Queueing Models

CS1538: Introduction to Simulations

Introduction to Queueing Theory

- Many simulations involve using one or more queues
  - People waiting in line to be served
  - Jobs in a process or print queue
  - Cars at a toll booth
  - Orders to be shipped from a company

- Queueing Theory
  - Characteristics of queues
  - Relationships between performance measures
  - Estimation of mean measures from a simulation
  - Effect of varying input parameters
  - Mathematical solution of a few basic queueing models

Components of a Queueing Model

- The calling population
  - Finite or infinite (often approx. "very large")
  - Infinite pool: arrival rate not affected by the number of customers who have left the calling population and joined the queueing system.
  - Finite pool: can affect arrival process
- The system capacity
- The arrival process
  - Infinite pool:
    - At random: interarrival times (e.g., Poisson Arrival Process)
    - At scheduled times (e.g., for flights)
    - At least one customer is assumed to be always present
  - Finite pool:
    - Distinguish between *pending* and *not pending*
- Queue behavior and queue discipline
  - Will customers balk, renege, jockey?
  - FIFO, LIFO, shortest processing time first, by priority, random?
- Service times and Service mechanism

Example Queueing System

- People waiting for a bank teller

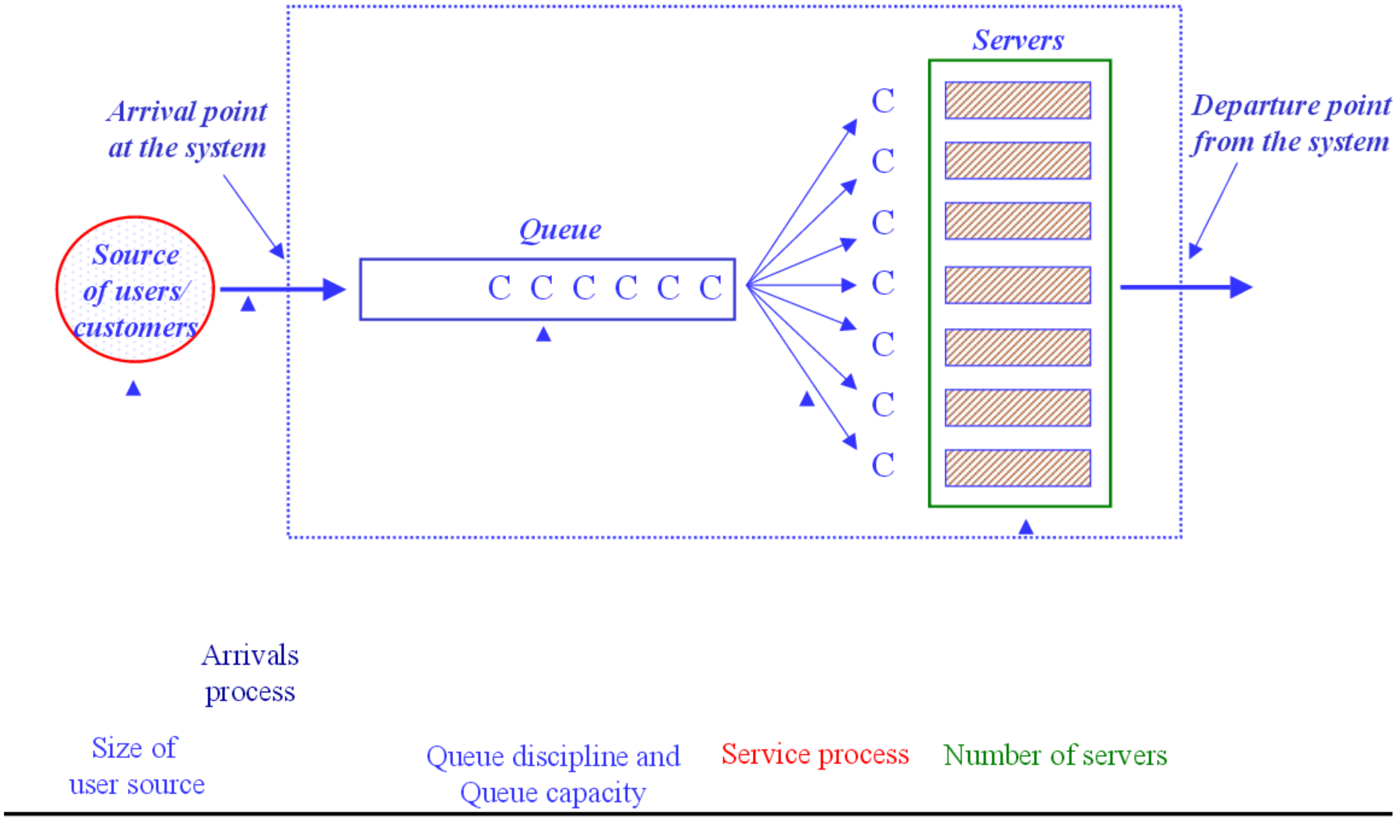
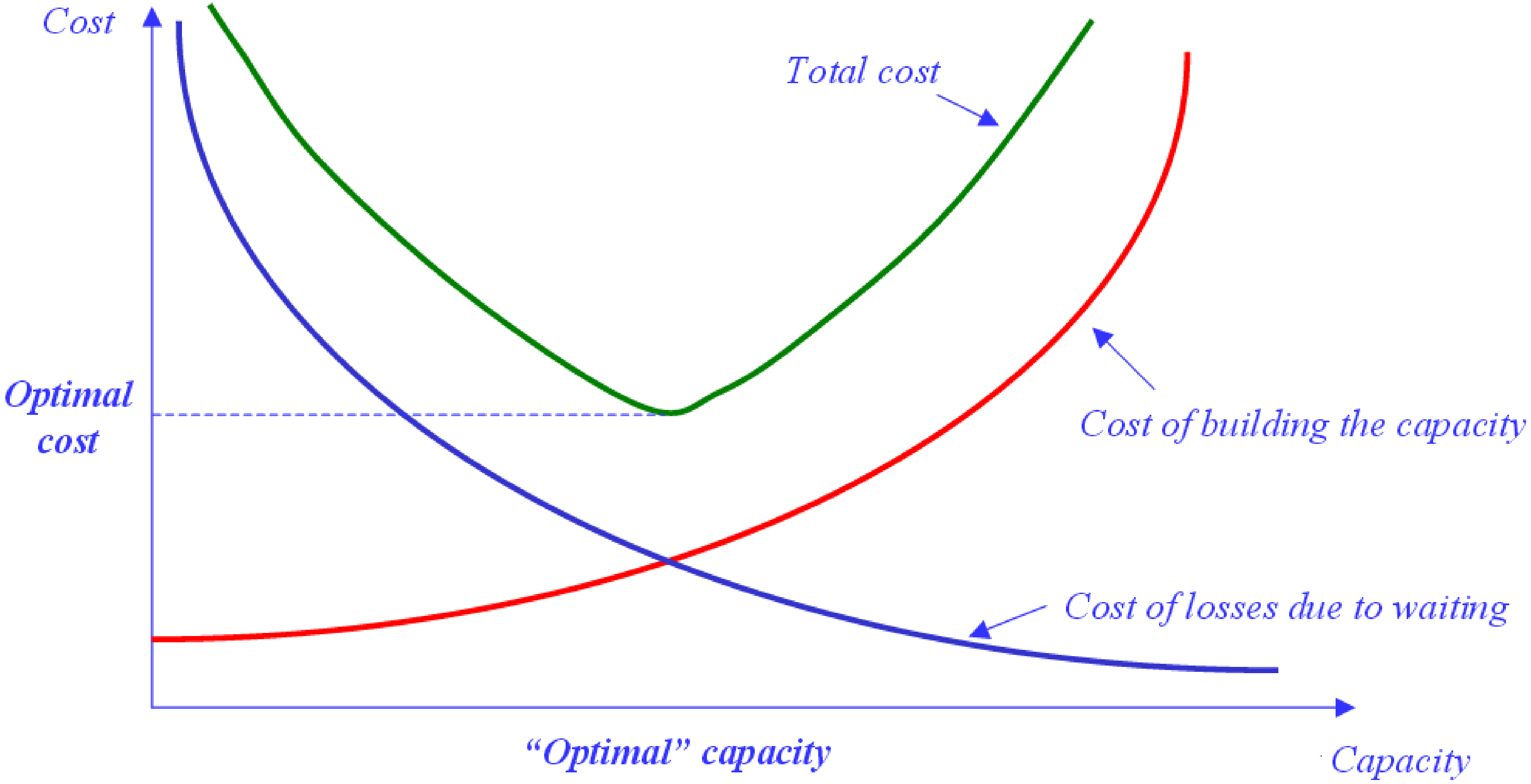


