What is Queuing Theory?

- Mathematical analysis of queues and waiting times in stochastic systems.
  - Used extensively to analyze production and service processes exhibiting random variability in market demand (arrival times) and service times.
- Queues arise when the short term demand for service exceeds the capacity
  - Most often caused by random variation in service times and the times between customer arrivals.
  - If long term demand for service > capacity the queue will explode!

Why is Queuing Analysis Important?

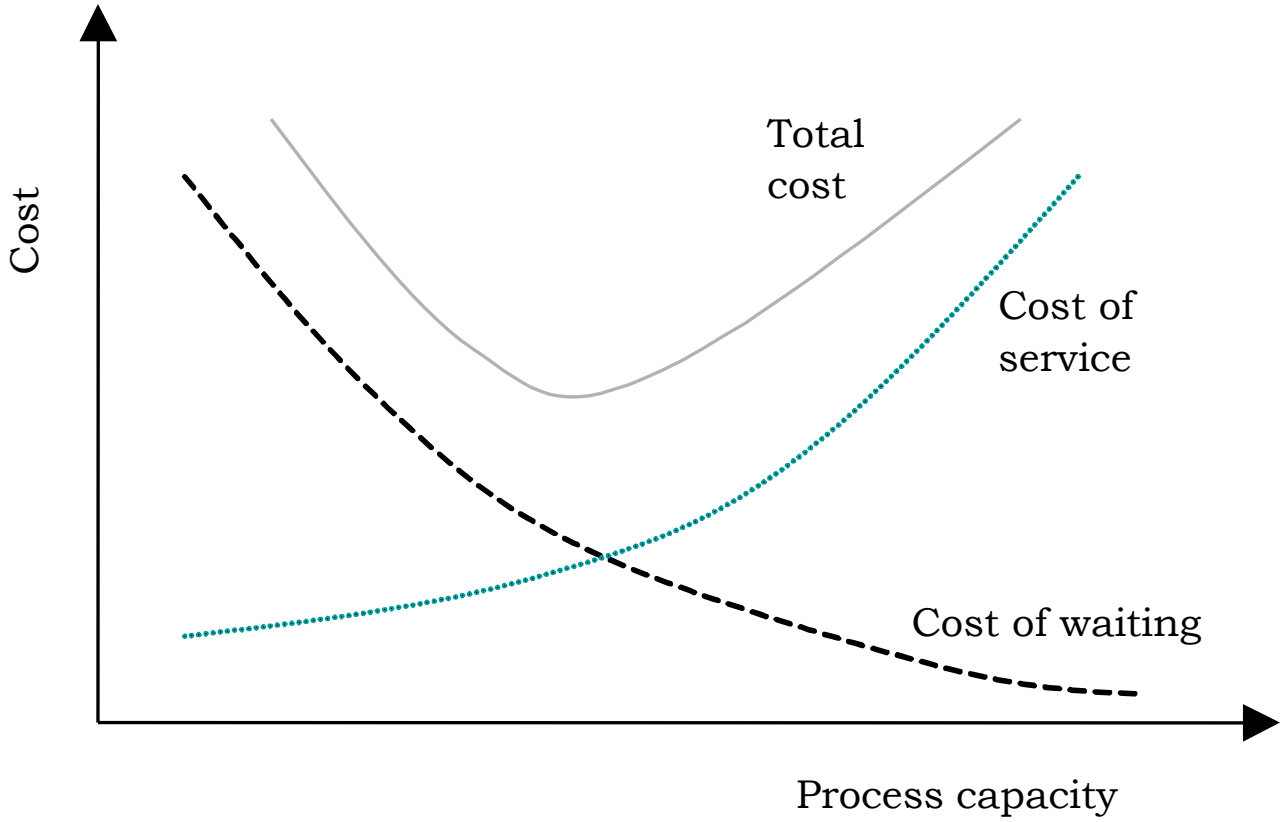
- Capacity problems are very common in industry and one of the main drivers of process redesign
  - Need to balance the cost of increased capacity against the gains of increased productivity and service
- Queuing and waiting time analysis is particularly important in service systems
  - Large costs of waiting and of lost sales due to waiting

Prototype Example – ER at County Hospital

- Patients arrive by ambulance or by their own accord
- One doctor is always on duty
- More and more patients seeks help ⇒ longer waiting times

