Introduction to Queuing and Simulation

Supervisor: Dr. Shaista Rais

DCS-UOK

Overview (I)

- What is queuing/ queuing theory?
  - Why is it an important tool?
  - Examples of different queuing systems
- Components of a queuing system
- The exponential distribution & queuing
- Stochastic processes
  - Some definitions
  - The Poisson process
- Terminology and notation
- Little's formula
- Birth and Death Processes

Overview (II)

- Important queuing models with FIFO discipline
  - The M/M/1 model
  - The M/M/c model
  - The M/M/c/K model (limited queuing capacity)
  - The M/M/c/∞/N model (limited calling population)
- Priority-discipline queuing models
- Application of Queuing Theory to system design and decision making

Overview (III)

- Simulation – What is that?
  - Why is it an important tool?
- Building a simulation model
  - Discrete event simulation
- Structure of a BPD simulation project
- Model verification and validation
- Example – Simulation of a M/M/1 Queue

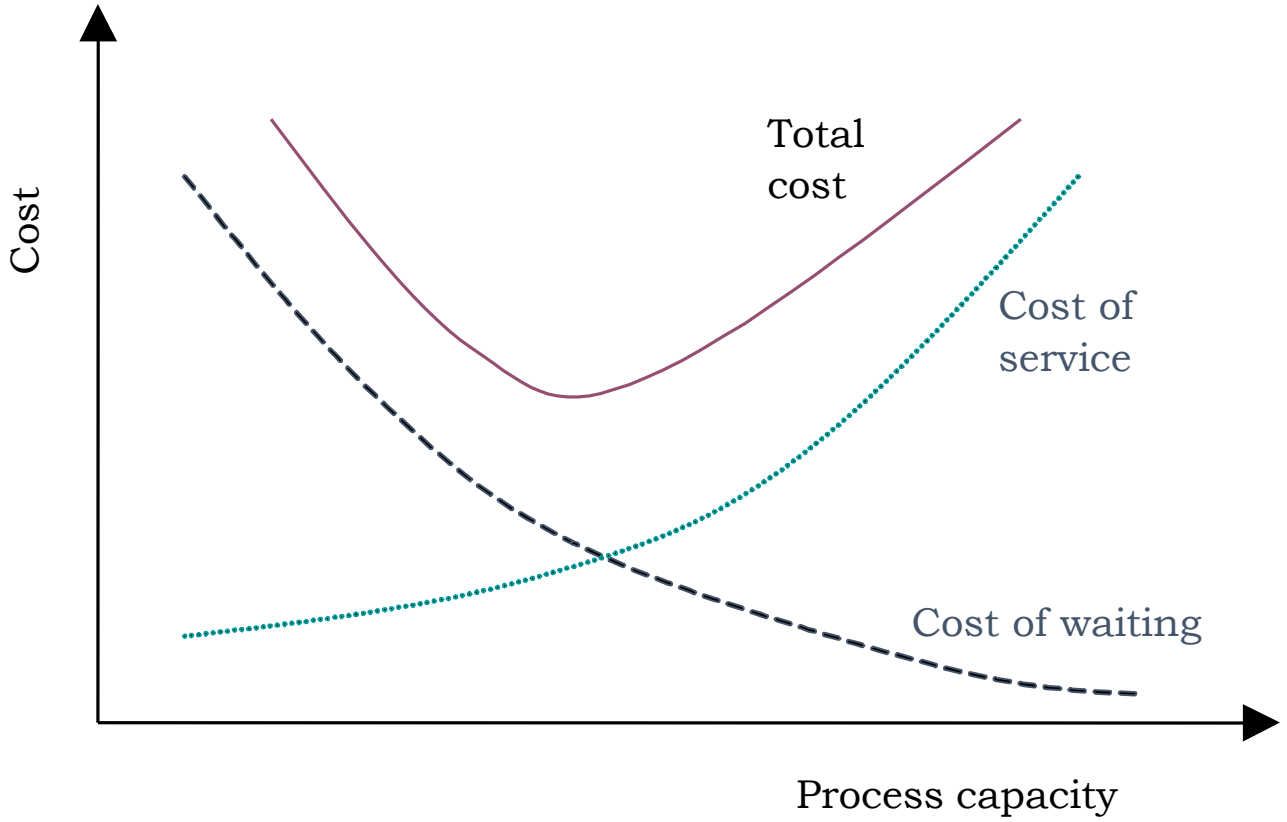
What is Queuing Theory?

