UNIVERSITY OF PATRAS

DEPARTMENT OF BUSINESS ADMINISTRATION

FURTHER OPERATIONAL RESEARCH TECHNIQUES

Lecture 4: Queuing Theory – An Introduction

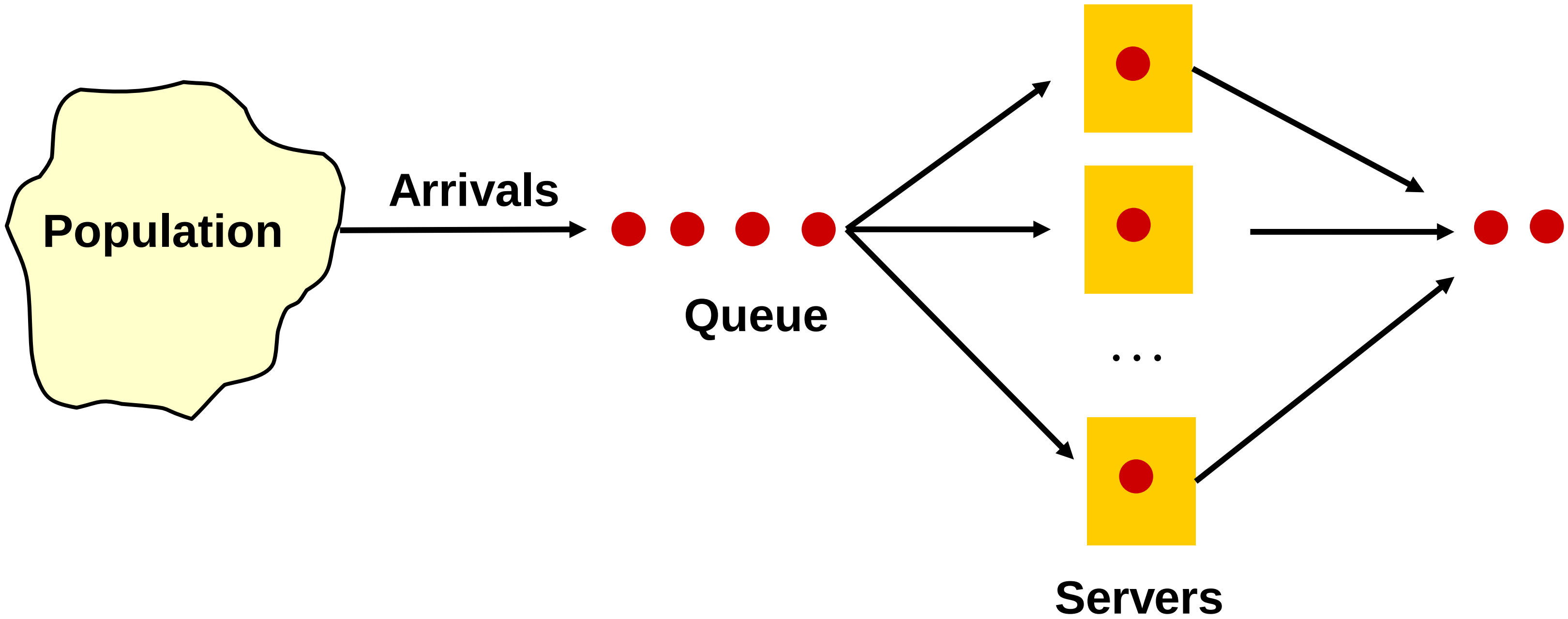
Patras 2022

Queues (Waiting Lines)

- Appear in nearly every system serving customers

- Elements
  - Arrivals
    - From where?
    - How often?
  - Service
    - How many service stations?
    - At what time?
  - Queue
    - Capacity?
    - Priorities?

Queues (continued)



- Assumptions regarding the statistical distribution of arrivals and service times
- Condition for a system to be sustainable: Total arrival rate < Total service rate
- Queues occur because the time between successive arrivals is variable

Cost elements

- Cost of establishing and operating service stations
  - Risk of under-utilization
- Waiting Cost
  - Cost of loss of time in the queue
  - <u>Cost of lost customers</u>
  - More serious consequences
- Problem:
  - How many service stations are required?
  - How should they operate?