➢ ***<u>Question</u>*: *Should another MD position be instated?***

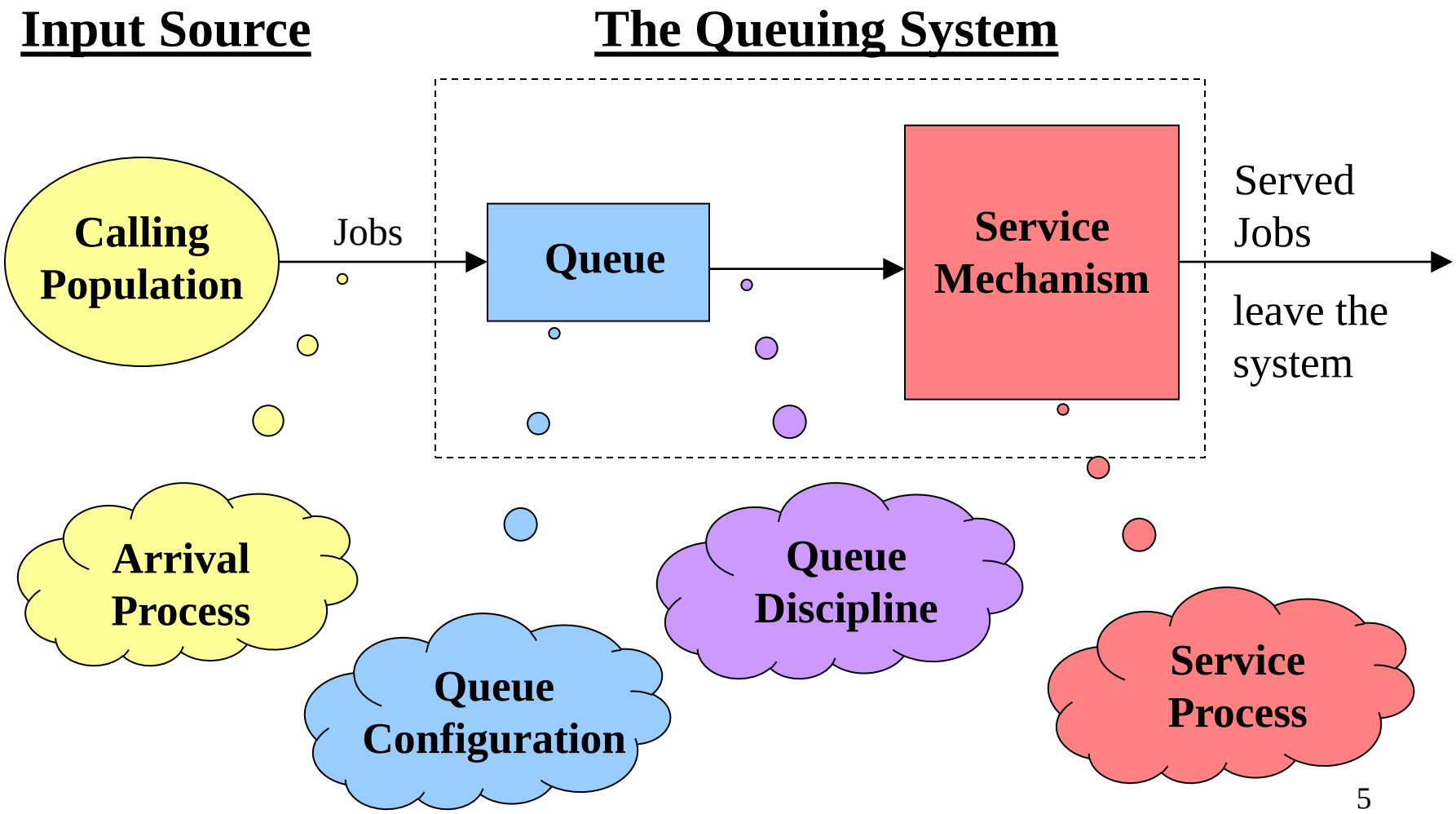
A Cost/Capacity Tradeoff Model



Examples of Real World Queuing Systems?

- Commercial Queuing Systems
  - Commercial organizations serving external customers
  - Ex. Dentist, bank, ATM, gas stations, plumber, garage …
- Transportation service systems
  - Vehicles are customers or servers
  - Ex. Vehicles waiting at toll stations and traffic lights, trucks or ships waiting to be loaded, taxi cabs, fire engines, elevators, buses …
- Business-internal service systems
  - Customers receiving service are internal to the organization providing the service
  - Ex. Inspection stations, conveyor belts, computer support …
- Social service systems
  - Ex. Judicial process, the ER at a hospital, waiting lists for organ transplants or student dorm rooms …

Components of a Basic Queuing Process



Components of a Basic Queuing Process (II)

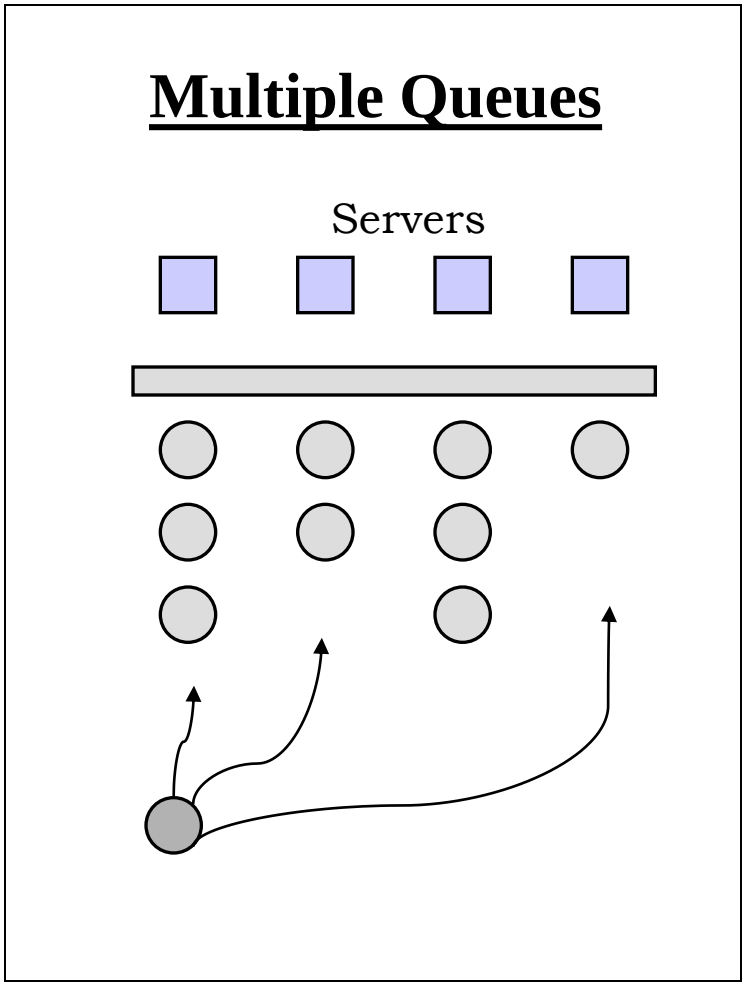
- The calling population
  - The population from which customers/jobs originate
  - The size can be finite or infinite (the latter is most common)
  - Can be homogeneous (only one type of customers/ jobs) or heterogeneous (several different kinds of customers/jobs)

