Operations & Logistics Management in Air Transportation

Professor David Gillen (University of British Columbia) &
Professor Benny Mantin (University of Waterloo)

Istanbul Technical University
Air Transportation Management
M.Sc. Program

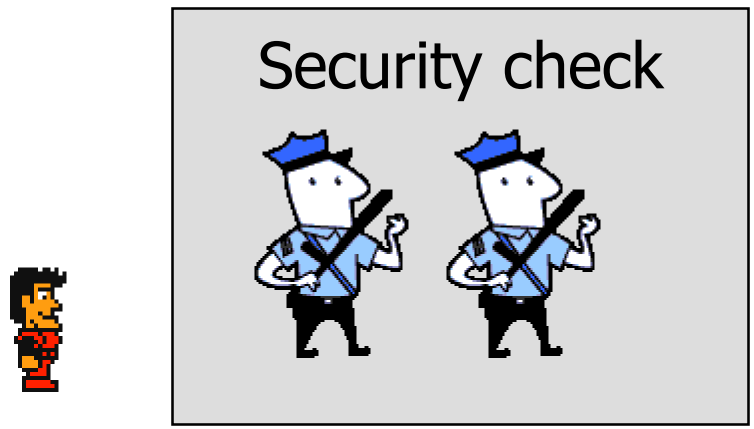
Air Transportation Systems and Infrastructure
Strategic Planning
Module 5-6 : 11 June 2014

VARIABILITY
IN PROCESSES AND QUEUES

LEARNİNG OBJECTİVES

- Variability and Process Analysis
  - What is *variability*?
  - What *impact* does variability have on processes?
  - How can we *quantify* the impact of variability on processes?
  - How can we *manage* variability in processes?

WHAT İS VARİABİLİTY?



Variability comes from …

Variable input

Variable Capacity

TYPES OF VARIABILITY

Predictable Variability	Unpredictable Variability
… refers to "knowable" changes in input and/or capacity rates	… refers to "unknowable" changes in input and/or capacity rates
Demand of pumpkins will go up during Thanksgiving	Supply of pumpkins will go down ***if*** the crop fails

- Both types of variability exist simultaneously
  - Pumpkin sales will go up during Thanksgiving, but we do not know the exact sales of pumpkins

Predictable Variability

Can be *controlled* by making changes to the system

- We could increase or decrease the demand for pumpkins by increasing or decreasing the price
- Restaurants will add staff during peak demand (lunch, dinner, etc.)

Unpredictable Variability

Is the result of the *lack* of knowledge or information

- Usually can be expressed with a probability distribution
- E.g., Express the probability that the pumpkin crop will fail using a probability distribution

Can be *reduced* by gaining more knowledge or collection information

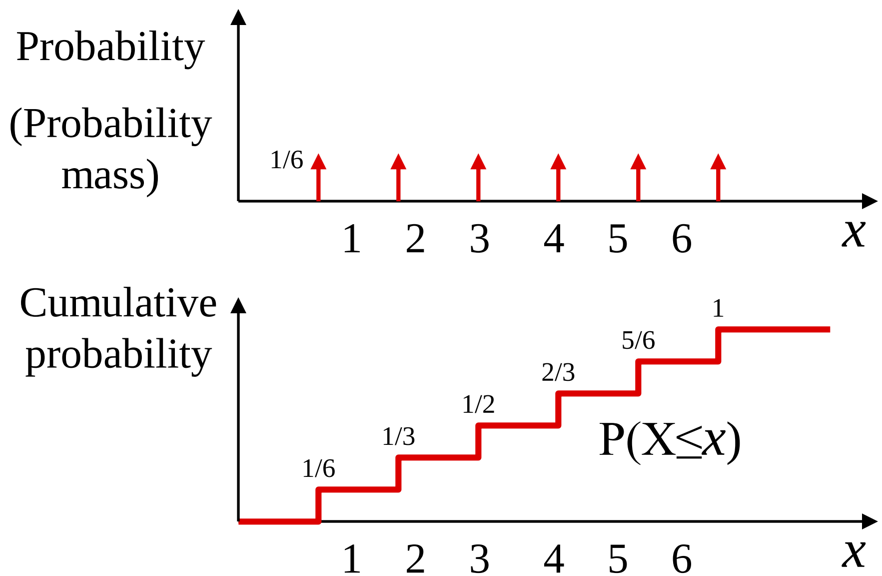
- By paying close attention to weather patterns, we could increase the accuracy of our prediction that the pumpkin crop will fail

SHORT REVİEW ON PROBABİLİTY (1)

Discrete Random Variable and Probability

Throw a dice; the number you get is a discrete random variable:

$$X = \begin{cases} 1, & \text{w.p. } 1/6 \\ 2, & \text{w.p. } 1/6 \\ 3, & \text{w.p. } 1/6 \\ 4, & \text{w.p. } 1/6 \\ 5, & \text{w.p. } 1/6 \\ 6, & \text{w.p. } 1/6 \end{cases}$$



$P\{X = 2\} = 1/6$

$P\{X \leq 2\} = P\{X=1 \text{ or } X=2\} = 1/3$

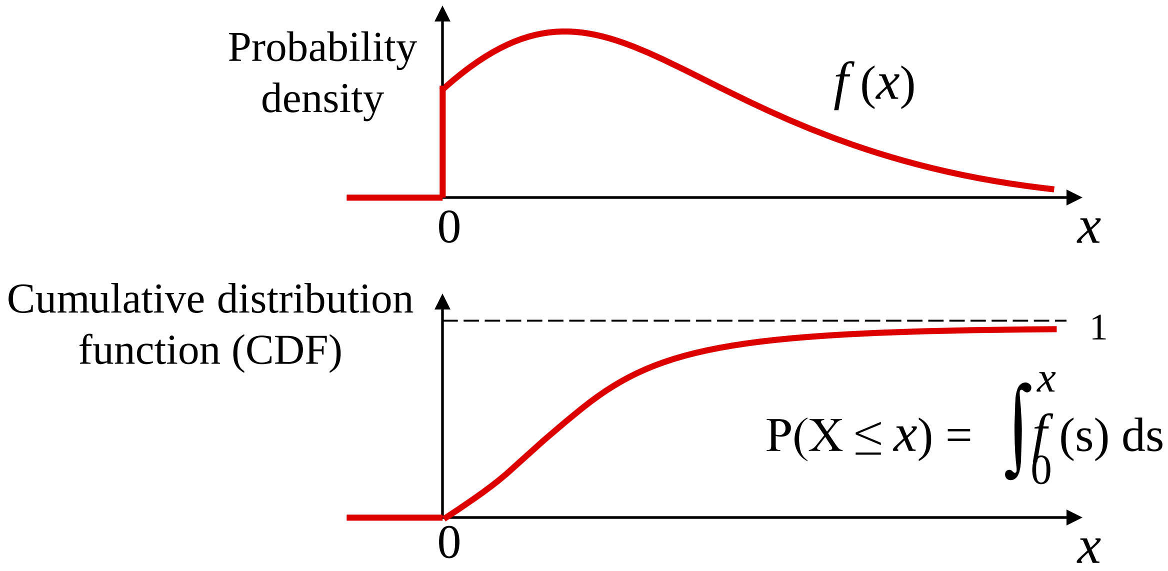
$P\{X \leq 2.1\} = P\{X=1 \text{ or } X=2\} = 1/3$

