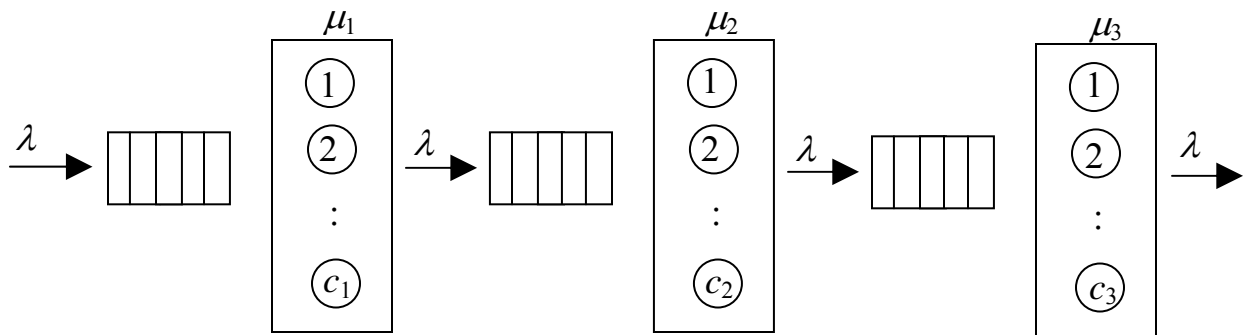


Queueing Theory (4)

- **Series Queues**

- Consider n queueing stations in series, where each station can be modeled as $M/M/c_i$, where c_i is the number of servers in station i , $i = 1, 2, \dots, n$.
- Customers arrive to the system according to a Poisson process with rate λ . All customers are served in series in stations 1 to n .
- Queueing could occur at any station. Assume that there is ample waiting space at all stations.
- The service time at station i , is exponential with rate μ_i .



- E.g.,
 - A manufacturing assembly line.
 - Traffic lights.
 - Clinic physical examination procedure.
 - Shopping at a grocery store.

- This series system can be analyzed based on the following fact.

Fact. *The output (departure) process from an $M/M/c$ queue is Poisson with the same parameter λ as the arrival process.*¹

- Then, each station can be analyzed as an *independent* $M/M/c_i$ with arrival rate λ and service rate μ_i .

- **Example 14.**

- Customers arrive to a supermarket at a Poisson rate of 40/hour during peak hours. It takes a customer on the average 3/4 hour to fill his shopping cart, the filling time being exponentially distributed. Upon filling their shopping cart customers move to a check-out line staffed by c cashiers, where they wait in a single line if all cashiers are busy. There is enough space for any number of waiting customers. Check-out time is exponentially distributed with mean 4 min.
- What is the minimum number of cashiers required during peak hours?
- This system can be modeled as two stations in series, with the first station as $M/M/\infty$ with $\lambda_1 = 40$ and $\mu_1 = 4/3$ and the second station as $M/M/c$ with $\lambda_2 = 40$ and $\mu_2 = 15$.

¹ This fact does *not* hold for an $M/G/c$ queue with non-exponential service times.

- In order for the check-out station to be stable,

$$\rho_2 = \lambda_2 / (c_2 \mu_2) < 1 \Rightarrow c > \lambda / \mu = 40 / 15 = 2.667 \Rightarrow c_{min} = 3 .$$

- Suppose management decided to add one more than the minimum number of cashiers needed.

- What is the mean delay at the checkout line?

- Applying the $M/M/4$ results, with $a = \lambda / \mu = 2.667$, and

$$\rho = a / 4 = 0.667 .$$

$$P_0^2 = \left(\sum_{n=0}^{c_2-1} \frac{a_2^n}{n!} + \frac{a_2^{c_2}}{c_2! (1-\rho_2)} \right)^{-1} = \left(1 + 2.667 + \frac{2.667^2}{2} + \frac{2.667^3}{6} + \frac{2.667^4}{4! (1-0.667)} \right)^{-1} \\ = 0.06$$

$$W_q^2 = \frac{a_2^{c_2}}{c_2! (c_2 \mu_2) (1-\rho_2)^2} P_0^2 = \frac{2.667^2}{4! (4 \times 15) (1-0.667)^2} 0.06 \\ = 0.019 \text{ hours} = 1.14 \text{ mins}$$

- What is the mean number of people at the check-out line and in the entire supermarket?