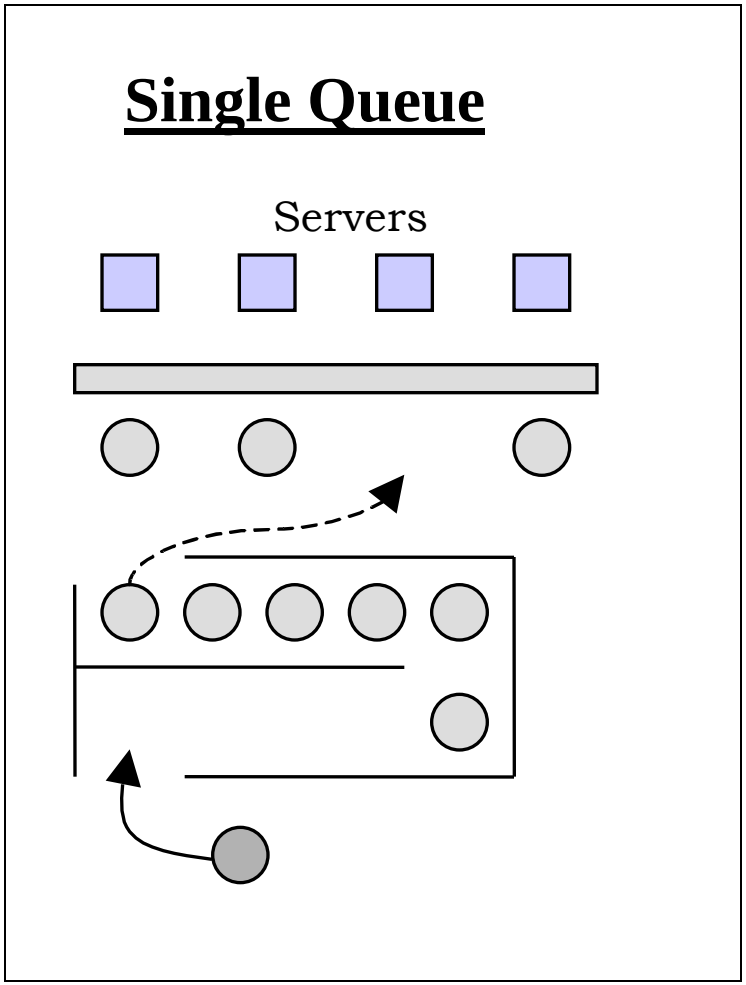
- The Arrival Process
  - Determines how, when and where customer/jobs arrive to the system
  - Important characteristic is the customers'/jobs' inter-arrival times
  - To correctly specify the arrival process requires data collection of interarrival times and statistical analysis.

Components of a Basic Queuing Process (III)

- The queue configuration
  - Specifies the number of queues
    - Single or multiple lines to a number of service stations
  - Their location
  - Their effect on customer behavior
    - Balking and reneging
  - Their maximum size (# of jobs the queue can hold)
    - Distinction between infinite and finite capacity

Example – Two Queue Configurations





Multiple v.s. Single Customer Queue Configuration

Multiple Line Advantages

1. **The service provided can be differentiated**
   - Ex. Supermarket express lanes
2. **Labor specialization possible**
3. **Customer has more flexibility**
4. **Balking behavior may be deterred**
   - Several medium-length lines are less intimidating than one very long line

Single Line Advantages

1. **Guarantees fairness**
   - FIFO applied to all arrivals
2. **No customer anxiety regarding choice of queue**
3. **Avoids "cutting in" problems**
4. **The most efficient set up for minimizing time in the queue**
5. **Jockeying (line switching) is avoided**

