Operations Management

Session 4: Waiting Lines

Industries with Queues

The Call Center Industry

- All Fortune 500 companies have at least one call center.
- Each firm has an average of 4,500 agents across all their sites.
- North American call centers employ 2.9 million of agents in 55,000 facilities.
- Worldwide, $300 billion is spent annually on call centers.

Source: Gilson and Khandelwal, "Getting more from call centers," *The McKinsey Quarterly, web exclusive, April 2005.*

Waiting for Service

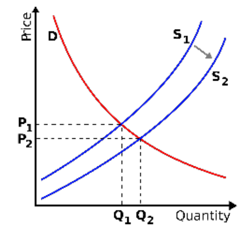
- Why are people willing to wait for service?
- What is the cost of waiting (For customers? For the firm?)
- What are the benefits of waiting? (For customers? For the firm?)
- Why do waiting lines form?

What is the Source of Waiting Lines?

Dice Game

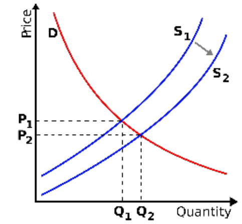
Period	Arrivals	Capacity	Wasted capacity	Queue
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				

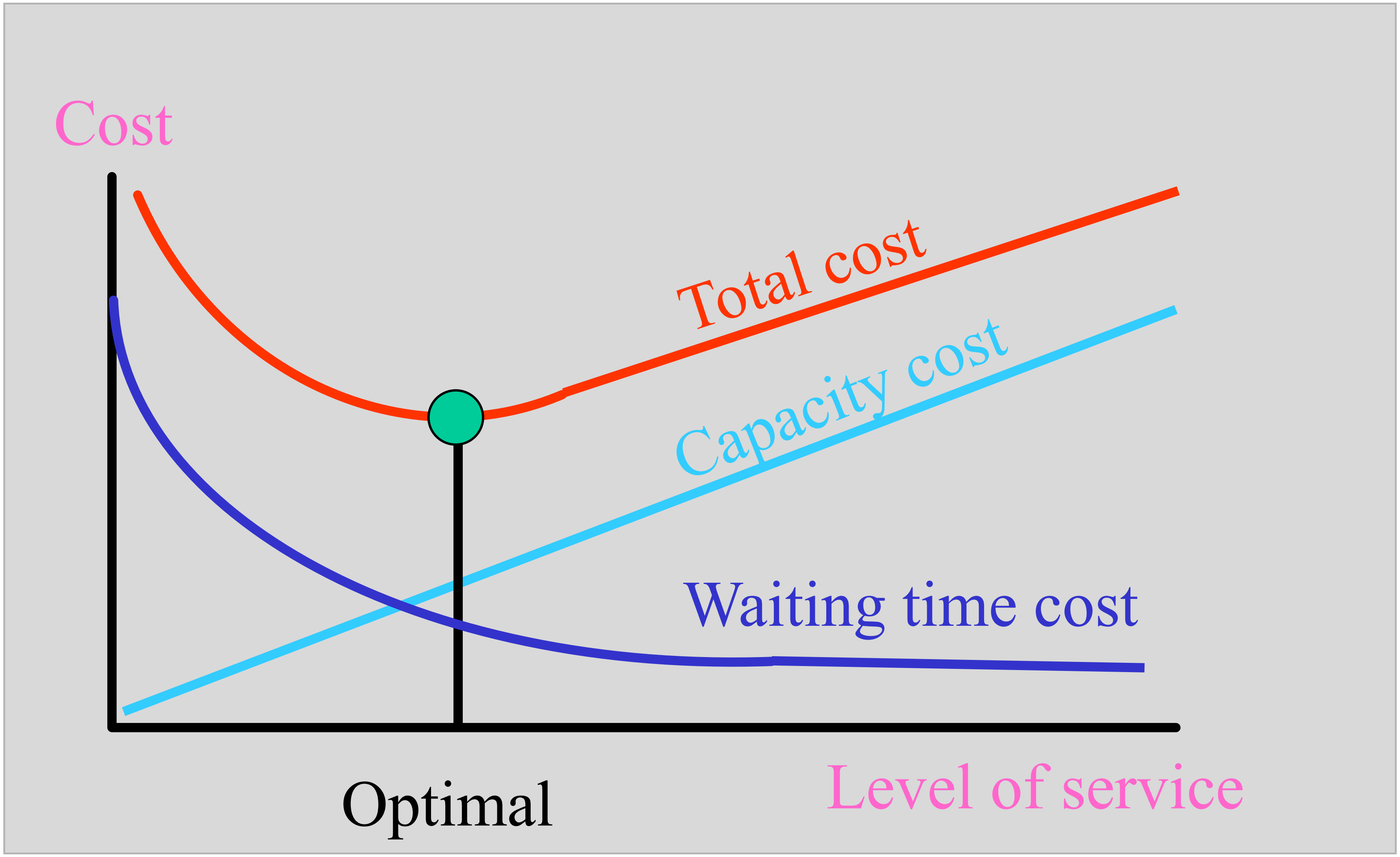
How to Reduce Delay?