$P\{X \leq x\}$ is a function of x, called the **cumulative distribution function (CDF)**

Short Review on Probability (2)

Continuous Random Variable and Probability

The time between two customers' arrival times is a continuous random variable



BASİC QUESTİONS

- What are the effects of variability on processes
  - In particular, how does variability affect

Average Throughput Rate

Average Inventory

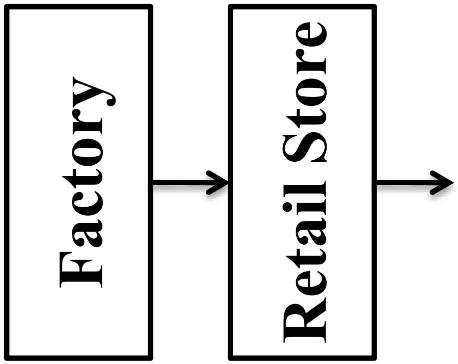
Average Flow Time

- If the effects are negative, how can we deal with it?

THE "MAKE-TO-ORDER" DICE GAME

- Retail store will roll dice first to observe demand, which will be communicated to the factory
- Factory will roll dice to observe capacity
- Factory will satisfy retailer demand, but is constrained by realized capacity
  - For example, if demand is 3 and capacity is 4, then factory will give the retail store 3 units
  - But if demand is 5 and capacity is 4, then factory will give the retail store only 4 units
  - No backlog
- Assume 1 roll of demand and capacity

  = 1 day



What is the average demand?

What is the average capacity?

CONSİDER A PROCESS WİTH NO VARİABİLİTY

Input
(1 person/min)

ATM

Throughput Rate?

Service time
(exactly 1 min)

- Assume that all customers are **identical**
- Customers arrive exactly 1 minute apart
- The service time is exactly 1 minute for all the customers

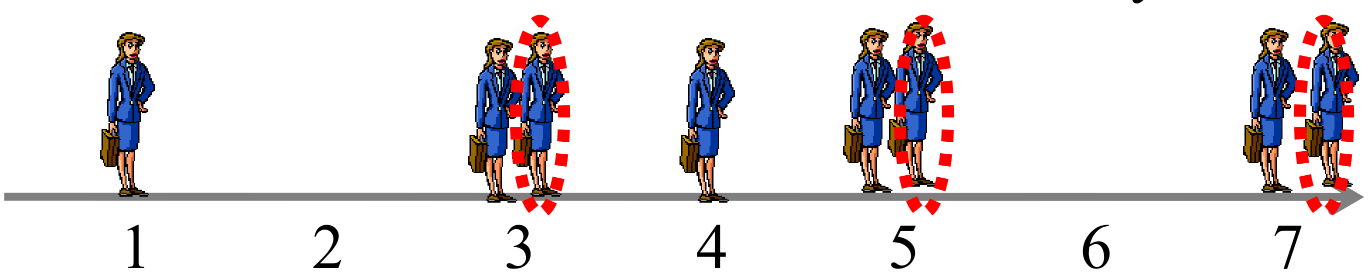
EFFECT OF INPUT VARIABILITY (NO BUFFER)

Random Input
0, 1, 2
customers/min
(with equal probability)

→ ATM → **Throughput Rate?**

Service time
(exactly 1 min)

- Assume that customers who find the ATM busy <u>do not wait</u>



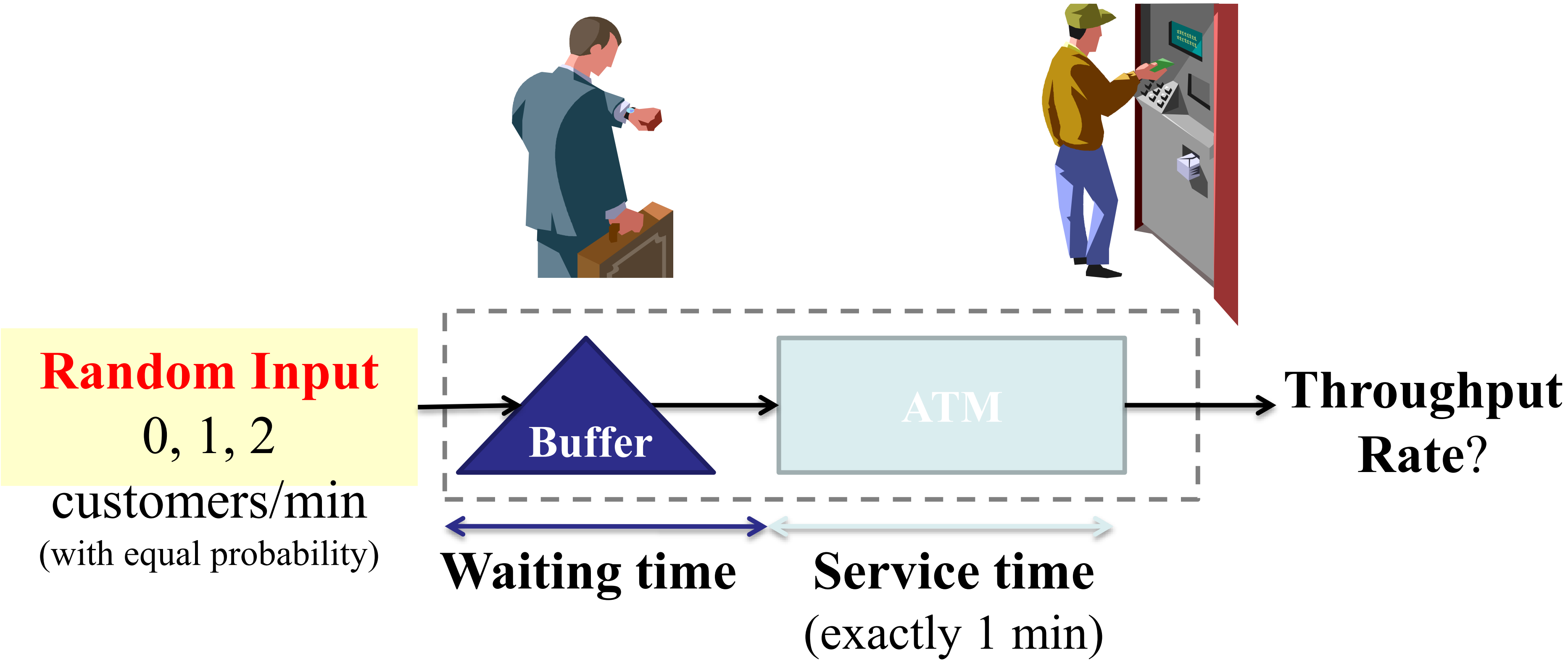
EFFECT OF INPUT VARİABİLİTY (**NO BUFFER**)

- When a process faces input variability, and a buffer cannot be built, some input may get lost
- Input variability ***can*** reduce the throughput
- Lower throughput means
  - Lost customers; lost revenue
  - Customer dissatisfaction
  - Less utilization of resources
- Little's Law holds

DEALİNG WİTH VARİABİLİTY

- When the arrival rate of customers is unpredictable, what could you do to increase throughput?

