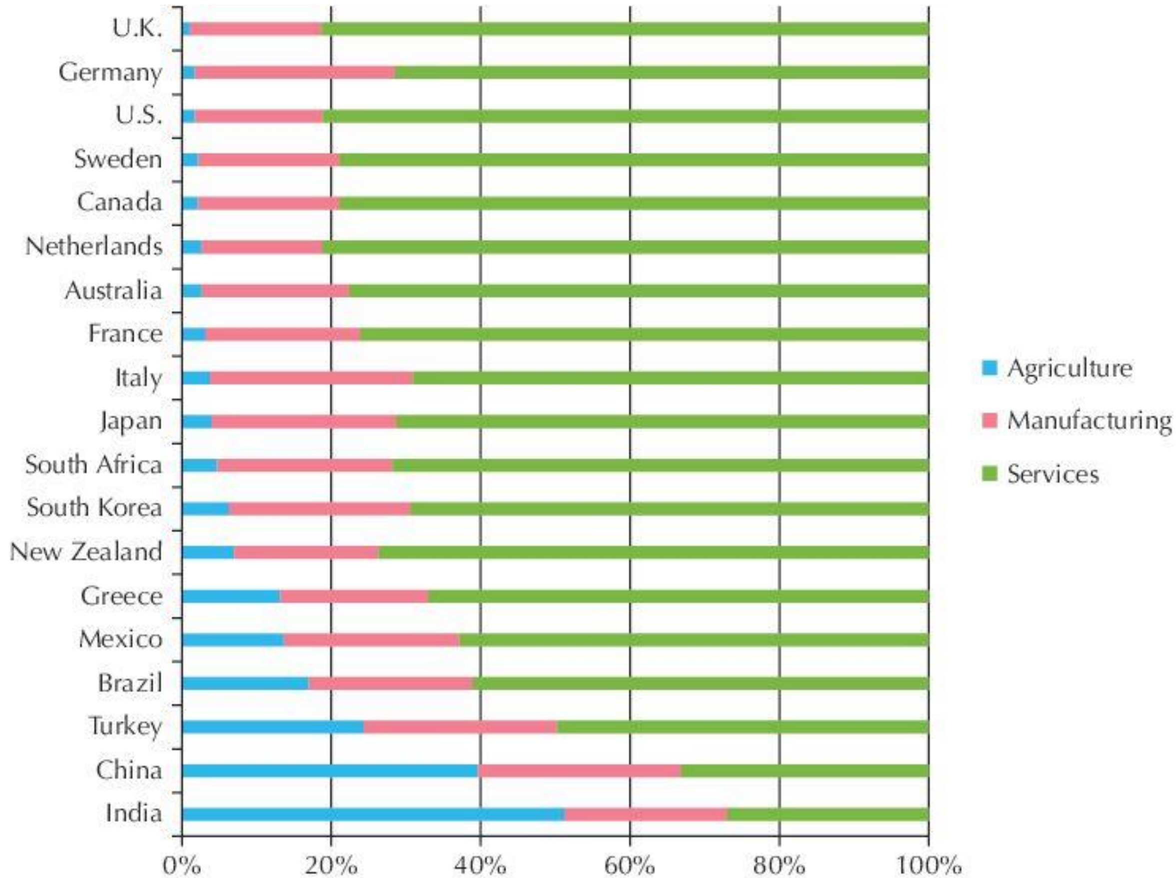
Lecture Outline

- Service Economy
- Characteristics of Services
- Service Design Process
- Tools for Service Design
- Waiting Line Analysis for Service Improvement

Service Economy

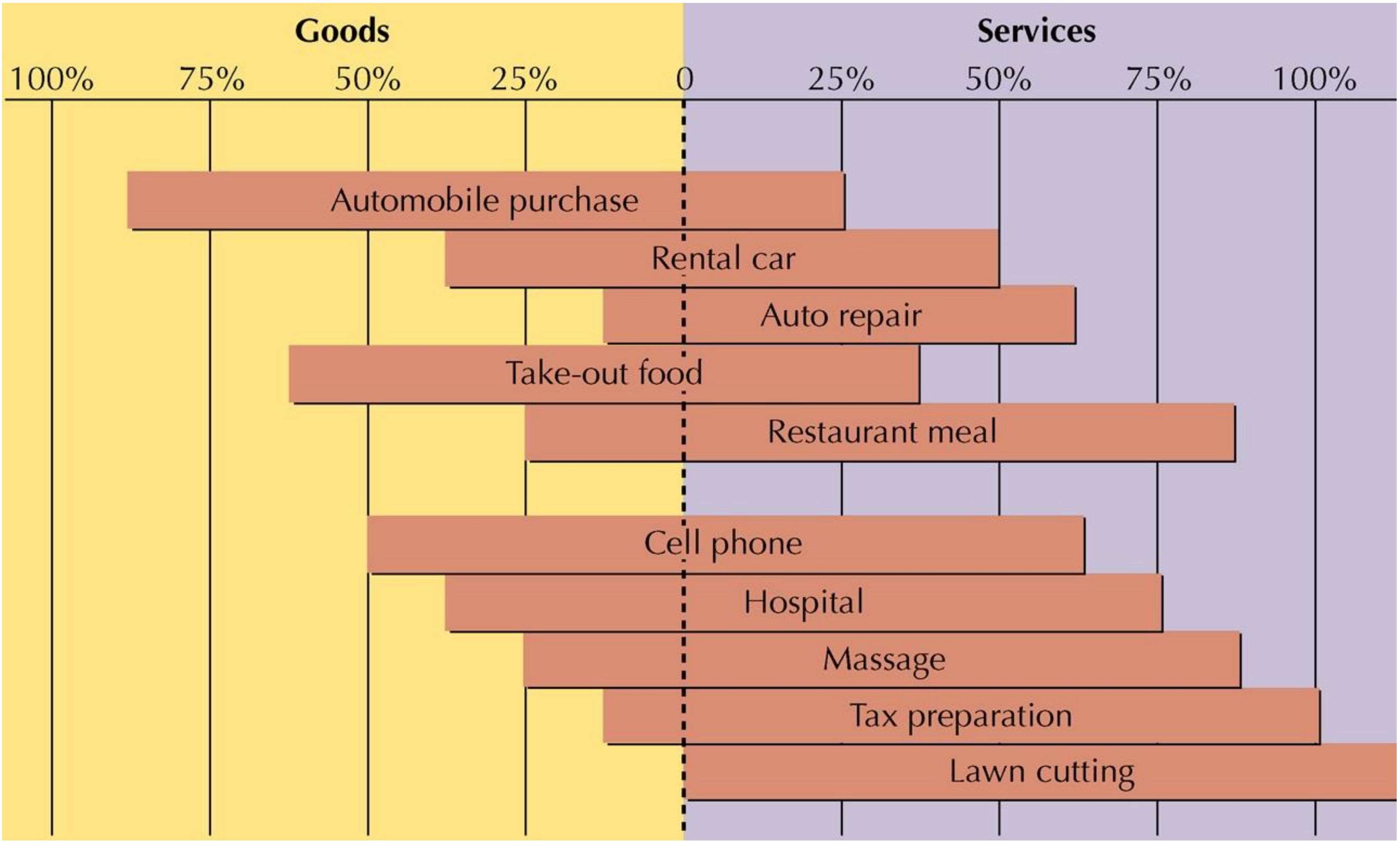
International Employment by Industry Sector