Figure fromhttp://ocw.mit.edu/courses/civil-and-environmental-engineering/1-203j-logistical-and-transportation-planning-methods-fall-2004/lecture-notes/qlec1.pdf

Applications of Queueing Theory

- Familiar Queues:
  - Check-ins at airports or hotels
  - ATMs
  - Fast food restaurants
  - Phone center lines (and telephone exchanges…)
  - Toll booths
  - Busy street intersections
  - Spatially-distributed services (e.g. police, fire)
- Economic Analyses
  - Tradeoff between customer satisfaction & system utilization

Application in Economic Analysis



Queueing Systems

- Road network
  - Customers: cars
  - Server(s): traffic lights
- Warehouse
  - Customers: orders
  - Server(s): order-picker

  - Customers: ?
  - Server(s): ?
- Buses
  - Customers: ?
  - Server(s): ?

Standard Queueing Notation:

A/B/c/N/K

- A represents the *interarrival time* distribution
- B represents the *service-time* distribution
- c represents number of parallel servers
- N represents the system capacity
- K represents the size of the calling population

- Common types of distr. for A/B:
  - M: represents exponential or Markov (more about this later)
  - D: deterministic/constant (not random)
  - Ek: Erlang of order k
  - G: general (arbitrary)

Example Queueing System

- People waiting for a bank teller

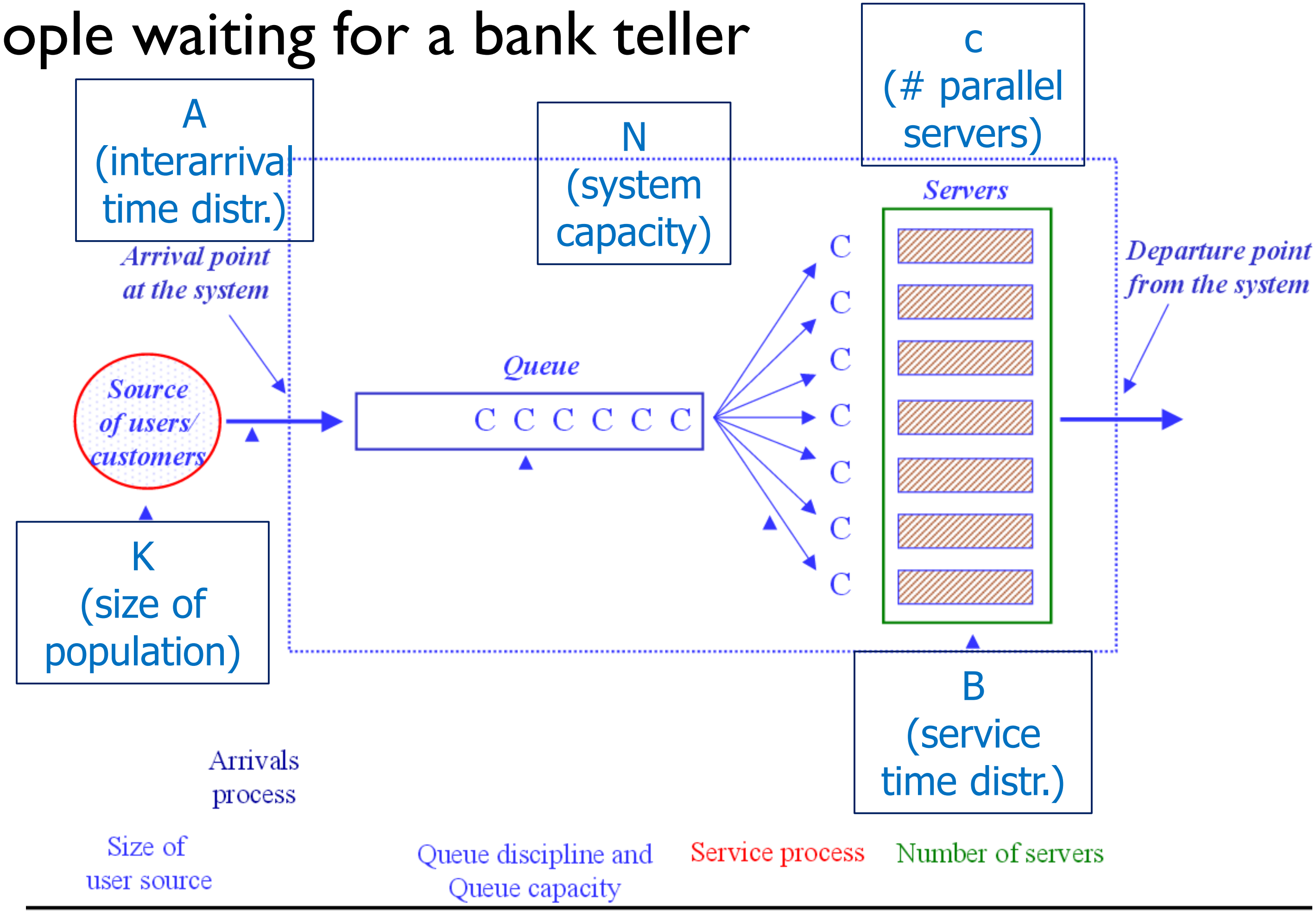


Figure fromhttp://ocw.mit.edu/courses/civil-and-environmental-engineering/1-203j-logistical-and-transportation-planning-methods-fall-2004/lecture-notes/qlec1.pdf

Measures of Performance?

- When running a simulation, we often want to know how well the hypothetical system (i.e. the model) is performing.
  - Useful for:
    - Comparing models for implementation
    - Identify problems (e.g. bottlenecks) in system

- What metrics should we record?

Long-Run Measures of Performance

- Some important queueing measurements
  - L = long-run average number of customers in the system
  - L_Q = long-run average number of customers in the queue
  - w = long-run average time spent in system
  - w_q = long-run average time spent in queue
  - ρ = server utilization (fraction of time server is busy)
  - *Others:*
    - Long-run proportion of customers who were delayed in queue longer than some threshold amount of time
    - Long-run proportion of customers who were turned away due to capacity constraints
    - Long-run proportion of time the waiting line contains more than some threshold number of customers

Measure of Performance
General case (G/G/c/N/K)

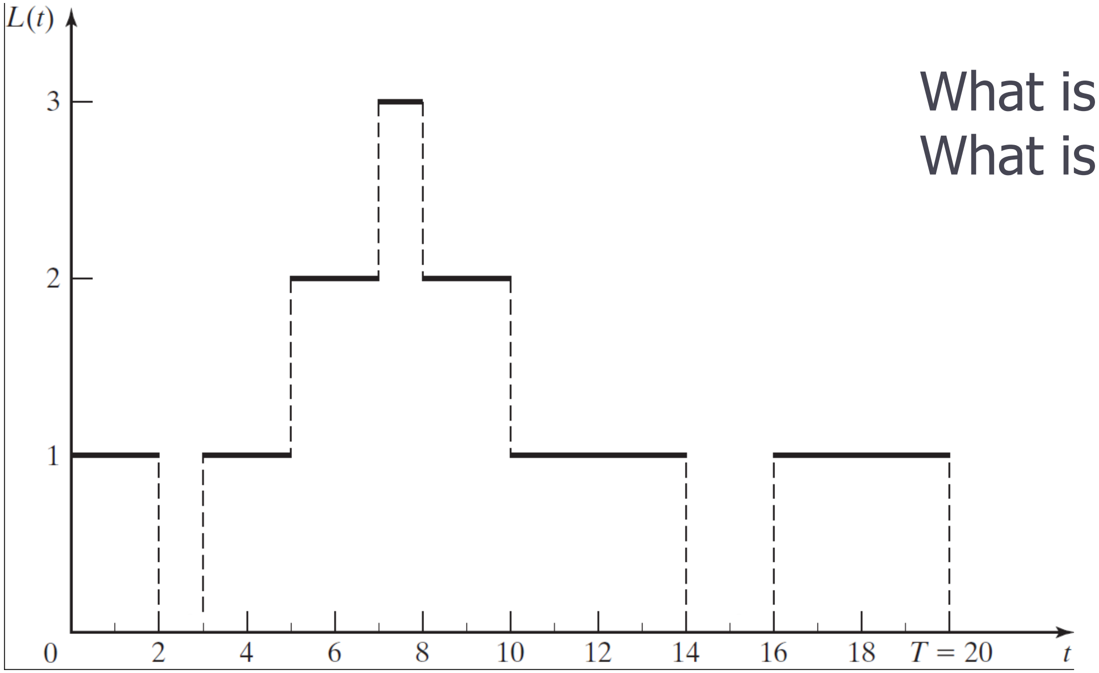
- There are some general relationships between the measures