Cost elements(diagram)

- **As the level of service increases, the service cost increases and the waiting cost decreases**

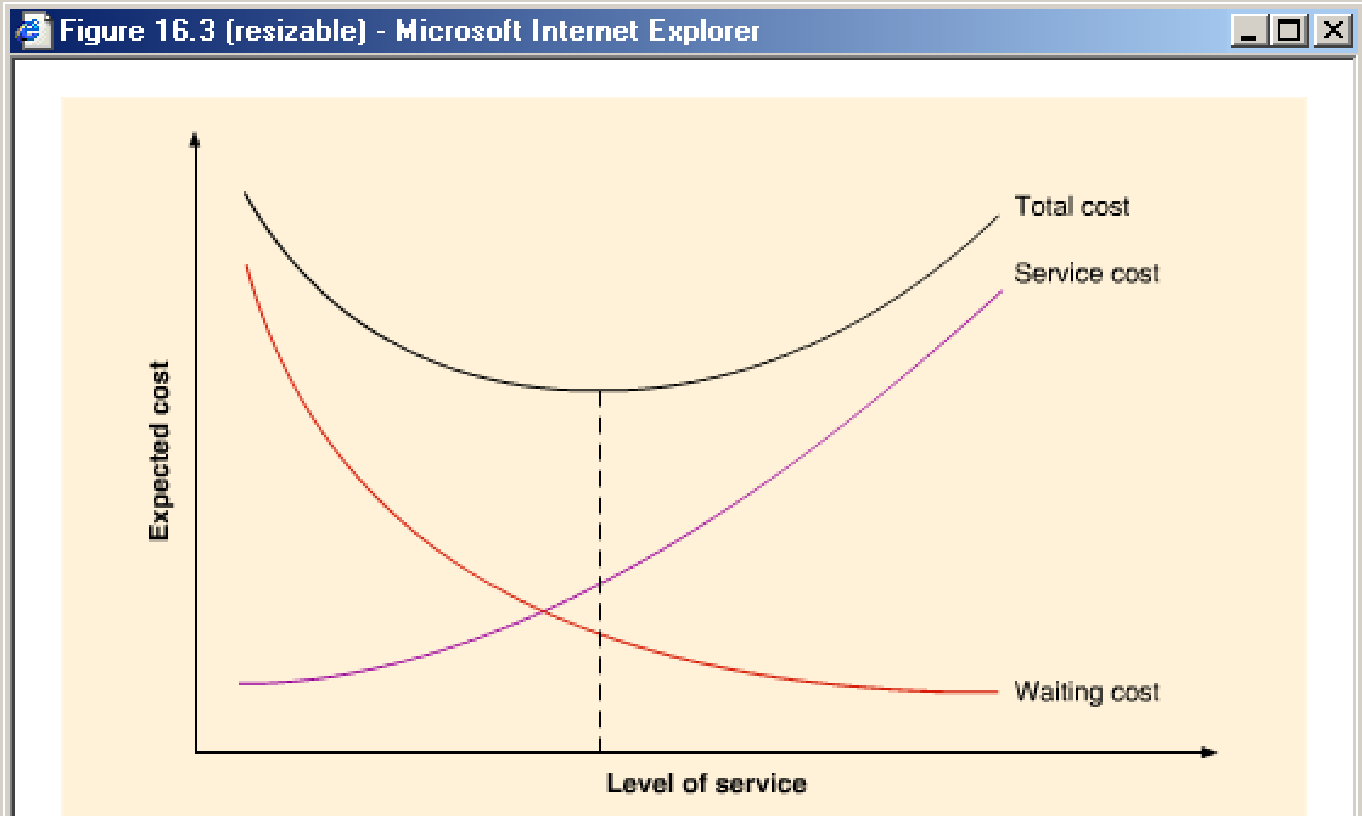


FIGURE 16.3 The Cost Relationship in Waiting Line Analysis

Notation

- In most systems we make assumptions regarding the statistical distribution of:
  - The time between successive arrivals
  - The service times
- Such systems are symbolically represented as: _/_/_
  - The 1st position represents the distribution of time between successive arrivals
  - The 2nd position represents the distribution of service time
  - The 3d position represents the number of servers
  - Symbol M denotes the Exponential distribution, E_k the Erlang distribution etc
- Example: Notation M/M/3 denotes a system with 3 servers, exponential time between arrivals and exponential service time

Notation (continued)

- $N(t)$ — number of customers in the system at time t
- $P_n(t)$ — probability that n customers are in the system at time t
- s — number of service stations
- λ_n — arrival rate when n customers are in the system
- μ_n — service rate when n customers are in the system