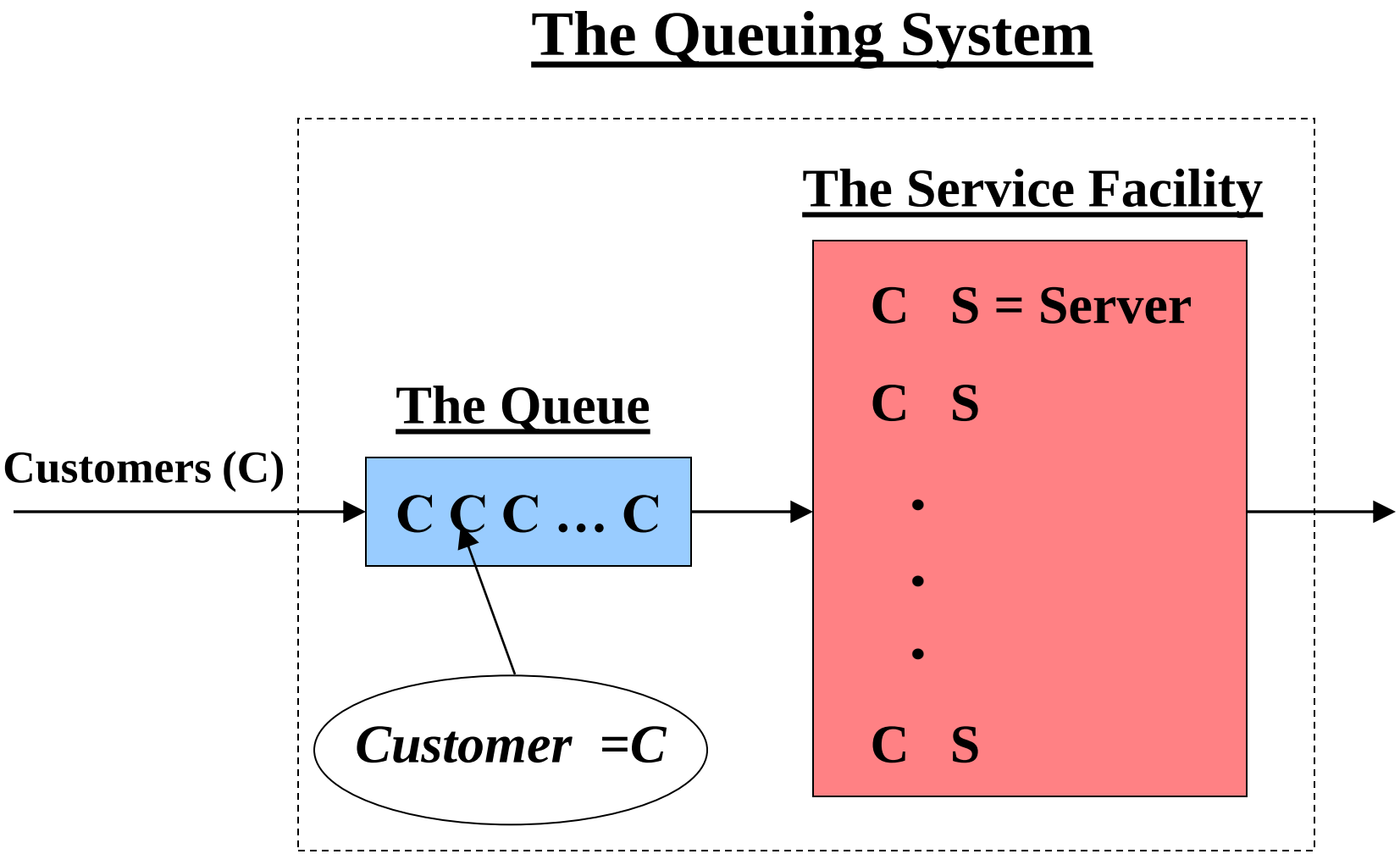
Components of a Basic Queuing Process (IV)

- The Service Mechanism
  - Can involve one or several service facilities with one or several parallel service channels (**servers**) - Specification is required
  - The service provided by a server is characterized by its service time
    - Specification is required and typically involves data gathering and statistical analysis.
    - Most analytical queuing models are based on the assumption of exponentially distributed service times, with some generalizations.
- The queue discipline
  - Specifies the order by which jobs in the queue are being served.
  - Most commonly used principle is FIFO.
  - Other rules are, for example, LIFO, SPT, EDD…
  - Can entail prioritization based on customer type.

Mitigating Effects of Long Queues

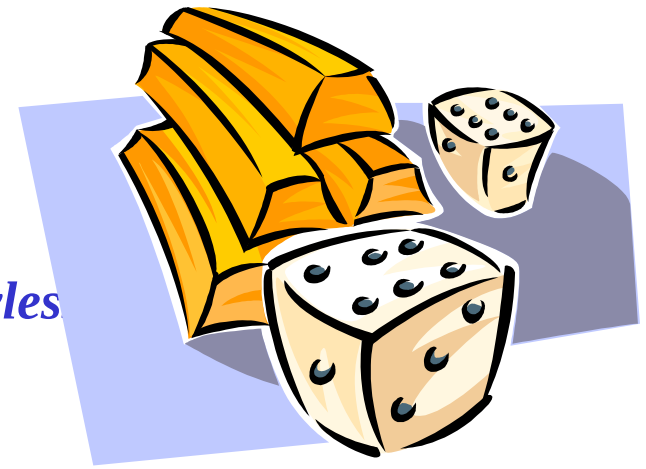
1. Concealing the queue from arriving customers
   - Ex. Restaurants divert people to the bar or use pagers, amusement parks require people to buy tickets outside the park, banks broadcast news on TV at various stations along the queue, casinos snake night club queues through slot machine areas.
2. Use the customer as a resource
   - Ex. Patient filling out medical history form while waiting for physician
3. Making the customer's wait comfortable and distracting their attention
   - Ex. Complementary drinks at restaurants, computer games, internet stations, food courts, shops, etc. at airports
4. Explain reason for the wait
5. Provide pessimistic estimates of the remaining wait time
   - Wait seems shorter if a time estimate is given.
6. Be fair and open about the queuing disciplines used

A Commonly Seen Queuing Model (I)



A Commonly Seen Queuing Model (II)

- Service times as well as interarrival times are assumed independent and identically distributed
  - If not otherwise specified
- Commonly used notation principle: A/B/C
  - A = The interarrival time distribution
  - B = The service time distribution
  - C = The number of parallel servers
- Commonly used distributions
  - M = Markovian (exponential) - ***Memoryless***
  - D = Deterministic distribution
  - G = General distribution



- Example: M/M/c
  - Queuing system with exponentially distributed service and inter-arrival times and c servers

The Exponential Distribution and Queuing

- The most commonly used queuing models are based on the assumption of exponentially distributed service times and interarrival times.