- At the checkout line,

$$L_2 = L_q^2 + a_2 = \lambda_2 W_q^2 + a_2 = 40 \times 0.019 + 2.667 = 3.43 .$$

- At the entire store the mean number is

$$L_1 + L_2 = \lambda_1 / \mu_1 + 3.43 = 40 / (4 / 3) + 3.43 = 33.43 .$$

- What is the probability that 25 people are in the store and 4 people are at check-out line?

- The required probability is

$$P_{25}^1 \times P_4^2 = \left(e^{-a_1} \frac{a_1^{25}}{25!} \right) \left(\frac{a_2^4}{4!} P_0^2 \right) = \left(\frac{30^{25}}{25!} \right) \left(\frac{2.667^4}{4!} 0.06 \right) = 0.006 .$$

- **The $M/GI/1$ queue**

- This is a single server-queue with Poisson arrivals with rate λ and general (non-exponential) service times, S_1, S_2, \dots , which are iid.
- This can be seen as a generalization of $M/M/1$ with general service times.
- As in $M/M/1$, the stability condition is $\rho = \lambda/\mu < 1$.
- Because of the non-exponential service times, birth death analysis cannot be used.
- However, an “imbedded” discrete time MC can be defined as the number in the system at customer departure epochs.
- Solving the discrete time MC leads to the following (Pollaczek-Kintchine) formula for the mean delay

$$W_q(M / GI / 1) = \frac{\lambda E[S^2]}{2(1 - \rho)} .$$

- Other measures of performance can be found from Little’s formula as usual.
- It is useful to write the delay in $M/GI/1$ as a function of the delay in $M/M/1$ with the same arrival and service rate.
- It can be shown that

$$W_q(M / GI / 1) = \frac{1 + C_s^2}{2} \frac{\rho^2}{\lambda(1 - \rho)} = \frac{1 + C_s^2}{2} W_q(M / M / 1) ,$$

where $C_s^2 = \text{var}[S]/(E[S])^2 = E[S^2]/(E[S])^2 - 1$, is the

coefficient of variation of service times.

- This implies that waiting time in $M/GI/1$ is proportional to service time variability measured in terms of C_s^2 .
- Note that for exponential service times $C_s^2 = 1$.
- When service time variability is higher (lower) than that of a “similar” $M/M/1$, the delay is higher (lower) in $M/GI/1$.
- For example, in a $M/GI/1$ with deterministic service times (known as $M/D/1$), $C_s^2 = 0$, and

$$W_q(M/D/1) = \frac{W_q(M/M/1)}{2}.$$

- **Example 16.**

- Suppose that failed machines are sent to a repair facility staffed by one repairman according to a Poisson process with rate 6/hour. A machine could fail due to two types of defects. Type 1 failure requires an exponentially distributed repair time with mean 7 minutes, while Type 2 failure requires an exponentially distributed repair time with mean 20 minutes. Suppose that the probability that a failure is of Type 1 is 0.9 (and that of Type 2 is 0.1). In this case, the overall repair time is said to have a hyperexponential distribution.
- What is the mean delay at the repair facility?