Effect of Input Variability (with buffer)



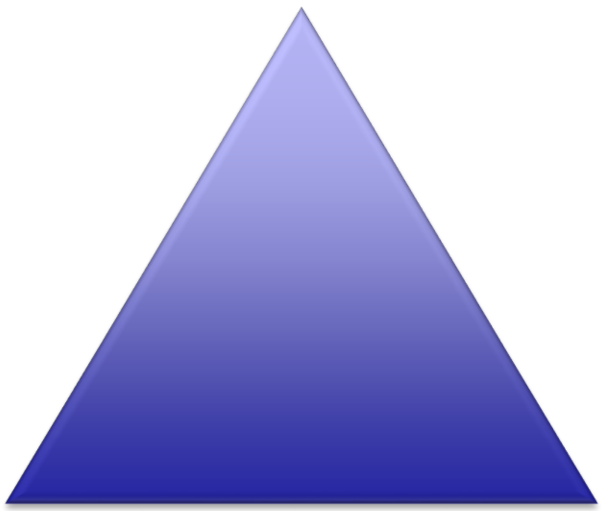
- Now assume that **customers wait**

  We can **build-up an inventory buffer**

EFFECT OF INPUT VARİABİLİTY (**WİTH BUFFER**)

- *If we can build up an inventory buffer*,
  variability leads to
  - An increase in the **average inventory** in the process
  - An increase in the **average flow time**
- Little’s Law holds

THE OM TRIANGLE

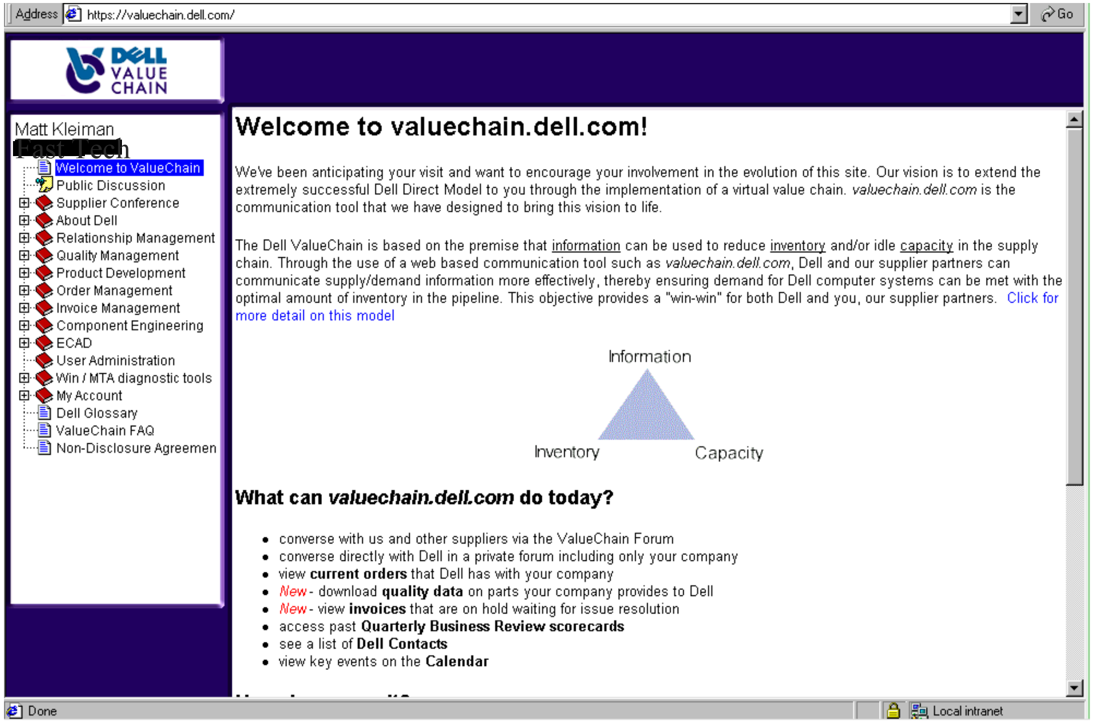


If a firm is striving to meet the *random* demand, then it can use **capacity**, **inventory** and **information** (variability reduction) as substitutes

You cannot have low inventory, low capacity, and low information acquisition effort at the same time. This is a trade-off.

OPERATIONS AT DELL

- **Inventory as “the physical embodiment of bad information”** (a senior exec at Dell)
- Substitute information for inventory
- Less inventory =>higher inventory turns



QUANTİFYİNG VARİABİLİTY

- So far, we focused on **qualitative** effect of variability
  - Without buffer, input may get lost and throughput may decrease
  - With buffer, queue may build up, flow time may increase

- But …
  - How long is the queue on average?
  - How long does a customer have to wait?

Why is it important to quantify variability and its impact?

These quantitative measures of process performance are important to any functions of a company

Marketing

Wants to use the short waiting time as a selling point

Finance

Wants to attract investors based on excellent operations performance

Accounting

Wants to know how much money is tied up in the queue

Operations

Wants to shorten the queue, and wants to quantify the trade-offs between capacity, inventory and variability

What is the impact (on inventory and flow time) of increasing/decreasing capacity by 10%?

First Steps in Quantifying Variability

- Probability Statements

  $P(X=4)$

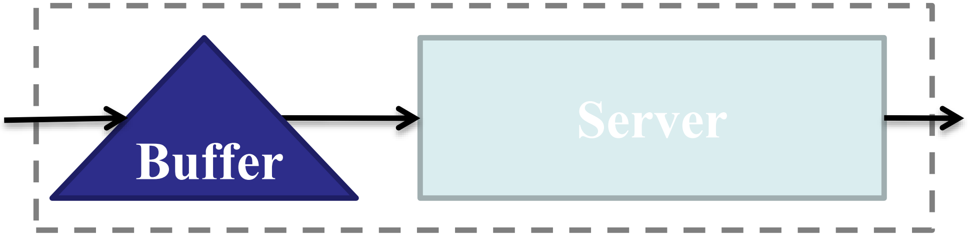
  $P(20 < T \leq 30)$

- Variances and Standard deviation
  - These lead to probability statements

- Coefficient of variation (CV)