Characteristics of Services

- Services
  - acts, deeds, or performances
- Goods
  - tangible objects
- Facilitating services
  - accompany almost all purchases of goods
- Facilitating goods
  - accompany almost all service purchases

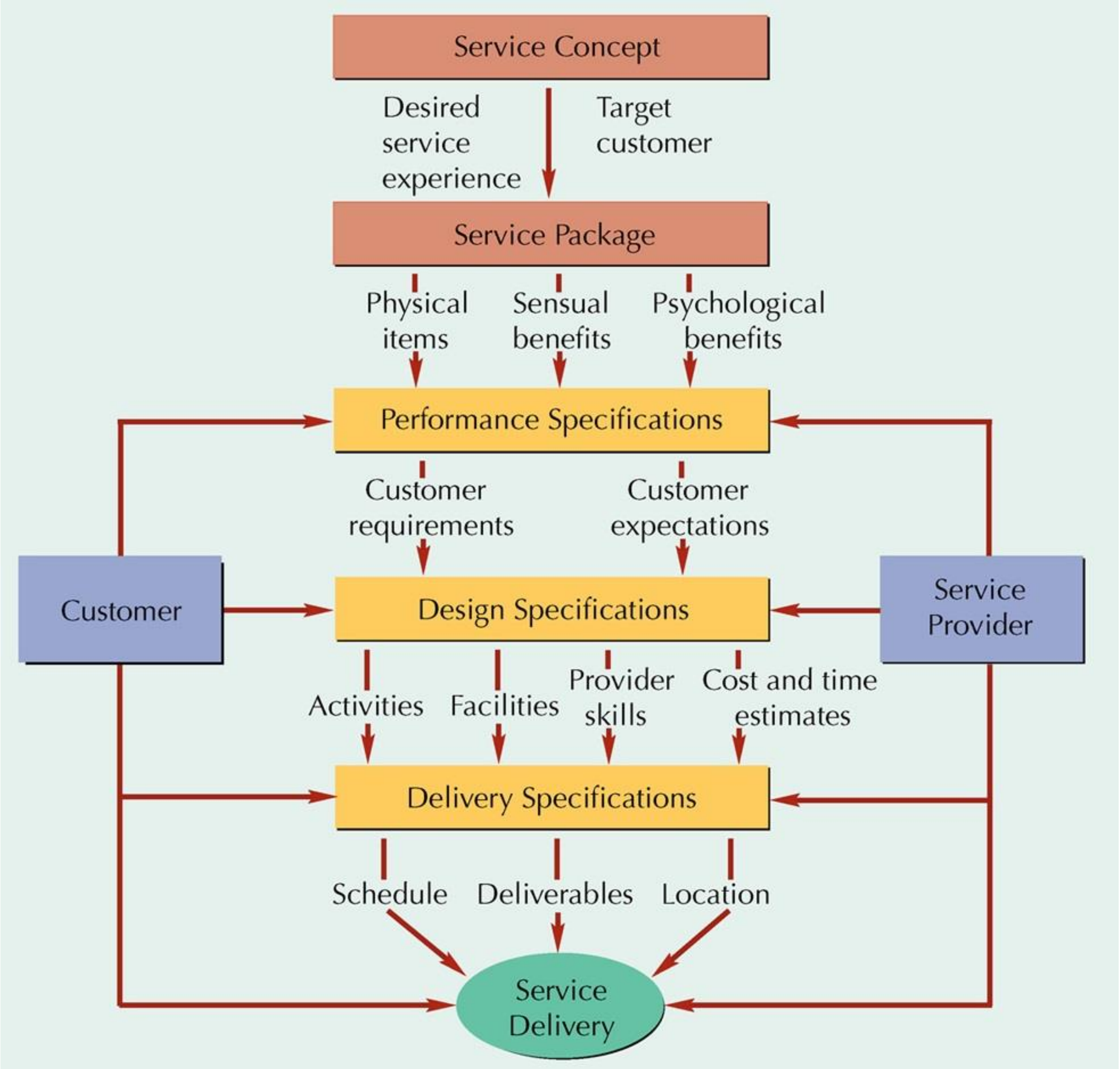
Continuum From Goods to Services



Characteristics of Services

- Services are intangible
- Service output is variable
- Services have higher customer contact
- Services are perishable
- The service and the service delivery are inseparable
- Services tend to be decentralized and geographically dispersed
- Services are consumed more often than products
- Services can be easily emulated