- By conditioning on the type of failure, the first two moments of the repair time, S , are given by

$$\begin{aligned} E[S] &= E[S \mid \text{Type 1}]P\{\text{Type I}\} + E[S \mid \text{Type 1}]P\{\text{Type I}\} \\ &= 7 \times 0.9 + 20 \times 0.1 = 8.3 \text{ min.} \end{aligned}$$

$$\begin{aligned} E[S^2] &= E[S^2 \mid \text{Type 1}]P\{\text{Type I}\} + E[S^2 \mid \text{Type 1}]P\{\text{Type I}\} \\ &= (2 \times 7^2) \times 0.9 + (2 \times 20^2) \times 0.1 = 168.2 \text{ min}^2. \end{aligned}$$

- Then, $C_s^2 = E[S^2]/(E[S])^2 - 1 = 168.2/8.3^2 - 1 = 1.442$.
- The mean delay in a $M/M/1$ with the same service and arrival rates is found as follows. In this case $\lambda = 6$ and $\mu = 60/8.3 = 7.23$. Then, $\rho = 0.83$, and

$$W_q(M / M / 1) = \frac{\rho^2}{\lambda(1 - \rho)} = \frac{0.83^2}{6(1 - 0.83)} = 0.675 \text{ hours.}$$

- Finally, the mean delay in the repair facility is

$$W_q(M / GI / 1) = \frac{1 + C_s^2}{2} W_q(M / M / 1) = 0.824 \text{ hours.}$$

- Waiting time is high here because of high service time variability.
- What is the probability that the repairman is idle?

$$P\{\text{server is idle}\} = 1 - \rho = 1 - 0.83 = 0.17.$$

- **A Queuing Cost Model**

- In some situations, management has control over queueing systems parameters.
- In the following, we assume that the number of servers c and/or the service rate μ are *decision variables*.
- Determining “optimal” values for c and μ is done in a way as to minimize expected cost per unit time.
- The cost function has two components:
 - Service cost per unit time, SC ;
 - Waiting cost per unit time, WC.
- The expected service cost per unit time is given by

$$E[SC] = C_s c \mu ,$$

where C_s (\$/unit service rate/server/unit time) is the unit service cost.

- Note that if μ is not a decision variable, then $C_s \mu$ can be replaced by $C'_s = C_s \mu$ (\$/server/unit time).
- In addition, the expected waiting cost is

$$E[WC] = C_w L ,$$

where L is the mean number in the system and, C_w (\$/customer/unit time) is the unit waiting cost.

- **Example 18.**

- Jobs arrive at machine shop according to a Poisson process at the rate of 80 jobs per week. An automatic machine represents the bottleneck in the shop. It is estimated that a unit increase in the production rate of the machine will cost \$250 per week. Delayed jobs result in lost business, which is estimated to be \$500 per job per week.
- Determine the optimum production rate of the automatic machine.
- The automatic machine can be modeled as an $M/M/1$ queue with $\lambda = 80$ and μ being a decision variable. The unit service cost is $C_s = \$250$ and the unit waiting cost is $C_w = \$500$.
- The expected weekly cost as a function of μ is given by

$$EC(\mu) = C_s\mu + C_wL = C_s\mu + C_w \frac{\lambda}{\mu - \lambda}.$$

- The optimal value of μ that minimizes $EC(\mu)$, μ^* , is obtained by differentiating $EC(\mu)$ as follows.

$$\frac{\partial EC(\mu)}{\partial \mu} = C_s - C_w \frac{\lambda}{(\mu - \lambda)^2},$$

$$\frac{\partial EC(\mu)}{\partial \mu} = 0 \Rightarrow C_s - C_w \frac{\lambda}{(\mu^* - \lambda)^2} = 0 \Rightarrow C_s = C_w \frac{\lambda}{(\mu^* - \lambda)^2}$$

$$\Rightarrow (\mu^* - \lambda)^2 = C_w \frac{\lambda}{C_s} \Rightarrow \mu^* = \lambda \pm \sqrt{C_w \frac{\lambda}{C_s}}.$$