- How do we estimate the measures from a simulation run?

- Types of estimators we might use:
  - Ordinary sample average
    - Ex: total time customers spent in system / total customers
  - Time-weighted sample average

Time-Average Number in System

- Consider a queueing system over period T
- Let L(t)=# of customers in the system at time t.
- Let T_i denote the total time during [0,T] s.t. the system contained exactly i customers



What is the value of T_1?
What is the value of T_3?

Time-Average Number in System

- We can calculate $\hat{L}$ the time-weighted-average number in a system:

$$\hat{L} = \frac{1}{T}\sum_{i=0}^{\infty} iT_i$$

  - It is an estimator for the long-run average number of customers in the system

- From the figure, we see that $\hat{L}$ is the area under the curve of T_i over time, so:

$$\hat{L} = \frac{1}{T}\sum_{i=0}^{\infty} iT_i = \frac{1}{T}\int_0^T L(t)\,dt$$

- What is $\hat{L}$ for the previous example?

Time-Average Number in System

- For many stable systems, as T →∞, $\hat{L}$ approaches ***L*,** which is known as the *long-run time-average number of customers in the system*
  - The estimator $\hat{L}$ is said to be strongly consistent for L
  - The longer we run the simulation, the closer it gets to L

Time-Average Number in Queue

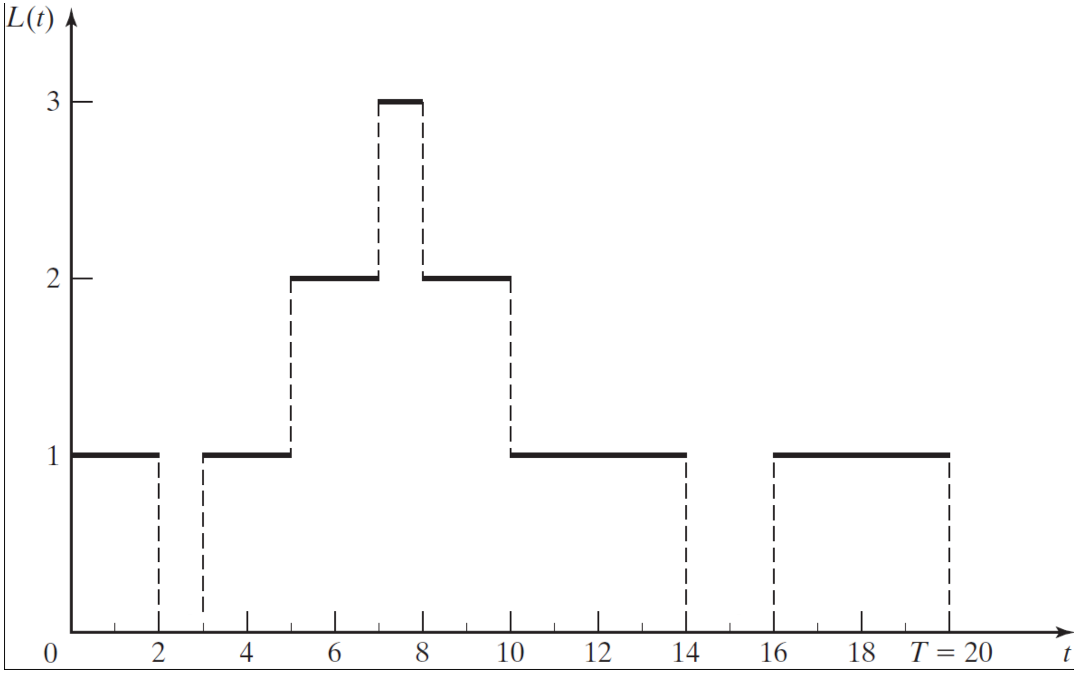
- The same principles can be applied to $\widehat{L_Q}$, the time-average number in the queue, and the corresponding ***L*$_Q$**, the long-run time average number in the queue: as T →∞,

$$\hat{L}_Q = \frac{1}{T}\sum_{i=0}^{\infty} i T_i^Q = \frac{1}{T}\int_0^T L_Q(t)\,dt \rightarrow L_Q$$

- T_i^Q denotes the total time during [0,T] in which exactly *i* customers are waiting in the queue
  - Note that you are not raising T_i to the Q power

Time-Average Number in Queue

- Suppose the figure below represents a single-server queue (G/G/1/N/K, N >= 3, K >= 3). What is the observed time-average number of customers waiting in the queue?



Average Time Spent in the System

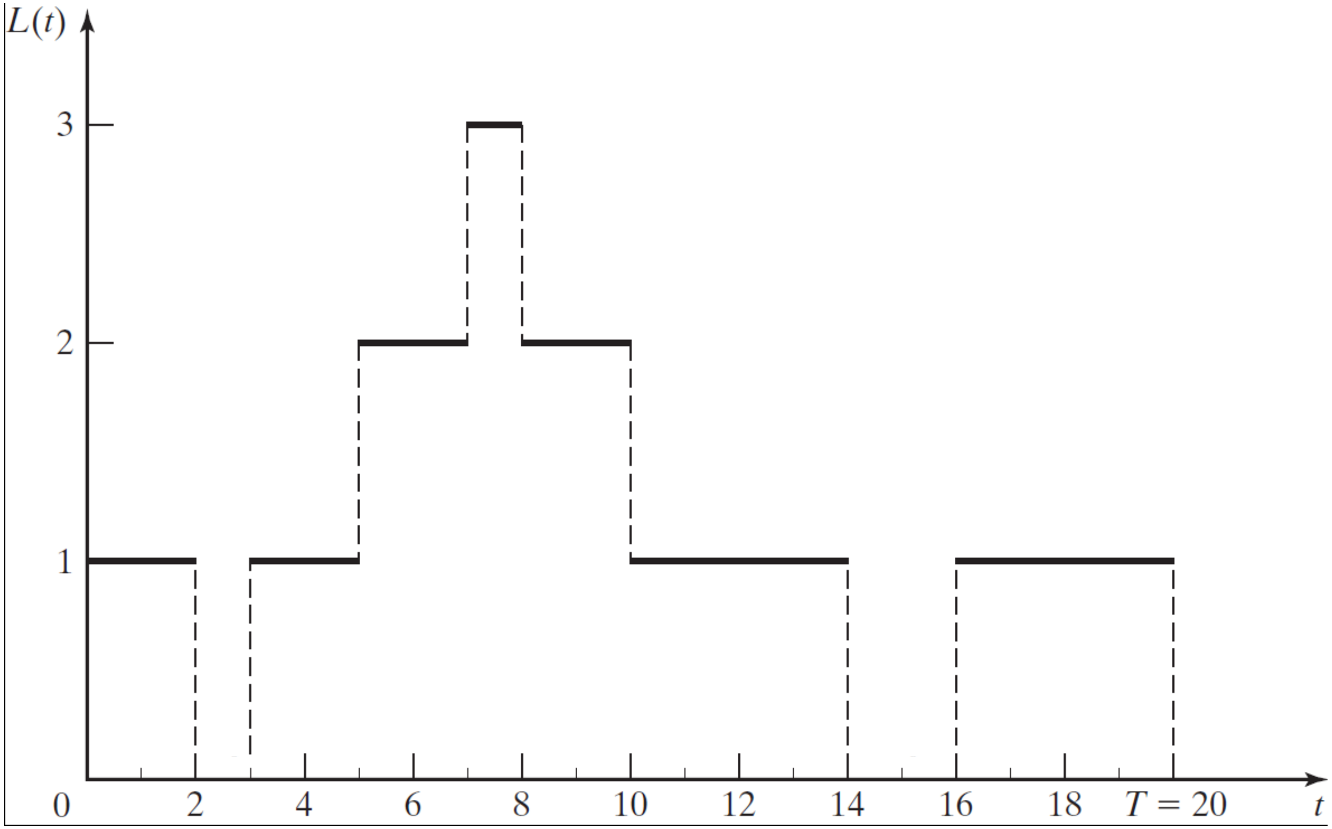
- Let W_i be the amount of time that the *i*th customer spent in the system during [0,T]. If there were N customers, the average time spent in a system per customer is:

$$\hat{w} = \frac{1}{N}\sum_{i=1}^{N} W_i$$

  - For stable systems, as N → ∞ , $\widehat{w} \rightarrow w$, where *w* is called the long-run average system time
  - Again, similar calculations can be done for just the queue part (i.e., estimating w_Q with $\widehat{w}_Q$)
    - □ We can think of these values as the observed delay and the long-run average delay per customer

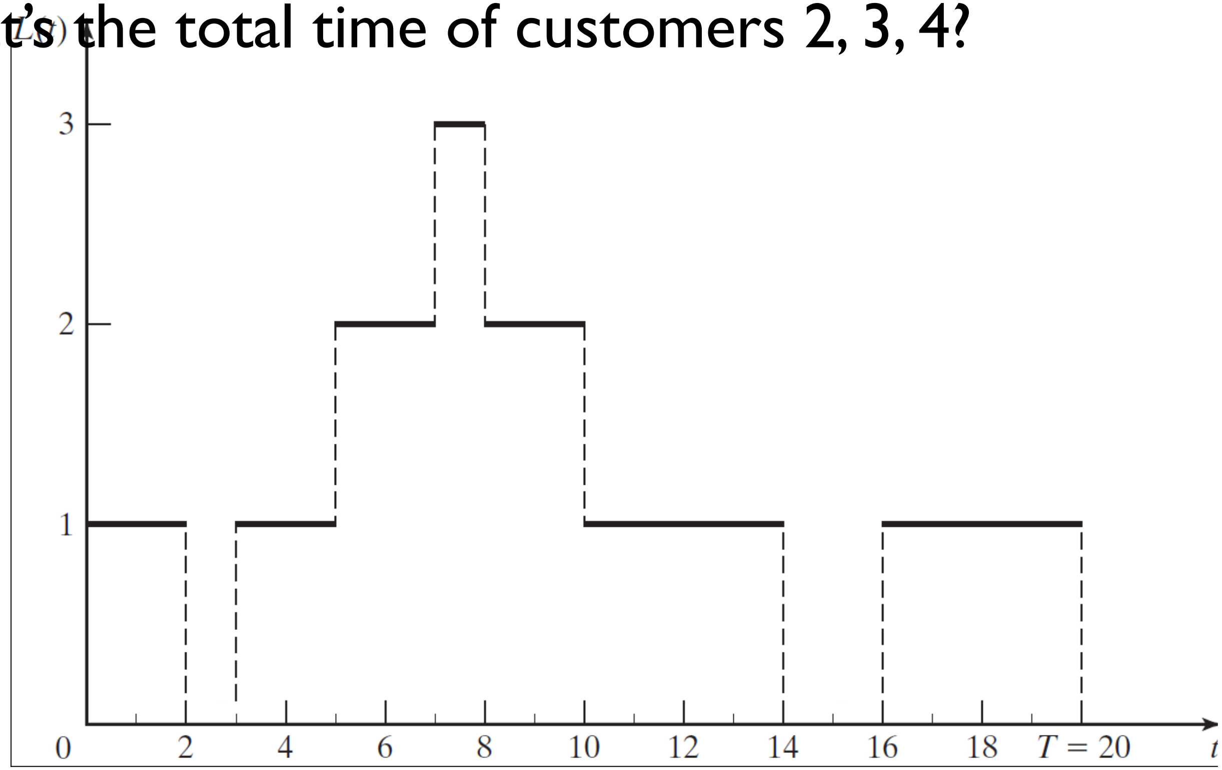
Average Time Example

- How many customers arrive in the system?



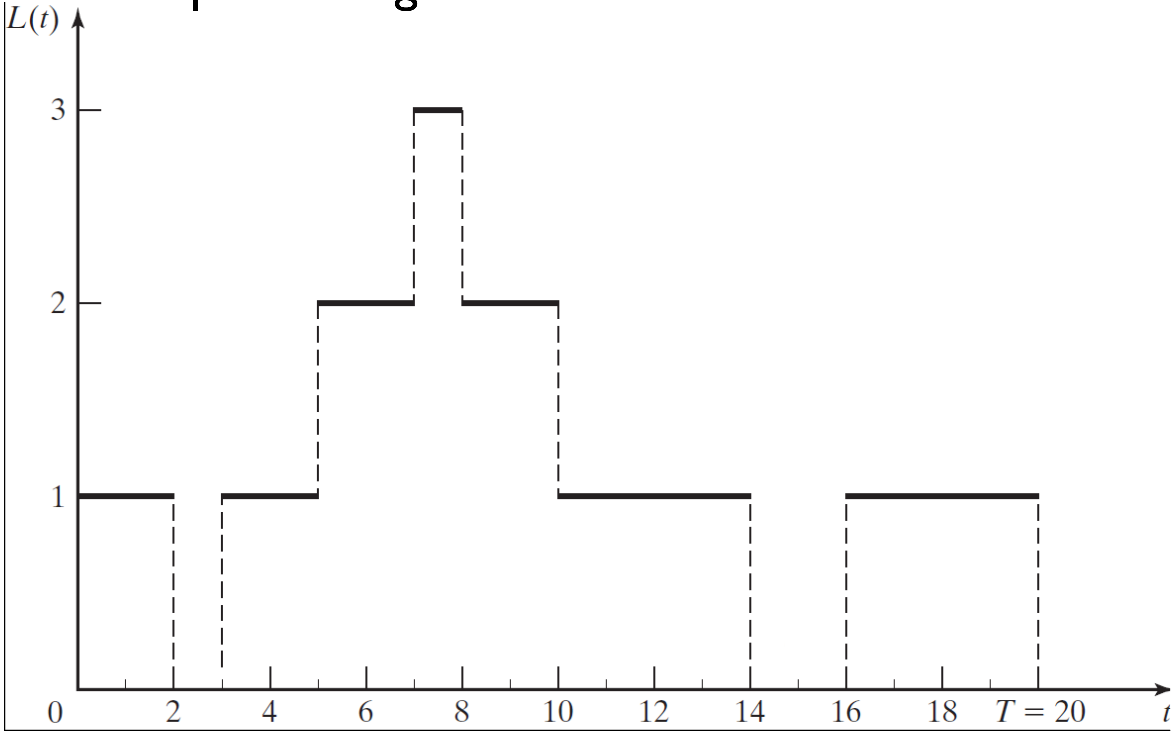
Average Time Example

- What's the total time customer 1 spent in the system
  - i.e. what's W_1?
- What's the total time of customer 5 (i.e. what's W_5?)
- What's the total time of customers 2, 3, 4?