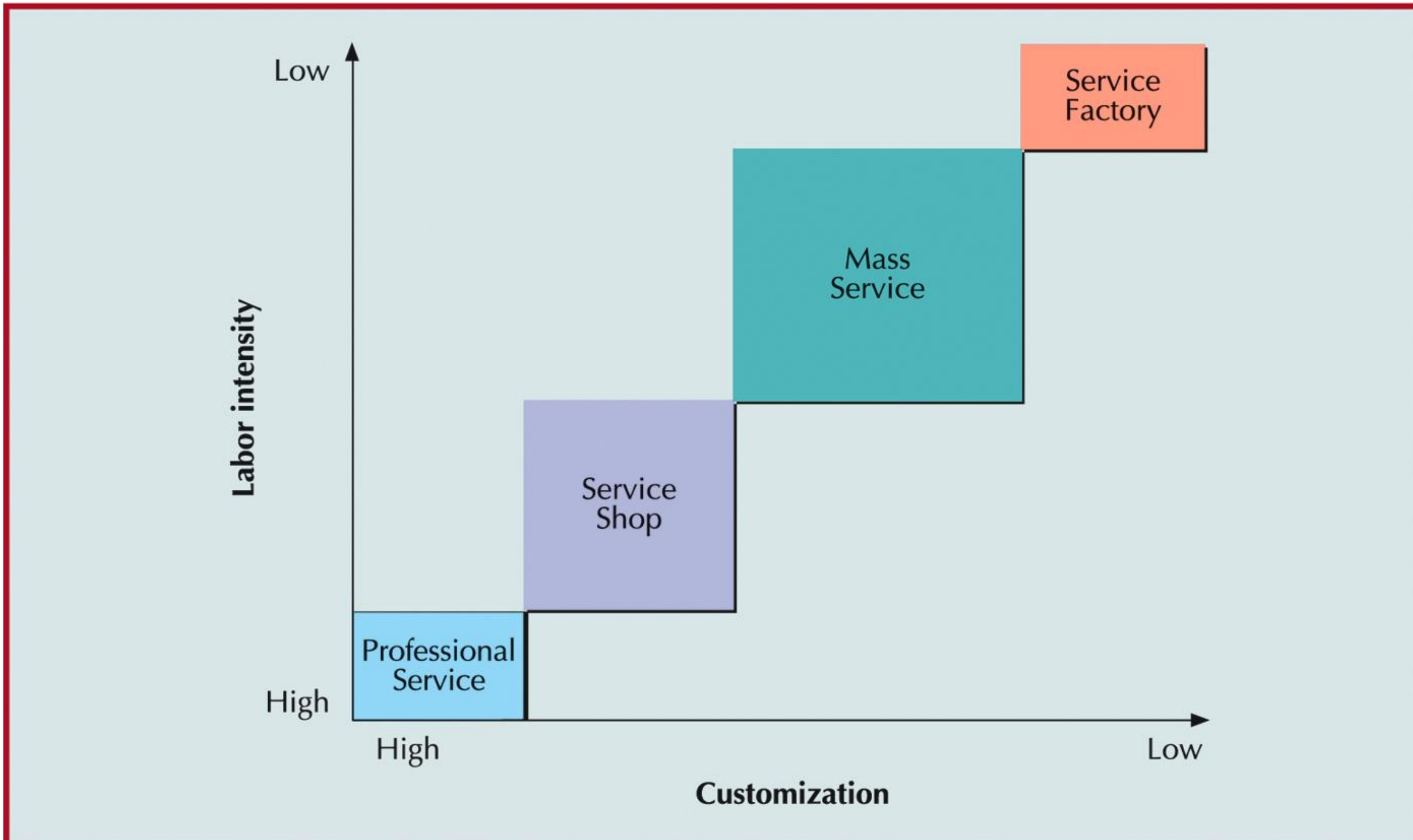
Service Design Process



Service Design Process

- Service concept
  - purpose of a service; it defines target market and customer experience
- Service package
  - mixture of physical items, sensual benefits, and psychological benefits
- Service specifications
  - performance specifications
  - design specifications
  - delivery specifications

Service Process Matrix



High vs. Low Contact Services

Design Decision	High-Contact Service	Low-Contact Service
▪ Facility location	▪ Convenient to customer	▪ Near labor or transportation source
▪ Facility layout	▪ Must look presentable, accommodate customer needs, and facilitate interaction with customer	▪ Designed for efficiency

High vs. Low Contact Services

Design Decision	High-Contact Service	Low-Contact Service
▪ Quality control	▪ More variable since customer is involved in process; customer expectations and perceptions of quality may differ; customer present when defects occur	▪ Measured against established standards; testing and rework possible to correct defects
▪ Capacity	▪ Excess capacity required to handle peaks in demand	▪ Planned for average demand

High vs. Low Contact Services

Design Decision	High-Contact Service	Low-Contact Service
▪ Worker skills	▪ Must be able to interact well with customers and use judgment in decision making	▪ Technical skills
▪ Scheduling	▪ Must accommodate customer schedule	▪ Customer concerned only with completion date

