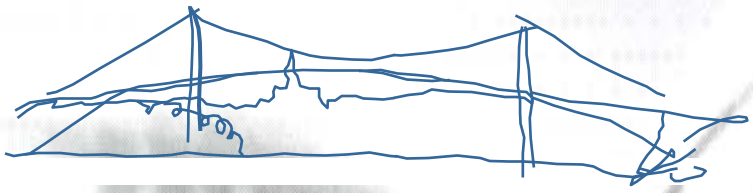
UNIVERSITÄT
DUISBURG
ESSEN
Campus Duisburg

Department of Engineering Sciences
Division of Mechanical Engineering
Transport Systems and Logistics
Lotharstraße 1 - 21
47057 Duisburg

Phone: 0203 379-2785
Fax: 0203 379-3048
eMail: b.noche@uni-duisburg.de



Prof. Dr.-Ing. Bernd Noche

Simulation in der Logistik

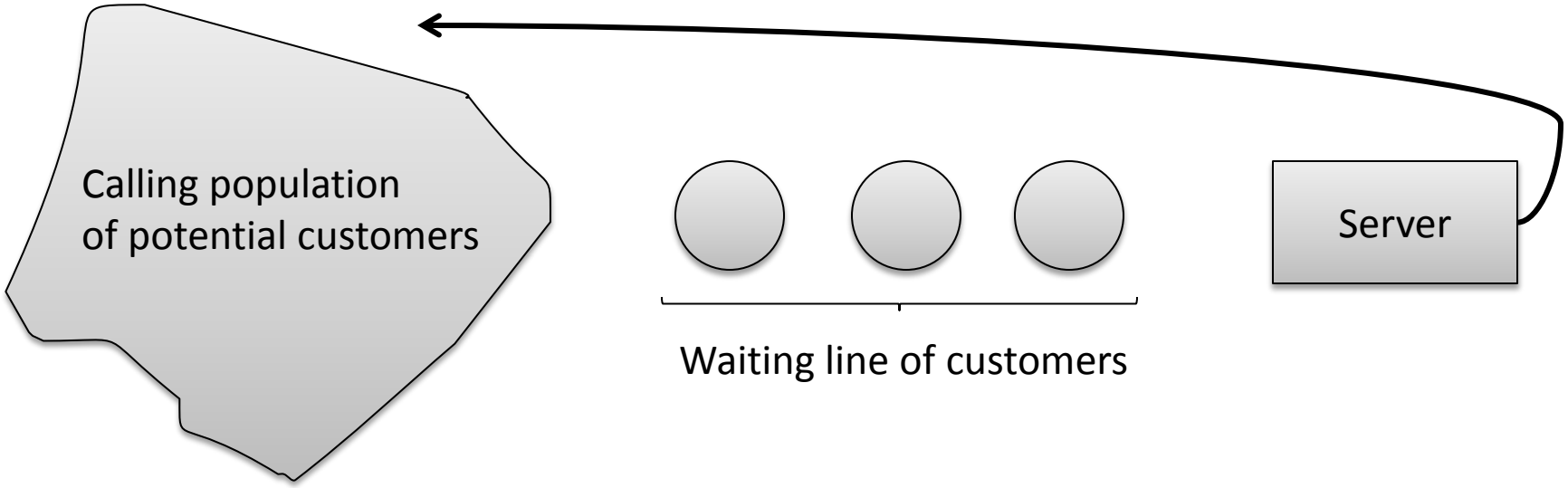
ehem.: **Simulation in Logistics I**

Queueing Models and its application in Dosimis-3

Lecturer: Prof. Dr.-Ing. Bernd Noche
M.Sc. Nan Liu

Queueing Models

- In this lecture, we will discuss some of the well-known models for developing an understanding of the dynamic behavior of queueing systems, the general characteristics of queues, the meaning and relationships between various performance measures.



Queueing Models

Queueing theory and simulation analysis are used to the *measures* of *system* performance as a function of the *input* parameters.

- **System**: service facilities, production systems, repair and maintenance facilities, and transprot and material handling systems.
- **Measures of system performance**: server utilization, length of waiting lines, and delays of customers.
- **Input parameters**: the arrival rate of customers, the service demands of customers, the rate at which a server works, and the number and arrangements of servers.

1. Characteristics of Queueing Systems

the key elements of a queueing system are the customers and the servers. The term "**customer**" can refer to people, machines, trucks, patients, pallets, airplanes, orders – anything that arrives at a facility and requires service. The term "**server**" may refer to receptions, repairpersons, AS/RS, order pickers – any resource which provides the requested service.

System	Customers	Servers
Repair facility	Machines	Repairperson
Warehouse	Pallets	Crane or forklift
Production line	Cases	Case packer
Warehouse	Orders	Order picker
Job shop	Jobs	Machines/workers
Mass transit	Riders	Buses, trains