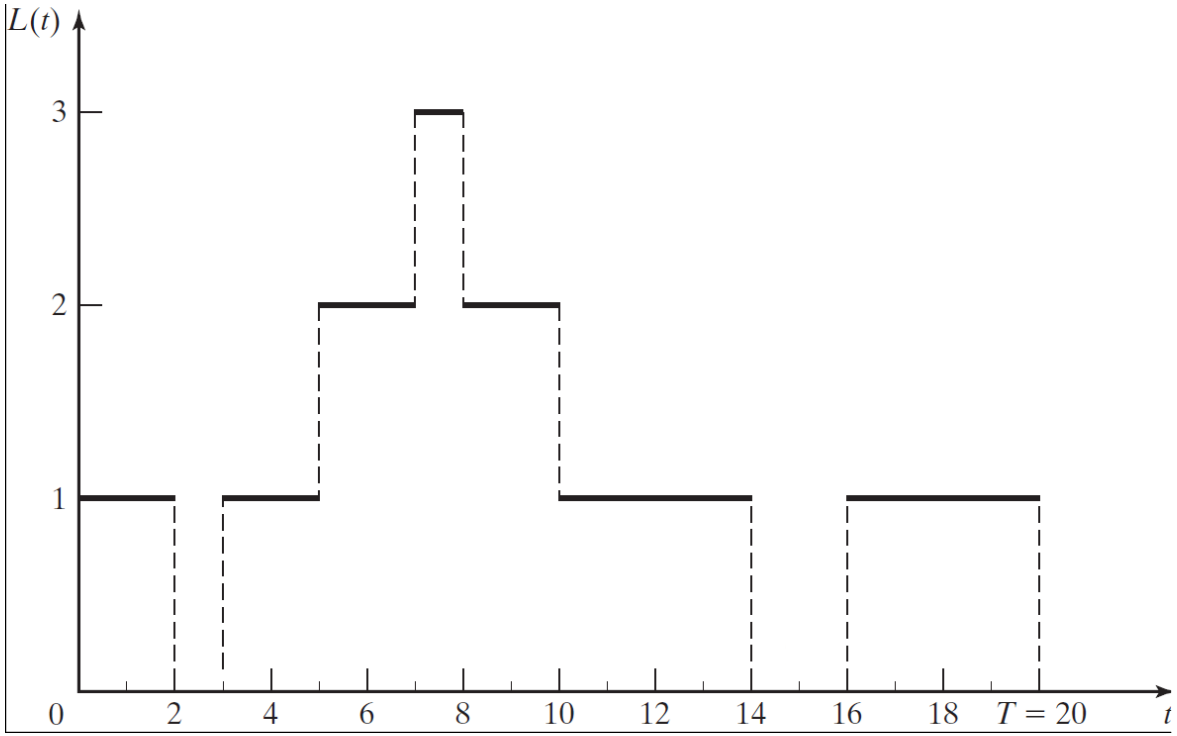
Average Time Example

- What's the total system time of customers 2, 3, 4?
  - Can't tell without more information
    - How many servers? Assume c = 1
    - What's the processing order? Assume FIFO



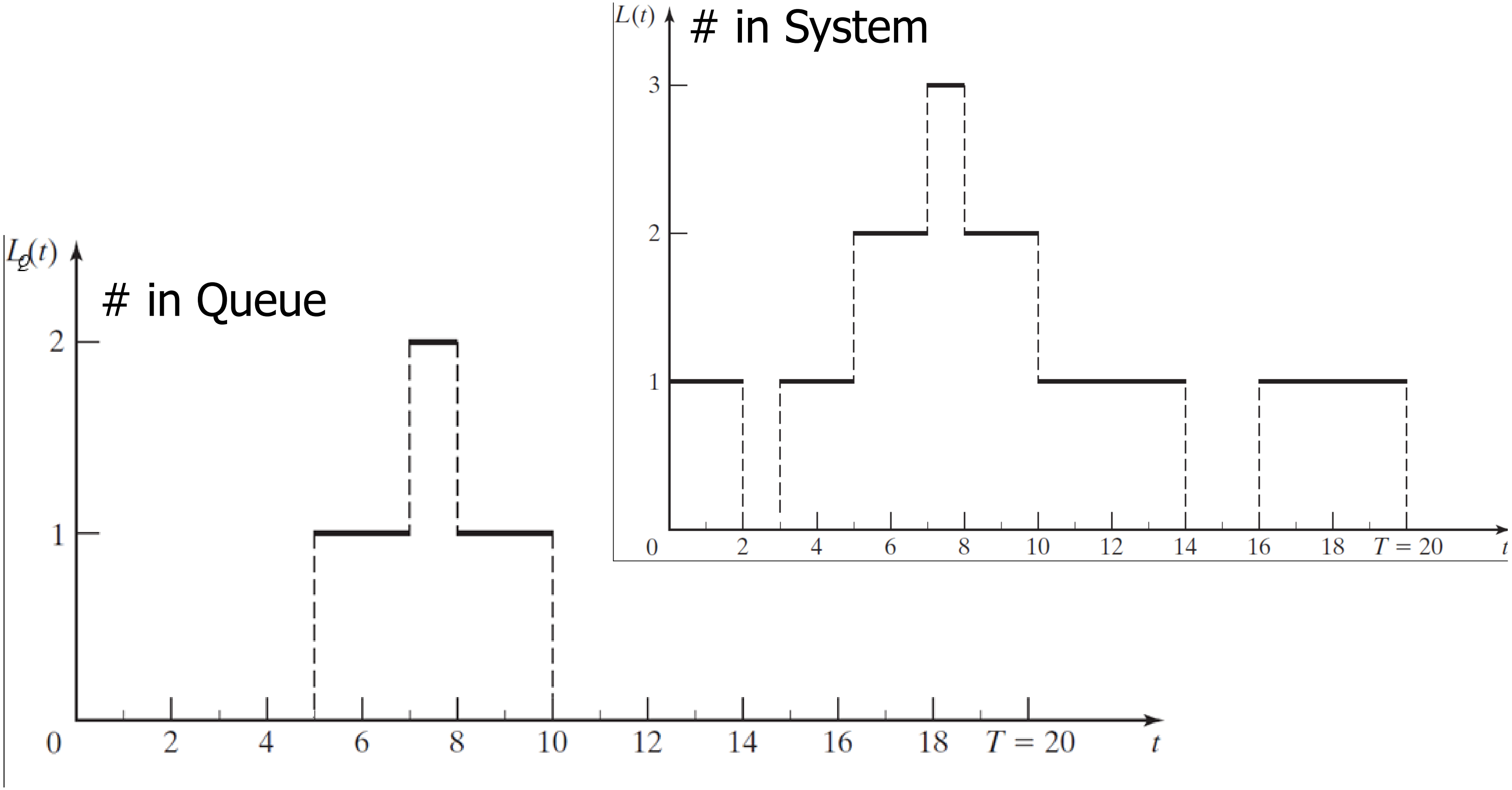
Average Time Example

- What’s the average time of customers in the system?



Average Time Example

- What's the average system time of customers in the queue?



Conservation Law

- Often called "Little's Equation"

  $\hat{L} = \hat{\lambda}\hat{w}$ and as T, N → ∞, $L = \lambda w$

  - where L is the long-run number in the system, λ is the arrival rate and *w* is the long-run time in the system
  - This holds for most queueing systems
- Sketch of derivation for a single server FIFO queueing model:
  - Total system time of all customers is also given by the total area under the number-in-system function, L(t):

$$\sum_{i=1}^{N} W_i = \int_0^T L(t)dt$$

  - Therefore,
    - □ $\hat{L} = \frac{1}{T}\int_0^T L(t)dt = \frac{1}{T}\frac{N}{N}\sum_{i=1}^{N} W_i = \frac{N}{T}\frac{1}{N}\sum_{i=1}^{N} W_i = \hat{\lambda}\hat{w}$
- Requirements: system is non-preemptive and stable

System Stability

- What's a "stable system"?
  - Informally, one that doesn't spiral out of control with too many people waiting indefinitely in line
- For the relatively simple systems that we've seen so far:
  - The arrival rate (λ) must be less than the service rate
    - □ i.e. customers must arrive with less frequency than they can be served
  - Consider a simple single queue system with a single server (G/G/1/∞/∞)
    - □ Let the service rate (# customers served per time unit) be μ
    - □ This system is stable if $\lambda < \mu$
    - □ If $\lambda > \mu$
      - □ This will lead to increase in the number in the system (L(t)) without bound as t increases
      - □ The wait line will grow at a rate of $(\lambda - \mu)$ customers per time unit
    - □ If $\lambda == \mu$ some systems (ex: deterministic) may be stable while others may not