Definition: A stochastic (or random) variable $T \in \exp(\alpha)$, i.e., is exponentially distributed with parameter α, if its frequency function is:

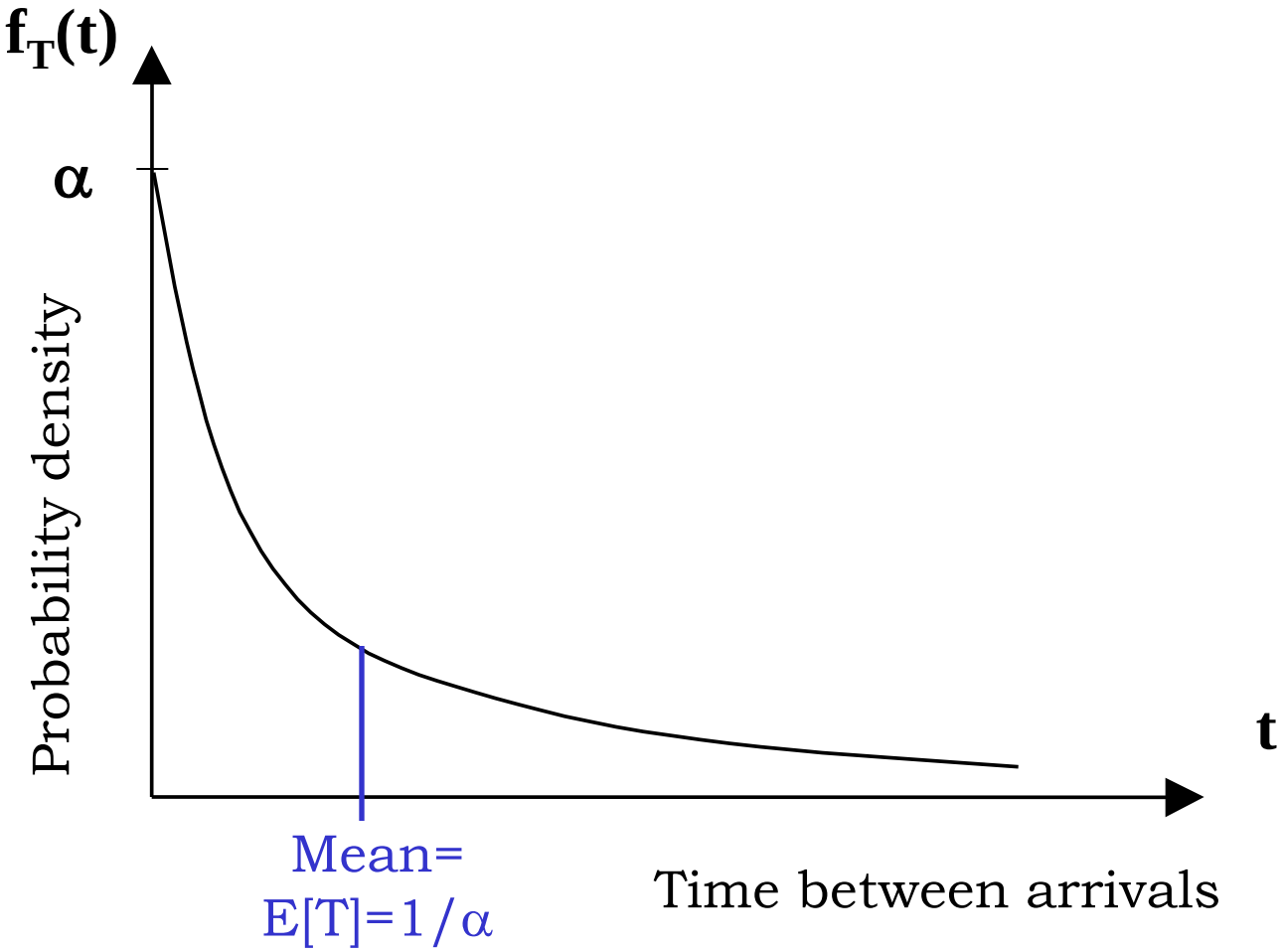
$$f_T(t) = \begin{cases} \alpha e^{-\alpha t} & \text{when } t \geq 0 \\ 0 & \text{when } t < 0 \end{cases}$$

$\Rightarrow$ The Cumulative Distribution Function is: $F_T(t) = 1 - e^{-\alpha t}$

$\Rightarrow$ The mean = **E[T] = 1/α**

$\Rightarrow$ The Variance = **Var[T] = 1/ α^2**

The Exponential Distribution



Properties of the Exp-distribution (IV)

- Relationship to the Poisson distribution and the Poisson Process

Let X(t) be the number of events occurring in the interval [0,t]. If the time between consecutive events is T and $T \in \exp(\alpha)$

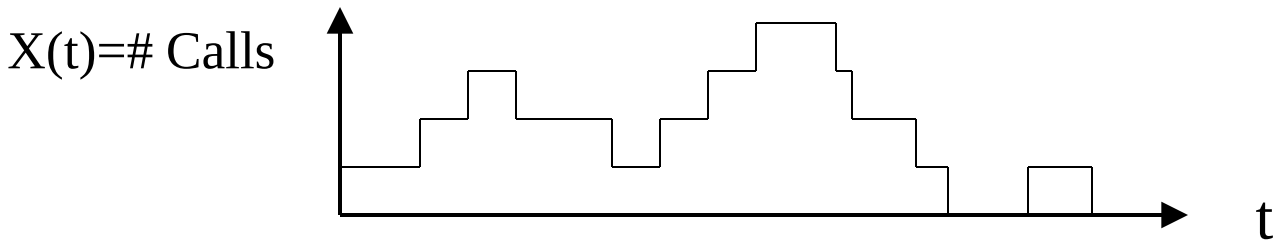
$$\Rightarrow \quad P(X(t) = n) = \frac{(\alpha t)^n e^{-nt}}{n!} \qquad \text{for } n = 0, 1, \ldots$$

$\Leftrightarrow X(t) \in Po(\alpha t) \Leftrightarrow \{X(t), t \geq 0\}$ constitutes a Poisson Process

Stochastic Processes in Continuous Time

- ***Definition:*** A stochastic process in continuous time is a family {X(t)} of stochastic variables defined over a continuous set of t-values.

