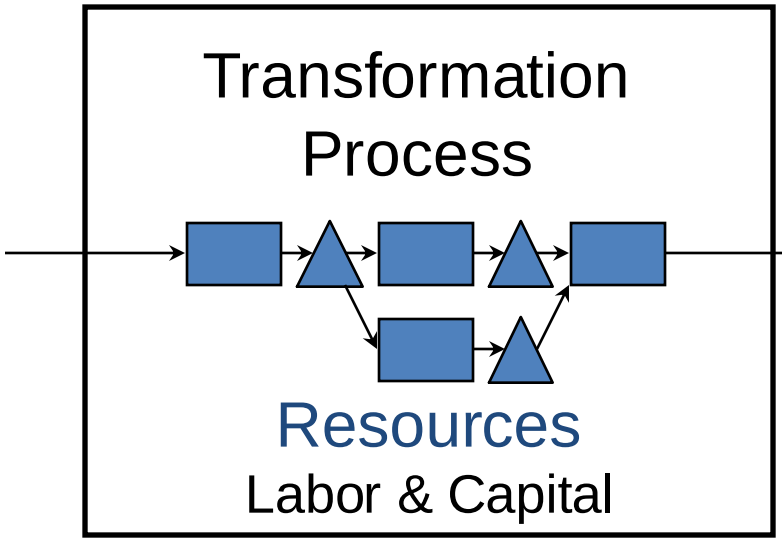
MSBA 7004
Operations Analytics

Class 4-1: Variability Analysis (I)
Safety Capacity – Queueing Analysis
2024

Learning Objectives

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

Sources of Uncertainties



Input from Supply Side

Predictable:

Unpredictable:

Predictable:

Unpredictable:

Output to Demand Side

Predictable:

Unpredictable:

Types of Variability

Predictable Variability

… refers to "knowable" changes in input and/or capacity rates

Ex. Demand of pumpkins will go up during Thanksgiving

Unpredictable Variability

… refers to "unknowable" changes in input and/or capacity rates

Ex. 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

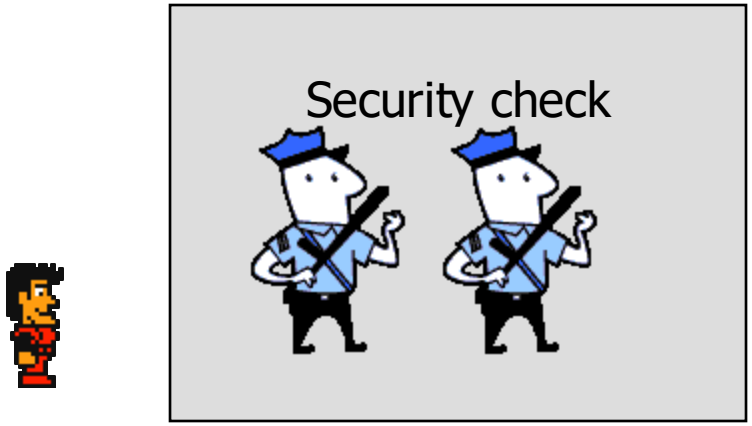
Is the result of the *lack* of knowledge or information

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

Deal with Unpredictable Variability: collecting more knowledge and/or information

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

What is Variability?