1. Characteristics of Queueing Systems

1. **The calling population** -- the population of potential customers – may be assumed to be finite or infinite. The main difference between the finite and infinite population models is how the arrival rate is defined. In an **infinite-population model**, the arrival rate is not affected by the number of customers who left the calling population and joined the queueing system. For **finite calling population models**, the arrival rate depends on the number of customers being served and waiting.
2. **System Capacity** -- In many queueing systems there is a limit to the number of customers that may be in the waiting line or system. An arriving customer who finds the system full does not enter but return to the calling population. When a system has limited capacity, a distinction is made between the arrival rate and the effective arrival rate
3. **The arrival process** of infinite population models is usually characterized in terms of interarrival times of successive customers. The most important model for arrivals is the **Poisson arrival process.** The second important class of arrivals is the **scheduled arrivals**, e.g. the scheduled airline flight arrivals to an airport. In this case, the interarival times may be constant. The third situation occurs when at least one customer is assumed to be always present in the queue, the that the server is never idle because of lacking of customers. E.g. the raw material for a production line.
4. **Queue behavior and Queue discipline. Queue behavior** refers to customer actions while in queue waiting for service to begin. **Queue discipline** refers to the logical ordering of customers in a queue and determines which customer will be chosen for service when the a server becomes free. **Common queue disciplines** include first-in, first-out (**FIFO**); last-in, last-serve (**LIFO**); service in random order (**SIRO**); shortest processing time first (**SPT**); and service according to priority (**PR**). In a job shop, queue disciplines are sometimes based on due dates and on expected processing time for a given type of job.
5. **Service Time and Service Mechanism**. **the service times** of successive arrivals are denoted by S1, S2, S3, …. They may be constant or of random duration. In the latter case, {S1, S2, S3, …} is usually characterized as a sequence of independent and identically distributed random variables. E.g. Exponential, weibull, gamma, lognormal, truncated normal distribution. A queueing system consists of a number of service centers and interconnecting queues. Each service center consists of some number of servers, c, working in parallel; that is, the customer takes the first available server. Parallel service mechanism are either single server (c=1), multiple server ($1<c<\infty$), or unlimited servers ($c=\infty$).

1. Poisson Process

- N(t) is a counting function that represents the number of events occurred in [0,t]. N(t) is said to be a Poisson process with mean rate λ if the following assumptions are fulfilled:

  1. Arrivals occur one at a time.
  2. {N(t), t ≥ 0} has stationary increments.

  The number of arrivals N(t) in a finite interval of length t obeys the Poisson(λt) distribution:

$$P\{N(t) = n\} = \frac{(\lambda t)^n}{n!} e^{-\lambda t}$$

  3. {N(t), t ≥ 0} has independent increments.

  The number of arrivals N(t_1, t_2) and N(t_3, t_4) in the non-overlapping intervals ($t_1 < t_2 < t_3 < t_4$) are independent.

- **An alternative definition of a Poisson process**:

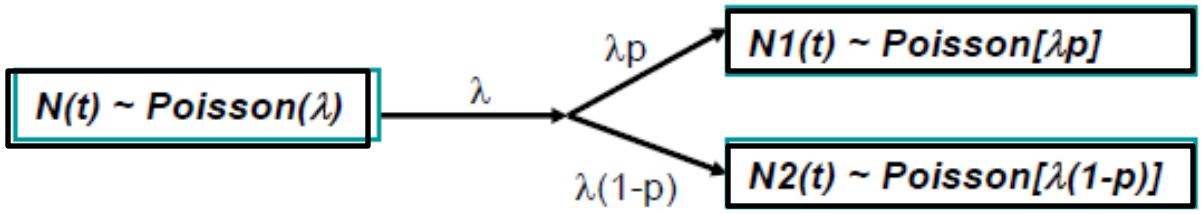
  if interarrival times are distributed exponentially and independently, then the number of arrivals by time t, say N(t), meets the three assumption and therefore is a Possion distribution.

1. Poisson Process

- **Splitting**:

  • *Suppose each event of a Poisson process can be classified as Type I, with probability p and Type II, with probability 1-p.*

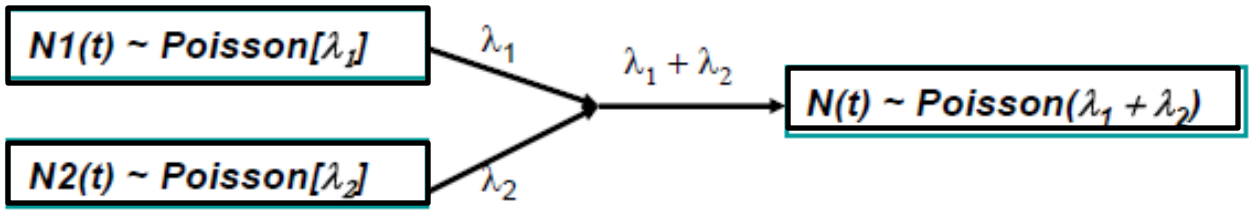
  • *N(t) = N1(t) + N2(t), where N1(t) and N2(t) are both Poisson processes with rates λp and λ(1-p)*



- **Pooling (Superposition)**:

  • *Suppose two Poisson processes are pooled together*

  • *N1(t) + N2(t) = N(t), where N(t) is a Poisson processes with rates $\lambda_1 + \lambda_2$*



1. Poisson Process

- Random selection
  - If a random selection is made from a Poisson process such that each arrival is selected with probability p, independently of the others, the resulting process is a Poisson process with rate pλ
- PASTA
  - The Poisson process has the so called PASTA property (Poisson Arrival See Time Averages): arriving customers find on average the same situation in the queueing system as an outside observer looking at the system at an arbitrary point in time. More precisely, the fraction of customers finding on arrival the system in some state A is exactly the same as the fraction of time the system is in state A.

The hitchhiker's Paradox

- Cars are passing a certain point of a road according to a Poisson process. The mean interval between the cars is 10 min. A hitchhiker arrives to this roadside point at random instant of time. what is the mean waiting time until the next car?

- The interarrival times in a Poisson process are exponentially distributed. From the memoryless property of the exponential distribution if follows that the time to the next arrival has the same Exp(λ) distribution had the expected time is thus ???

2. Kendall Queueing Notation

D. Kendall

(* 15. Januar 1918 Yorkshire, UK)

- In 1953, Kendall proposed a notation system for parallel server systems:

A / B / c / N / K

1) **A** represents the interarrival time distribution.
2) **B** represents the service time distribution.
3) **c** represents the number of parallel servers.
4) **N** represents the system capacity.
5) **K** represents the size of the calling population.