$$CV = \frac{\text{Standard Deviation}}{\text{Mean}}$$

A SINGLE SERVER PROCESS



Process Boundary

A queue forms in a buffer

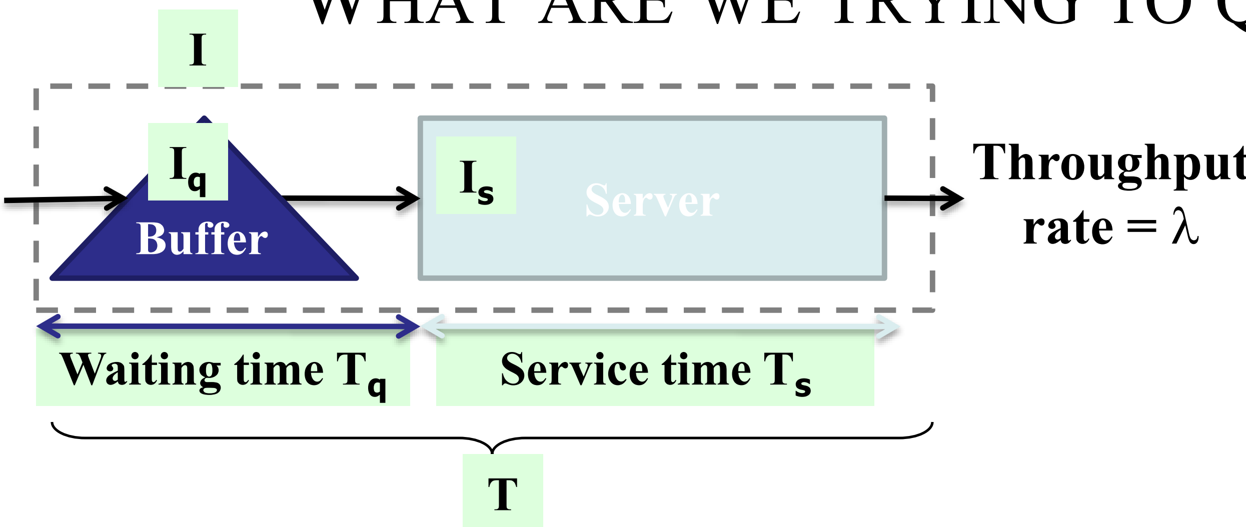
Note: We are focusing on long-run averages, ignoring the predictable variability that may be occurring in the short run. In reality, we should be concerned with both types of variability

λ	Long-run average input rate
$1/\lambda$	(Average) Customer inter-arrival time
μ	Long-run average processing rate of a single server
$1/\mu$	Average processing time by one server

A single phase service system is stable whenever $\lambda < \mu$

K	Buffer capacity (for now, let $K = \infty$)
c	Number of servers in the resource pool (for now, let c=1)

WHAT ARE WE TRYİNG TO QUANTİFY?



Little's Law holds

$$I_q = \lambda T_q$$
$$I_s = \lambda T_s$$
$$I = \lambda T$$

System Characteristics

Utilization	ρ (In a stable system, $\rho = \lambda/\mu < 100\%$)
Safety Capacity	$\mu - \lambda$

Performance Measures

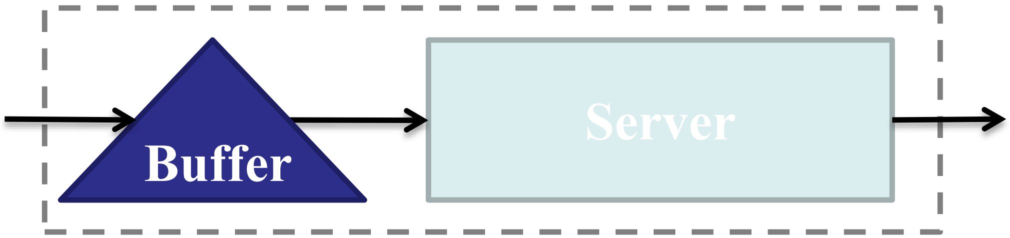
T_q	Average waiting time (in queue)
I_q	Average queue length
T_s	Average time spent at the server
I_s	Average number of customers being served
$T=T_q+T_s$	Average flow time (in process)
$I=I_q+I_s$	Average number of customers in the process

QUICK "QUIZ"

Assumption: $\lambda \leq \mu$

Service rate: μ
persons/min
(average capacity rate)

Arrival rate:
λ persons/min
(average input rate)



Average throughput rate
λ persons/min

- Average number of persons in the system:

$$I = I_q + I_s$$

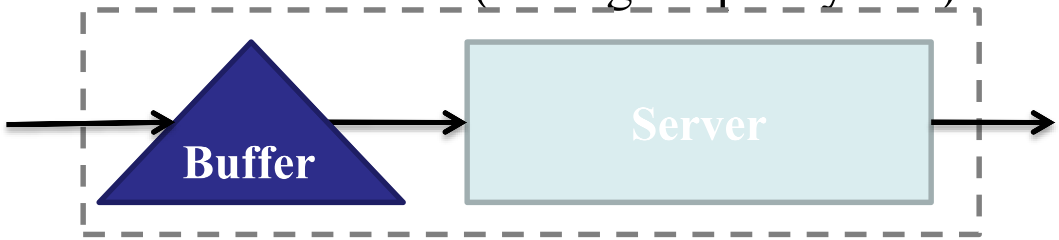
- Question: I_s=??? (Express I_s in terms of λ and μ)

SİNGLE-SERVER QUEUİNG MODEL

Assumption: $\lambda \leq \mu$

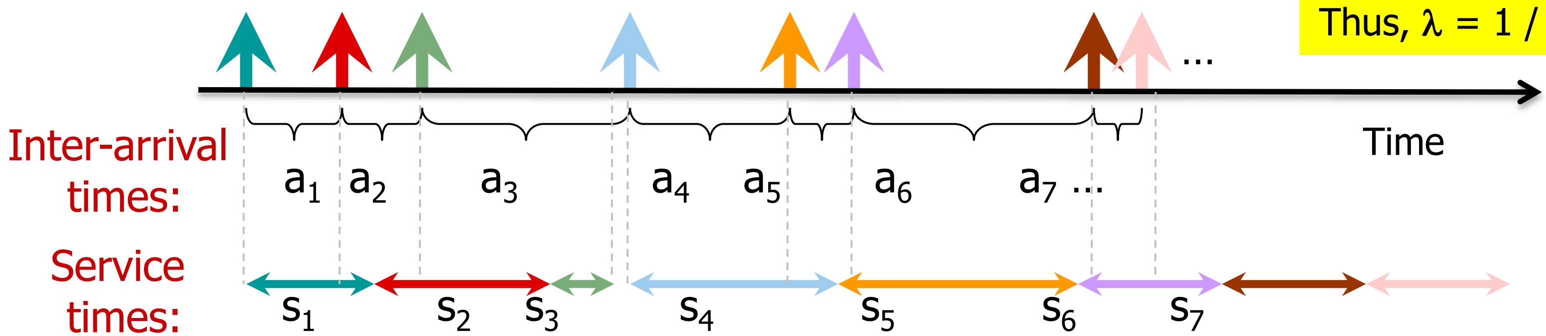
Service rate: μ persons/min
(average capacity rate)

Arrival rate:
λ persons/min
(average input rate)



Average throughput rate
λ persons/min

On average, 1 person arrives every E{a} min. Thus, $\lambda = 1 / E\{a\}$



On average, 1 person can be served every E{s} min. Thus, $\mu = 1 / E\{s\}$

Pollaczek-Khinchin (PK) Formula

$$I_q \cong \frac{\rho^2}{1-\rho} \times \frac{C_a^2 + C_s^2}{2}$$

"=" for special cases

"≈" in general

I_q	Average queue length (excl. the one in service)
ρ	(Long run) Average utilization = Average Throughput / Average Capacity = λ / μ
$C_a = \sigma\{a\}/E\{a\}$	Coefficient of variation (CV) of inter-arrival times
$C_s = \sigma\{s\}/E\{s\}$	Coefficient of variation (CV) of service times