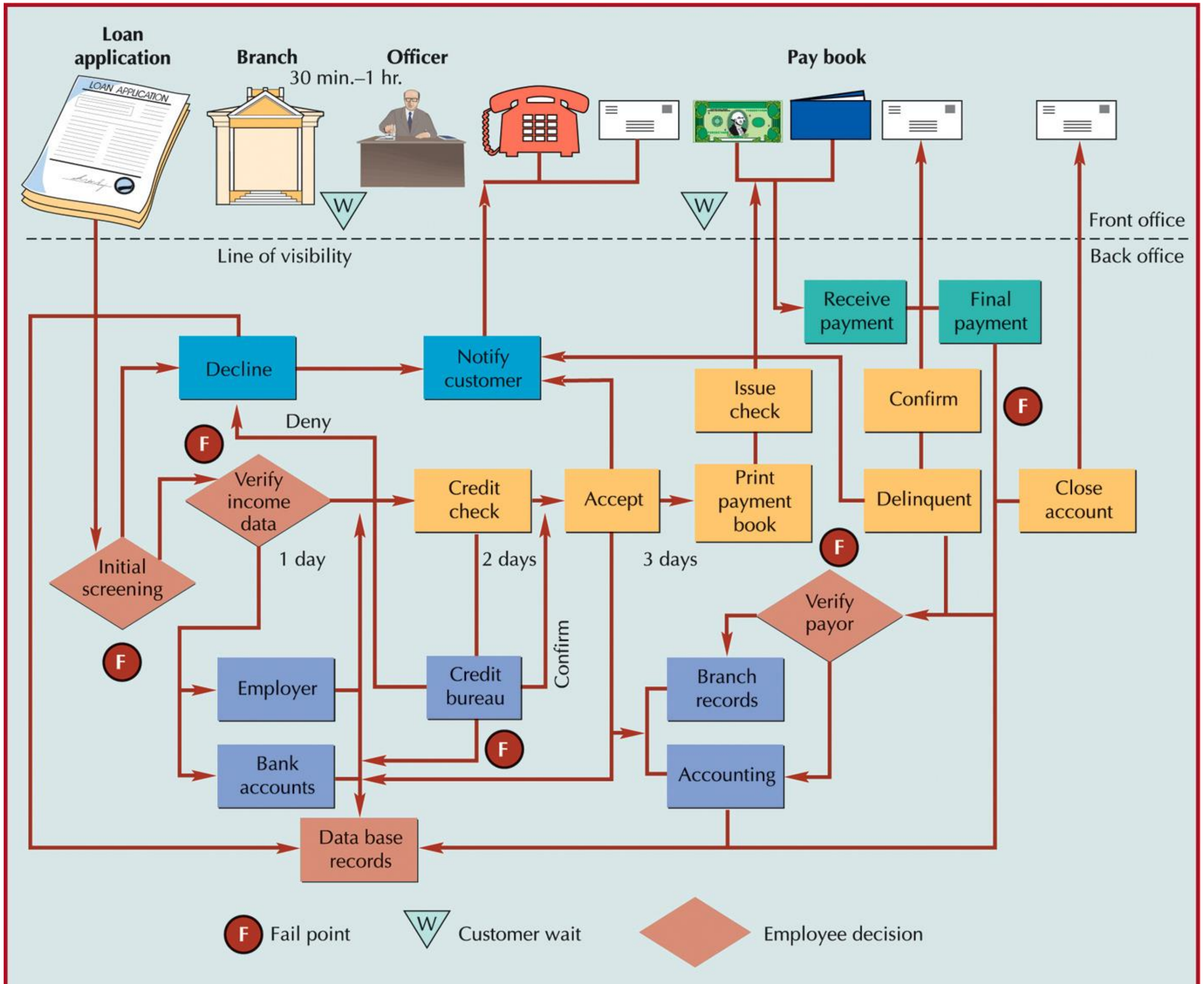
High vs. Low Contact Services

Design Decision	High-Contact Service	Low-Contact Service
▪ Service process	▪ Mostly front-room activities; service may change during delivery in response to customer	▪ Mostly back-room activities; planned and executed with minimal interference
▪ Service package	▪ Varies with customer; includes environment as well as actual service	▪ Fixed, less extensive

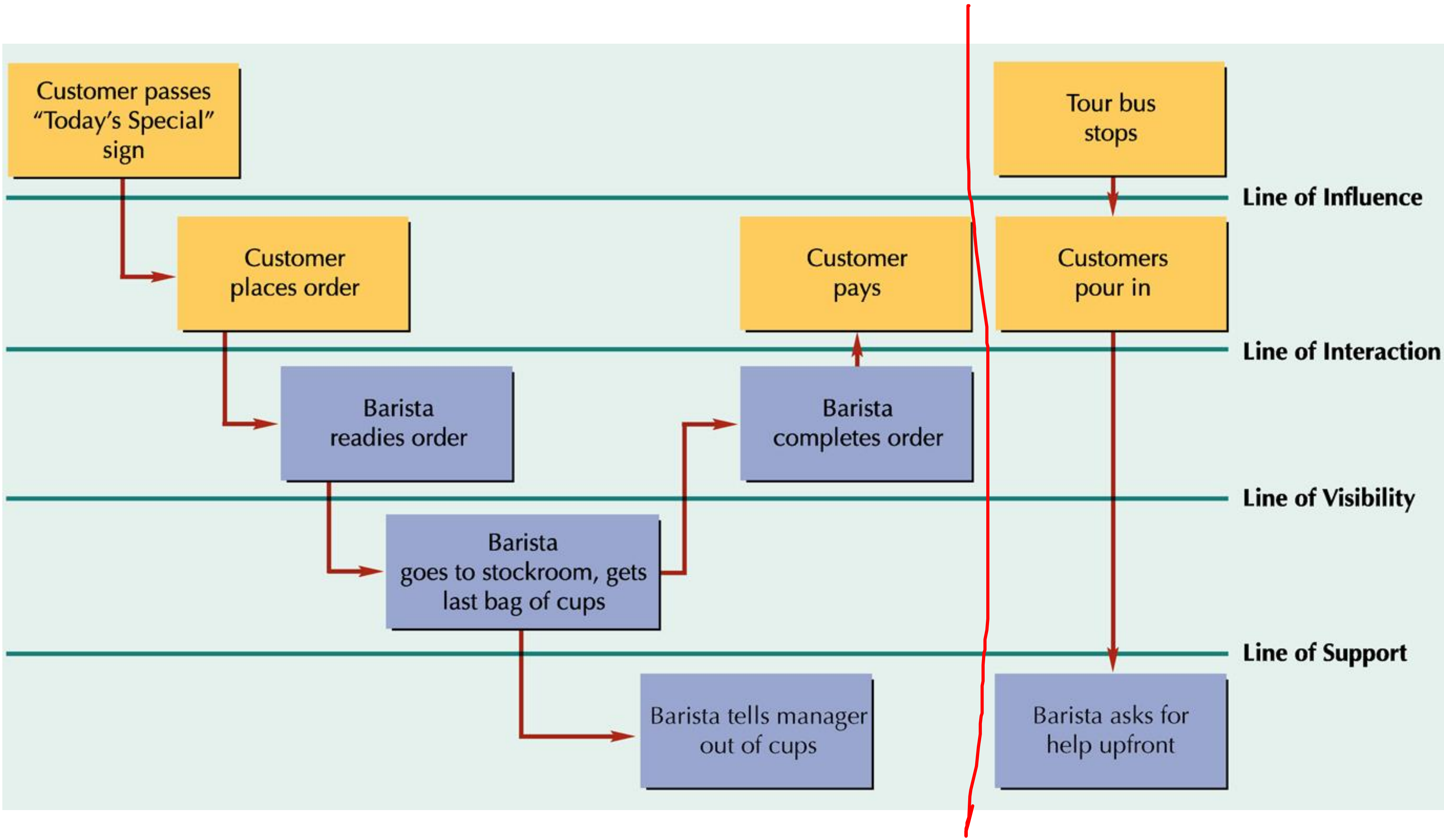
Tools for Service Design

- Service blueprinting
  - line of influence
  - line of interaction
  - line of visibility
  - line of support
- Front-office/Back-office activities
- Servicescapes
  - space and function
  - ambient conditions
  - signs, symbols, and artifacts
- Quantitative techniques