Common symbols for A and B include:

- M: exponential or Markov,
- D: constant or deterministic,
- E_k : Erlang of order k,
- H: hyperexponential,
- G: arbitrary or general,
- GI: general independent.

e.g., M / M / 1 / ∞ / ∞ indicates a single server system that has unlimited queue capacity and an infinite population of potential arrivals and the interarrival times and service times are exponentially distributed.

3. Long-Run Measures of Performance of Queueing Systems

The primary long-run measures of performance of queueing systems are:

- **L** : long-run time-average number of customers in the system
- L_Q : long-run time-average number of customers in the queue
- **L(t)** : the number of customers in system at time t
- $L_Q(t)$: the number of customers in queue at time t
- **w**: long-run average time spent in system
- w_Q : long-run average time spent in queue
- W_n : total time spent in system by the nth customer
- W_n^Q: total time spent in queue by customer n
- **ρ**: server utilitzation
- **λ**: arrival rate
- λ_e : effective arrival rate

3.1 Little's Law

The average number of customers in the system at an arbitrary point in time is equal to the average number of arrivals per time unit, **times** the average time spent in the system:

$$L = \lambda w$$

This equation works for almost all queueing system or subsystems, regardless of the number of servers, the queue discipline, or any other special circumstances.

3.2 Server Utilitzaion

- Server utilization is defined as the proportion of time that a server is busy. Observed server utilization, denoted by $\hat{\rho}$ is defined over a specified time interval [0, T]. And the long run server utilization is ρ. For system with long-run stability $\hat{\rho} \rightarrow \rho$ as $T \rightarrow \infty$.

3.2.1 Server Utilization in G/G/1/∞/∞ Queues

- Any single-server queueing system with average arrival rate λ customers per time unit, where average service time E(S) = 1/μ time units, infinite queue capacity and calling population.
- The average number of customers in the server is

$$\hat{L}_s = \frac{1}{T}\int_0^T \left(L(t) - L_Q(t)\right)dt = \frac{T - T_0}{T}$$

- In general, for a single-server queue:

$$\rho = \lambda E(S) = \lambda/\mu$$

3.2.1 Server Utilization in G/G/1/∞/∞ Queues

- For a single-server queue to be stable, the arrival rate λ must be less than the service rate μ, that is, $\lambda < \mu$ or

$$\rho = \lambda/\mu < 1$$

- For an unstable queue ($\lambda > \mu$), long-run server utilization is__? And the waiting line growth rate is __?

3.2.2 Server Utilization in G/G/c/ ∞/∞ Queues

- A system with *c identical servers in parallel.* If an arriving customer finds more than one server idle, the customer chooses a server without favoring any particular server. For systems in statistical equilibrium, the average number of busy servers, Ls, is:

$$Ls = \lambda E(s) = \lambda/\mu.$$

- The long-run average server utilization is:

$$\rho = Ls/c = \lambda/c\mu.$$

where $\lambda < c\mu$ for stable systems

4. Steady-State Behavior of Infinite-Population Markovian Models

- For the infinite-population models, If the arrival process is exponentially distributed with rate λ arrivels per time unit (mean 1/ λ), and the service time may be exponetially disributed (M) or arbitrary (G). The queue discipline is FIFO. Then these models are called the **markovian models**.
- A transient mathematical analysis, or more likely a simulation model, would be chosen tool of analysis for the transient behavior of a queue over a relatively short period of time and given some specific initial conditions(such as idle and empty). The method used in this lecture here are inappropriate for them.

4. Long-Run Measures of Performance of Queueing Systems

The primary long-run measures of performance of queueing systems are:

- **L** : long-run time-average number of customers in the system
- L_Q : long-run time-average number of customers in the queue
- **L(t)** : the number of customers in system at time t
- $L_Q(t)$: the number of customers in queue at time t
- **w**: long-run average time spent in system
- w_Q : long-run average time spent in queue
- W_n : total time spent in system by the nth customer
- W_n^Q : total time spent in queue by customer n
- **ρ**: server utilitzation
- **λ**: arrival rate
- λ_e : effective arrival rate
- P_n : steady state probability of having n customers in system
- $P_n(t)$: probability of n customers in system at time t

4.1 Properties of processes with statistical equilibrium

- A queueing system is said to be in **statistical equilibrium**, or **steady state**, when the system is in a given state that is not time dependent. That is

$$P(L(t)) = n) = P_n(t) = P_n$$

  is independent of time t.

Two Properties of processes with statistical equilibrium:

1. The state of statistical equilibrium is reached from any starting state.
2. The process remain in statistical equilibrium once it has reached it.

4.2 M/G/1 Queue

- A queue with only one server and the service time have mean $1/\mu$ and variance σ^2, then the M/G/1 queue has the steady-state parameters:

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

$$P_0 = 1 - \rho$$

$$L = \rho + \frac{\rho^2(1+\sigma^2\mu^2)}{2(1-\rho)}$$

$$w = \frac{1}{\mu} + \frac{\lambda(1/\mu^2 + \sigma^2)}{2(1-\rho)}$$

$$L_Q = \frac{\rho^2(1+\sigma^2\mu^2)}{2(1-\rho)}$$

$$w_Q = \frac{\lambda(1/\mu^2 + \sigma^2)}{2(1-\rho)}$$

Example 1

- Welding machines malfunctions occur according to a Possion process at rate $\lambda = 1.5$ per hour. Obervation over several months has found that repairs by a single mechanic take an average time of 30 minutes with a standard deviation of 20 minutes.
  - The service rate is $\mu = 2$ per hour,
  - The arrival rate is $\lambda = 1.5$ per hour,
  - The proportion of time the mechanic is busy is $\rho = \lambda / \mu = 1.5 / 2 = 0.75$
  - The steady-state time average number of broken machines is:

    $L = 0.75 + \{(1.5)^2 [(0.5)^2 + 1/9]/(2(1-0.75))\} = 0.75 + 1.625 = 2.375$ machines.