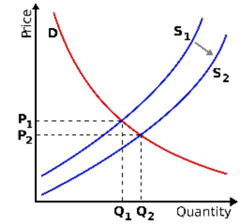
Demand Side	Both	Supply Side

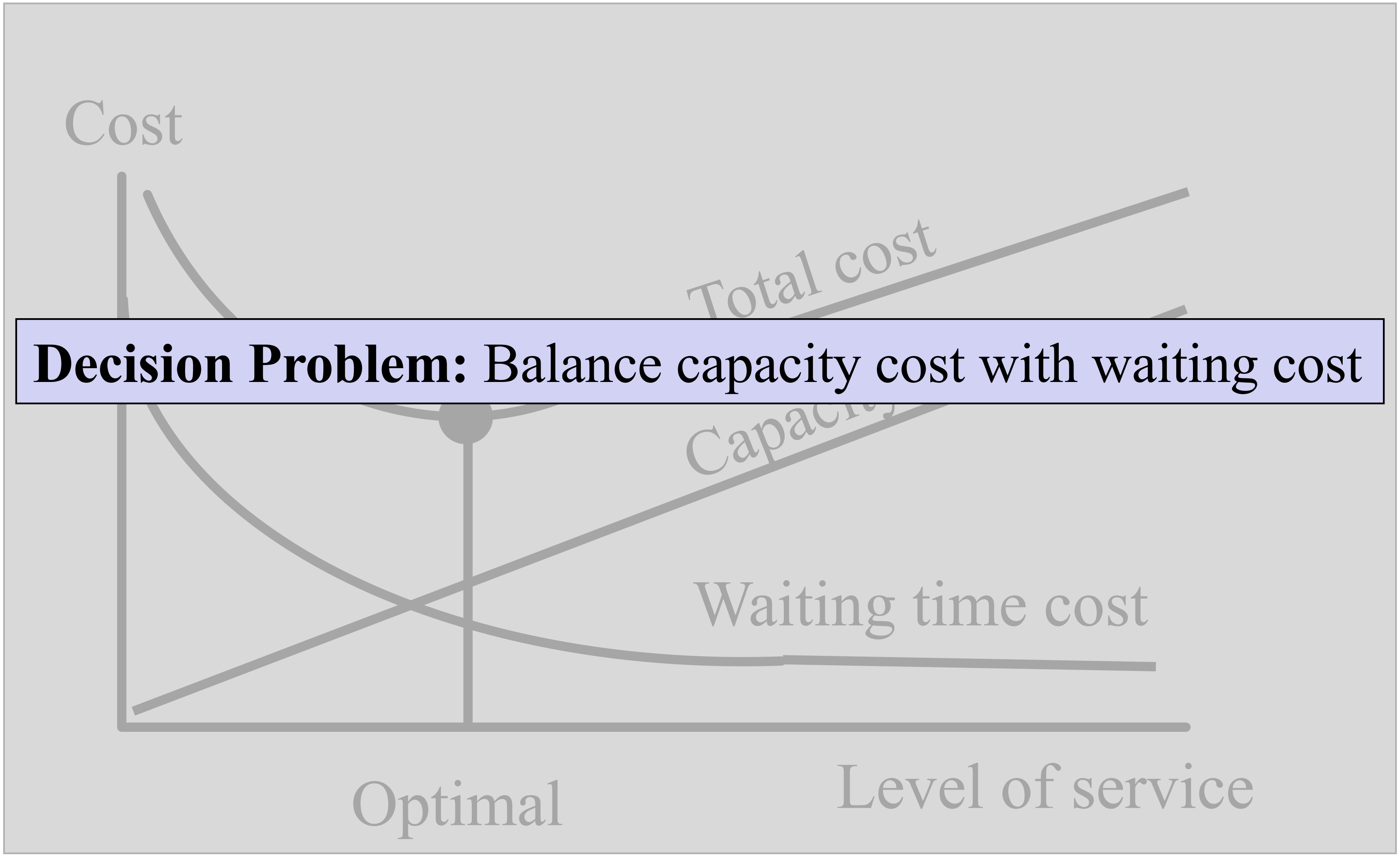
Waiting Line Costs





Waiting Line Costs





Decision Problem: Balance capacity cost with waiting cost

Performance Measures

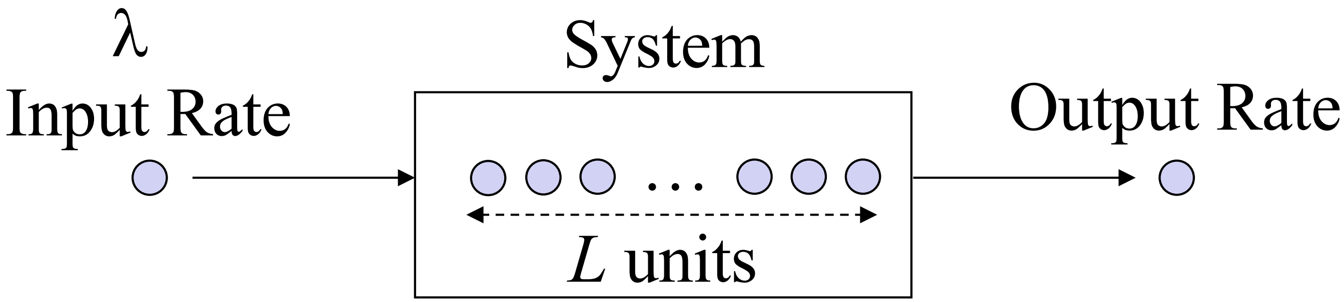
- L_Q = Queue Length
- W_Q = Waiting time in Queue
- L_S = Number in System
- W_S = Time in System (Sojourn Time)
- ρ = Server Utilization

*All are average measures

Questions:

1. What performance measure matters the most?