- *Example: The number of phone calls connected through a switch board*



- ***Definition:*** A stochastic process {X(t)} is said to have independent increments if for all disjoint intervals (t_i, t_i+h_i) the differences $X_i(t_i+h_i)-X_i(t_i)$ are mutually independent.

The Poisson Process

- The standard assumption in many queuing models is that the arrival process is Poisson

Two equivalent definitions of the Poisson Process

1. The times between arrivals are independent, identically distributed and exponential
2. X(t) is a Poisson process with arrival rate λ.

Terminology and Notation

- The state of the system = the number of customers in the system
- Queue length = (The state of the system) – (number of customers being served)

N(t) = Number of customers/jobs in the system at time t

$\mathbf{P_n(t)}$ = The probability that at time t, there are n customers/jobs in the system.

λ_n = Average arrival intensity (= # arrivals per time unit) at n customers/jobs in the system

μ_n = Average service intensity for the system when there are n customers/jobs in it. (Note, the total service intensity for all ***occupied*** servers)

ρ = The utilization factor for the service facility. (= The expected fraction of the time that the service facility is being used)

Example – Service Utilization Factor

- Consider an M/M/1 queue with arrival rate = λ and service intensity = μ
- λ = Expected capacity demand per time unit
- μ = Expected capacity per time unit

$$\Rightarrow \quad \rho = \frac{\text{Capacity Demand}}{\text{Available Capacity}} = \frac{\lambda}{\mu}$$

- Similarly if there are c servers in parallel, i.e., an M/M/c system but the expected capacity per time unit is then c*μ

$$\Rightarrow \quad \rho = \frac{\text{Capacity Demand}}{\text{Available Capacity}} = \frac{\lambda}{c * \mu}$$

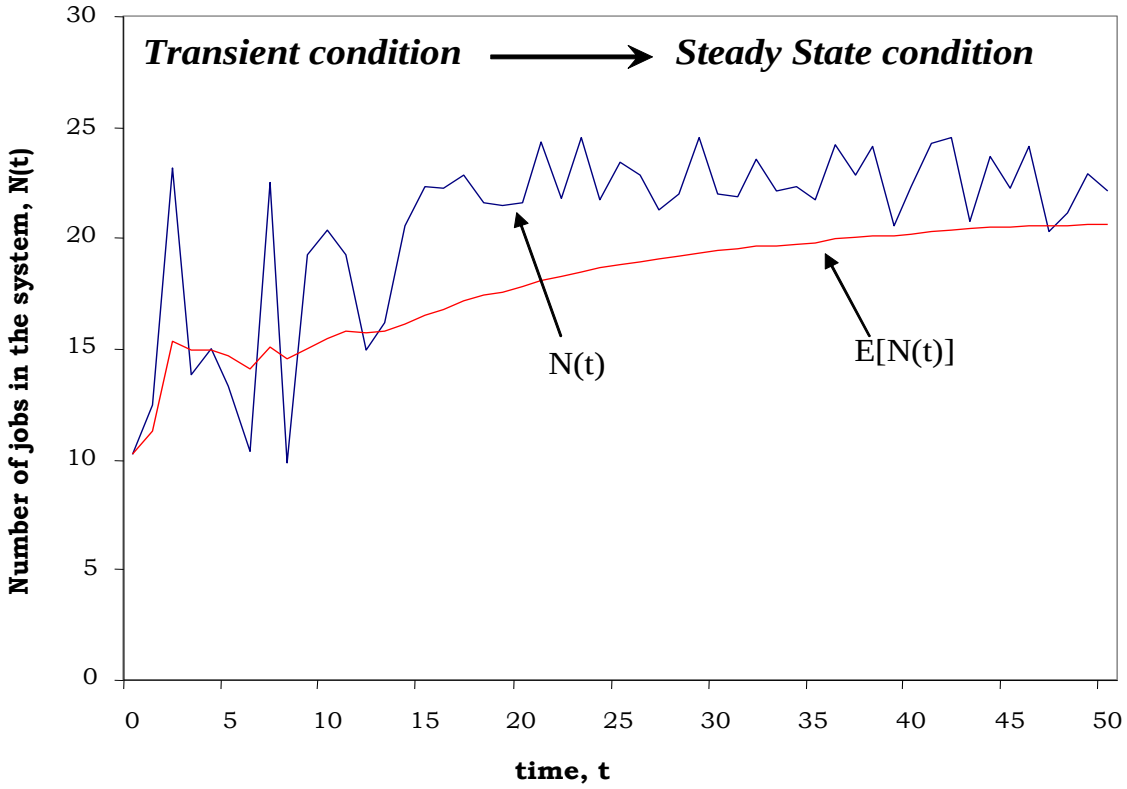
Queuing Theory Focus on Steady State

- Steady State condition
  - Enough time has passed for the system state to be independent of the initial state as well as the elapsed time
  - The probability distribution of the state of the system remains the same over time (is stationary).
- Transient condition
  - Prevalent when a queuing system has recently begun operations
  - The state of the system is greatly affected by the initial state and by the time elapsed since operations started
  - The probability distribution of the state of the system changes with time

With few exceptions Queuing Theory has focused on analyzing steady state behavior

Transient and Steady State Conditions

- Illustration of transient and steady-state conditions
  - N(t) = number of customers in the system at time t,
  - E[N(t)] = represents the expected number of customers in the system.