4.2 M/G/1 Queue

- The source of waiting line and delay L_Q:

$$L_Q = \frac{\rho^2(1+\sigma^2\mu^2)}{2(1-\rho)}$$

Can be rewritten as :

$$L_Q = \frac{\rho^2}{2(1-\rho)} + \frac{\lambda^2\sigma^2}{2(1-\rho)}$$

Variability

- The first term involves only the ratio of mean arrival rate λ to mean service rate μ. If λ and μ are held constant, the length of the waiting line (L_Q) will only depend on the *variability*, σ^2, of the service time.

4.2 M/G/1 Queue

- If two systems have identical mean service times and mean interarrival times, the one with the more variable service times will tend to have longer lines on the average.
- Do not confuse “steady state” with low variability or short waiting lines, a system in steady state or statistical equilibrium can be highly variable and can have long waiting lines.

Example 2

- Two worker are competing for a job. Alex is faster than Ben on average, but Ben is more consistent. The arrival occur according to a Poisson process at a rate of λ = 2 per hour (1/30 per minute). Alex's statistics are an average service time of 24 minutes with a standard deviation of 20 minutes. Ben's average service time is 25 minutes with a standard deviation of 2 minutes. If the average length of queue is the criterion of hiring, which worker should be hired?
  - For Alex, λ = 1/30 minutes, 1/μ = 24 minutes, σ = 20² = 400 minutes², ρ = 24 / 30 = 4/5, and the average queue length:

$$L_Q = \frac{(1/30)^2[24^2 + 400]}{2(1-4/5)} = 2.711 \text{ customers}$$

  - For Ben, λ = 1/30 minutes, 1/μ = 25 minutes, σ = 2² = 4 minutes², ρ = 25 / 30 = 5/6, and the average queue length:

$$L_Q = \frac{(1/30)^2[25^2 + 4]}{2(1-5/6)} = 2.097 \text{ customers}$$

  - Alex's greater service variability resutls in an average queue length about 30 % longer than baker.. On this basisiof average queue length, Ben wins.

4.3 M/M/1 Queue

- Suppose the service times in an M/G/1 queue are exponentially distributed with mean $1/\mu$, then the variance is $\sigma^2 = 1/\mu^2$. It becomes a M/M/1 queue. The M/M/1 queue is a useful approximate model when service times have standard deviation approximately equal to their means.

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

$$P_n = (1-\rho)\rho^n$$

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

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

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

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

Example 3

- The interarrival time and the service time at a single-chair unisex barbershop are exponential distributed. The values of λ and μ are 2 per hour and 3 per hour.

$$\rho = \frac{\lambda}{\mu} = \frac{2}{3}$$

$$P_0 = 1 - \rho = \frac{1}{3}$$

$$P_1 = \frac{1}{3} \cdot \left(\frac{2}{3}\right)^1 = \frac{2}{9}$$

$$P_2 = \frac{1}{3} \cdot \left(\frac{2}{3}\right)^2 = \frac{4}{27}$$

$$P_{\geq 4} = 1 - \sum_{n=0}^{3} P_n = \frac{16}{81}$$

$$L = \frac{\lambda}{\mu - \lambda} = \frac{2}{3-2} = 2 \text{ Customers}$$

$$w = \frac{L}{\lambda} = \frac{2}{2} = 1 \text{ hour}$$

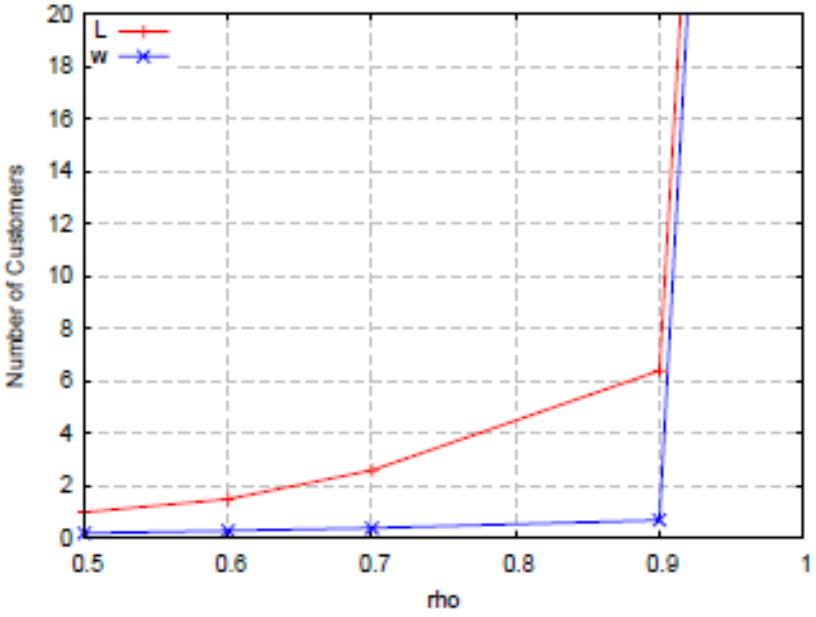
$$w_Q = w - \frac{1}{\mu} = 1 - \frac{1}{3} = \frac{2}{3} \text{ hour}$$

$$L_Q = \frac{\lambda^2}{\mu(\mu - \lambda)} = \frac{4}{3(3-2)} = \frac{4}{3} \text{ Customers}$$

$$L = L_Q + \frac{\lambda}{\mu} = \frac{4}{3} + \frac{2}{3} = 2 \text{ Customers}$$

Example 4

- A M/M/1 queue with servicer rate μ = 10 customers per hour. How the L and w increase as the arrival rate, λ , increase from 5 to 8.64 by increments of 20%? If then the arrival rate increase to 10?
  - For any M/G/1 queue, if λ/μ ≥ 1, waiting lines tend to continually grow in length;
  - Increase in average system time w and average number in system L is highly nonlinear as a function of ρ.