Remarks

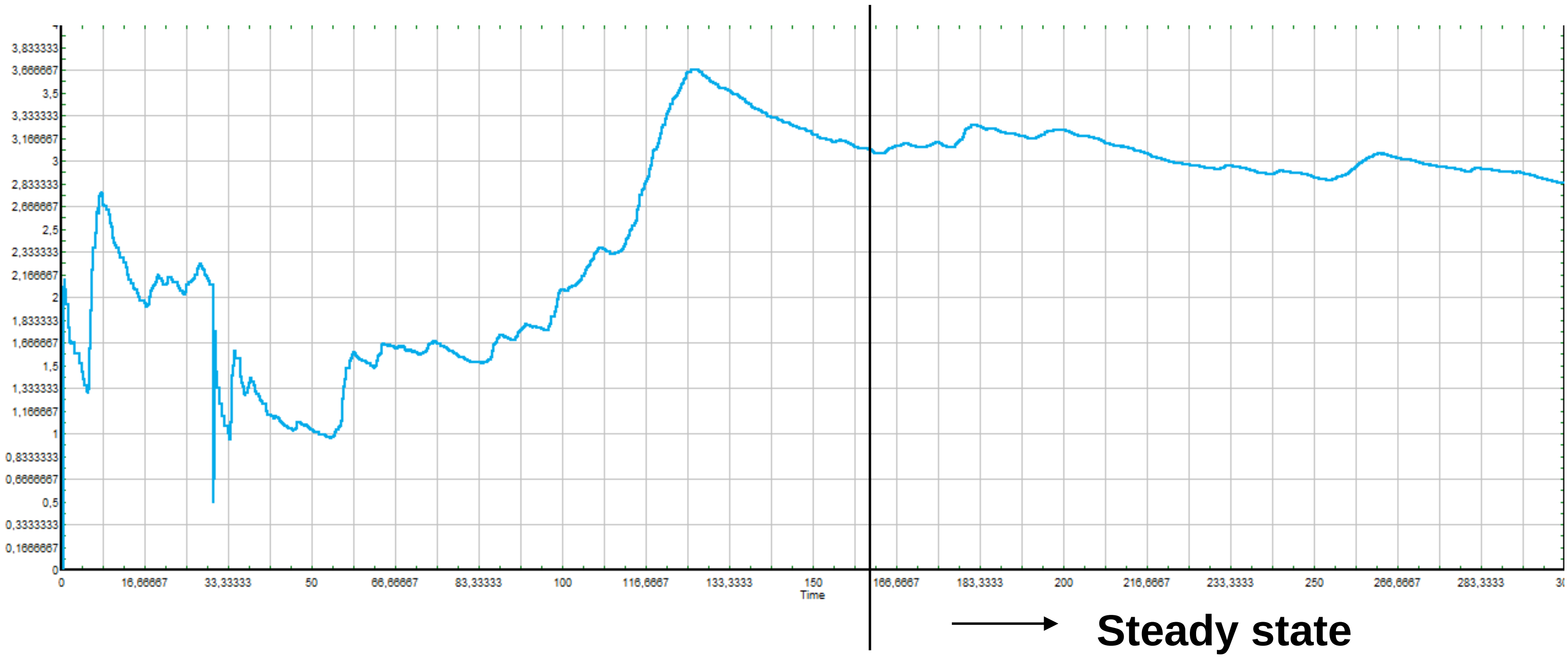
- In many situations the arrival rate is (assumed to be) constant ($\lambda_n = \lambda$ for all n)
- A similar assumption is often made for the service rate ($\mu_n = \mu$ for all n)
- $\rho = \frac{\lambda}{s\mu}$ is the utilization rate of the servers
- We are interested in the behavior of the system at its steady state, i.e. after the initial period of operation

Operating Characteristics

- **L: Average number of customers in the system**
- **L_q: Average number of customers in the queue**
- $W = E(\mathcal{W})$: **Average time spent in the system**
- $W_q = E(\mathcal{W}_q)$: **Average waiting time in the queue**
- **P_n: Probability that there are n customers in the system**
- **P_0: Probability that the system is empty**

Steady State

- This is a graph of L_q in a particular system



- Steady state: the state where the probability of the number of customers in the system (or in the queue) is independent of time

Fundamental Formulas

- It can be shown that: $$L = \lambda \cdot W$$
- This is known as Little's Formula
- Similarly, it may be shown that: $$L_q = \lambda \cdot W_q$$
- If the arrival rate λ is not constant, then we can use the mean (average) arrival rate in the formulas above
- The following also holds: $$W = W_q + \frac{1}{\mu}$$

(Time spent in the system is Time waiting in the Queue plus Service Time)

Cost Elements

- **Let**
  - **c_s: the cost of establishing and operating each server per unit time**
  - **c_w: the waiting cost of each customer per unit time**
- **The total system cost per unit time is:**

$$TC = c_s \cdot s + c_w \cdot L$$

- **If the waiting cost concerns only customers in the queue then**

$$TC = c_s \cdot s + c_w \cdot L_q$$

Arrivals

- <u>Assumption</u>: The number of arrivals follows a Poisson distribution with mean rate λ
- It can be shown that:
  - When the number of arrivals follows a Poisson distribution with mean rate λ, the time between successive arrivals follows an exponential distribution with mean 1/λ and vice-versa.
- Main consequence
  - The time until the next arrival is independent of the time of the previous arrival
- The process is "memoryless"
  - This basically means that <u>arrivals are totally random</u>