PK Formula and OM Triangle

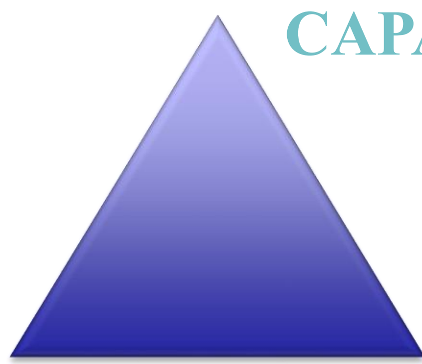
$$I_q \cong \frac{\rho^2}{1-\rho}\times\frac{C_a^2+C_s^2}{2}=\frac{\lambda}{\mu}\times\frac{\lambda}{\mu-\lambda}\times\frac{C_a^2+C_s^2}{2}$$

μ= Capacity Rate

λ = Input Rate

Variability

CAPACITY

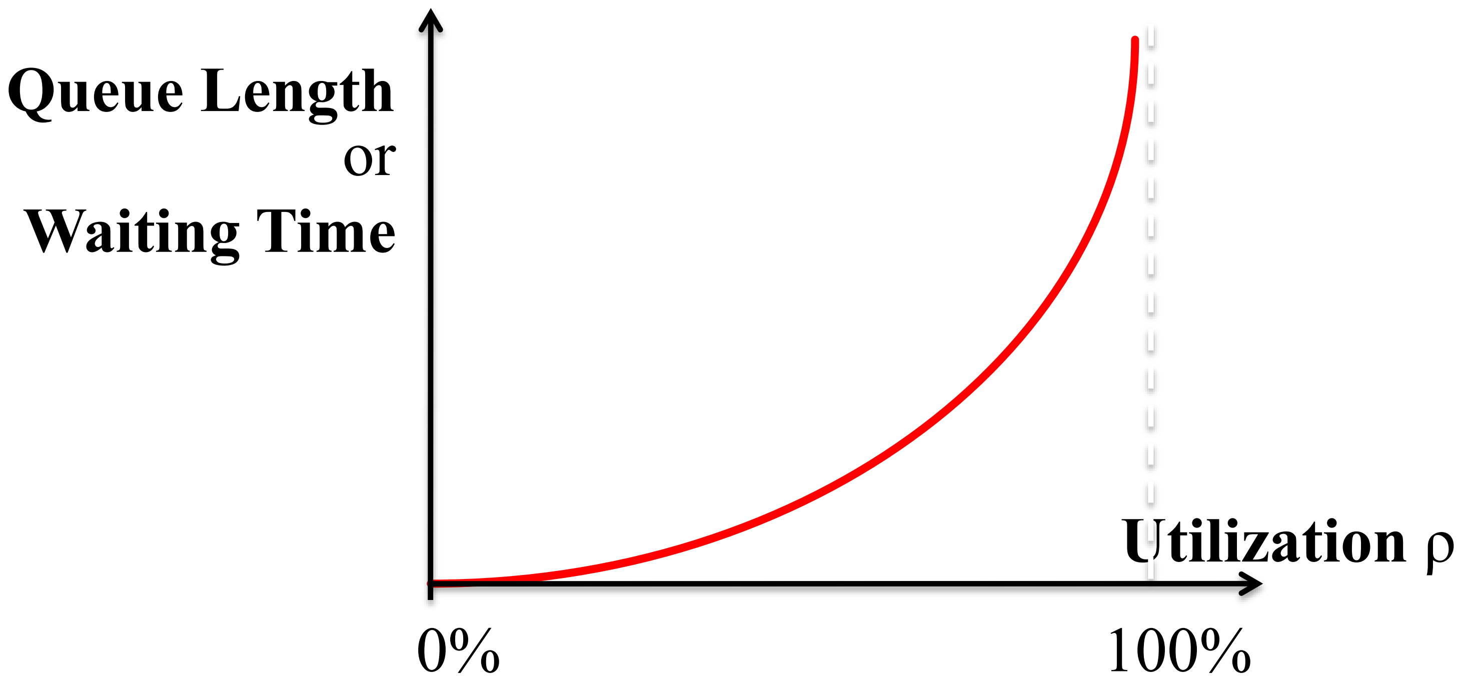


INVENTORY

INFORMATION

IMPACT OF UTİLİZATİON ($\rho = \lambda/\mu$)

Impact on Queue Length (Inventory)	Impact on Waiting Time (Flow Time)
$I_q \cong \frac{\rho^2}{1-\rho} \times \frac{C_a^2 + C_s^2}{2}$	$T_q = I_q / \lambda$ **Little's Law**



UTİLİZATİON

$$\text{Utilization} = \frac{\text{Throughput Rate}}{\text{Capacity Rate}} = \frac{\text{Actual output rate}}{\text{maximum output rate}} \leq 100\%$$

- Utilization gives us information about "excess capacity"
- The utilization of each resource in a process can be presented with a ***utilization profile***

Resource	Capacity Rate (units/hour)	Input Rate (units/hour)	Utilization
1	6	4	66.67%
2	7	4	57.14%
3	8	4	50.00%
4	6	4	66.67%
5	5	4	80.00%

- What is the optimal utilization of a resource?