λ	5.0	6.0	7.2	8.64	10.0
ρ	0.50	0.60	0.72	0.864	1.0
L	1.00	1.50	2.57	6.35	∞
w	0.20	0.25	0.36	0.73	∞

4.4 The effect of Utilization and service Variability

- For almost all queues, if lines are too long, they can be reduced by decreasing server utilization ρ or by decreasing the service time variability σ^2.
- The utilization factor ρ can be reduced by decreasing the arrival rate λ, increasing the service rate μ, or increasing the number of servers.

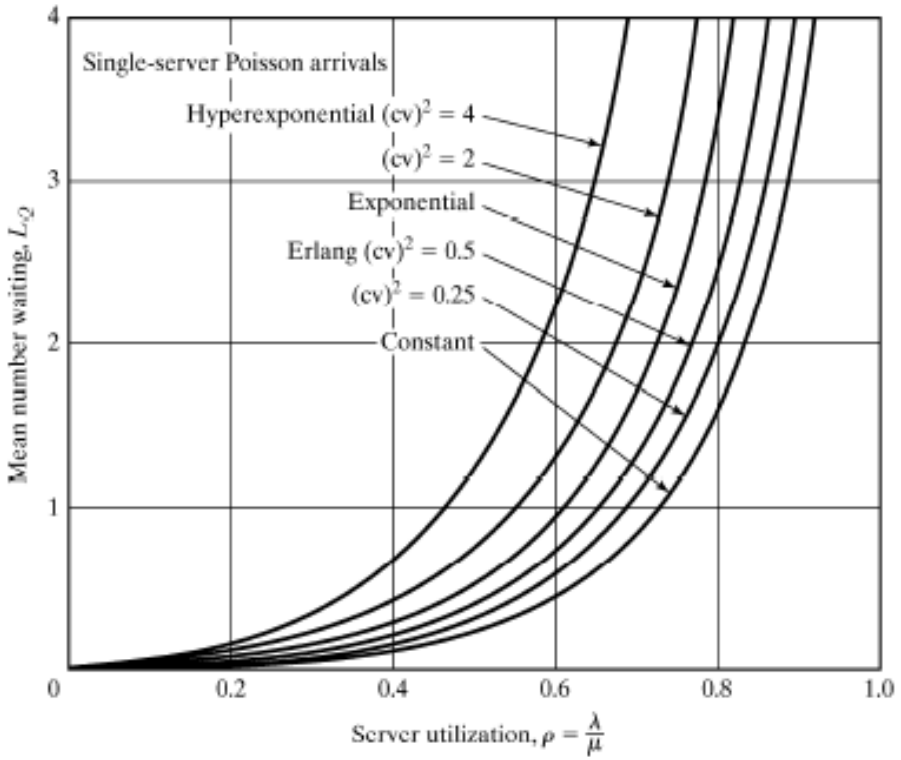
4.4 The effect of Utilization and service Variability

- A measure of the variability of a distribution is the **<u>coefficient of variation</u>** (cv) of a positive random variable X, defined by :

$$(cv)^2 = \frac{V(X)}{[E(X)]^2}$$

The larger its value, the more variable is the distribution relative to its expected value.

For deterministic service times, V(X) = 0, so cv = 0. for Erlang service times of order k, V(X) = 1/kμ², and E(X) = 1/μ , so that cv = 1/√K, For exponential service times at service rate μ, the mean service time is E(X) = 1/μ, and the variance is V(X) = 1/μ², so cv = 1.



4.4 The effect of Utilization and service Variability

- The formula for L_Q for any M/G/1 queue can be rewritten in terms of the coefficient of variation by noticing that $(cv)^2 = \sigma^2/(1/\mu)^2 = \sigma^2\mu^2$. Therefore:

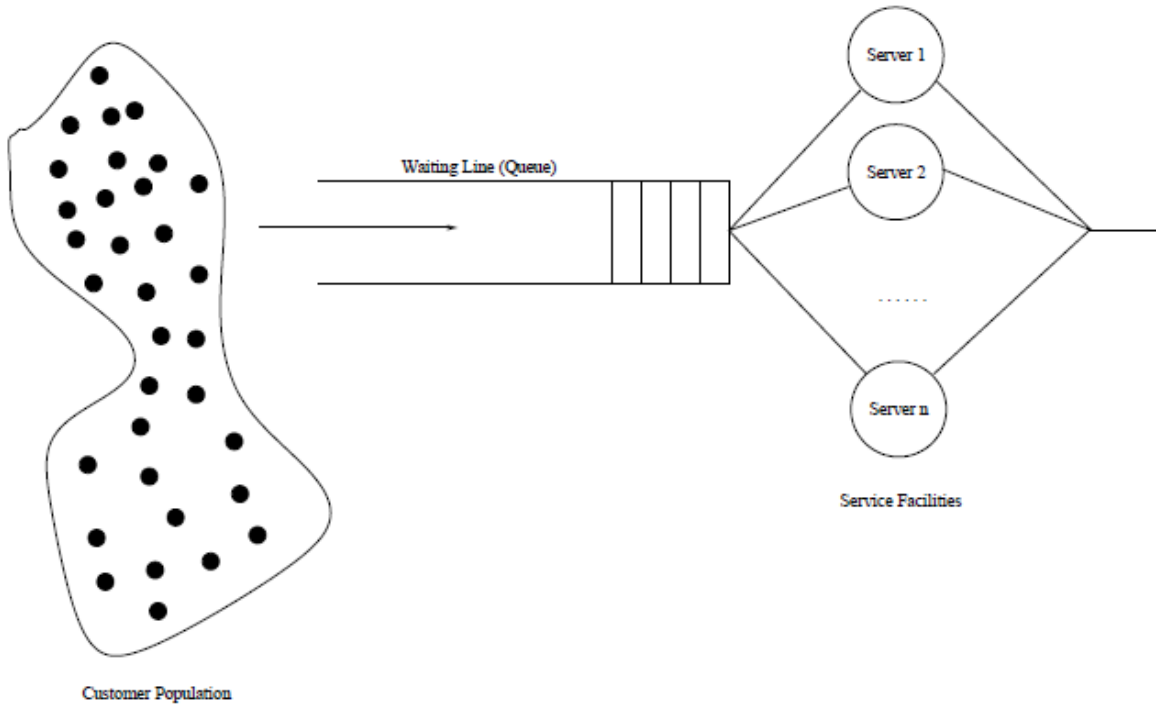
$$L_Q = \frac{\rho^2(1+\sigma^2\mu^2)}{2(1-\rho)} = \left(\frac{\rho^2}{1-\rho}\right)\left(\frac{1+(cv)^2}{2}\right)$$