- 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

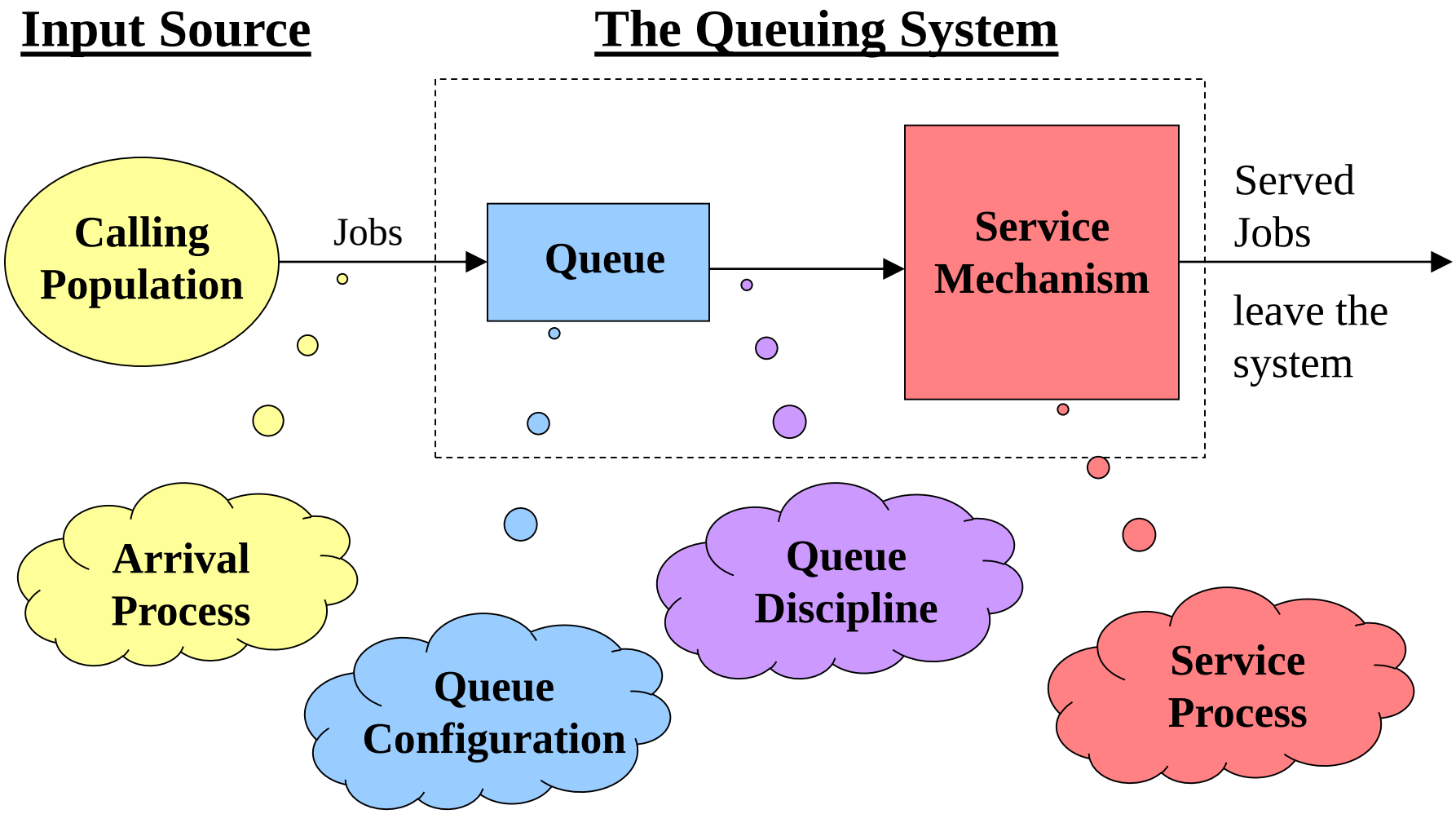
A Cost/Capacity Tradeoff Model



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



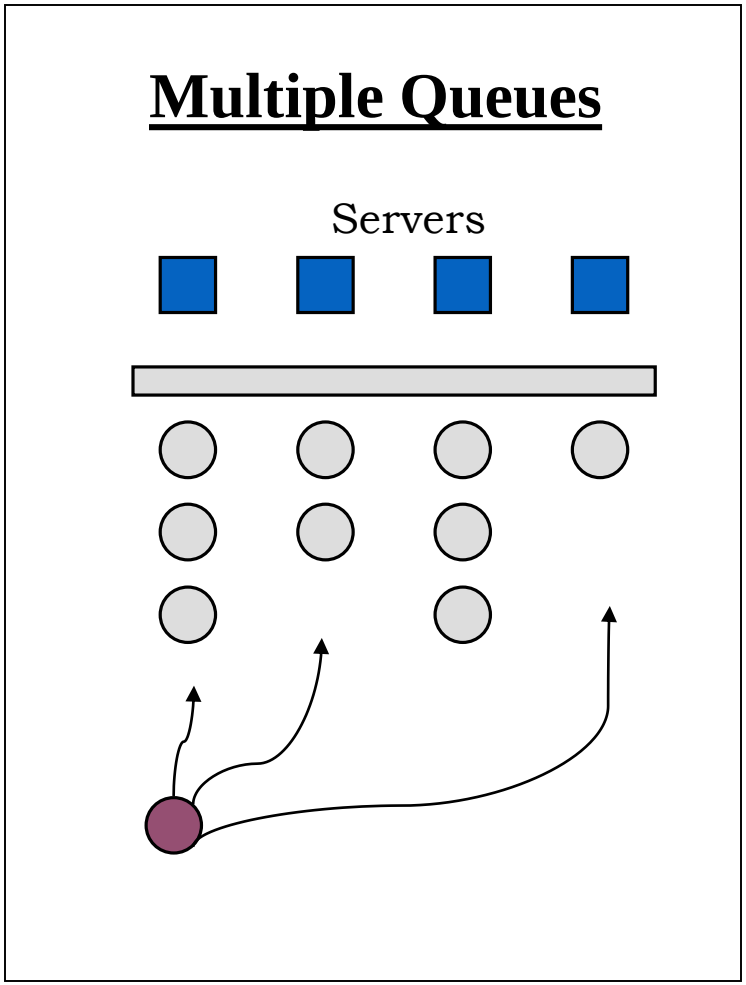
Components of a Basic Queuing Process (II)

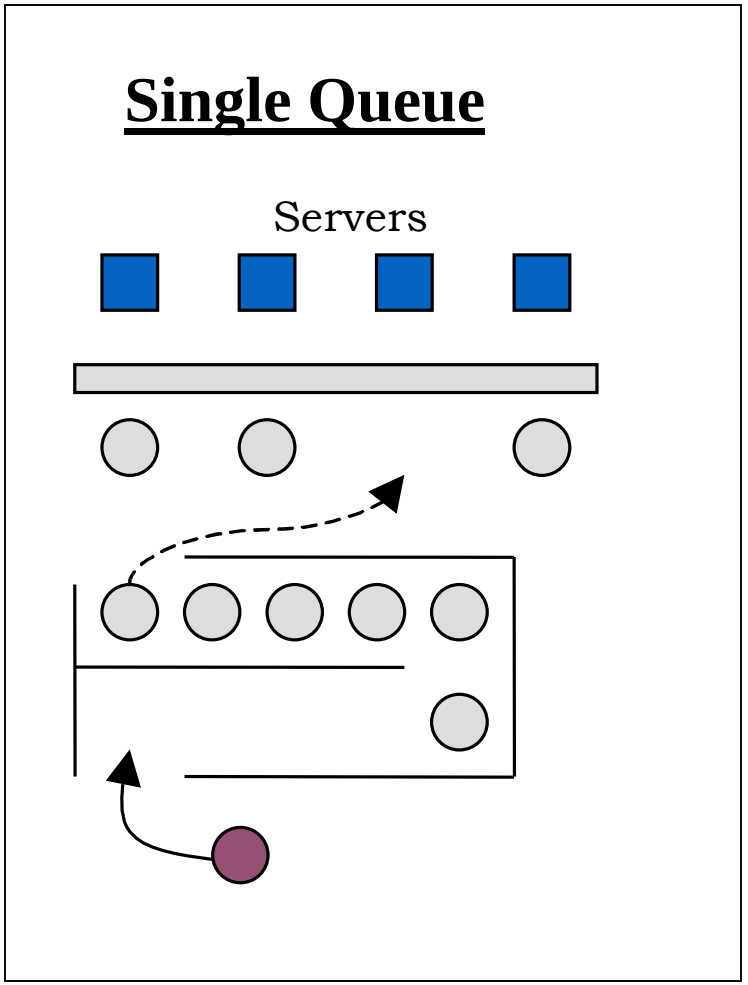
- The calling population
  - The population from which customers/jobs originate
  - The size can be finite or infinite (the latter is most common)
  - Can be homogeneous (only one type of customers/ jobs) or heterogeneous (several different kinds of customers/jobs)
- The Arrival Process
  - Determines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