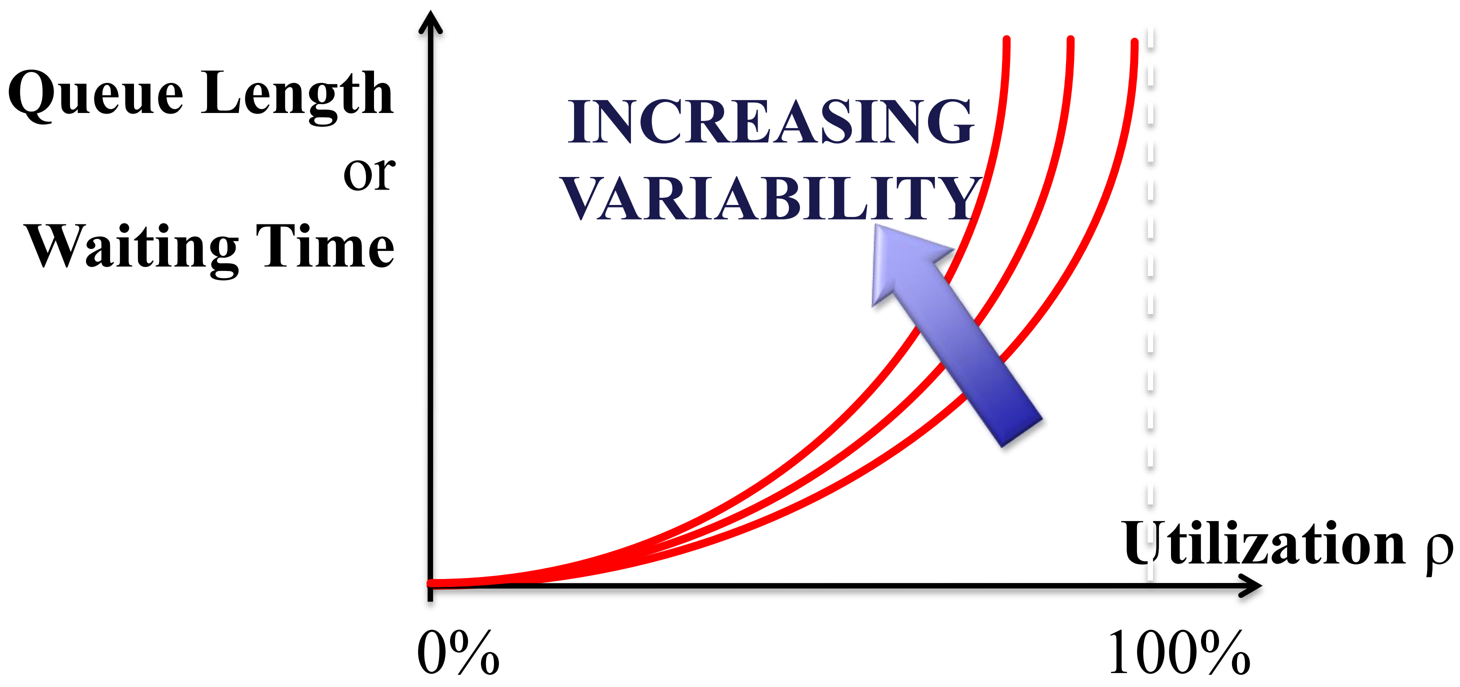
Utilization: An Important Insight

With No Variability	With Variability
• Maximizing utilization is a good idea in a process with no variability	• Maximizing utilization is a very bad idea in a process with variability • What is the correct utilization for a resource when variability is present? • It depends … on the amount of variability, the sensitivity to delay, etc.

Impact of Variability

$$I_q \cong \frac{\rho^2}{1-\rho} \times \frac{C_a^2 + C_s^2}{2}$$

$$T_q = I_q / \lambda$$ **Little's Law**



QUEUİNG THEORY

The PK formula given above comes from "queuing theory", the study of queues

The version of PK formula we used above makes the following assumptions

Assumptions

- Single server
- Single queue
- No limit on queue length
- All units that arrive enter the queue system stays in the queue till served
  (No units "balk" at the length of the queue)
- First-in-first-out (FIFO)
- All units arrive independently of each other

QUEUİNG NOTATİON: G/G/1 QUEUE

- The queue we studied above is called a

G/G/1 queue

The first **"G"** refers to the fact that the "**arrivals**" follows a "general" (probability) distribution
The second **"G"** refers to the fact that the "**service time**" follows a "general" (probability) distribution
The **"1"** refers to the fact that there is a **single server**

- Using observed data, get estimates for C_a and C_a

$C_a = \sigma\{a\}/E\{a\}$	Coefficient of variation of inter-arrival times
$C_s = \sigma\{s\}/E\{s\}$	Coefficient of variation of service times

SİMPLE EXAMPLE

- Customers arrive at rate 4/hour, and mean service time is 10 minutes
- Assume that standard deviation of inter-arrival times equals 5 minutes, and the standard deviation of service time equals 3 minutes
- What is the average size of the queue? What is the average time that a flow unit spends in the queue?

$$\lambda = 4 \qquad E[a] = 1/4 \text{ hour}$$

$$\mu = 6 \qquad E[s] = 1/6 \text{ hour}$$

$$\rho = \lambda/\mu = 4/6 = 2/3$$

$$\sigma[a] = 1/12 \text{ hour} \qquad C_a = \frac{\sigma[a]}{E[a]} = \frac{1/12}{1/4} = \frac{1}{3}$$

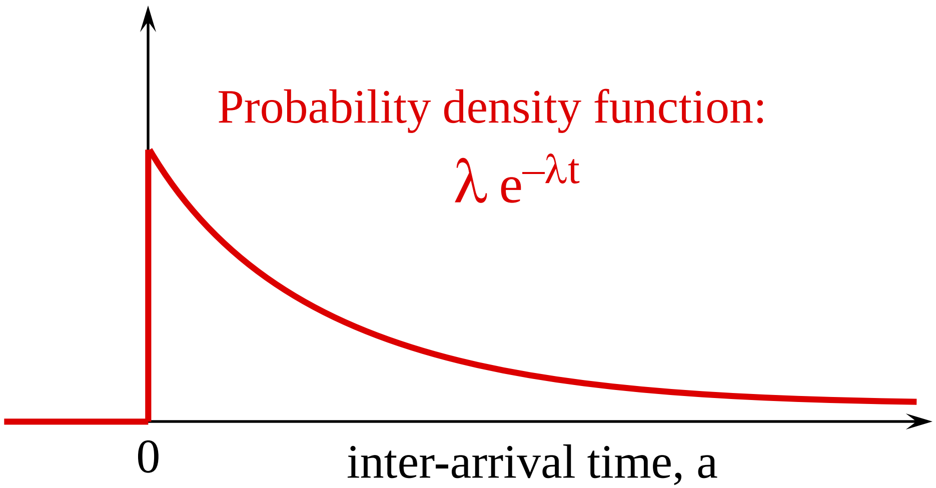
$$\sigma[s] = 1/20 \text{ hour} \qquad C_s = \frac{\sigma[s]}{E[s]} = \frac{1/20}{1/6} = \frac{3}{10}$$

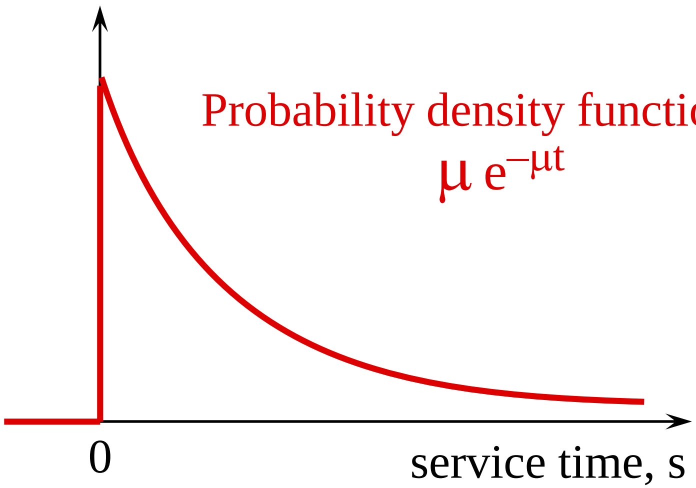
$$I_q \cong \frac{\rho^2}{1-\rho} \times \frac{C_a^2 + C_s^2}{2} = \frac{(2/3)^2}{1/3} \times \frac{(1/3)^2 + (3/10)^2}{2}$$

$$T_q = I_q/\lambda = I_q/4$$

WHAT İF WE DON'T HAVE DATA ABOUT THE PROCESS?

- Suppose you start a service business. You haven't seen the actual customers arrival process, but you want to have some idea about the queue you will be facing.
- Need to make some **assumptions** about the customer arrival process, and service time distribution
- A mostly commonly used distribution is the **exponential distribution**





WHY USE THESE ASSUMPTİONS?