2. What is the relationship between W_Q and W_S?
3. What is the relationship between L_S and W_S?
4. Which value do we want ρ to be?

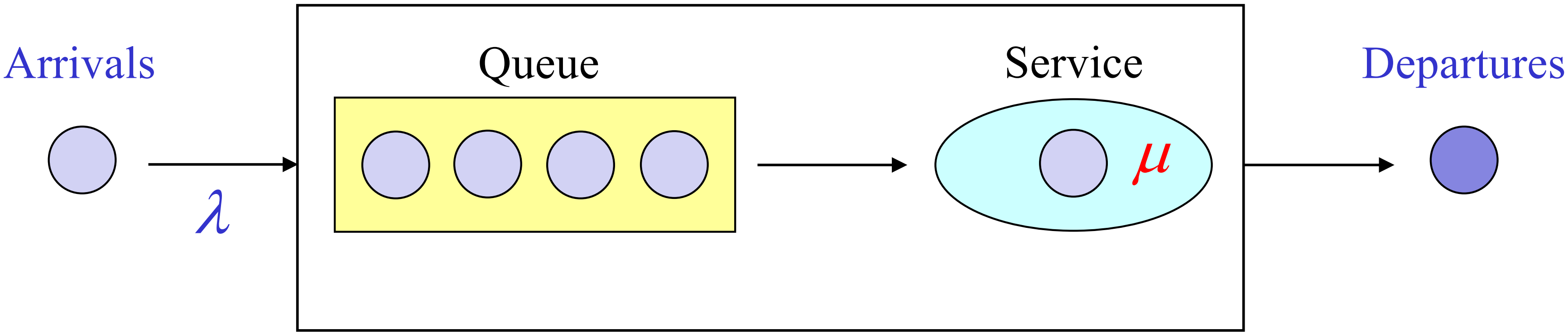
Little's Law



Waiting Time = W

$$L = \lambda \times W$$

Single Server Queue



A single server queue is defined by:

1. Jobs arrival: $\longrightarrow$ λ : rate of arrivals (e.g. cust/hour)
2. Queue discipline: FCFS
3. Jobs processing: $\longrightarrow$ μ : service rate (e.g. cust/hour)

The M/M/1 Queueing System

The simplest queueing model.