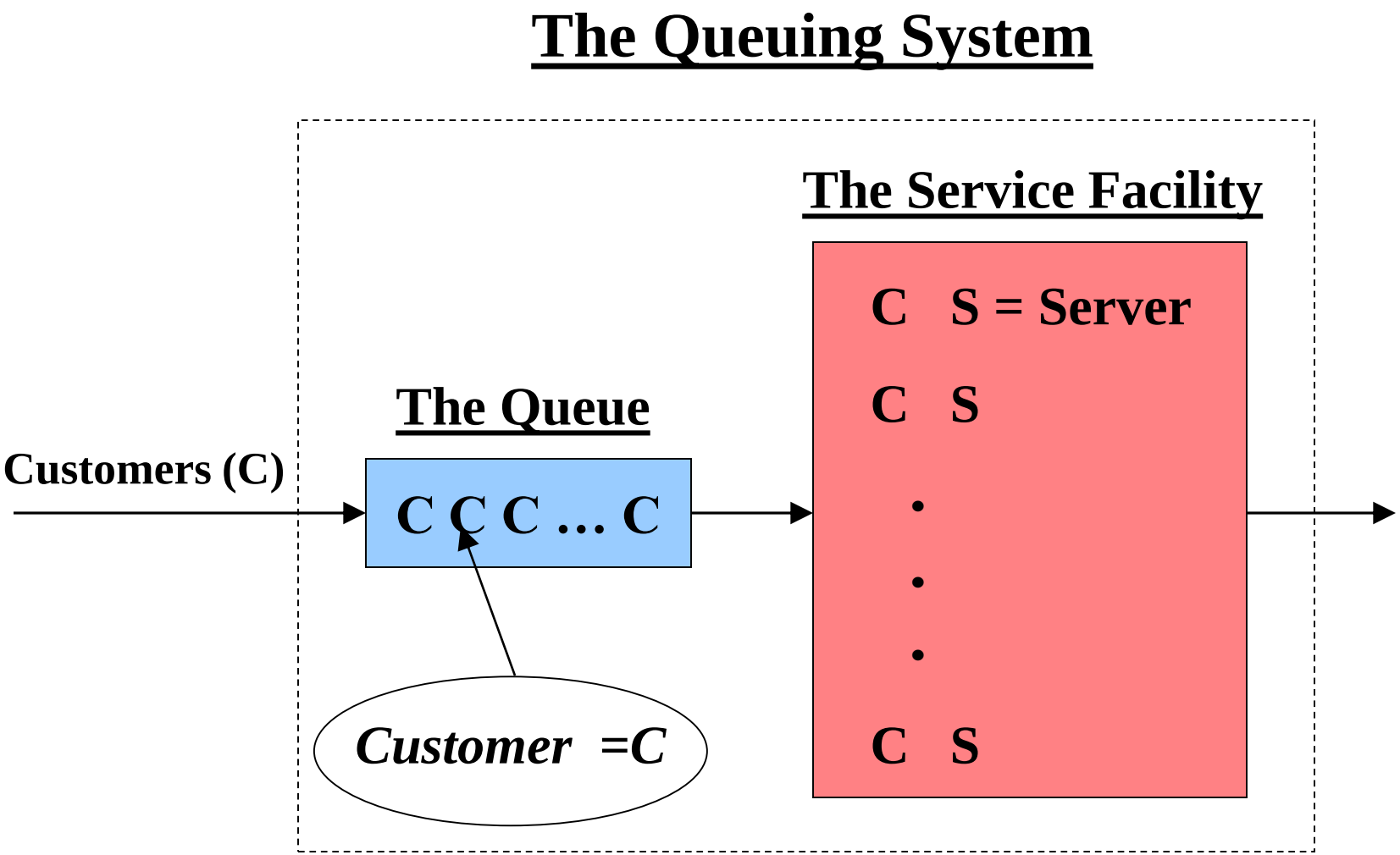
Components of a Basic Queuing Process (IV)

- 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

Customer Behaviour in a Queue

Various customers while in queue behave differently. Their behavior pattern can fall in one of the following categories.

1. Balking

A customer may not like to wait in a que due to lack of space or otherwise. They do not join the queue at their correct position and attempt to jump the queue and reach the service centre by passing other ahead of them. This is known as balking.

2. Reneging

A customer may leave the queue due to impatience. This is called reneging.

3. Collusion

Some customers may collaborate and only one of them may join the queue. As at the cinema ticket window one person may join the queue and purchase tickets for his friends.

4. Jockeying

If there are more than one queues then one customer may leave one queue and join the other. This occurs generally in the super market or shopping malls.

The Exponential Distribution and Queuing

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

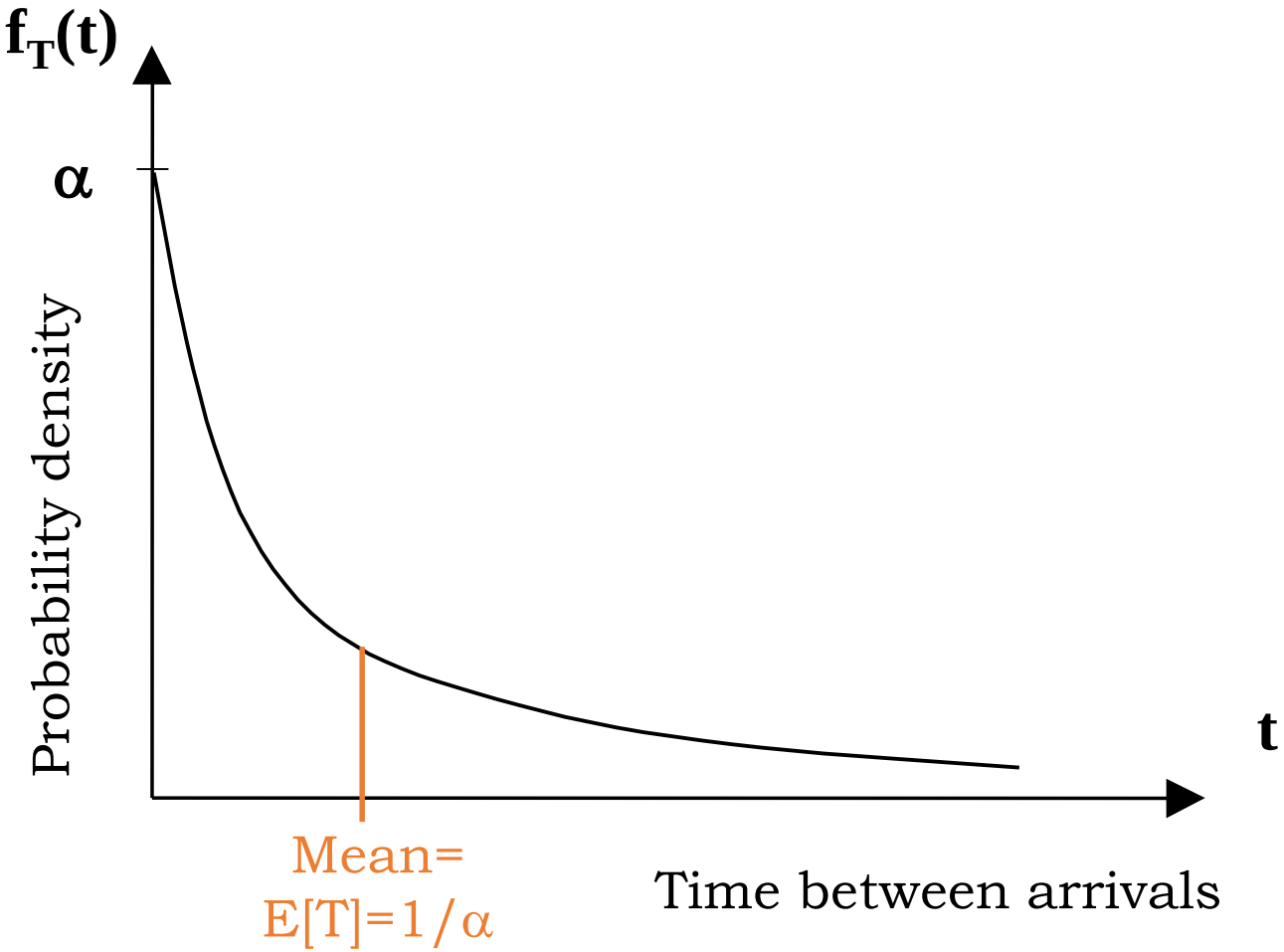
$$f_T(t) = \begin{cases} \alpha e^{-\alpha t} & \text{when } t \geq 0 \\ 0 & \text{when } t < 0 \end{cases}$$

$\Rightarrow$ 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 (I)

- **Property 1:** $f_T(t)$ is strictly decreasing in t

  $\Leftrightarrow$ $P(0 \le T \le \Delta t) > P(t \le T \le t+\Delta t)$ for all t, $\Delta t \ge 0$

- **Implications**
  - Many realizations of T (i.e.,values of t) will be small; between zero and the mean
  - Not suitable for describing the service time of standardized operations when all times should be centered around the mean
    - Ex. Machine processing time in manufacturing
  - Often reasonable in service situations when different customers require different types of service
  - Often a reasonable description of the time between customer arrivals

Properties of the Exp-distribution (II)

- **Property 2:** Lack of memory

  $\Leftrightarrow$ $P(T>t+\Delta t \mid T>t) = P(T>\Delta t)$ for all $t, \Delta t \geq 0$

- **Implications**
  - It does not matter when the last customer arrived, (or how long service time the last job required) the distribution of the time until the next one arrives (or the distribution of the next service time) is always the same.
  - Usually a fair assumption for interarrival times
  - For service times, this can be more questionable. However, it is definitely reasonable if different customers/jobs require different service

Properties of the Exp-distribution (III)

- **Property 3:** The minimum of independent exponentially distributed random variables is exponentially distributed

  Assume that $\{T_1, T_2, \ldots, T_n\}$ represent n independent and exponentially distributed stochastic variables with parameters $\{\alpha_1, \alpha_2, \ldots, \alpha_n\}$.