- In many situations, the exponential distribution assumption is a good approximation for what really happens
- Easy to analyze because coefficient of variation (CV) is 1 for exponential distributions

$$\mathrm{CV} = \frac{\text{Standard Deviation}}{\text{Mean}}$$

- Recall the P-K formula

$$I_q \cong \frac{\rho^2}{1-\rho} \times \frac{C_a^2 + C_s^2}{2} = \boxed{????}$$

M/M/1 QUEUE

The first **"M"** indicates that the **inter-arrival times** are **exponentially** distributed
The second **"M"** indicates that the **service times** are **exponentially** distributed
The **"1"** refers to the fact that there is a **single server**

- Assume First-Come First-Serve (FCFS) rule
- For M/M/1 queue, the P-K formula is *exact* (=, not ≈)

$$I_q = \frac{\rho^2}{1-\rho} = \frac{\lambda^2}{\mu(\mu-\lambda)}$$

- Average waiting time in queue

(Little's Law) $T_q = I_q / \lambda$

$I = I_q + I_s = ???$
$T = T_q + T_s = ???$

SİMPLE EXAMPLE

- Customers arrive at rate 4/hour, and mean service time is 10 minutes

- What is the average size of the queue? What is the average time that a flow unit spends in the queue?

$$\lambda = 4 \qquad E[a] = 1/4 \text{ hour}$$

$$\mu = 6 \qquad E[s] = 1/6 \text{ hour}$$

$$\rho = \lambda/\mu = 4/6 = 2/3$$

PRACTICE PROBLEM

Professor Longhair holds office hours everyday to answer students' questions.

Students arrive at an average rate of 50 per hour.

Professor Longhair can process students at an average rate of 60 per hour.

What is the average number of students waiting outside Professor Longhair's office, and how long do they wait on average?

Assume the inter-arrival time and the service time are both exponentially distributed

(We can also say that the arrival rate follows a **Poisson** distribution)

$$\lambda = 50$$

$$\mu = 60$$

$$\rho = \lambda/\mu = 50/60 = 5/6$$

$$I_q = \frac{\lambda^2}{\mu(\mu-\lambda)} = \frac{50^2}{60(60-50)} = \frac{25}{6}$$

$$T_q = \frac{I_q}{\lambda} = \frac{25/6}{50} = \frac{1}{12}$$

M/D/1 QUEUE

The first **"M"** indicates that the **inter-arrival times** are **exponentially** distributed
The second **"D"** indicates that the **service times** are a **constant**
The **"1"** refers to the fact that there is a **single server**

- Assume First-Come First-Serve (FCFS) rule
- For M/D/1 queue, the P-K formula gives

$$I = I_q + I_s = ???$$

$$I_q = \frac{\rho^2}{1-\rho} \times \frac{1}{2} = \frac{\lambda^2}{2\mu(\mu-\lambda)}$$

- Average waiting time in queue

(Little's Law) $$T_q = I_q / \lambda$$

$$T = T_q + T_s = ???$$

SİMPLE EXAMPLE

- Customers arrive at rate 4/hour, and mean service time is ***exactly*** 10 minutes

- What is the average size of the queue? What is the average time that a flow unit spends in the queue?

$$\lambda = 4 \qquad E[a] = 1/4 \text{ hour}$$

$$\mu = 6 \qquad E[s] = 1/6 \text{ hour}$$

$$\rho = \lambda/\mu = 4/6 = 2/3$$

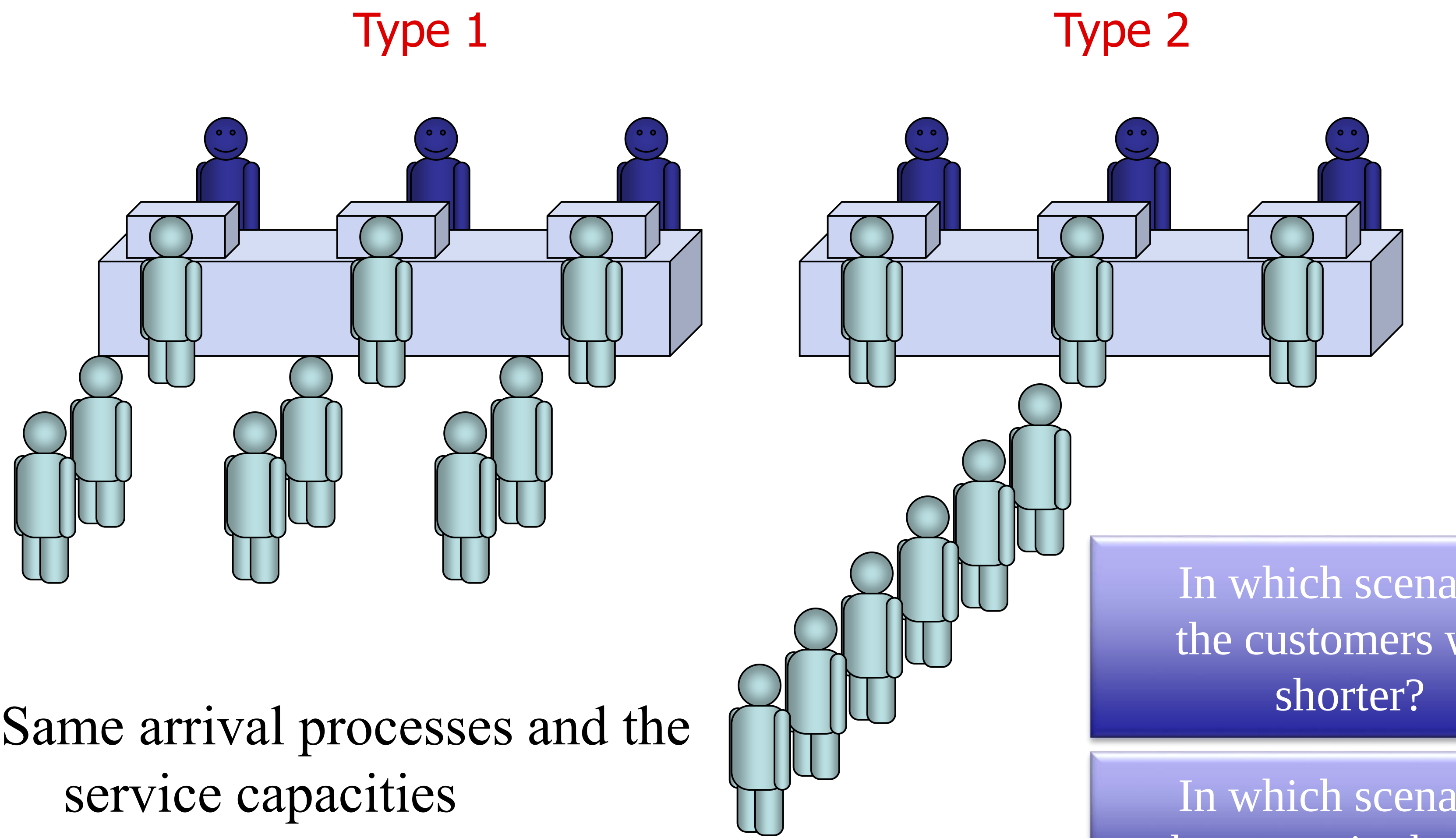
$$I_q = \frac{\lambda^2}{2\mu(\mu-\lambda)} = \frac{4^2}{2\times 6(6-4)} = \frac{2}{3}$$

$$T_q = \frac{I_q}{\lambda} = \frac{2/3}{4} = \frac{1}{6}$$

OTHER TYPES OF QUEUES

- Multiple servers
- Limited buffer size

WHİCH TYPE OF QUEUE DO YOU PREFER?



