions using the line. Each session generates Poisson traffic with a mean of 2 packets/sec. The packet lengths are exponentially distributed with mean of 2000 bits. The system designers must choose between giving each session a dedicated 8 kbps piece of bandwidth or having all packets compete for a single 64 kbps shared channel. Which alternative gives better response time (i.e. W)?
  - We need to compare two alternative $M/M/1$ models.
  - The first model has four channels each operating as a $M/M/1$ with $\lambda_1 = 2$ packets/sec and $\mu_1 = 8000/2000 = 4$ packets/sec. Therefore, $W^1 = 1/(\mu_1 - \lambda_1) = 1/2 = 0.5$ sec.
  - The second model has one channel with $\lambda_2 = 8 \times 2 = 16$ packets/sec, and $\mu_2 = 64000/2000 = 32$ packets/sec. Therefore, $W^2 = 1/(\mu_2 - \lambda_2) = 1/16 = 0.063$ sec.

- The response time in the shared system is 8 times less. This is the better alternative.
- (Note that both systems have the same $\rho = 0.5$.)
- Why this result?
- In the dedicated system, some channels could be idle, while other channels have long queues. This does not happen in the shared system.

- **The *M*/*M*/*c* queue**
  - This is a queue with a Poisson arrival process with rate λ, exponential service times with rate μ and c servers.
  - It is a generalization of $M/M/1$ with multi-servers.

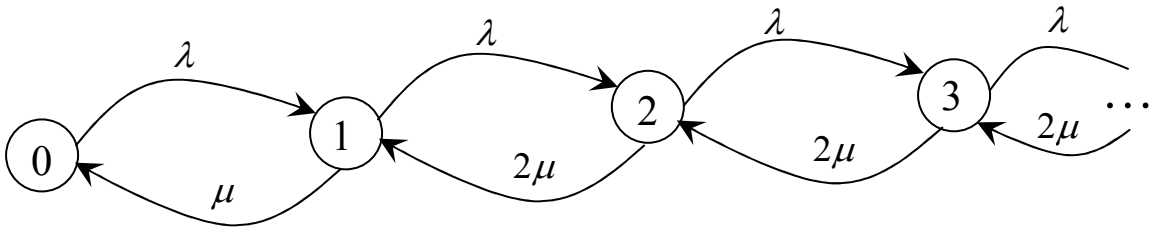


  - The number of customers in the $M/M/c$ system $L(t)$ is a birth death process with $\lambda_n = \lambda$, and

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

  - The expression for μ_n follows since the minimum of n exponential rvs with rate μ is exponential with rate $n\mu$.

- The transition diagram for $c = 2$ is shown below.



- Recall that $\rho = \lambda/(c\mu)$ is the traffic intensity.
- Define $a = \lambda/\mu$. This is the mean number of busy servers.
- In the following we assume $\rho < 1$.
- Applying the general flow balance equation for a birth-death process, the limiting probabilities are given by

$$P_0 = \left(1 + \sum_{n=1}^{\infty} \frac{\lambda_{n-1}\lambda_{n-2}\ldots\lambda_1\lambda_0}{\mu_n\mu_{n-1}\ldots\mu_2\mu_1} \right)^{-1}$$

$$= \left(1 + \sum_{n=1}^{c-1} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n + \sum_{n=c}^{\infty} \frac{1}{c!c^{n-c}}\left(\frac{\lambda}{\mu}\right)^n \right)^{-1}$$

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

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

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

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

- Then,

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

- Then, the mean number in queue is

$$\boxed{L_q} = \sum_{n=c}^{\infty} (n-c) P_n = \sum_{n=c}^{\infty} (n-c) \frac{a^n}{c!c^{n-c}} P_0 = \frac{a^c P_0}{c!} \sum_{n=c}^{\infty} (n-c) \frac{a^{n-c}}{c^{n-c}}$$

$$= \frac{a^c P_0}{c!} \sum_{m=0}^{\infty} m\rho^m = \boxed{\frac{a^c \rho}{c!(1-\rho)^2} P_0}$$

since $\sum_{m=0}^{\infty} m\rho^m = \dfrac{\rho}{(1-\rho)^2}$

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

$$W_q = \frac{L_q}{\lambda} = \frac{a^c \rho}{\lambda c!(1-\rho)^2} P_0 = \frac{a^c}{c!(c\mu)(1-\rho)^2} P_0$$

- In addition, the mean number in the system is

$$L = L_q + a = a + \frac{a^c \rho}{c!(1-\rho)^2} P_0$$

- And the mean time in the system, is

$$W = W_q + \frac{1}{\mu} = \frac{1}{\mu} + \frac{a^c}{c!(c\mu)(1-\rho)^2} P_0$$

- The probability that all servers are busy is

$$P_{c+} = \sum_{n=c}^{\infty} P_n = \sum_{n=c}^{\infty} \frac{a^c}{c!c^{n-c}} P_0 = \frac{a^c}{c!(1-\rho)} P_0 = \frac{P_c}{1-\rho}$$

- Let T_q be the waiting time in queue of a customer. Then,

$$P\{T_q > t\} = \frac{a^c P_0}{c!(1-\rho)} e^{-c\mu(1-\rho)t}$$

- This can be shown similar to the $M/M/1$ case.