L_Q for *M/M/1* queue

Corrects the *M/M/1* formula to account for a non-exponential service time dist'n

4.5 M/M/c Queue

- There are c channels operating in parallel. Each of these channels has an independent and identical exponential service-time distribution with mean μ. The arrival rate is Poisson with rate λ . Arrivals will join a single queue and enter the first available service channel.



4.5 M/M/c Queue

- To achieve statistical equilibrium, the offered load λ/μ must satisfy $\lambda/\mu < c$, where $\lambda/(c\mu) = \rho$ is the server utilization.
- The steady-state parameters for *M/M/c :*

$$\rho = \frac{\lambda}{c\mu}$$

$$P_0 = \left\{ \left[\sum_{n=0}^{c-1} \frac{(\lambda/\mu)^n}{n!} \right] + \left[\left(\frac{\lambda}{\mu} \right)^c \left(\frac{1}{c!} \right) \left(\frac{c\mu}{c\mu - \lambda} \right) \right] \right\}^{-1}$$

$$P\left(L(\infty) \geq c\right) = \frac{(c\rho)^c P_0}{c!(1-\rho)}$$

$$L = c\rho + \frac{(c\rho)^{c+1} P_0}{c(c!)(1-\rho)^2} = c\rho + \frac{\rho \cdot P\left(L(\infty) \geq c\right)}{1-\rho}$$