Same arrival processes and the service capacities

In which scenario, the customers wait shorter?

In which scenario, the queue is shorter?

MULTİ-SERVER QUEUİNG MODEL

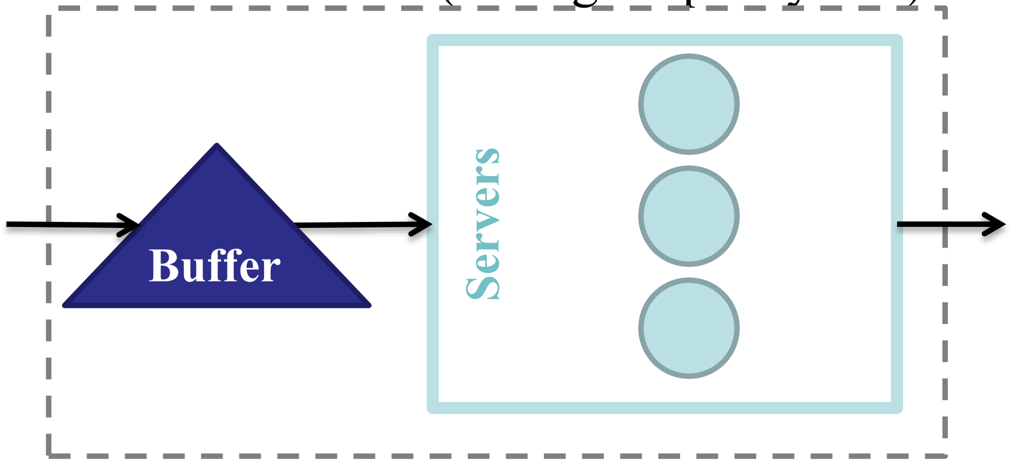
c = number of servers

Assumption: $\lambda \leq c\mu$

$\rho = \lambda / c\mu$

Service rate (per server):
μ persons/min
(average capacity rate)

Arrival rate:
λ persons/min
(average input rate)



Average throughput rate
λ persons/min

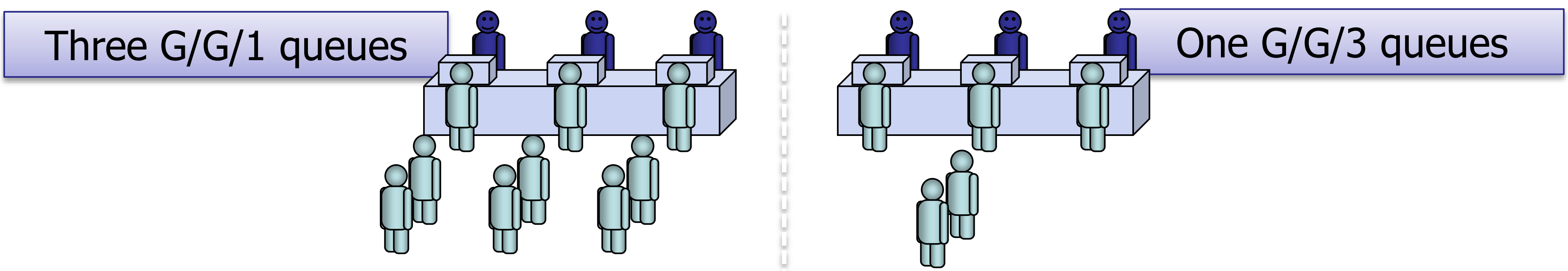
- Customers only form one queue
- The first customer in the queue will be served by the next empty server

Question: I_s=???

MULTİ-SERVER QUEUE: P-K FORMULA

$$\boxed{I_q \cong \frac{\rho^{\sqrt{2(c+1)}}}{1-\rho} \times \frac{C_a^2 + C_s^2}{2}} \qquad \rho = \lambda / c\mu$$

- All other things being equal, if the number of servers c increases, then I_q decreases



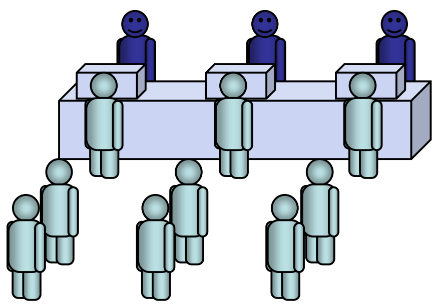
$$3 \times \frac{\rho^2}{1-\rho} \times \frac{C_a^2 + C_s^2}{2} \qquad\qquad \frac{\rho^{2\sqrt{2}}}{1-\rho} \times \frac{C_a^2 + C_s^2}{2}$$

- “Risk Pooling” decreases queue length dramatically

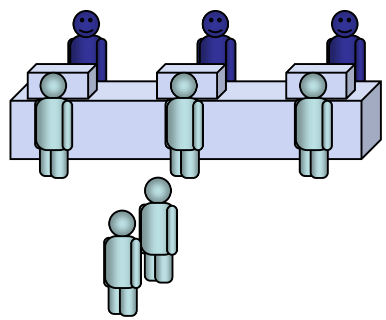
Recommended reading: “A long line for a shorter wait at the supermarket” *New York Times*, June 23, 2007.

Risk pooling or Demand aggregation

Type 1



Type 2



The inventory in queue and wait time is reduced in an G/G/c queue (as compared to c number of G/G/1 queues)

Why does it make sense?

Independent demand streams impose greater variability when compared to a “pooled” demand stream

Approach: Adding independent random variables

Example Applications:

- Component commonality in product design
- Portfolio effects in finance
- Safety stock

DİCE GAME REVİSİTED

M/M/C QUEUE

The first **"M"** indicates that the **inter-arrival times** are **exponentially** distributed
The second **"M"** indicates that the **service times** are **exponentially** distributed
The last **"c"** indicates c **servers**

- Assume First-Come First-Serve (FCFS) rule
- For M/M/1 queue, the P-K formula is

$$I_q \cong \frac{\rho^{\sqrt{2(c+1)}}}{1-\rho}$$

- Note: C_q and C_s are equal to 1 because of the exponential distribution assumption

SUMMARY

- In systems with variability, averages do not tell the whole story
- Unpredictable variability can cause loss of throughput rate
- Inventory buffers or increased capacity may be needed to deal with variety
- In variable systems, inventory and flow time increase non-linearly with utilization (see the P-K formula)
- The impact of variability (on inventory and flow time) can be quantified using the P-K formula, Little's Law, and assumptions about the probability distributions of variability
- "Risk pooling" reduces queue length and wait times