Arrival Rates, Service Rates and Stability

- For a system with multiple servers (ex: G/G/c/∞/∞)
  - Stable if the net service rate of all servers together is greater than the arrival rate
    - If all servers have the same rate μ, then the system is stable if $\lambda < c\mu$
- For a system with finite system capacity or calling pool the system can be stable even if the arrival rate exceeds the service rate
    - Ex: G/G/c/N/∞
      - The system is unstable until it "fills" up to N. At this point, excess arrivals are not allowed into the system, so it is stable from that point on
    - Ex: G/G/c/∞/k
      - With a fixed calling population, we are in effect restricting the arrival rate

Server Utilization

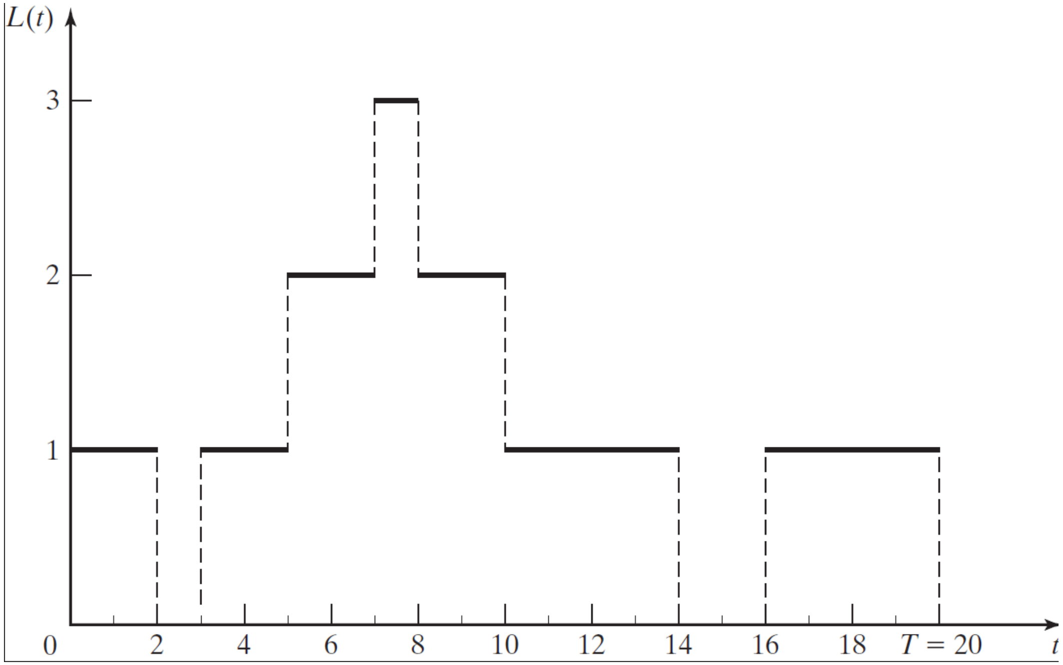
- Calculates the fraction of the time that the server is busy
  - Observed server utility is denoted as $\hat{\rho}$
  - Long-run server utilization is denoted as ρ

- For a G/G/1/∞/∞ system:
  - $\hat{\rho} = \frac{1}{T}\sum_{i=1}^{\infty} T_i$

  - Alternatively, we can think about ρ in terms of customer arrival rate, λ, and the service rate, μ (# of customers served per time unit)
    - $\rho = \frac{\lambda}{\mu}$

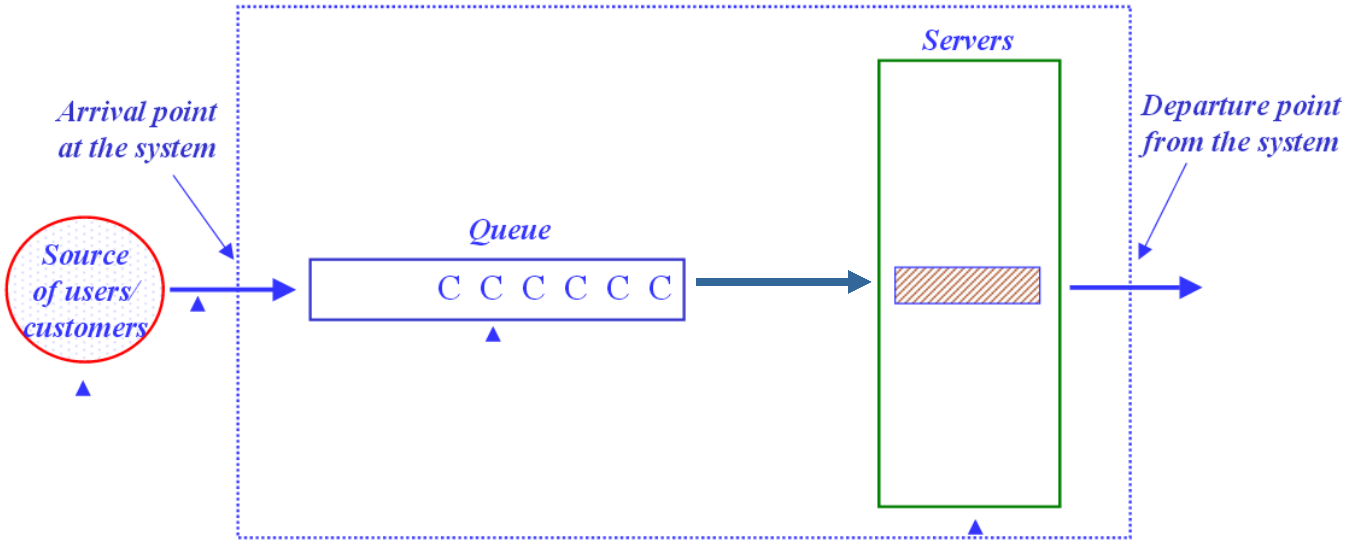
Server Utilization Example

- What is the observed server utilization ($\hat{\rho}$) in the situation depicted below?
  - What assumptions needed to be made?



Server Utilization in G/G/1/∞/∞

- People waiting for a bank teller



- What's the server utilization?

Figure fromhttp://ocw.mit.edu/courses/civil-and-environmental-engineering/1-203j-logistical-and-transportation-planning-methods-fall-2004/lecture-notes/qlec1.pdf

Single-Server Subsystem of G/G/1/∞/∞

- N_S: Capacity of server subsystem
  - 1
- $\hat{L}_S$: # customers in server subsystem
  - 0 or 1
  - $\hat{L}_S = \hat{\rho}$
- ρ: Server utilization
  - $\rho = \frac{\lambda_S}{\mu}$
- Stable as long as:
  - $\lambda_S < \mu$ or $\rho = \frac{\lambda_S}{\mu} < 1$
  - $\lambda_S = \lambda$
    - Why not $\lambda_S > \lambda$