Notation For Steady State Analysis

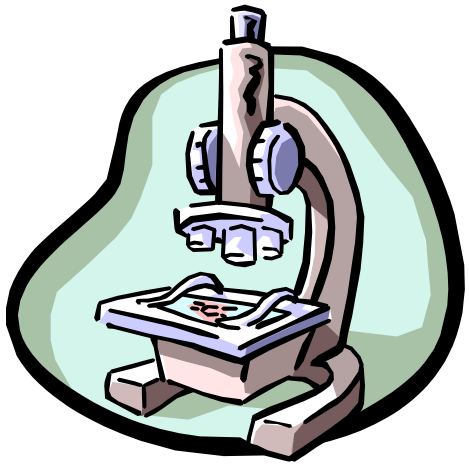
$\mathbf{P_n}$ = The probability that there are exactly n customers/jobs in the system (in steady state, i.e., when $t \to \infty$)

L = Expected number of customers in the system (in steady state)

$\mathbf{L_q}$ = Expected number of customers in the queue (in steady state)

W = Expected time a job spends in the system

$\mathbf{W_q}$ = Expected time a job spends in the queue



Birth-and-Death Processes

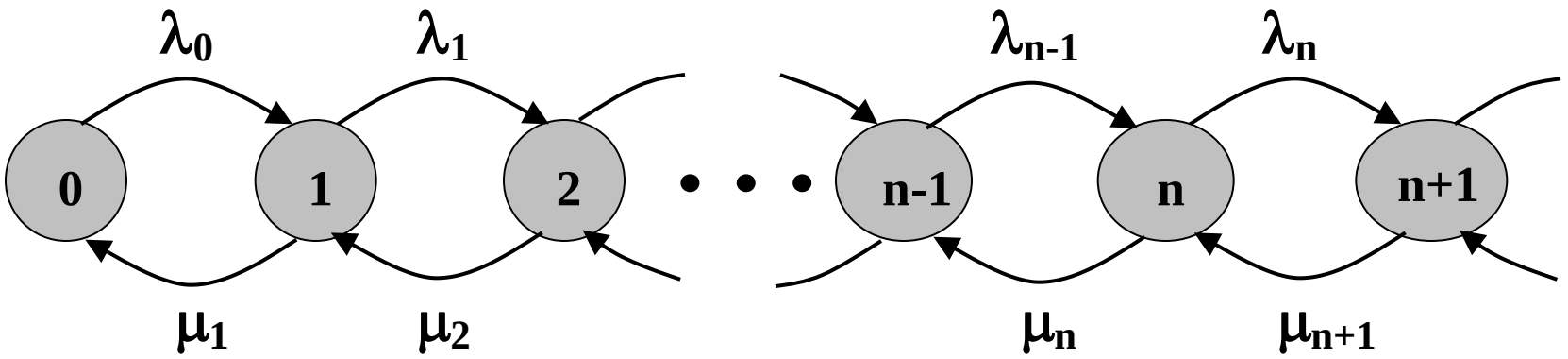
- The foundation of many of the most commonly used queuing models
  - ✓ Birth – equivalent to the arrival of a customer or job
  - ✓ Death – equivalent to the departure of a served customer or job

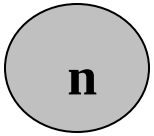
Assumptions

1. Given N(t)=n,
   - The time until the next birth (T_B) is exponentially distributed with parameter λ_n (Customers arrive according to a Po-process)
   - The remaining service time (T_D) is exponentially distributed with parameter μ_n
2. T_B & T_D are mutually independent stochastic variables and state transitions occur through exactly one *Birth* ($n \rightarrow n+1$) or one *Death* ($n \rightarrow n-1$)

A Birth-and-Death Process Rate Diagram

- Excellent tool for describing the mechanics of a Birth-and-Death process





= *State n, i.e., the case of n customers/jobs in the system*

The M/M/1 - model

Assumptions - the Basic Queuing Process

- ✓ Infinite Calling Populations
  - Independence between arrivals
- ✓ The arrival process is Poisson with an expected arrival rate λ
  - Independent of the number of customers currently in the system
- ✓ The queue configuration is a single queue with possibly infinite length
  - No reneging or balking
- ✓ The queue discipline is FIFO
- ✓ The service mechanism consists of a single server with exponentially distributed service times
  - μ = expected service rate when the server is busy

Example – ER at County Hospital

- Situation
  - Patients arrive according to a Poisson process with intensity λ ($\Leftrightarrow$ the time between arrivals is exp(λ) distributed.
  - The service time (the doctor's examination and treatment time of a patient) follows an exponential distribution with mean $1/\mu$ (=exp(μ) distributed)

  $\Rightarrow$ *The ER can be modeled as an M/M/c system where c=the number of doctors*

- Data gathering

  $\Rightarrow$ λ = 2 patients per hour

  $\Rightarrow$ μ = 3 patients per hour

- Questions
  - Should the capacity be increased from 1 to 2 doctors?
  - How are the characteristics of the system (ρ, W_q, W, L_q and L) affected by an increase in service capacity?

Summary of Results – County Hospital

- Interpretation
  - To be in the queue = to be in the waiting room
  - To be in the system = to be in the ER (waiting or under treatment)

Characteristic	One doctor (c=1)	Two Doctors (c=2)
ρ	2/3	1/3
P_0	1/3	1/2
$(1-P_0)$	2/3	1/2
P_1	2/9	1/3
L_q	4/3 patients	1/12 patients
L	2 patients	3/4 patients
W_q	2/3 h = 40 minutes	1/24 h = 2.5 minutes
W	1 h	3/8 h = 22.5 minutes

- Is it warranted to hire a second doctor ?