We impose 4 assumptions:

1. Single server,
2. FCFS discipline,
3. Exponential interarrival time,
4. Exponential service time.

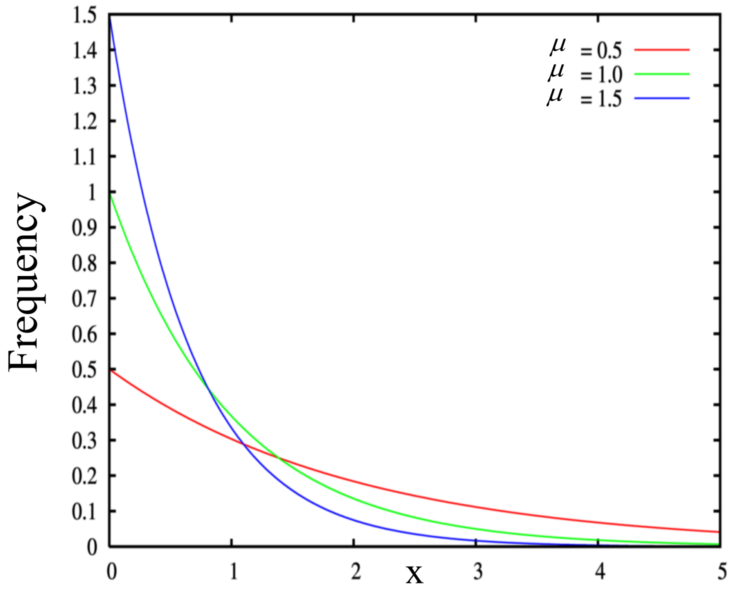
<u>Discussion:</u> Are these assumptions realistic?

Exponential Distribution

<u>Exponential service time</u>

The time it takes to serve a customer follows an **Exponential distribution** with parameter μ.

$\mu e^{-\mu x}$



Similarly, the time between two successive customer arrivals follows an **Exponential distribution** with parameter λ.

Exponential Distribution

Service time probability:

$$\Pr(\text{Service time} > x) = e^{-\mu x} \qquad (e = 2.7181)$$

Average service rate = μ. Average service time = $1/\mu$.

Similarly:

$$\Pr(\text{Interarrival time} > x) = e^{-\lambda x}$$

Average arrival rate = λ.