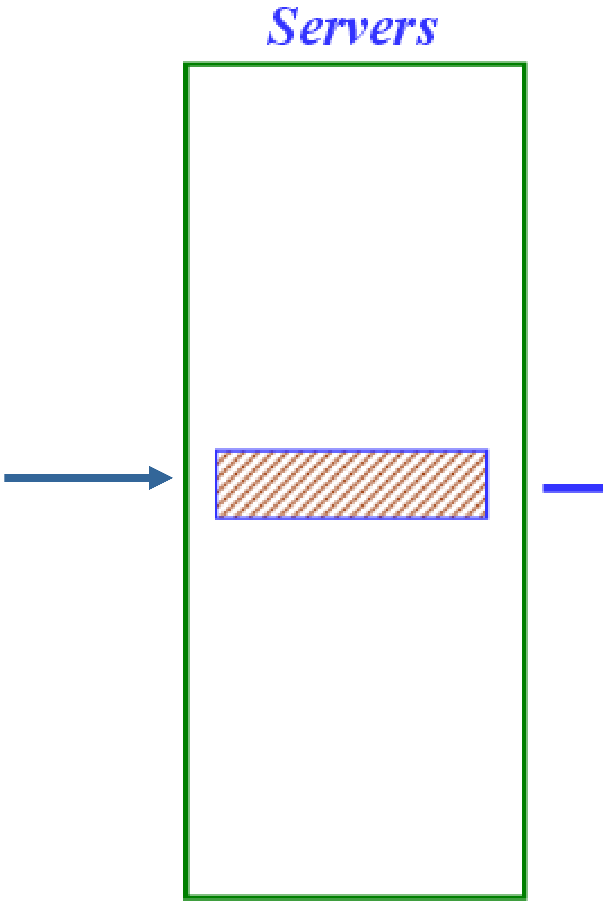


Figure fromhttp://ocw.mit.edu/courses/civil-and-environmental-engineering/1-203j-logistical-and-transportation-planning-methods-fall-2004/lecture-notes/qlec1.pdf

Server Utilization for G/G/1/∞/∞ Systems

- For a single server, we can consider the server portion as a "system" (w/o the queue)
  - This means L_s, the average number of customers in the "server system," equals ρ
  - The average system time w_s is the same as the average service time

    $w_s = 1/\mu$
  - From the conservation equation, we know $L_s = \lambda_s w_s$
  - For the system to be stable, $\lambda_s = \lambda$, since we cannot serve faster than customers arrive and if we serve more slowly the line will grow indefinitely
  - Therefore: $\rho = L_s = \lambda_s w_s = \lambda\ (1/\mu) = \lambda/\mu$
- This shows: a stable queueing system must have a server utilization of less than 1

Server Utilization for $G/G/c/\infty/\infty$ systems

- Similar analysis holds for a stable system with c servers
- Since the system is stable, we must have

$$\lambda < c\mu$$

- And so the server utilization generalizes to

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

- Example: Customers arrive at random to a license bureau at a rate of 50 customers/hour. Currently, there are 20 clerks, each serving 5 customers/hour on average.
  - What is the average server utilization?
  - What is the average number of busy servers?
  - What is the minimum # of clerks needed to keep the system stable?

Steady-State Behavior of M/M/c/∞/∞ Systems

- A queueing system is said to be in statistical equilibrium, or steady state, if the probability that the system is in a given state is not time dependent
  - e.g., the prob. of having n people in the system doesn't depend on time – Pr(L(t)=n) is some value P_n for all time t

- For relatively simple queueing models, some of the long-run steady state performance measures can be calculated analytically
  - May enable us to avoid a simulation for simple systems
  - May give us a good starting point even if the actual system is more complicated

Queueing Systems with Markov properties

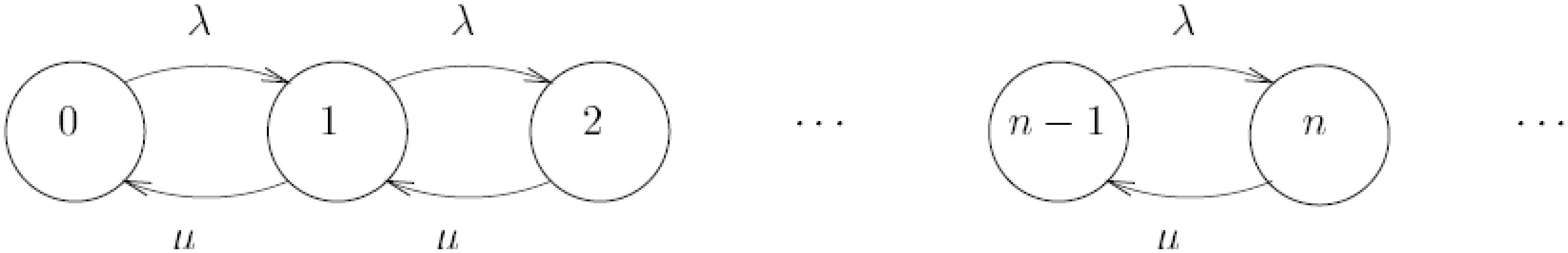
- Why is an Exponential Distribution called "Markov"?
  - Short answer: Has to do with its memoryless property

- Markov Chains (discrete)
  - A set of random variables ("states") $X_1, X_2, \ldots$ forms a Markov Chain if the probability of transitioning from X_i to X_{i+1} does not depend on any of the previous $X_1, \ldots X_{i-1}$

  - $\Pr(X_{i+1} \mid X_1, \ldots, X_i) = \Pr(X_{i+1} | X_i)$

Markov Process

- Consider a (continuous) random variable Y that describes how long a system will be in one state before transitioning to a different state
  - For example, in a queueing system, Y could model duration before another arrival into the system (which changes the system state)
  - This time should not depend on how long the process has been in the current state
  - Thus, when arrivals or services times are exponentially distributed, they are often called Markovian



Steady-State Behavior of M/M/c/∞/∞ Systems

- Notation:
  - P_n is the probability that there are *n* people in the system
    - Since we are at steady state, this probability doesn't change over time
    - $Pr(L(t) = n) = P_n(t) = P_n$
- Invariants
  - $L = \sum_{n=0}^{\infty} n * P_n$
    - L is computed just like any other expectation over a probability distr.
  - $w = L/\lambda$
    - From the conservation equation
  - $w_Q = w - 1/\mu$
    - μ is the rate of service so 1/μ is the average time to serve one customer
  - $L_Q = \lambda w_Q$
    - From the conservation equation again

M/M/1 in steady state

- Arrivals follow a Poisson distribution (λ arrivals per time unit)
- Service times are exponentially distributed with mean 1/μ (and variance 1/μ²)
- Its performance measures at steady state are listed on the right:

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

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

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

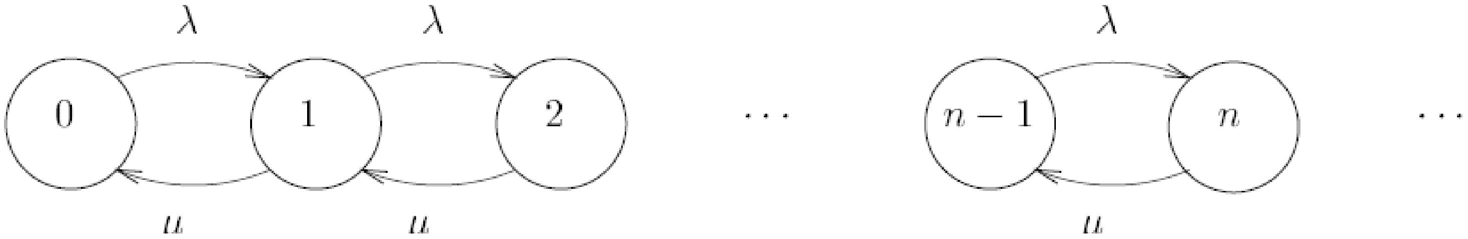
$$L_Q = L - \rho = \frac{\rho^2}{1-\rho}$$

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

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

M/M/1 steady state measures

- At steady-state, the system should converge to:
  - $0 = -\lambda P_0 + \mu P_1$
  - $0 = \lambda P_{n-1} - (\lambda+\mu) P_n + \mu P_{n+1}$



- From the first eq: we have $P_1 = \lambda/\mu \, P_0 = \rho P_0$
- For the second eq (n=1): $P_2 = -\lambda/\mu P_0 + (\lambda+\mu)/\mu P_1 = \rho P_1 = \rho^2 P_0$
  - More generally $P_n = \rho^n P_0$
- We also know that $\Sigma \, P_i = 1$, so we know $P_0 \Sigma \rho^i = 1$
  - Summing over the inf. series, we get $P_0/(1-\rho) = 1$, so $P_0 = (1-\rho)$
  - so $P_n = \rho^n (1-\rho)$

Derivation cf. Adan and Resing (2001) “Queueing Theory”