Queuing Modeling and System Design (I)

- Design of queuing systems usually involve some kind of capacity decision
  - The number of service stations
  - The number of servers per station
  - The service time for individual servers

  $\Rightarrow$ ***The corresponding decision variables are λ, c and μ***

- Examples:
  - The number of doctors in a hospital,
  - The number of exits and cashiers in a supermarket,
  - The choice of machine type at a new investment decision,
  - The localization of toilets in a new building, etc…

Queuing Modeling and System Design (II)

- Two fundamental questions when designing (queuing) systems
  - ***Which service level should we aim for?***
  - ***How much capacity should we acquire?***
- The cost of increased capacity must be balanced against the cost reduction due to shorter waiting time
  - ⇒ Specify a waiting cost or a shortage cost accruing when customers have to wait for service or…
  - ⇒ … Specify an acceptable service level and minimize the capacity under this condition
- The shortage or waiting cost rate is situation dependent and often difficult to quantify
  - Should reflect the monetary impact a delay has on the organization where the queuing system resides

Different Shortage Cost Situations

1. External customers arrive to the system
   - **Profit organizations**

     ⇒ The shortage cost is primarily related to lost revenues – "Bad Will"
   - **Non-profit organizations**

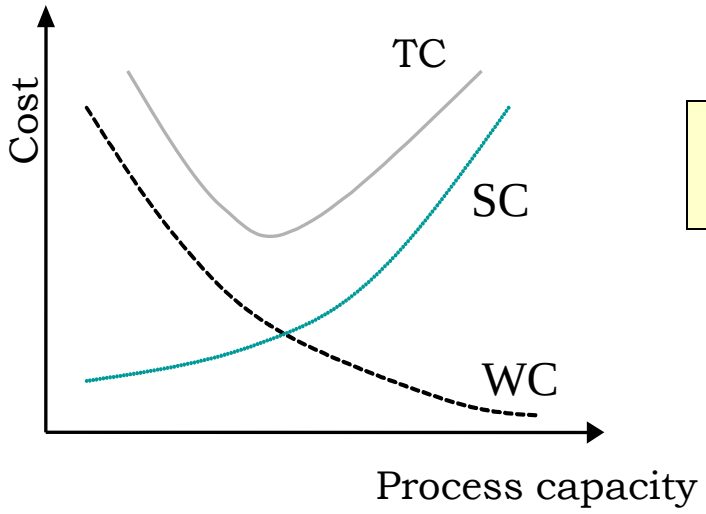
     ⇒ The shortage cost is related to a societal cost
2. Internal customers arrive to the system

   ⇒ The shortage cost is related to productivity loss and associated profit loss

- Usually it is easier to estimate the shortage costs in situation 2. than in situation 1.

Analyzing Design-Cost Tradeoffs

- Given a specified shortage or waiting cost function the analysis is straightforward
- Define
  - WC = Expected Waiting Cost (shortage cost) per time unit
  - SC = Expected Service Cost (capacity cost) per time unit
  - TC = Expected Total system cost per time unit
- The objective is to minimize the total expected system cost



Min TC = WC + SC

Analyzing Linear Waiting Costs

- Expected Waiting Costs as a function of the number of customers in the system
  - C_w = Waiting cost per customer and time unit
  - C_wN = Waiting cost per time unit when N customers in the system

$$WC = C_w \sum_{n=0}^{\infty} nP_n = C_w L$$

- Expected Waiting Costs as a function of the number of customers in the queue

$$WC = C_w L_q$$

Analyzing Service Costs

- The expected service costs per time unit, SC, depend on the number of servers and their speed

- Definitions
  - c = Number of servers
  - μ = Average server intensity (average time to serve one customer)
  - $C_S(\mu)$ = Expected cost per server and time unit as a function of μ

$$SC = c * C_S(\mu)$$

A Decision Model for System Design

Determining μ and c

- Both the number of servers and their speed can be varied
  - Usually only a few alternatives are available
- Definitions
  - A = The set of available μ - options

$$\min_{\mu \in A, c=0,1,\ldots} TC = c \cdot C_s(\mu) + WC$$

- Optimization
  - Enumerate all interesting combinations of μ and c, compute TC and choose the cheapest alternative

From a structural point of view, a few fast servers are usually better than several slow ones with the same maximum capacity

Example – "Computer Procurement"

- A university is about to lease a super computer
- There are two alternatives available
  - The M computer which is more expensive to lease but also faster
  - The C computer which is cheaper but slower
- Processing times and times between job arrivals are exponential $\Rightarrow$ M/M/1 model
  - $\lambda = 20$ jobs per day
  - $\mu_M = 30$ jobs per day
  - $\mu_C = 25$ jobs per day
- The leasing and waiting costs:
  - Leasing price: $C_M = \$500$ per day, $C_C = \$350$ per day
  - The waiting cost per job and time unit is estimated to \$50 per job and day
- Question:
  - Which computer should the university choose in order to minimize the expected costs?

Example of a M/M/1 Queue

- Assume a small branch office of a local bank with only one teller.
- Empirical data gathering indicates that inter-arrival and service times are exponentially distributed.
  - The average arrival rate = λ = 5 customers per hour
  - The average service rate = μ = 6 customers per hour
- Using our knowledge of queuing theory we obtain
  - ρ = the server utilization = $5/6 \approx 0.83$
  - L_q = the average number of people waiting in line
  - W_q = the average time spent waiting in line

  $L_q = 0.83^2/(1-0.83) \approx 4.2$ $\quad$ $W_q = L_q/\lambda \approx 4.2/5 \approx 0.83$
- How do we go about simulating this system?
  - How do the simulation results match the analytical ones?