  Let $U=\min\{T_1, T_2, \ldots, T_n\} \Rightarrow \quad U \in \exp(\sum_{i=1}^{n} \alpha_i)$

- **Implications**
  - Arrivals with exponentially distributed interarrival times from n different input sources with arrival intensities $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ can be treated as a homogeneous process with exponentially distributed interarrival times of intensity $\lambda = \lambda_1 + \lambda_2 + \ldots + \lambda_n$.
  - Service facilities with n occupied servers in parallel and service intensities $\{\mu_1, \mu_2, \ldots, \mu_n\}$can be treated as one server with service intensity $\mu = \mu_1 + \mu_2 + \ldots + \mu_n$

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 $T \in \exp(\alpha)$

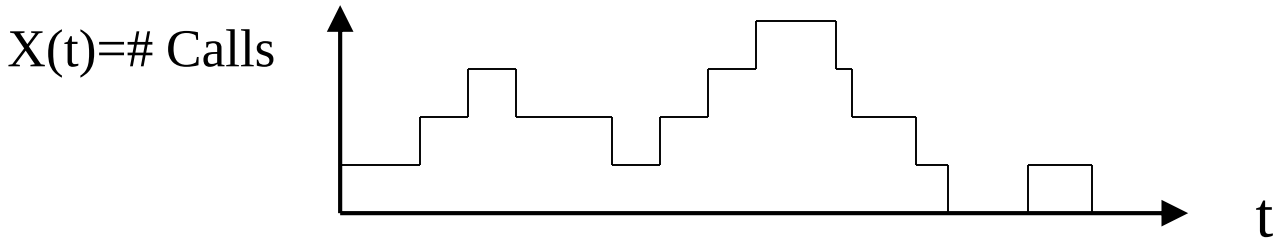
$$\Rightarrow \quad P(X(t) = n) = \frac{(\alpha t)^n e^{-nt}}{n!} \qquad \text{for } n = 0, 1, \ldots$$

$\Leftrightarrow X(t) \in Po(\alpha t) \Leftrightarrow \{X(t), t \geq 0\}$ 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 iff.
   a) X(t) have independent increments
   b) For a small time interval h it holds that
      - P(exactly 1 event occurs in the interval [t, t+h]) = $\lambda h + o(h)$
      - P(more than 1 event occurs in the interval [t, t+h]) = $o(h)$

Properties of the Poisson Process

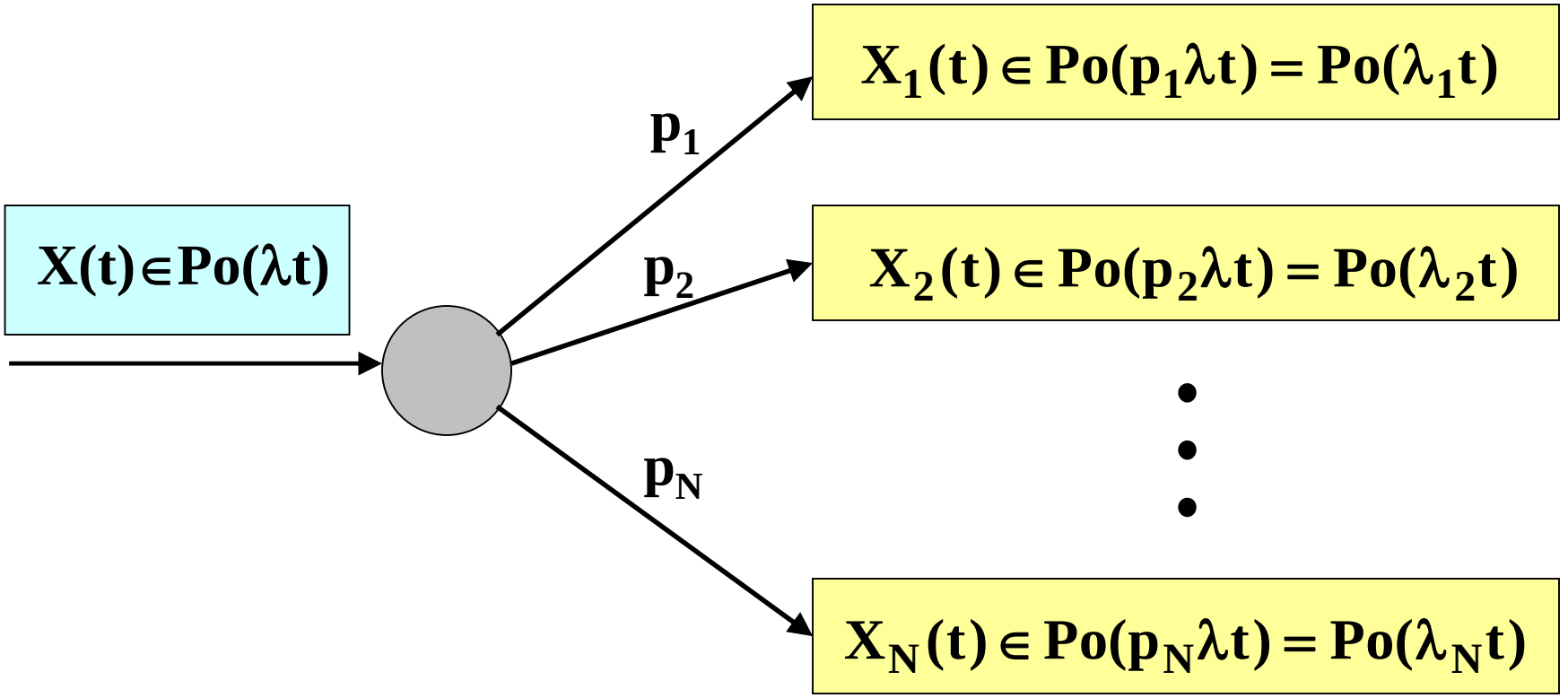
- Poisson processes can be aggregated or disaggregated and the resulting processes are also Poisson processes

  a) Aggregation of N Poisson processes with intensities $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ renders a new Poisson process with intensity $\lambda = \lambda_1 + \lambda_2 + \ldots + \lambda_n$.

  b) Disaggregating a Poisson process $X(t) \in Po(\lambda t)$ into N sub-processes $\{X_1(t), X_2(t), , \ldots, X_3(t)\}$ (for example N customer types) where $X_i(t) \in Po(\lambda_i t)$ can be done if

     - For every arrival the probability of belonging to sub-process i = p_i
     - $p_1 + p_2 + \ldots + p_N = 1$, and $\lambda_i = p_i \lambda$