- **Example 5 (enhanced bank service)**
  - In the $M/M/1$ bank model with $\lambda = 9$ customers/hour and $\mu = 10$ customers/hour. We find that the mean waiting time is $W_q = 54$ mins.
  - Management is considering adding more servers to bring the mean waiting time to less than 5 minutes. How many more servers should be added?
  - Try adding another server. For this $M/M/2$ system, $a = 0.9$ and $\rho = a/2 = 0.45$. Then,

$$P_0 = \left(\sum_{n=0}^{c-1} \frac{a^n}{n!} + \frac{a^c}{c!(1-\rho)} \right)^{-1} = \left(1 + 0.9 + \frac{0.9^2}{2!(1-0.45)} \right)^{-1} = 0.379$$

$$W_q = \frac{a^c}{c!(c\mu)(1-\rho)^2} P_0 = \frac{0.9^2}{2!(2\times 10)(1-0.45)^2} 0.379$$
$$= 0.025 \text{ hours} = 1.5 \text{ mins}$$

  - Adding one more server achieves the desired service level.

- **Example 6.**
  - An airline is planning a new telephone reservation center. Each agent will have a reservations terminal and can serve a typical caller in 5 minutes, the service time being exponentially distributed. Calls arrive randomly and the system has a large message buffering system to hold calls that arrive when no agent is free. An average of 36 calls per hour is expected during the peak period of the day. The design criterion for the new facility is that the probability a caller will find all agents busy must not exceed 0.1 (10%).
  - How many terminals should be provided?
  - This can be modeled as an $M/M/c$ with $\lambda = 36$ calls /hour, $\mu = 60/5 = 12$ calls /hour, and c is to be determined such that $P_{c+} < 0.1$.
  - The minimum number of terminals, c, needed is one that achieves stability. That is,

$$\rho = \lambda/(c\mu) < 1 \Rightarrow 36/(12c) < 1 \Rightarrow c > 3.$$

  - Try $c = 4$, then $\rho = 36/48 = 0.75$ and $a = 4\rho = 3$.

$$P_0 = \left(\sum_{n=0}^{c-1} \frac{a^n}{n!} + \frac{a^c}{c!(1-\rho)} \right)^{-1} = \left(1 + 3 + \frac{3^2}{2} + \frac{3^3}{6} + \frac{3^4}{4!(1-0.75)} \right)^{-1}$$
$$= 0.037736$$
$$P_{c+} = \frac{a^c}{c!(1-\rho)} P_0 = \frac{3^4}{4!(1-0.75)} 0.037736 = 0.509$$

  - So, four terminals won't do it.

- Try $c = 5$. Repeating the same computations yields

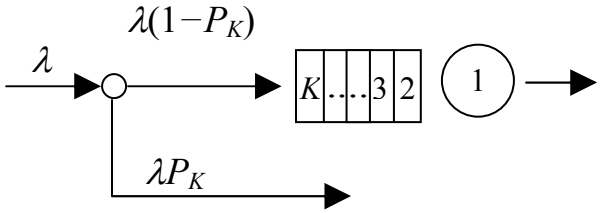
$$P_{c+} = 0.232 .$$

- Try $c = 6$. Repeating the same computations yields

$$P_{c+} = 0.0991 .$$

- So, six terminals are needed to achieve the desired service level.

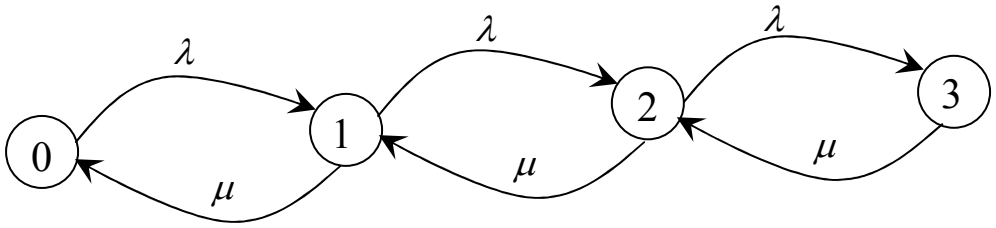
- **The *M*/*M*/1/*K* queue**
  - This is a single-server queue with Poisson arrivals at rate λ, and exponential service times with rate μ.
  - However, the system can accommodate at most K customers (i.e., there are only $K - 1$ waiting spaces).
  - K is known as the buffer size.
  - E.g.,
    - A bank 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 7**
  - 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%. Generalize $M/M/1/K$ to multi-server.