Service Blueprinting



Service Blueprinting

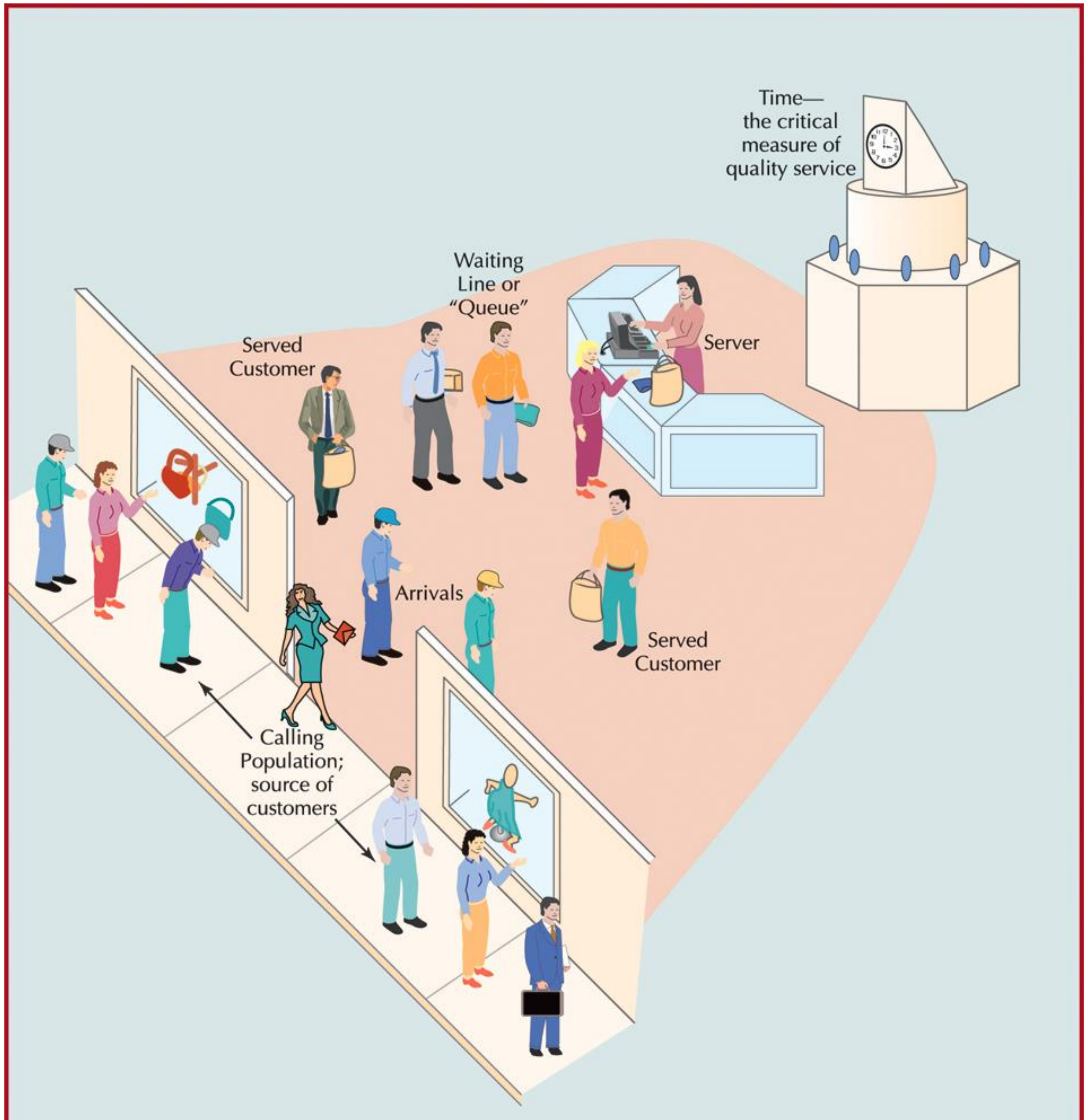


Elements of Waiting Line Analysis

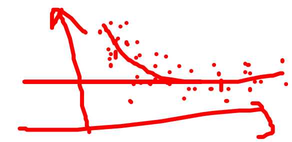
- Operating characteristics
  - average values for characteristics that describe performance of waiting line system
- Queue
  - a single waiting line

Hop

- Waiting line system
  - consists of arrivals, servers, and waiting line structure
- Calling population
  - source of customers; infinite or finite


Time—
the critical
measure of
quality service
Waiting
Line or
"Queue"
Served
Customer
Server
Arrivals
Served
Customer
Calling
Population;
source of
customers

Elements of Waiting Line Analysis



- Arrival rate (λ)
  - frequency at which customers arrive at a waiting line according to a probability distribution, usually Poisson
- Service rate (μ)
  - time required to serve a customer, usually described by negative exponential distribution
- Service rate must be higher than arrival rate ($\lambda < \mu$)
- Queue discipline
  - order in which customers are served
- Infinite queue
  - can be of any length; length of a **finite** queue is limited

Elements of Waiting Line Analysis

(b) Multiple-server waiting line

(a) Single-server waiting line

- Channels
  - number of parallel servers for servicing customers
- Phases
  - number of servers in sequence a customer must go through