Illustration – Disaggregating a Poisson Process



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

$P_n(t)$ = 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 $c*\mu$

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

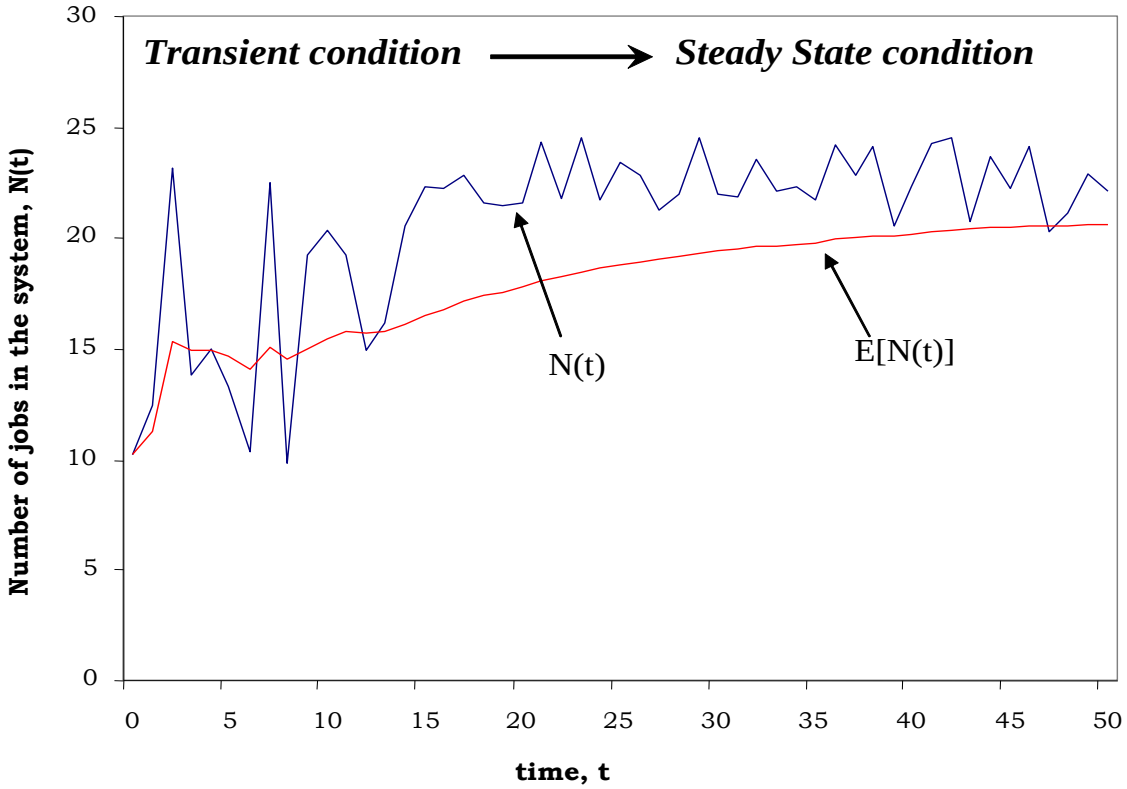
Queuing Theory Focus on Steady State

- 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

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)

L_q = Expected number of customers in the queue (in steady state)

W = Expected time a job spends in the system

W_q = Expected time a job spends in the queue

Little's Formula Revisited

- Assume that $\lambda_n = \lambda$ and $\mu_n = \mu$ for all n

$$L = \lambda W \qquad L_q = \lambda W_q$$

- Assume that λ_n is dependent on n

Let $\bar{\lambda} = \sum_{n=0}^{\infty} P_n \lambda_n$

$$L = \bar{\lambda} W \qquad L_q = \bar{\lambda} W_q$$

Birth-and-Death Processes

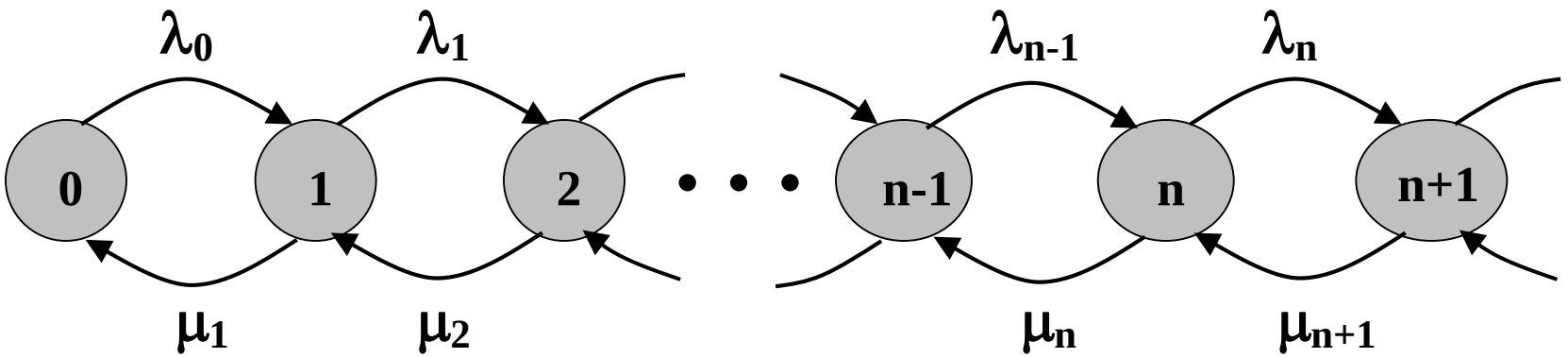
- The foundation of many of the most commonly used queuing models
  - ✓ Birth – equivalent to the arrival of a customer or job
  - ✓ Death – equivalent to the departure of a served customer or job

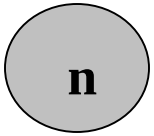
Assumptions

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*

Steady State Analysis of B-D Processes (I)

- In steady state the following balance equation must hold for every state (proved via differential equations)

The Rate In = Rate Out Principle:

Mean entrance rate = Mean departure rate

- In addition the probability of being in one of the states must equal 1

$$\sum_{i=0}^{\infty} P_i = 1$$

Steady State Analysis of B-D Processes (II)

State	Balance Equation		
0	$\mu_1 P_1 = \lambda_0 P_0$	⇒	$P_1 = \frac{\lambda_0}{\mu_1} P_0$
1	$\lambda_0 P_0 + \mu_2 P_2 = \lambda_1 P_1 + \mu_1 P_1$	⇒	$P_2 = \frac{\lambda_1}{\mu_2} P_1$
⋮	⋮		
n	$\lambda_{n-1} P_{n-1} + \mu_{n+1} P_{n+1} = (\lambda_n + \mu_n) P_n$	⇒	$P_n = \frac{\lambda_{n-1}}{\mu_n} P_{n-1}$
⋮	⋮		

Normalization: $\sum_{i=0}^{\infty} P_i = P_0 \left(1 + \frac{\lambda_0}{\mu_1} + \frac{\lambda_0 \lambda_1}{\mu_1 \mu_2} + \frac{\lambda_0 \lambda_1 \lambda_2}{\mu_1 \mu_2 \mu_3} + \cdots \right) = 1$

C_0 C_2

Steady State Analysis of B-D Processes (III)

- Steady State Probabilities

$$P_0 = 1 \Big/ \sum_{i=0}^{\infty} C_i \qquad P_n = C_n P_0$$

- Expected Number of Jobs in the System and in the Queue
  - Assuming c parallel servers

$$L = \sum_{i=0}^{\infty} i \cdot P_i \qquad L_q = \sum_{i=c}^{\infty} (i-c) \cdot P_i$$