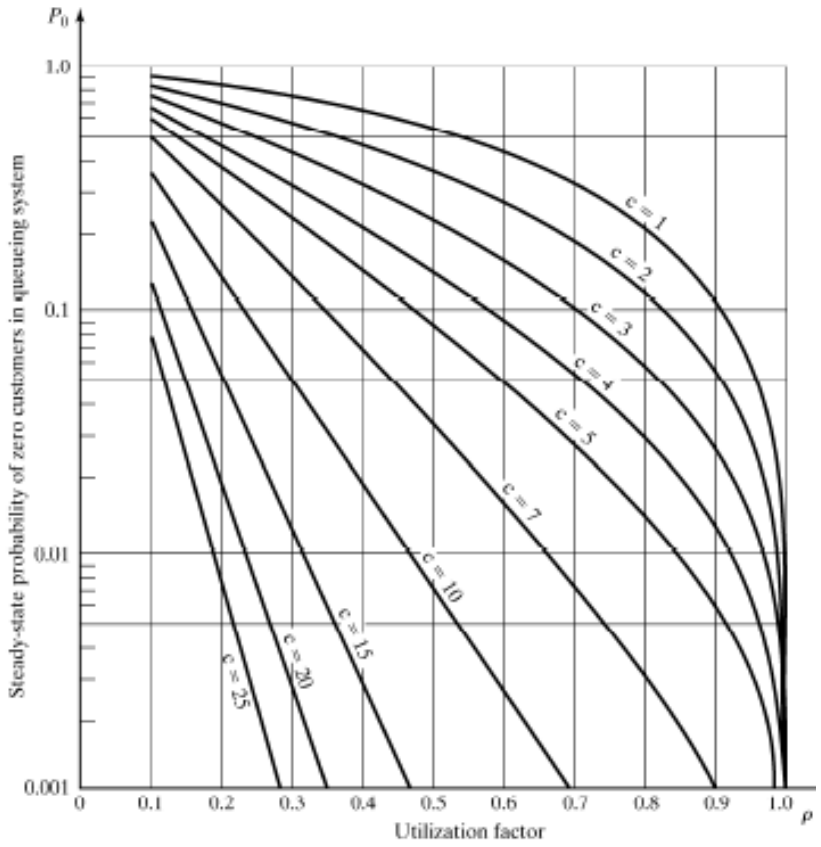
$$w = \frac{L}{\lambda}$$

$$L_Q = \frac{\rho \cdot P\left(L(\infty) \geq c\right)}{1-\rho}$$

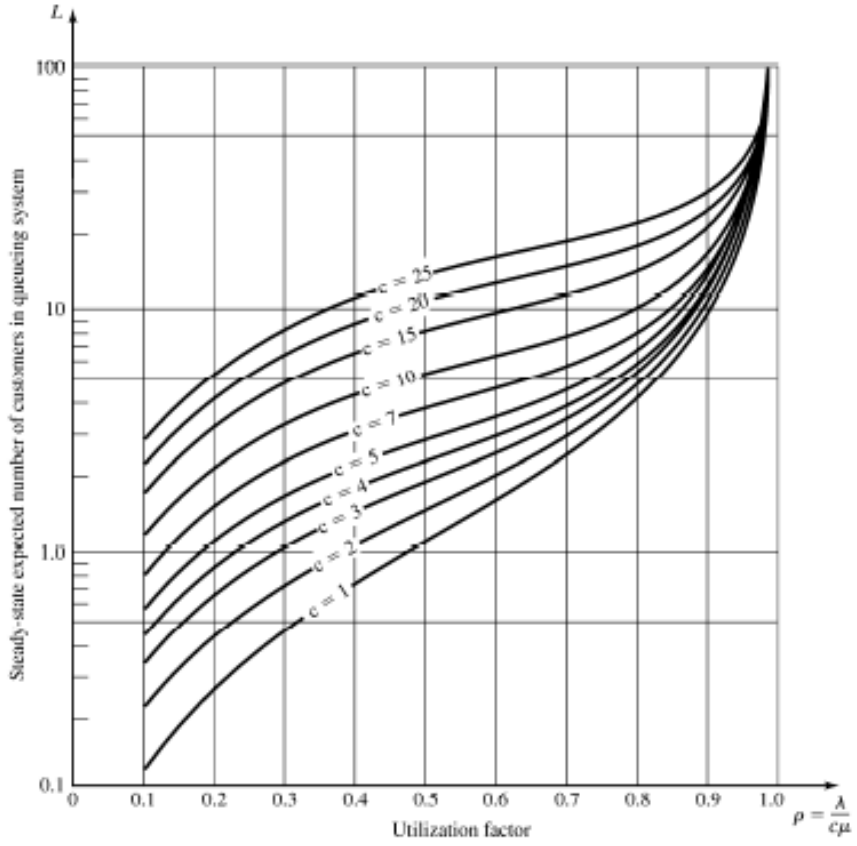
$$L - L_Q = c\rho$$

- P($L(\infty) \geq c$) means the probability that all servers are busy.
- $L(\infty)$ is a random variable representing the number in system in statistical equilibrium after a very long time.

4.5 M/M/c Queue



Values of P_o for M/M/c model



Values of L for M/M/c model

4.6 M/G/c Queue

- M/G/c queue: general service times and c parallel server. Recall that formulas L_Q and w_Q for the M/G/1 queue can be obtained from the corresponding M/M/1 formulas by multiplying them by the correction factor $(1 + (cv)^2)/2$.
- The approximate formulas for M/G/c can be approximated from those of the M/M/c model by applying the same correction factor for L_Q and w_Q. Little's law can be used to calculate L and w. unfortunately, there is no general method for correcting the steady-state probabilities, P_n.

4.7 M/G/∞ Queue

- There are three situations in which it is appropriate to treat the number of servers as infinite.
  1. When each customer is its own server, as self-service systems;
  2. When service capacity far exceeds service demand, a so-called ample-server system;
  3. When it wants to know how many servers are required so that customers are rarely delayed.

  The steady-state parameters for the M/G/∞ queue are:

$$P_n = \frac{e^{-\frac{\lambda}{\mu}}\left(\frac{\lambda}{\mu}\right)^n}{n!}, n = 0,1,\ldots \qquad w_Q = 0$$

$$P_0 = e^{-\frac{\lambda}{\mu}} \qquad L = \frac{\lambda}{\mu}$$

$$w = \frac{1}{\mu} \qquad L_Q = 0$$

4.8 M/M/c/N Queue

- M/M/c/N/∞: service times are exponentially distributed at rate μ and there are c servers where the total system capacity is N ≥ c customer. If an arrival occurs when the system is full, that arrival is turned away and does not enter the system. Effective arrival rate $\boldsymbol{\lambda_e}$ *is defined as the mean number of arrivals per* time unit who enter and remain in the system, the effective arrival rate is computed by:

$$\lambda_e = \lambda\,(1 - P_N)$$

$1 - P_N$ is the probability that a customer, upon arrival, will find space and be able to enter the system.

$$P_0 = \left[1 + \sum_{n=1}^{c} \frac{a^n}{n!} + \frac{a^c}{c!} \sum_{n=c+1}^{N} \rho^{n-c}\right]^{-1}$$

$$P_N = \frac{a^N}{c!\,c^{N-c}} P_0$$

$$L_Q = \frac{P_0 a^c \rho}{c!(1-\rho)}\left(1 - \rho^{N-c} - (N-c)\rho^{N-c}(1-\rho)\right)$$

$$w_Q = \frac{L_Q}{\lambda_e}$$

$$w = w_Q + \frac{1}{\mu}$$

$$L = \lambda_e w$$

$$a = \frac{\lambda}{\mu}$$

$$\rho = \frac{\lambda}{c\mu}$$

Steady-steate parameters for the M/M/c/N Queue

Example 5

- The unisex barbershop described in Example 3 can hold only **three** customers, on in service and two waiting. Additional customers are turned away when the system is full. The offered load is as previously determined, $\lambda/\mu = 2/3$.
- First determine P_0 : $$P_0 = \frac{1}{\left[1+\frac{2}{3}+\frac{2}{3}\sum_{n=2}^{3}\left(\frac{2}{3}\right)^{n-1}\right]} = 0.415$$
- The probability there are three customers in the system:
- Effective arrival rate:

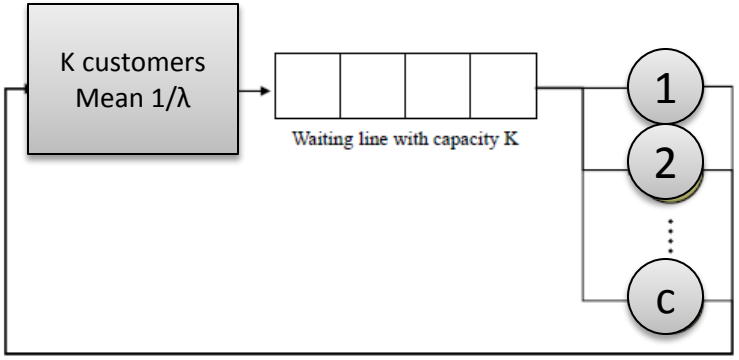
$$P_N = P_3 = \frac{\left(\frac{2}{3}\right)^3}{1!1^2} P_0 = \frac{8}{65} = 0.123$$

$$\lambda_e = 2\left(1-\frac{8}{65}\right) = \frac{114}{65} = 1.754$$

- Expected number of customer in the system: $L = \lambda_e w = \frac{66}{65} = 1.015$
- Server utilization: $1 - P_0 = \frac{\lambda_e}{\mu} = 0.585$
- The server utilization decrease when imposed a capacity constraint.