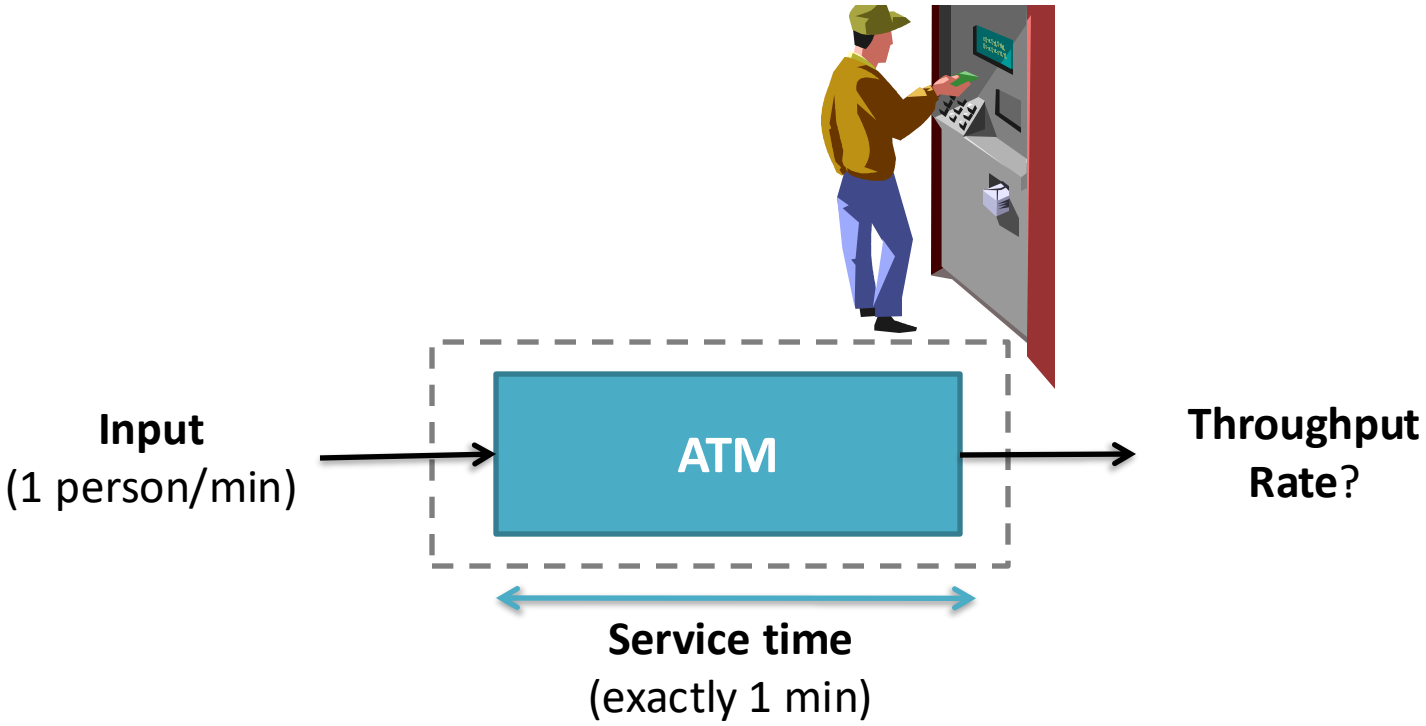
Variability comes from:

Fundamental Questions

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

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

Consider a process with no variability



- 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**)

Utilization?

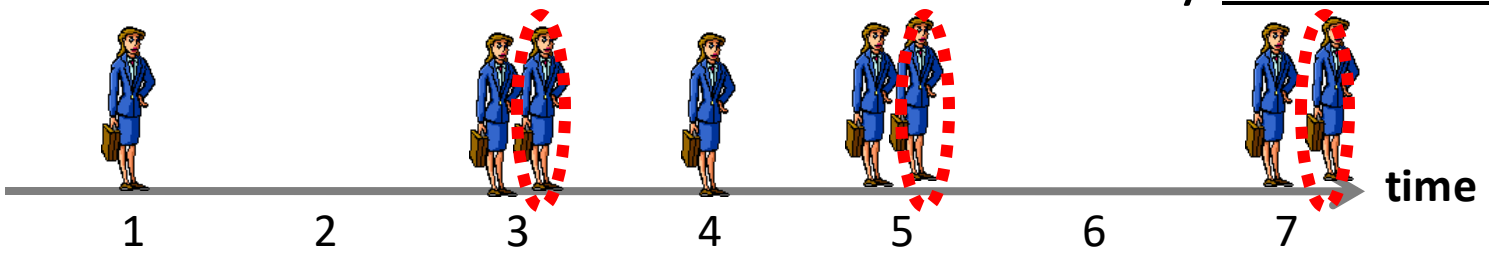
Random Input
0, 1, 2 customers/min
(with equal probability)

ATM

Throughput Rate?

Service time
(exactly 1 min)