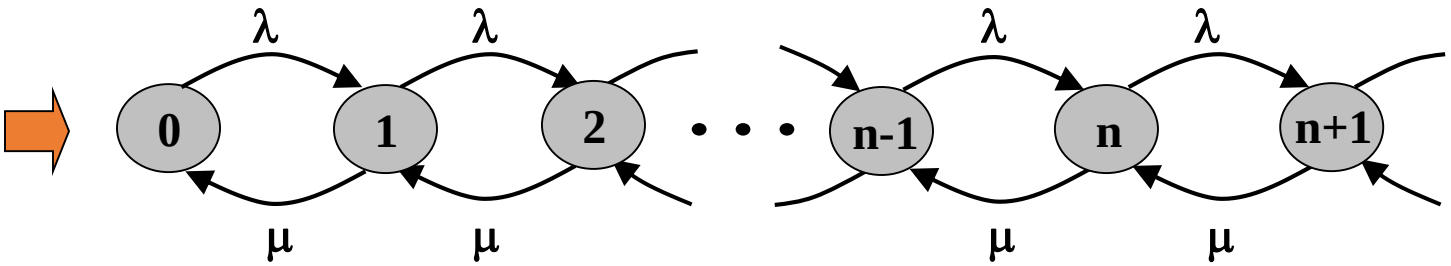
The M/M/1 - model

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

The M/M/1 Model

- $\lambda_n = \lambda$ and $\mu_n = \mu$ for all values of n=0, 1, 2, …



❖ Steady State condition: $\rho = (\lambda/\mu) < 1$

$P_0 = 1-\rho$

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

$P(n \geq k) = \rho^k$

$L = \rho/(1-\rho)$

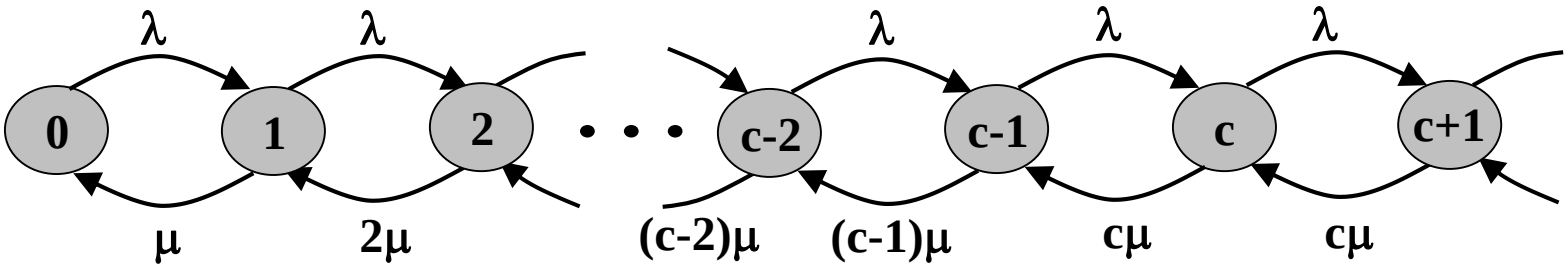
$W = L/\lambda = 1/(\mu-\lambda)$

$L_q = \rho^2/(1-\rho) = L-\rho$

$W_q = L_q/\lambda = \lambda/(\mu(\mu-\lambda))$

The M/M/c Model (I)

- Generalization of the M/M/1 model
  - Allows for c identical servers working independently from each other



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

$$P_n = \begin{cases} \dfrac{(\lambda/\mu)^n}{n!} P_0 & \text{for} \quad n = 1,2,\ldots,c \\ \dfrac{(\lambda/\mu)^n}{c!\,c^{n-c}} P_0 & \text{for} \quad n = c+1, c+2,\ldots \end{cases}$$

Steady State Condition:

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

The M/M/c Model (II)

- A Condition for existence of a steady state solution is that $\rho = \lambda/(c\mu) < 1$

$$L_q = \sum_{n=c}^{\infty}(n-c)P_n = \ldots = \frac{(\lambda/\mu)^c \rho}{c!(1-\rho)^2} P_0$$

Little's Formula $\Rightarrow$ $W_q = L_q/\lambda$

$$W = W_q + (1/\mu)$$

Little's Formula $\Rightarrow$ $L = \lambda W = \lambda(W_q + 1/\mu) = L_q + \lambda/\mu$

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

  $\Rightarrow$ $\lambda = 2$ patients per hour

  $\Rightarrow$ $\mu = 3$ 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 ?

The M/M/c/K – Model (I)

- An M/M/c model with a maximum of K customers/jobs allowed in the system
  - If the system is full when a job arrives it is denied entrance to the system and the queue.
- Interpretations
  - A waiting room with limited capacity (for example, the ER at County Hospital), a telephone queue or switchboard of restricted size
  - Customers that arrive when there is more than K clients/jobs in the system choose another alternative because the queue is too long (Balking)

The M/M/c/K – Model (II)

- Still a Birth-and-Death process but with a state dependent arrival intensity

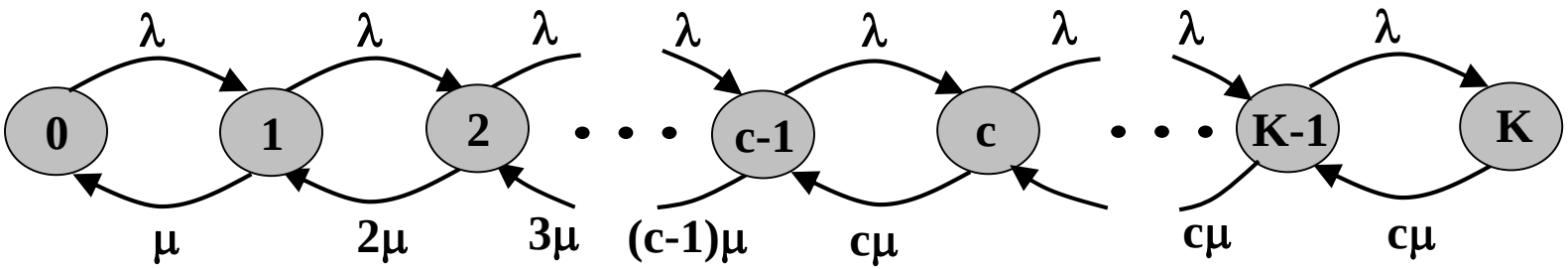
$$\lambda_n = \begin{cases} \lambda & \text{for } n = 0,1,2,\ldots,K-1 \\ 0 & \text{for } n \geq K \end{cases}$$

Observation

The M/M/c/K model always has a steady state solution since the queue can never "explode"

The M/M/c/K – Model (III)

- The state diagram has exactly K states provided that c<K



- The general expressions for the steady state probabilities, waiting times, queue lengths etc. are obtained through the balance equations as before (Rate In = Rate Out; for every state)