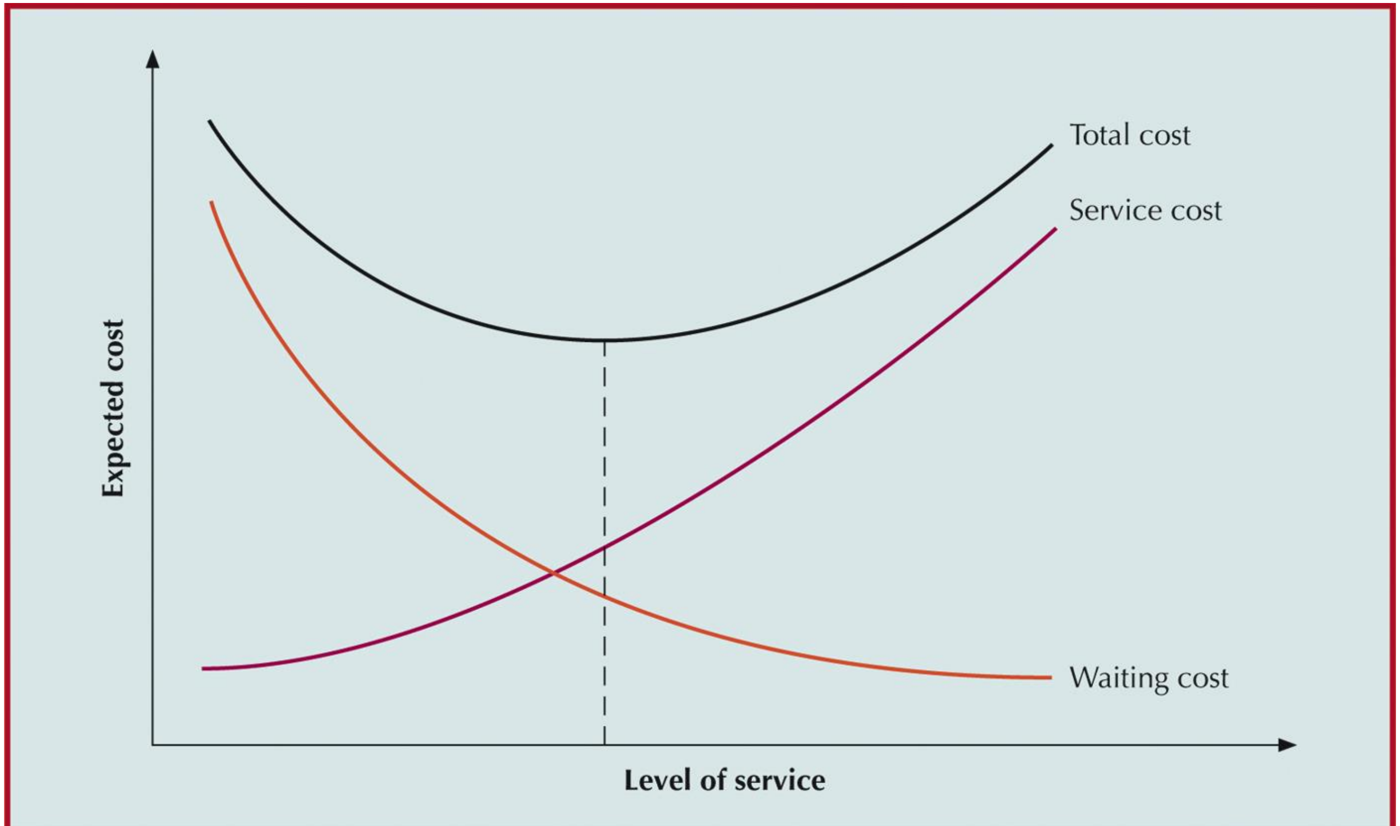
Operating Characteristics

- *Operating characteristics* are assumed to approach a *steady state*

Notation	Operating Characteristic
L	Average number of customers in the system (waiting and being served)
L_q	Average number of customers in the waiting line
W	Average time a customer spends in the system (waiting and being served)
W_q	Average time a customer spends waiting in line
P_0	Probability of no (i.e., zero) customers in the system
P_n	Probability of *n* customers in the system
ρ	Utilization rate; the proportion of time the system is in use

Traditional Cost Relationships

- As service improves, cost increases



Psychology of Waiting

- Waiting rooms
  - magazines and newspapers
  - televisions
- Bank of America
  - mirrors
- Supermarkets
  - magazines
  - “impulse purchases”

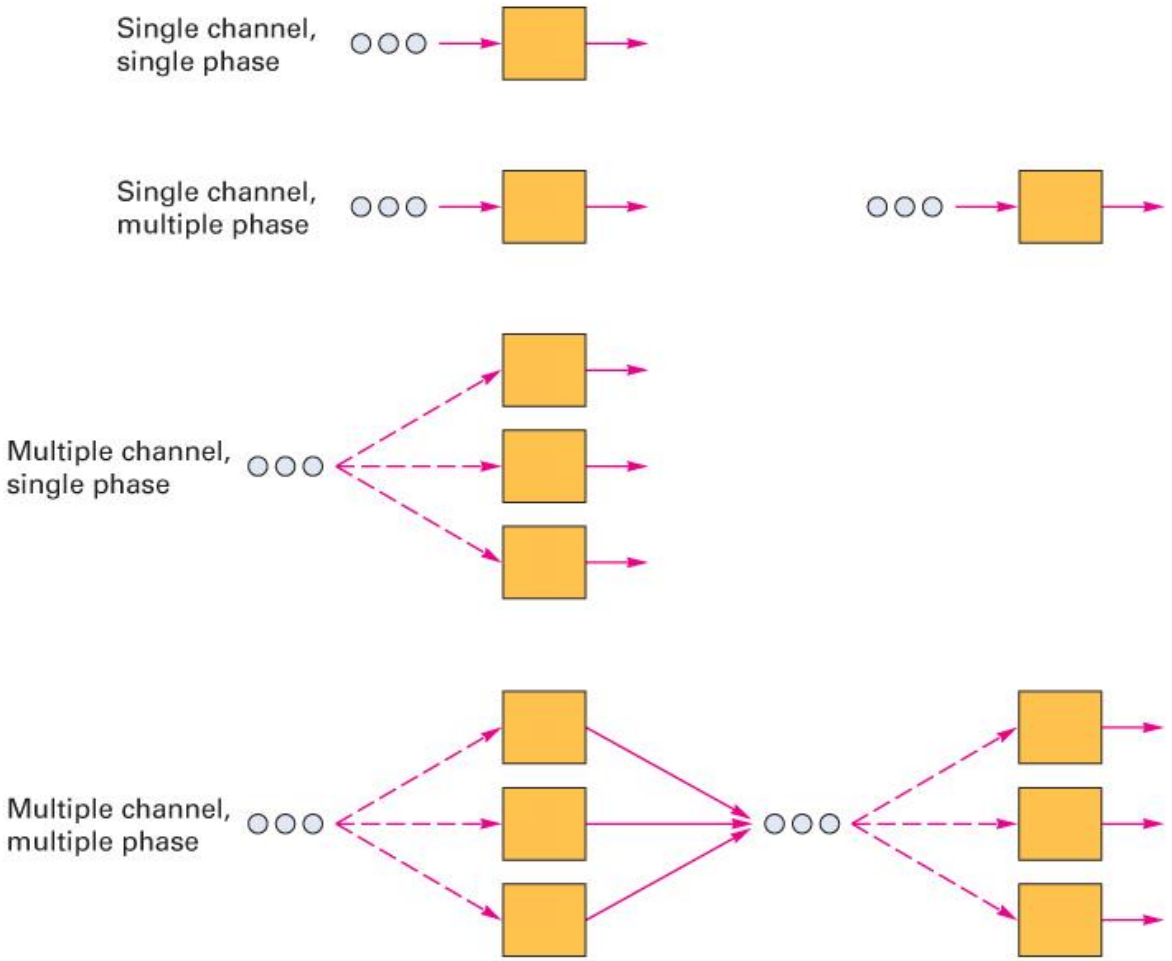
Psychology of Waiting

- Preferential treatment
  - Grocery stores: express lanes for customers with few purchases
  - Airlines/Car rental agencies: special cards available to frequent-users or for an additional fee
  - Phone retailers: route calls to more or less experienced salespeople based on customer's sales history
- Critical service providers
  - services of police department, fire department, etc.
  - waiting is unacceptable; cost is not important

Waiting Line Models

- *Single-server* model
  - simplest, most basic waiting line structure
- Frequent variations (all with Poisson arrival rate)
  - exponential service times
  - general (unknown) distribution of service times
  - constant service times
  - exponential service times with finite queue
  - exponential service times with finite calling population