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

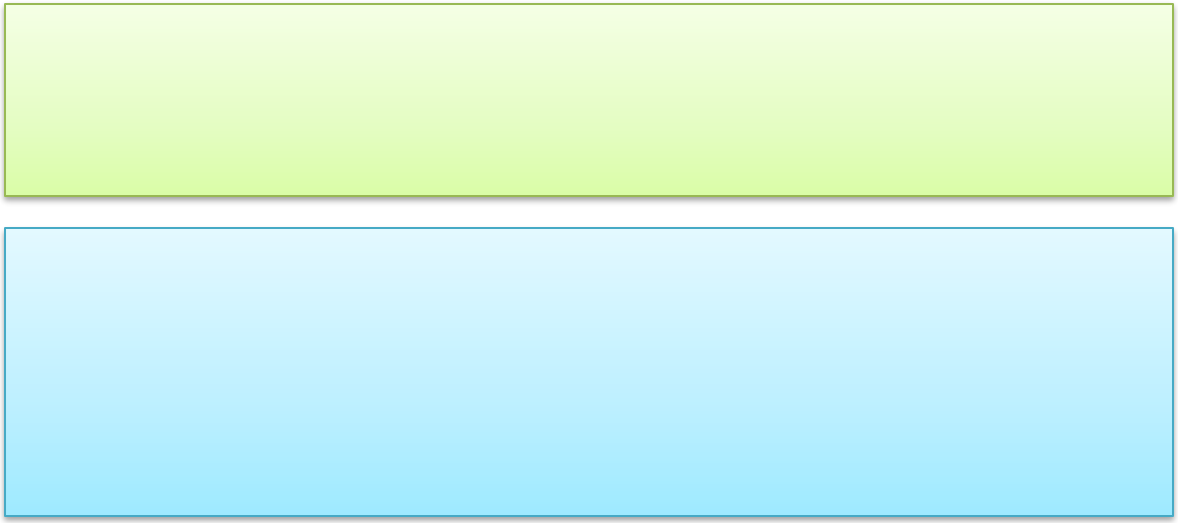


Effect of Input Variability (**no buffer**)

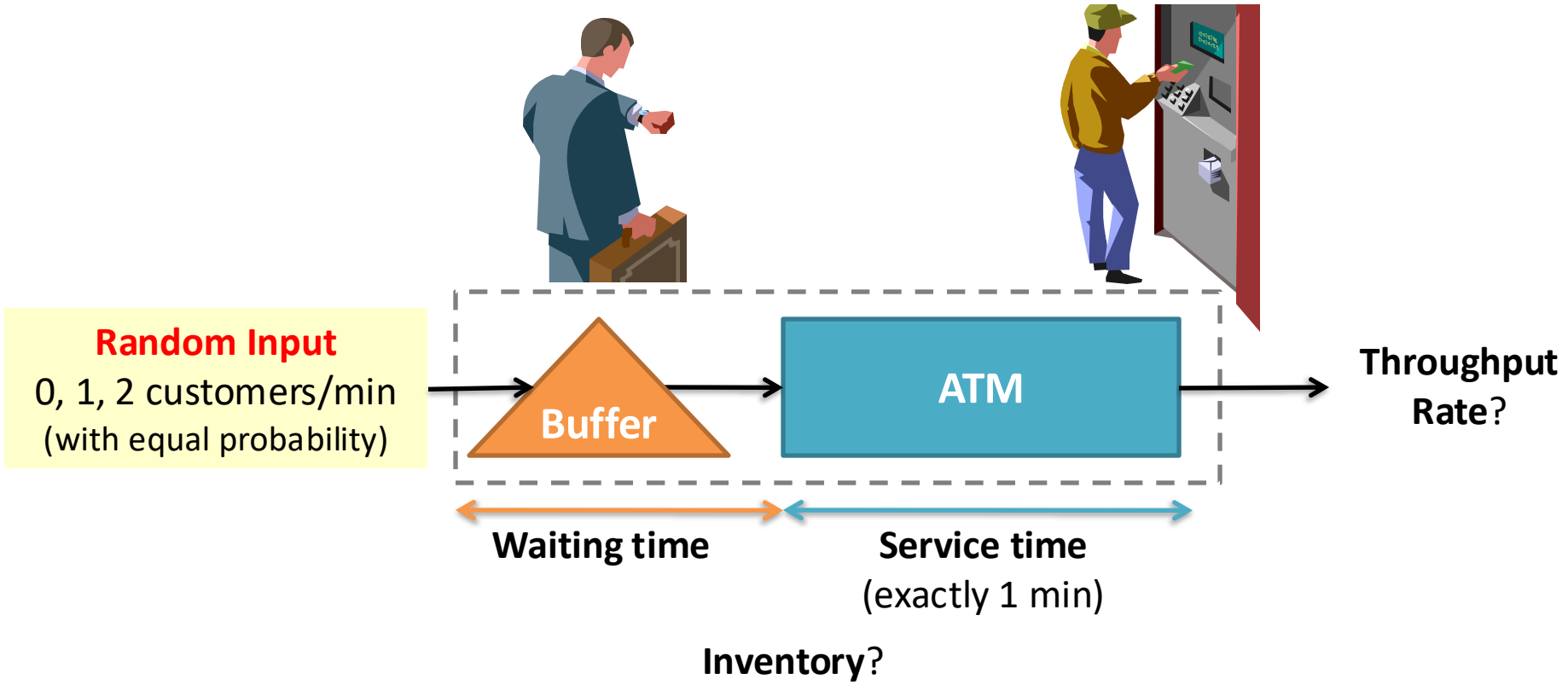
- When a process faces input variability, and a buffer cannot be built, some input may get lost
- Input variability ***can*** reduce the throughput
- Lower throughput means
  - Lost customers; lost revenue
  - Customer dissatisfaction
  - Less utilization of resources
- Little's Law holds