Results for the M/M/1/K – Model

- For $\rho = (\lambda/\mu) \neq 1$

$$P_0 = \frac{1-\rho}{1-\rho^{K+1}}$$

$$P_n = \rho^n P_0 = \frac{1-\rho}{1-\rho^{K+1}} \cdot \rho^n$$

$$L = \frac{\rho}{1-\rho} - \frac{(K+1)\rho^{K+1}}{1-\rho^{K+1}}$$

$$L_q = L - (1 - P_0)$$

$$W = L / \bar{\lambda}$$

$$W_q = L_q / \bar{\lambda}$$

Where $\bar{\lambda} = \sum_{n=0}^{\infty} \lambda_n P_n$

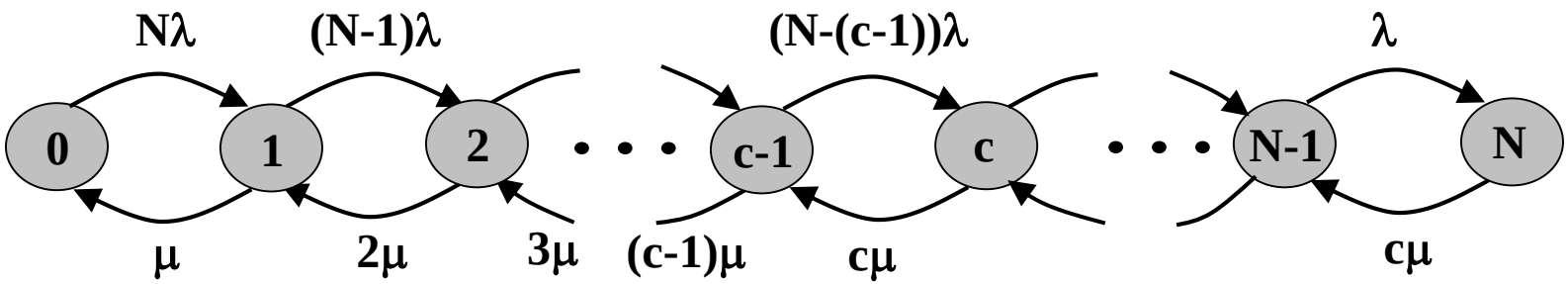
The M/M/c/∞/N – Model (I)

- An M/M/c model with limited calling population, i.e., N clients
- A common application: Machine maintenance
  - c service technicians is responsible for keeping N service stations (machines) running, that is, to repair them as soon as they break
  - Customer/job arrivals = machine breakdowns
  - Note, the maximum number of clients in the system = N
- Assume that (N-n) machines are operating and the time until breakdown for each machine i, T_i, is exponentially distributed ($T_i \in \exp(\lambda)$). If U = the time until the next breakdown

  $\Rightarrow$ $U = \text{Min}\{T_1, T_2, \ldots, T_{N-n}\} \Rightarrow U \in \exp((N-n)\lambda)$.

The M/M/c/∞/N – Model (II)

- The State Diagram (c service technicians and N machines)
  - λ = Arrival intensity per operating machine
  - μ = The service intensity for a service technician



- General expressions for this queuing model can be obtained from the balance equations as before

Priority-Discipline Queuing Models

- For situations where different customers have different priorities
  - For example, ER operations, VIP customers at nightclubs…
- Assuming a situation with N priority classes (where class 1 has the highest priority) there are two fundamental priority principles to consider.
  1. **Non-Preemptive priorities**
     - A customer being served cannot be ejected back into the queue to leave place for a customer with higher priority
  2. **Preemptive priorities**
     - A customer of lower priority that is being served will be thrown back into the queue to leave room for a higher priority customer
- Assuming that all customers experience independent exp(μ) service times and arrive according to Poisson processes $\Rightarrow$ both models can be analyzed as special case M/M/c models

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

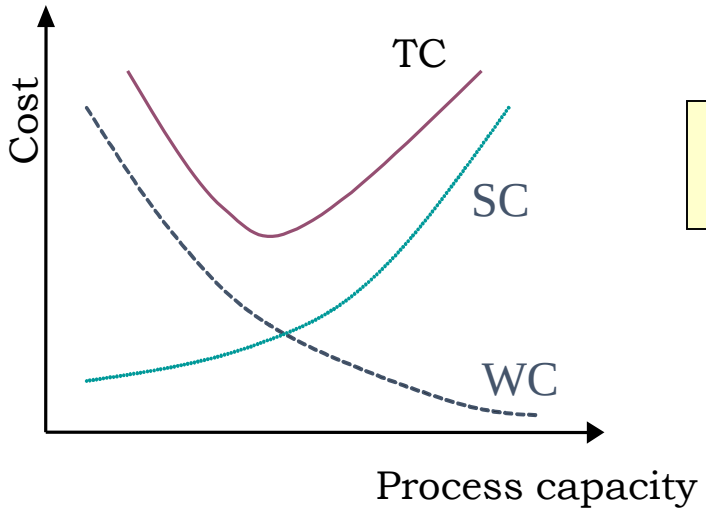
1. External customers arrive to the system
   - **Profit organizations**
     ⇒ The shortage cost is primarily related to lost revenues – "Bad Will"
   - **Non-profit organizations**
     ⇒ The shortage cost is related to a societal cost
2. Internal customers arrive to the system

   ⇒ 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
  - $C_w N$ = 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
  - λ = 20 jobs per day
  - μ_M = 30 jobs per day
  - μ_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 job is estimated to \$50 per job and day
- Question:
  - Which computer should the university choose in order to minimize the expected costs?

Simulation – What is it?

- Experiment with a model mimicking the real world system
  - Ex. Flight simulation, wind tunnels, …
- In BPD situations computer based simulation is used for analyzing and evaluating complex stochastic systems
  - Uncertain service and inter-arrival times

Simulation – Why use it?

- Cheaper and less risky than experimenting with the actual system.
- Stimulates creativity since it is easy to test the effect of new ideas
- A powerful complement to the traditional symbolical and analytical tools
- Fun tool to work with!

Simulation v.s. Symbolic & Analytical Tools

- **Strengths**
  - + Provides a quantitative measure
  - + Flexible – can handle any kind of complex system or statistical interdependencies
  - + Capable of finding inefficiencies otherwise not detected until the system is in operation
- **Weaknesses**
  - – Can take a long time to build
    - Usually requires a substantial amount of data gathering
  - – Easy to misrepresent reality and draw faulty conclusions
  - – Generally not suitable for optimizing system parameters

- A simulation model is primarily descriptive while an optimization model is by nature prescriptive