M/M/1 steady state measures

$$L = \sum_{i=0}^{\infty} iP_i = 0[1-\rho]\rho^0 + 1[1-\rho]\rho^1 + 2[1-\rho]\rho^2 + 3[1-\rho]\rho^3 + \cdots$$

$$= 0 + \rho - \rho^2 + 2\rho^2 - 2\rho^3 + 3\rho^3 - 3\rho^4 + \cdots$$

$$= \rho(1 + \rho + \rho^2 + \rho^3 + \cdots) = \frac{\rho}{1-\rho}$$

And from Little's Eq ($L = \lambda w$):

- $L = \rho / (1-\rho) = \lambda w$, so
- $w = L/\lambda = \rho /\lambda * 1/ (1-\rho) = 1/\mu(1-\rho)$

- As ρ → 1, L and w would grow towards ∞

Avg. # of people in line = avg # of people in sys – those being served:

- $L_Q = L - \rho = \rho/(1-\rho) - (\rho - \rho^{2)}/(1-\rho) = \rho^2/(1-\rho)$

Avg. time spent in line = avg. time in sys – service time

- $w_Q = w - 1/\mu = 1/\mu(1-\rho) -(1- \rho)/\mu(1-\rho) = \rho/\mu(1-\rho)$

M/M/1 example (adapted from ex 6.12)

- Suppose the customer arrival rate is 10 per hour, following a Poisson distribution.
- You have a choice of hiring either Alice or Bob. Alice works at a rate of 11 customers per hour, while Bob works at a rate of 12 customers per hour. However, Bob wants to be paid about twice as much as Alice. Should you consider hiring Bob?

M/G/1 in Steady-State

- Arrivals follow a Poisson distribution (λ arrivals per time unit)
- Service times follow an arbitrary distribution that has a mean of 1/μ and variance of σ^2
- The performance measures for this case is more complex b/c the service time is described by an arbitrary distribution
- In general, no simple expression for $P_1, P_2, \ldots$

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

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

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

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

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

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

$$P_0 = 1 - \rho$$

M/G/1 in Steady-State

- What if $\sigma^2 = 0$
  - i.e. the service times are all the same (= mean)
    - For example a deterministic distribution
  - In this case the equations for L and L_Q greatly simplified:

$$L_Q = \frac{\rho^2(1+0^2\mu^2)}{2(1-\rho)} = \frac{\rho^2}{2(1-\rho)}$$

  - In this case L_Q depends solely on the server utilization, ρ
    - □ Note as ρ → 0 (low server utilization) L_Q → 0
    - □ Note as ρ → 1 (high server utilization) L_Q → ∞

- If utilization is fixed, then as σ^2 increases, L_Q also increases

M/G/1 in Steady-State

- Other measures such as *w* and w_Q increase with σ^2 as well
  - This indicates that all other factors being equal, a system with a lower variance will tend to have better performance
    - (See Ex. 6.9) In some cases a lower μ will give shorter lines than a higher μ, if it has a lower σ^2
      - □ Two workers are competing for a job. Able claims an average service time that is faster than Baker's. Baker claims to be more consistent in speed, but slower on average. Customer arrivals occur according to a Poisson process at a rate of 2 per hour (1/30 per minute). Who should be hired if average queue length is the hiring criterion?
        - □ Able: avg service time = 24 minutes with standard deviation of 20 minutes
        - □ Baker: avg service time = 25 minutes with standard deviation of 2 minutes
      - □ Note that Able has a longer long-run queue length despite his faster rate
      - □ However, he also has a higher P_0, indicating that more people experience no delay

Coefficient of Variation

- We can generalize this idea, comparing various distributions using the coefficient of variation, *cv*:

  $(cv)^2 = Var(X)/(E[X])^2 = \sigma^2/(1/\mu)^2 = \sigma^2\mu^2$

- We can rewrite L_Q of M/G/1 using cv:

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

  - This highlights the relationship with L_Q in M/M/1
    - The second term modifies the M/M/1 formula to account for a nonexponential service-time distribution.
    - In an exponential distr. $\sigma^2 = 1/\mu^2$
    - Distributions that have a larger cv have a larger L_Q for a given server utilization, ρ

M/G/1 Example (*Exercise 6.6)*

- Patients arrive for a physical exam according to a Poisson process at the rate of 1/hr.
- The physical exam requires 3 stages, each one independently and exponentially distributed with a service time of 15min.
- A patient must go through all 3 stages before the next patient is admitted to the facility.
- Determine the average number of delayed patients, L_Q, for this system.

- Note: since a patient has to go through three stages, the process is better described by Erlang than Exponential

Multiple servers case

- M/M/c
  - The performance measures are a little more involved (see right)
  - Can be expressed in terms of the probability when the system is empty, P_0, and when all servers are busy, $\sum_{n=c}^{\infty} P_n$, which the textbook denotes as Pr(L(∞) ≥c).
- M/G/c
  - Approximate L_Q and w_Q by multiplying the M/M/c equations (see right) by the approximation factor $(1+cv^2)/2$

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

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

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

$$L - L_Q = c\rho = \frac{\lambda}{\mu}$$

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

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

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

Multi-servers example (Adapted from ex. 6.13)

- Poisson arrival at a rate of λ = 2 customers per minute
- Exponentially distributed service time of 40 seconds (so service rate of μ=1.5 customers per minute)
- The system wouldn't be stable if c=1. Why not?
- What if c=2:
  - What is the chance of having no one in the system?
  - What is the chance that both servers are busy?
  - What is the time-average length of the waiting line?
  - What is the time-average number in the system?
  - What is the average time a customer spent waiting in the system?

Infinite # of servers

- Special case when c=∞
  - This can model self service systems
  - It's appropriate for situations where service capacity far exceeds demands
  - It can be used to answer the question: how many servers are required so that customers will rarely be delayed?

M/G/∞ steady state

- No one waits in line, so performance measures having to do w/ the queue are all 0
- Avg. time spent in the system is just the avg service time
- The avg. # of customers in the system is the same as the server utilization
- The prob. of customers in the system is described by a Poisson distribution

$$w_Q = 0$$

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

$$L_Q = 0$$

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

$$P_0 = e^{-\lambda/\mu} = e^{-\rho}$$

$$P_n = \frac{e^{-\lambda/\mu}(\lambda/\mu)^n}{n!}$$

M/G/∞ Example (Ex. 6.15)

- Prior to introducing their new subscriber-only, online computer information service, The Connection must plan their system capacity in terms of the number of users that can be logged on simultaneously. If the service is successful, customers are expected to log on around 500 per hour and stay connected for an average of 3 hours.
  - What is the expected number of simultaneous users?
  - If they want to ensure adequate capacity 95% of the time, what capacity should they be prepared to handle?

M/M/c/N/∞

- There are c servers and the system capacity is N >= c customers
- If an arrival occurs while the system is full, that customer is turned away
  - So we need to determine the **effective arrival rate** λ_e
    - $\lambda_e = \lambda(1\text{-}P_N)$
    - When system is not capped: $\lambda_e = \lambda$

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

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

$$P_N = \frac{(c\rho)^N}{c!\,c^{N-c}}P_0$$

$$L_Q = \frac{P_0(c\rho)^c\rho}{c!(1-\rho)^2}\left[1-\rho^{N-c}-(N-c)\rho^{n-c}(1-\rho)\right]$$

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

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

$$L = L_Q + \frac{\lambda_e}{\mu} = \lambda_e w$$

M/M/c/K/K

- There is a finite set of K possible customers (a small number)
- The system can support up to all K customers.
- Effective arrival rate is $\lambda_e = \sum_{n=0}^{K} (K-n)\lambda P_n$

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

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

$$P_{n,n<c} = Choose(K,n) \frac{\lambda^n}{\mu^n} P_0$$

$$P_{n,c<=n<=K} = \frac{K!}{(K-n)!c!c^{n-c}} \frac{\lambda^n}{\mu^n} P_0$$

$$L = \sum_{n=0}^{K} nP_n$$

$$L_Q = \sum_{n=c+1}^{K} (n-c)P_n$$

$$L - L_Q = \frac{\lambda_e}{\mu} = c\rho$$

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

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