- Since ρ should be < 1 , i.e., $\mu > \lambda$,

$$\mu^* = \lambda + \sqrt{C_w \frac{\lambda}{C_s}}.$$

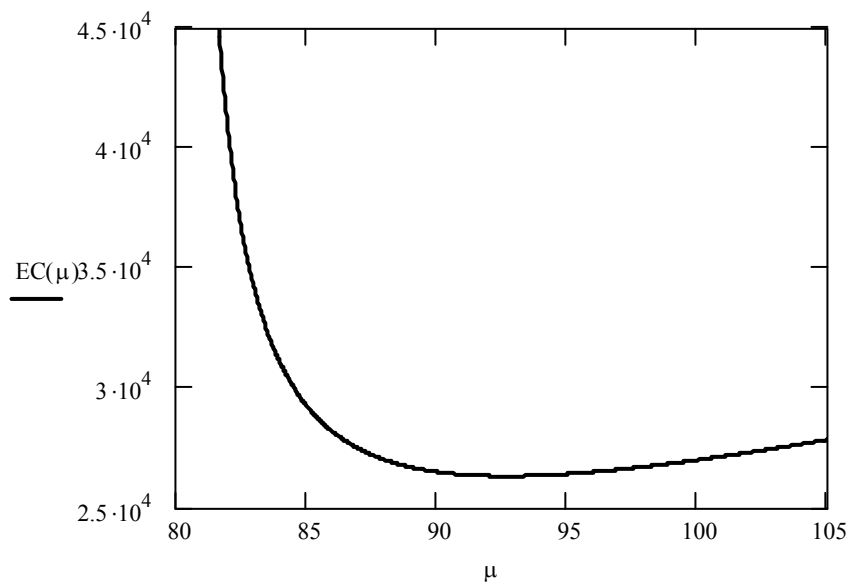
- We also need to check the second-order conditions to confirm that μ^* achieves the maximum value of $EC(\mu)$,

$$\frac{\partial^2 EC(\mu)}{\partial \mu^2} = 2C_w \frac{\lambda}{(\mu - \lambda)^3} > 0.$$

- For the automatic machine, Since ρ should be < 1 ,

$$\mu^* = \lambda + \sqrt{C_w \frac{\lambda}{C_s}} = 80 + \sqrt{500 \times \frac{80}{250}} = 92.65 \text{ jobs/week}$$

- Suppose that models of the machine available in the market have speeds, 80, 85, 90, 95, and 100 jobs/week. Which model should be chosen?
- The *convexity* of the cost function implies that models with speeds 90 and 95 are the most efficient. See figure.



- To see whether 90 or 95, we compute the expected cost for each. We find that $EC(90) = \$26,500$, and $EC(95) = \$26,417$.
- The model with speed 95 should be chosen.
- **Example 19.**
 - A repair facility has c repairmen. Broken machines arrive at a Poisson rate of 17.5 /hour. Each repairman can handle 10 machines per hour. Hiring a repairman costs \$12/hour. The cost of lost production per waiting machine is \$50/hour.
 - How many repairmen should be hired?
 - The repair facility can be modeled as an $M/M/c$ with $\lambda = 17.5$, $\mu=10$, and c being a decision variable.

- The unit service cost is $C_s' = \$12$ and the unit waiting cost is $C_w = \$50$.
- The expected hourly cost can then be written as a function of c as

$$EC(c) = C_s'c + C_w L(M/M/c) = 12c + 50L(M/M/c).$$

- Since no simple formula exists for $L(M/M/c)$, the optimal value of c , c^* , that minimizes $EC(c)$ is determined via a numerical search.
- One fact that facilitates the search is that $EC(c)$ is *convex*. Then, c^* can be determined is the minimum value of c such that $EC(c+1) > EC(c)$.
- The numerical search works as follows. First, the minimum value of c that achieves $\rho = \lambda/(c\mu) < 1$ is 2.

c	$L(M/M/c)$	$EC(c)$
2	7.467	\$397.5
3	2.217	\$146.85
4	1.842	\$140.10
5	1.769	\$148.45

- Therefore, $c^* = 4$ repairmen should be hired.