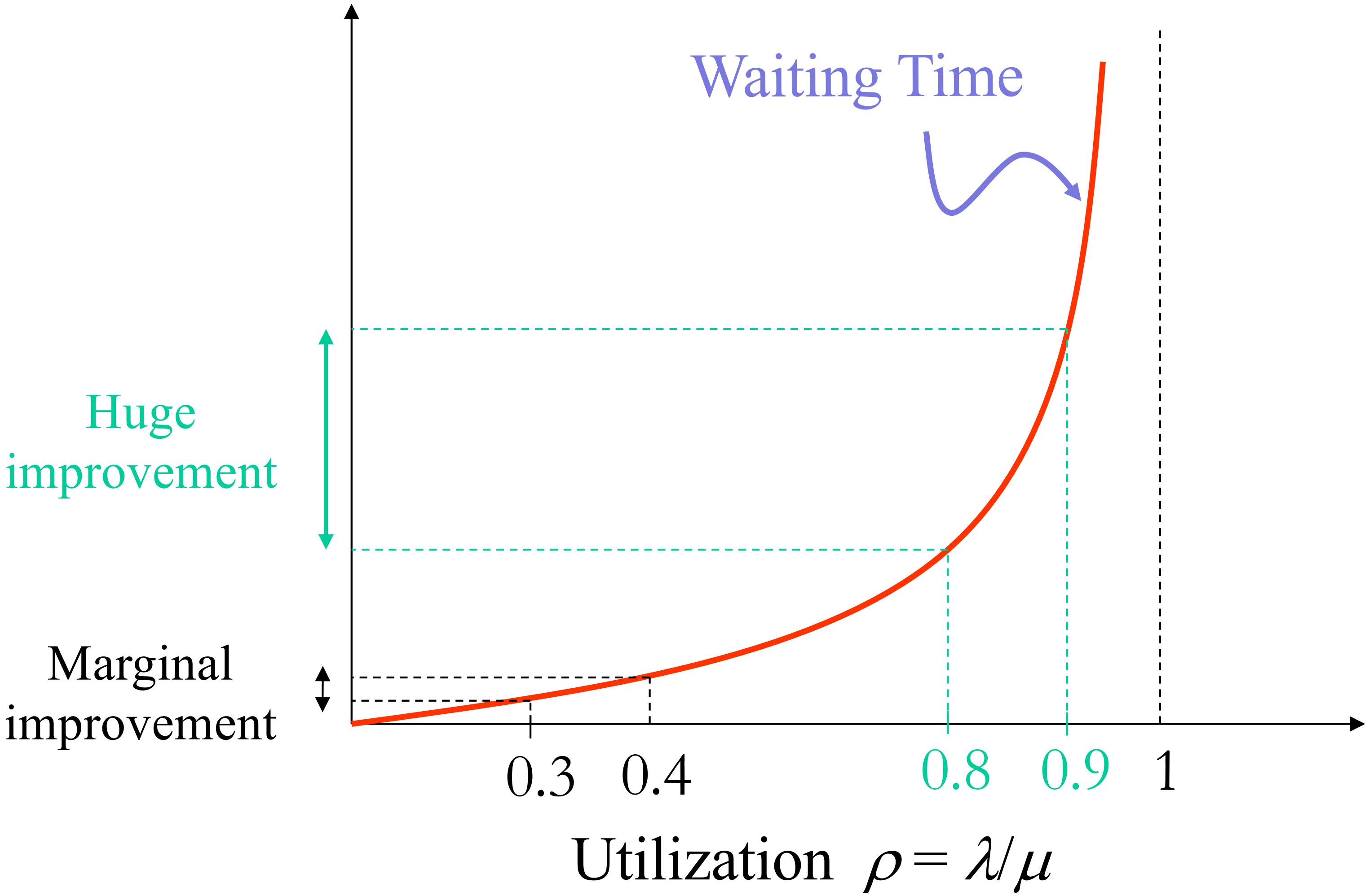
Model Equations

M/M/1

$\lambda<\mu$

Average number of units in System	$L_S = \dfrac{\lambda}{\mu - \lambda}$
Average Time in System	$W_S = \dfrac{1}{\mu - \lambda}$
Average Number of Units in Queue	$L_Q = \dfrac{\lambda^2}{\mu(\mu - \lambda)}$
Average Time in Queue	$W_Q = \dfrac{\lambda}{\mu(\mu - \lambda)}$
System Utilization	$\rho = \dfrac{\lambda}{\mu}$

Single Server Queueing System



Single Server Queueing System

Example: Consider an M/M/1 system with arrival rate $\lambda = 2$ customers/hour, and an average service time equal to 20 minutes per customer.

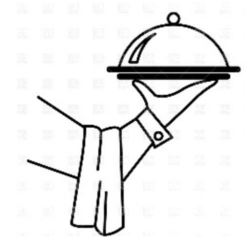
1) What is the average number of customers in the system and in the queue?

2) What is the average time in the system and the system utilization?

Single Server Queueing System

3) Suppose the cost of keeping a customer in the system is \$5 per minute, and the cost of having a server with capacity μ customers per hour, is equal to \$$(150\mu)$ per hour. What is the optimal level of capacity for this system?

Summary



- Waiting lines form due to variability.
- Basic tradeoff between cost and quality.
- Decisions:
  - Service capacity: service rate, (and number of servers).
  - System configuration.
- Performance measures
  - **M/M/1** model - Under Exponential service and interarrival times.
  - General universal relationships (Little's Law).