Dealing with Variability

- 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**

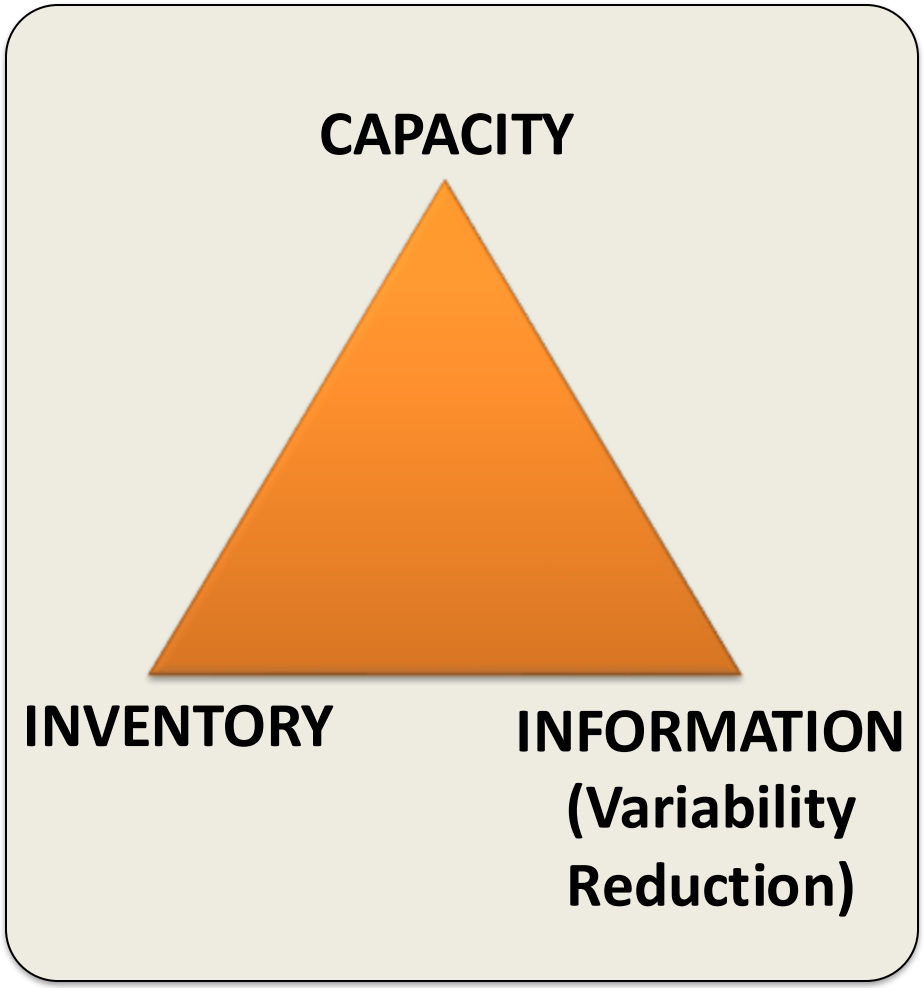
  We can **build-up inventory in the buffer**

Effect of Input Variability (**with 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

Effect of Service Variability

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, low information acquisition effort at the same time. This is a trade-off.

Quick Summary… so far

- In systems with variability, averages do not tell the whole story
- Variability leads to short-term mismatches between supply and demand
- Unpredictable variability can cause loss of throughput rate if there is no buffer
- Unpredictable variability can increase the inventory cost (or waiting time) when there is a buffer

Quantifying Variability

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

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

Learning Objectives

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

Why is it important to quantify variability and its impact?

Quantitative measures of process performance are important to *all* functions of a firm

Why is it important to quantify variability and its impact?