Common Queuing Systems

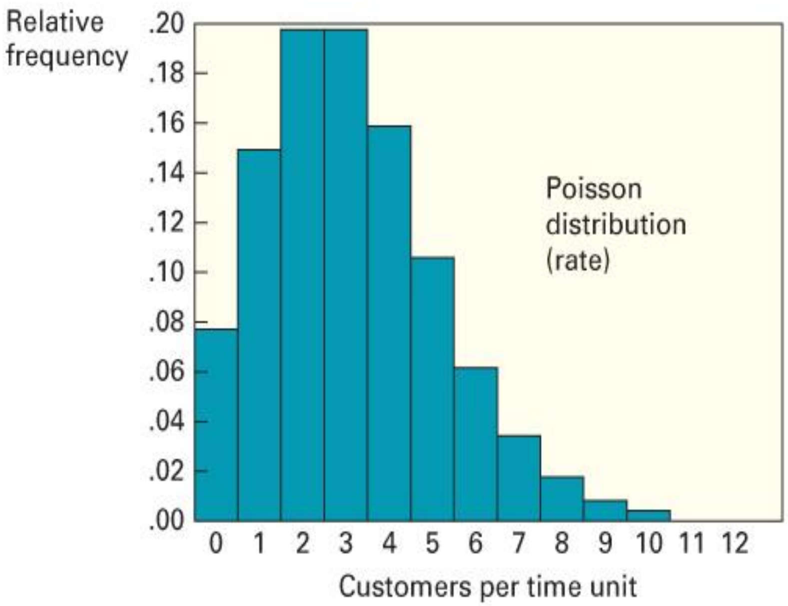


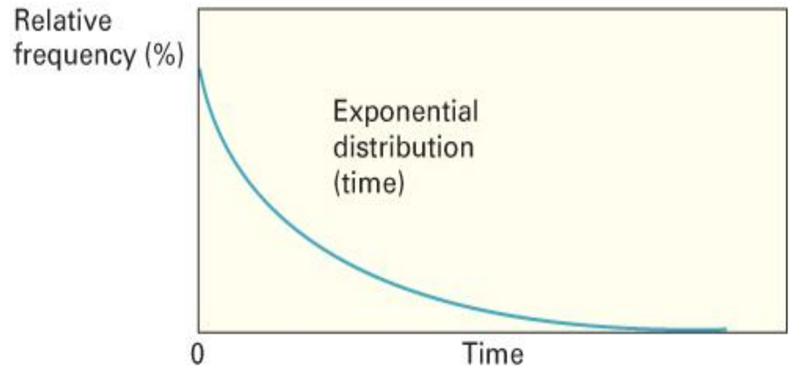
A few commercial service systems

Type of System	Customers	Servers
Barber Shop	People	Barber
Bank teller service	People	Teller
Gas station	Cars	Pump
Automobile repair shop	Car Owners	Mechanic
Production System	Jobs	Machine
Maintenance System	Machines	Repair Crew

- Customers and servers in a service system will be domain specific.
- Customers in one domain may be servers in the other domains.

Poisson and Exponential





$$p_{(x)} = \frac{\lambda^x \mathrm{e}^{-\lambda}}{x!}$$

Mean = λ

$$f(x) = \lambda e^{-\lambda x}$$

$$p_{(x \geq x_0)} = \mathrm{e}^{-\lambda x_0}$$

Mean = $1/\lambda$

- Average arrival rate of 3.2 customers per hour. The arrival rate can be modelled by a Poisson distribution.

1. What is the probability of 5 customers arriving?

2. Probability of having more than 7 customers?

- Average arrival rate of 3.2 customers per hour. The arrival rate can be modelled by a Poisson distribution.

1. What is the probability of 5 customers arriving?

   .1141

2. Probability of having more than 7 customers?

   .0169

Arrivals at a bank are poisson distributed with a mean of 1.2 customers every minute.

- What is the average time between arrivals?

- and what is the probability that at least 2 minutes will elapse between one arrival and the next arrival?

Arrivals at a bank are poisson distributed with a mean of 1.2 customers every minute.

- What is the average time between arrivals?

  .833

- and what is the probability that at least 2 minutes will elapse between one arrival and the next arrival?

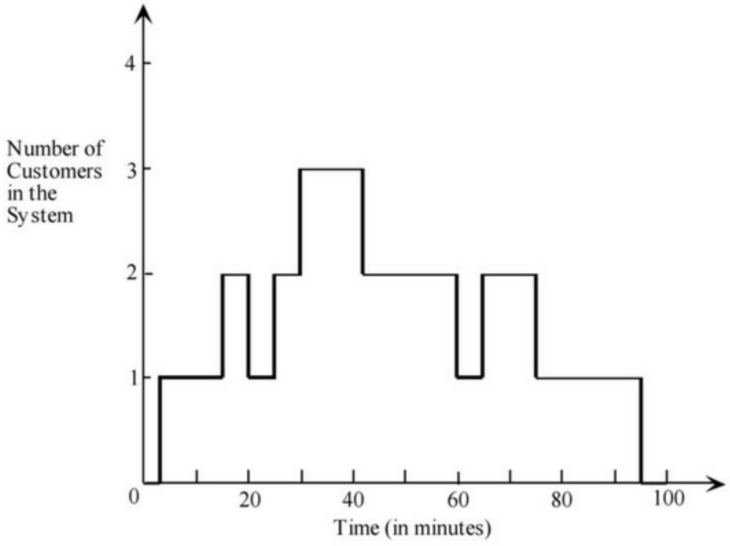
  0.907

Basic Single-Server Model

- Assumptions
  - Poisson arrival rate
  - exponential service times
  - first-come, first-served queue discipline
  - infinite queue length
  - infinite calling population
- Computations
  - λ = mean arrival rate
  - μ = mean service rate
  - n = number of customers in line

Elements of queuing systems

- Interarrival times
  - The time between consecutive arrivals
  - Most queueing models assume interarrival time to be exponential.
- Service time
  - When a customer enters service, the elapsed time from the beginning to the end of the service is referred to as the service time.
  - Exponential service times are easy to analyse.