5. Steady-State Behavior of Finite-Population Models M/M/c/K/K

- In many practical problems the assumption of an infinite calling population leads to invalid results, since the calling population is normally small. In this situation, the presence of one or more customers in the system has a strong effect on the distribution of future arrivals, and the use of an infinite-population model can be misleading.
- A finite-calling-population model with K customers. The time between the end of one service visit and the next call for service for each number of the population is exponentially distributed with mean time 1/λ; service time are also exponentially distributed with mean 1/μ; there are c parallel servers, and the system capacity is K, so that all arrivals remain for service.



5. Steady-State Behavior of Finite-Population Models M/M/c/K/K

- Some of the steady-state probabilities of M/M/c/K/K:

$$P_0 = \left[\sum_{n=0}^{c-1} \binom{K}{n} \left(\frac{\lambda}{\mu}\right)^n + \sum_{n=c}^{K} \frac{K!}{(K-n)!c!c^{n-c}} \left(\frac{\lambda}{\mu}\right)^n \right]^{-1}$$

$$P_n = \begin{cases} \binom{K}{n} \left(\frac{\lambda}{\mu}\right)^n P_0, & n = 0,1,\ldots,c-1 \\ \frac{K!}{(K-n)!c!c^{n-c}} \left(\frac{\lambda}{\mu}\right)^n, & n = c, c+1,\ldots K \end{cases}$$

$$L = \sum_{n=0}^{K} nP_n, \quad w = L/\lambda_e, \quad \rho = \frac{\lambda_e}{c\mu}$$

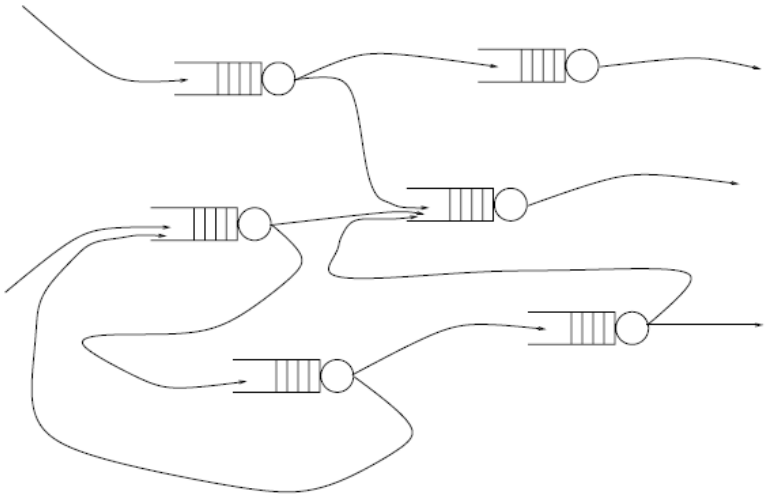
- It is better to use spreadsheet or calculation program for evaluate these complex formulas.

6. Network of Queues

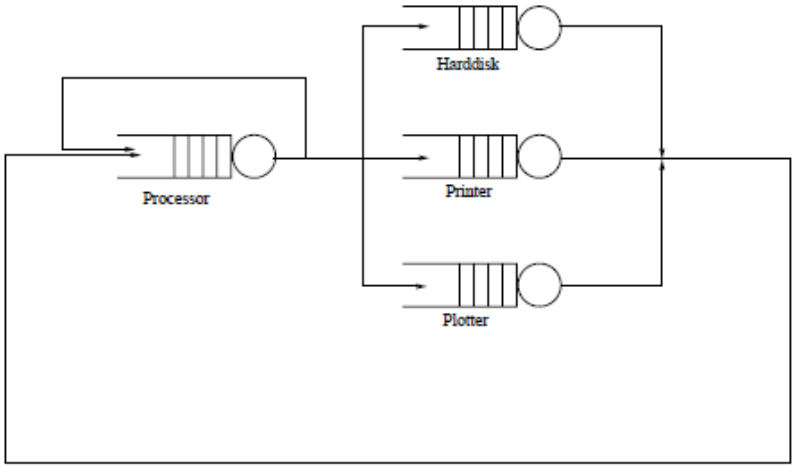
- So far we have only looked at a single standalone queueing system. However, most real systems are better represented as a network of queues. In a queueing network a customer finishing service in a service facility A is immediately proceeding to another service facility or he is leaving the system.
- One basic classification of queueing networks is the distinction between ***open*** and ***closed queueing* networks**.

Network of Queues

- In an **open network** new customers may arrive from outside the system (coming from a conceptually infinite population) and later on leave the system. In a **closed queueing network** the number of customers is fixed and no customer enters or leaves the system.



An Open Queueing Network



A Closed Queueing Network

Summary

- Characteristics of Queueing System
- Kendall Queueing Notation
- Long-Run Measures of Performance of Queueing Systems
- Little's Law
- G/G/1, G/G/c, M/G/1, M/M/1, M/M/c, M/G/c, M/G/∞
- M/M/c/N, M/M/c/K/K
- Network of Queues