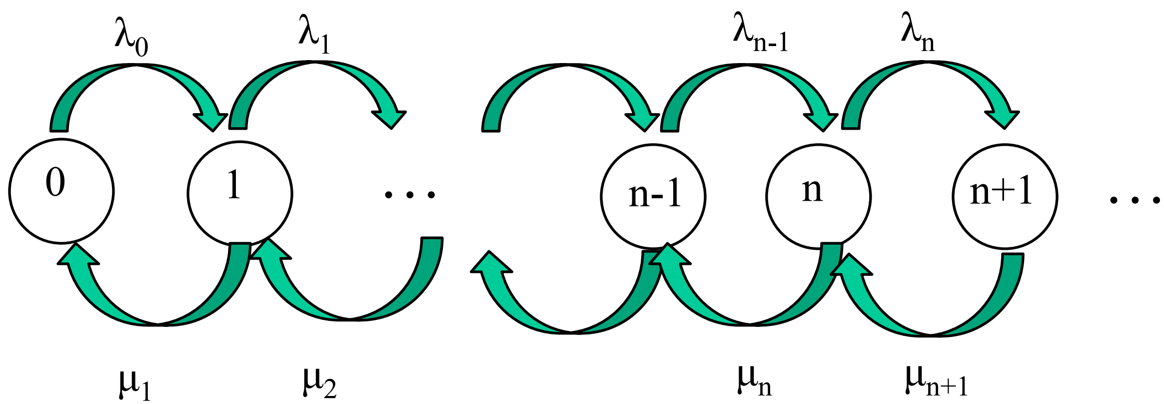
Services

- The service time may be
  - constant
  - random variable

- It is usually assumed that the service time is distributed exponentially with mean value $1/\mu$, where μ is the service rate

Representation as a Birth and Death Process

- Arrivals and departures can be represented as a birth and death process
- In such a representation, the state of the system is the number of customers in the system at any point in time



- It can be shown that when the system is in steady state, the total flow into any state n is equal to the total flow out of that state i.e.

$$\lambda_{n-1}P_{n-1} + \mu_{n+1}P_{n+1} = (\lambda_n + \mu_n)P_n \quad \textbf{(Balance equation)}$$

Calculation of Probabilities

- We can prove (by induction) that the following relationships hold:

$$P_n = \frac{\lambda_{n-1}}{\mu_n} \cdot P_{n-1} \quad \text{and} \quad P_n = \frac{\lambda_0 \lambda_1 \ldots \lambda_{n-1}}{\mu_1 \mu_2 \ldots \mu_n} \cdot P_0$$

- Based on these and the fact that $\sum_{j=0}^{\infty} P_j = 1$

  we can calculate the probabilities P_j

System with s=1 service station (M/M/1)

- **Assumptions**
  - **Poisson arrivals with mean rate λ per unit time**
  - **One service station**
  - **Service rate is μ per unit time**
  - **Service time is exponential with mean value 1/μ**

- **The system is sustainable i.e. it may reach a steady state only if**

$$\rho = \frac{\lambda}{\mu} < 1$$

Results

- **It can be shown that:**
  - $P_0 = 1 - \rho$ **(probability that the system is empty)**
  - $P_n = (1 - \tau) \cdot \rho^n$ **(probability that there are n customers)**
  - **Average number of customers in the system:**

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

  - **Average number of customers in the queue:**

$$L_q = \frac{\rho^2}{1 - \rho}$$

More Results

- The mean waiting time in the system may be calculated by

  Little's Formula

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

- The methodology is similar for systems with more service

  stations

Example

A small airport has only one runway for take-off and landing. During the busy summer season flights arrive at a rate of 30 flights per hour, following the Poisson distribution. Every flight needs an average time of 90 seconds for landing. If the runway is occupied, waiting planes have to circle overhead until the runway becomes available. Every flight that has to wait incurs a cost of 5000 Euros per hour due to increased fuel consumption. Calculate the following:

(a) The average number of planes waiting for their turn to land.

(b) The average waiting time for landing.

(c) The average number of flights in the system.

(d) The average waiting cost.

Exercise

- In an M/M/1 queueing system the arrival rate (λ) and the service rate (μ) are doubled. Calculate the effect on the following operating characteristics:
  - P_0
  - W
  - L_q