Modern Simulation Software Packages are Breaking Compromises

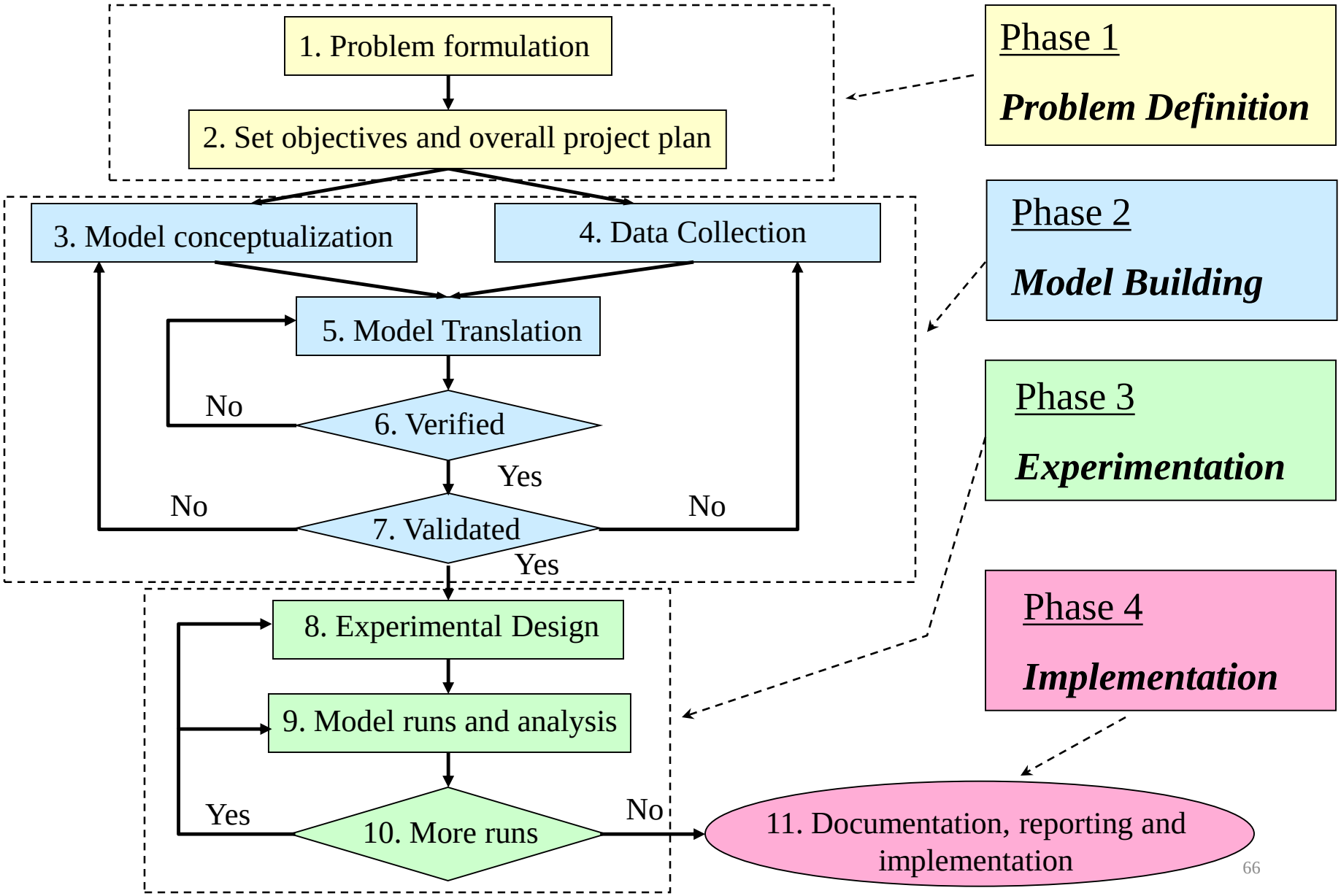
- Graphical interfaces
  - ⇒ Achieves the descriptive benefits of symbolic tools like flow charts
- Optimization Engines
  - ⇒ Enables efficient automated search for best parameter values

Building a Simulation Model

- General Principles
  - The system is broken down into suitable components or entities
  - The entities are modeled separately and are then connected to a model describing the overall system

  $\Rightarrow$ *A bottom-up approach!*
- The basic principles apply to all types of simulation models
  - Static or Dynamic
  - Deterministic or Stochastic
  - Discrete or continuous
- In BPD and OM situations computer based Stochastic Discrete Event Simulation (e.g. in Extend) is the natural choice
  - Focuses on events affecting the state of the system and skips all intervals in between

Steps in a BPD Simulation Project



Model Verification and Validation

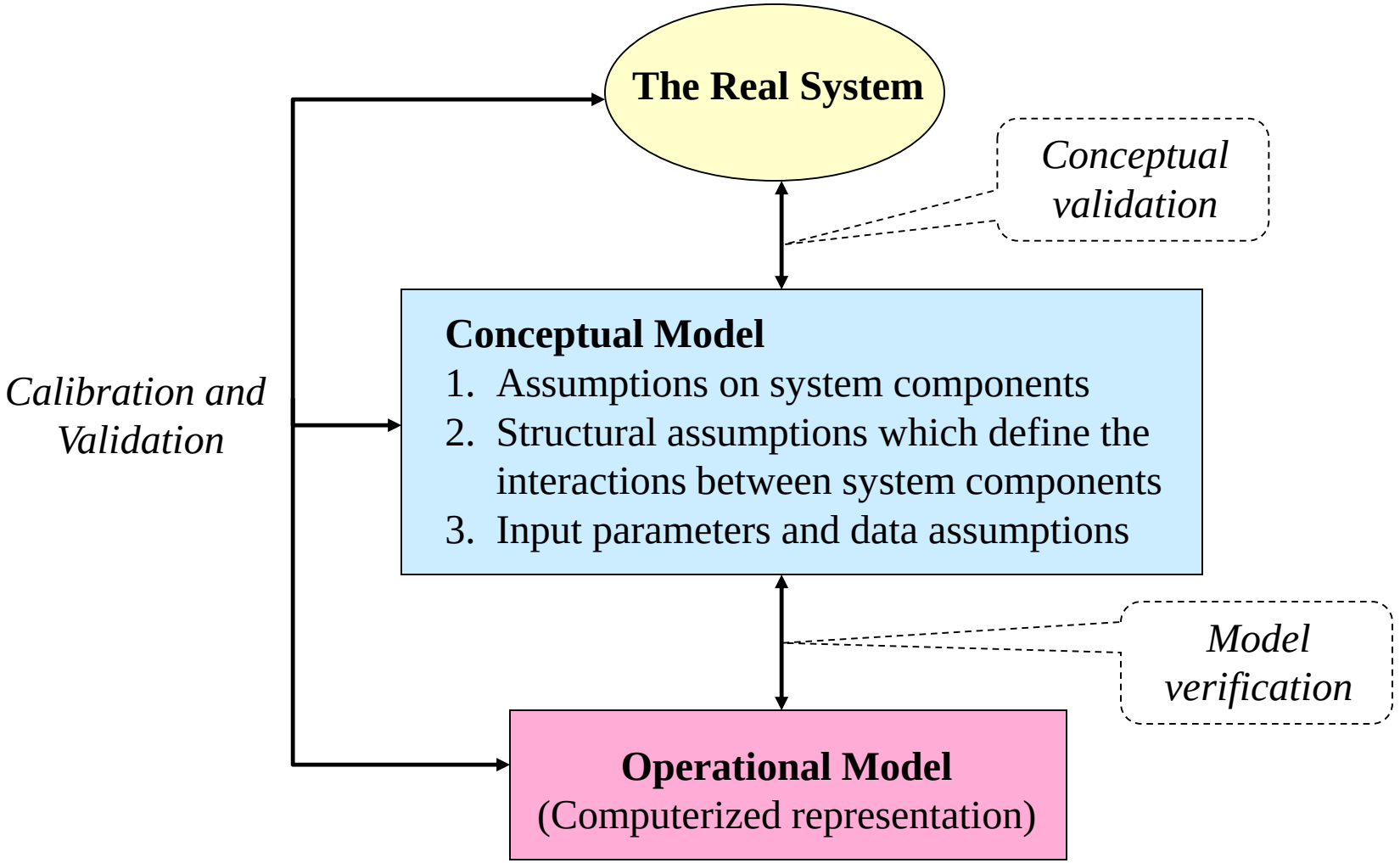
- Verification (efficiency)
  - Is the model correctly built/programmed?
  - Is it doing what it is intended to do?
- Validation (effectiveness)
  - Is the right model built?
  - Does the model adequately describe the reality you want to model?
  - Does the involved decision makers trust the model?

➢Two of the most important and most challenging issues in performing a simulation study

Model Verification Methods

- Find alternative ways of describing/evaluating the system and compare the results
  - Simplification enables testing of special cases with predictable outcomes
    - Removing variability to make the model deterministic
    - Removing multiple job types, running the model with one job type at a time
    - Reducing labor pool sizes to one worker
- Build the model in stages/modules and incrementally test each module
  - Uncouple interacting sub-processes and run them separately
  - Test the model after each new feature that is added
  - Simple animation is often a good first step to see if things are working as intended

Validation - an Iterative Calibration Process



Example – Simulation 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

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