Evolution of number of customers

Operating Characteristics

- *Operating characteristics* are assumed to approach a *steady state*

Notation	Operating Characteristic
L	Average number of customers in the system (waiting and being served)
L_q	Average number of customers in the waiting line
W	Average time a customer spends in the system (waiting and being served)
W_q	Average time a customer spends waiting in line
P_0	Probability of no (i.e., zero) customers in the system
P_n	Probability of n customers in the system
ρ	Utilization rate; the proportion of time the system is in use

Basic Single-Server Model

- probability that no customers are in queuing system

$$P_0 = \left(1 - \frac{\lambda}{\mu}\right)$$

- probability of n customers in queuing system

$$P_n = \left(\frac{\lambda}{\mu}\right)^n \cdot P_0 = \left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right)$$

- average number of customers in queuing system

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

- average number of customers in waiting line

$$L_q = \frac{\lambda^2}{\mu\,(\mu - \lambda)}$$

Operating Characteristics

- *Operating characteristics* are assumed to approach a *steady state*

Notation	Operating Characteristic
L	Average number of customers in the system (waiting and being served)
L_q	Average number of customers in the waiting line
W	Average time a customer spends in the system (waiting and being served)
W_q	Average time a customer spends waiting in line
P_0	Probability of no (i.e., zero) customers in the system
P_n	Probability of n customers in the system
ρ	Utilization rate; the proportion of time the system is in use

Basic Single-Server Model

- average time customer spends in queuing system

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

- average time customer spends waiting in line

$$W_q = \frac{\lambda}{\mu\,(\mu - \lambda)}$$

- probability that server is busy and a customer has to wait (utilization factor)

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

- probability that server is idle and customer can be served

$$I = 1 - \rho$$

$$= 1 - \frac{\lambda}{\mu} = P_0$$

Basic Single-Server Model Example

$\lambda = 24$

$\mu = 30$

$P_0 =$

$L =$

$L_q =$

Basic Single-Server Model Example

$\lambda = 24$

$\mu = 30$

$$P_0 = \left(1 - \frac{\lambda}{\mu}\right) = \left(1 - \frac{24}{30}\right)$$

$= 0.20$ probability of no customers in the system

$$L = \frac{\lambda}{\mu - \lambda} = \frac{24}{30 - 24}$$

$= 4$ customers on the average in the queuing system

$$L_q = \frac{\lambda^2}{\mu(\mu - \lambda)} = \frac{(24)^2}{30(30 - 24)}$$

$= 3.2$ customers on the average in the waiting line

Basic Single-Server Model Example

W =

Wq =

ρ =

I =

Basic Single-Server Model Example

$$W = \frac{1}{\mu - \lambda} = \frac{1}{30 - 24}$$

$= 0.167$ hour (10 minutes) average time in the system per customer

$$W_q = \frac{\lambda}{\mu(\mu - \lambda)} = \frac{24}{30(30 - 24)}$$

$= 0.133$ hour (8 minutes) average time in the waiting line per customer

$$\rho = \frac{\lambda}{\mu} = \frac{24}{30}$$

$= 0.80$ probability that the server will be busy and the customer must wait

$I = 1 - \rho = 1 - 0.80$

$= 0.20$ probability that the server will be idle and a customer can be served

Service Improvement Analysis

- Waiting time (8 min.) is too long
  - hire assistant for cashier?
    - increased service rate
  - hire another cashier?
    - reduced arrival rate
- Is improved service worth the cost?

Summary

IMPORTANCE OF QUEUES IN SERVICE DESIGN

ELEMENTS OF QUEUING SYSTEMS AND USUAL MODEL ASSUMPTIONS.

Other measures of performance

Other than averages: worst case scenarios

What will be the maximum number of customers in the system?

What will be the maximum waiting time of customers in the system?

To answer the above, we need to know

Steady-state probability of having exactly n customers in the system.

$P(W \leq t)$: Probability the time spent in the system will be no more than t.

Example of common goals

No more than 5% of customers wait more than 2 hours:

- $P(W \leq 2 \text{ hours}) \geq 0.95$

No more than three customers 95% of the time

Summary

Measuring Performance of a queueing system.

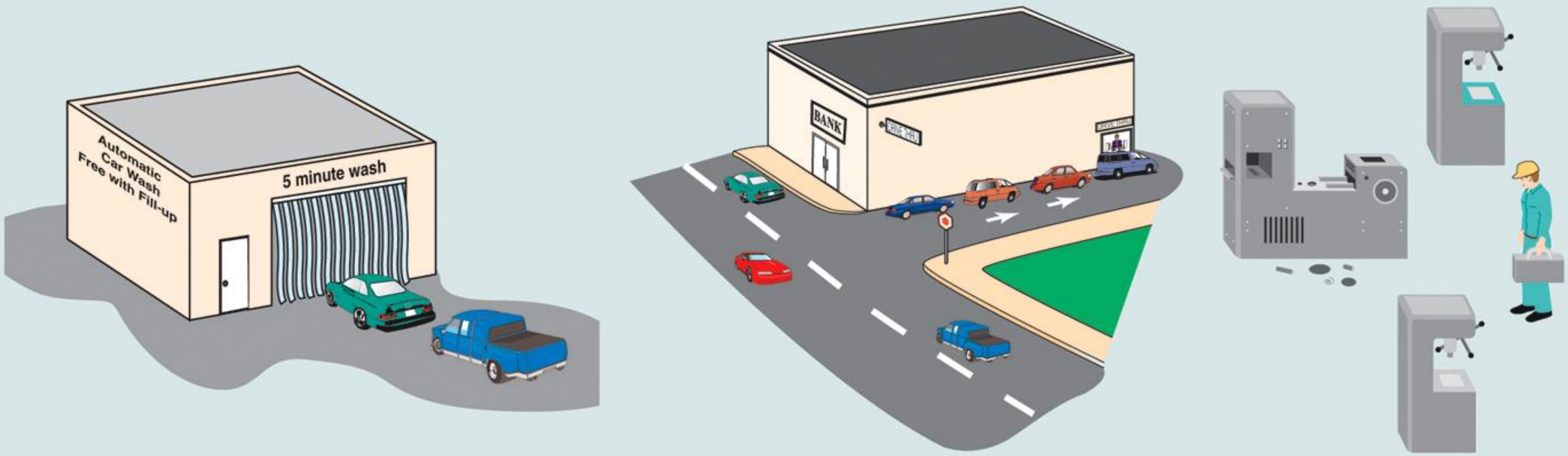
Expected measure of performance

Worst case measure of performance

Advanced Single-Server Models

- Constant service times
  - occur most often when automated equipment or machinery performs service
- Finite queue lengths
  - occur when there is a physical limitation to length of waiting line
- Finite calling population
  - number of “customers” that can arrive is limited

Advanced Single-Server Models



(*a*) Single-Server, Constant Time

(*b*) Single-Server, Finite Queue

(*c*) Single-Server, Finite Calling Population