Quantitative measures of process performance are important to *all* functions of a firm

Map view — Last situation as at 13/10/2024 10:45am

Hospital	Reference waiting time
Hong Kong Island	
Pamela Youde Nethersole Eastern Hospital	Over 6 hours
Queen Mary Hospital	Over 1 hour
Ruttonjee Hospital	Over 1 hour
Kowloon	
Caritas Medical Centre	Around 1 hour
Kwong Wah Hospital	Over 8 hours
Queen Elizabeth Hospital	Over 1 hour
United Christian Hospital	Over 8 hours
New Territories	
Alice Ho Miu Ling Nethersole Hospital	Over 1 hour
North District Hospital	Over 8 hours
North Lantau Hospital	Around 1 hour
Pok Oi Hospital	Over 1 hour
Prince of Wales Hospital	Over 8 hours
Princess Margaret Hospital	Over 6 hours
St John Hospital	Around 1 hour
Tin Shui Wai Hospital	Over 7 hours
Tseung Kwan O Hospital	Over 6 hours
Tuen Mun Hospital	Over 2 hours
Yan Chai Hospital	Over 1 hour

Why is it important to quantify variability and its impact?

Quantitative measures of process performance are important to *all* functions of a firm

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 (inventory)

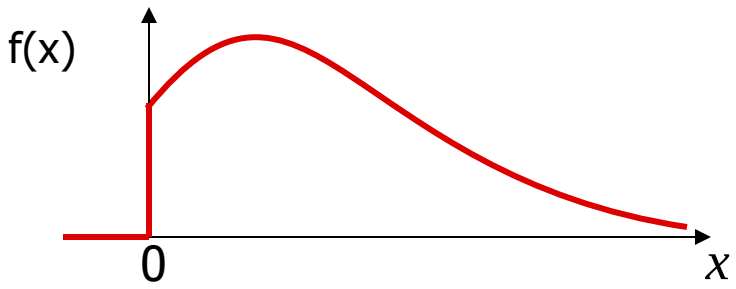
Operations

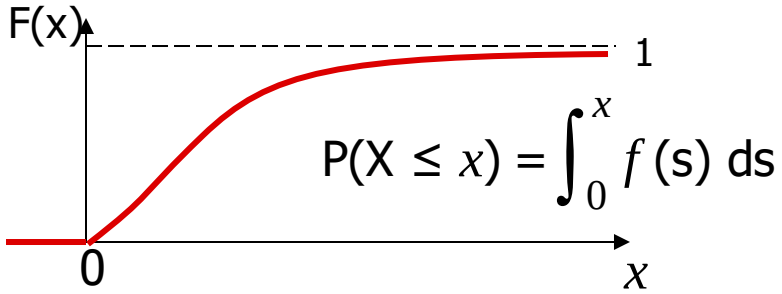
Wants to shorten the queue, and wants to quantify the trade-offs between capacity, inventory and variability

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

Review on Probabilities: **Continuous** Random Variable

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





F(x)=P{X ≤ x} is the **cumulative distribution function (CDF)**

f(x)=F′(x) is the **probability density function (PDF)**

Note that P{X =x}=0 for continuous random variable.

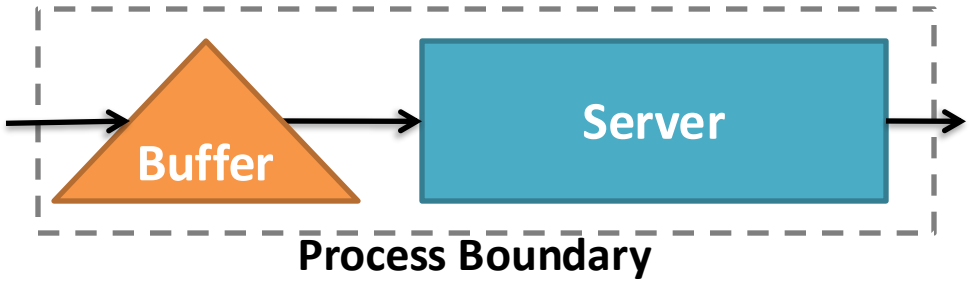
Basic Concepts for Quantifying Variability

- Expectation, Variances and Standard deviation
  - $\bar{X} = E(X) = \sum_{n=0}^{\infty} p(X = n)n \; or \; \int_{-\infty}^{\infty} f(x)xdx$
  - $Var(X) = E\,(X - \bar{X})^2$ STD(X)=$\sqrt{Var(X)}$
- Coefficient of variation (CV)

$$\mathrm{CV(X)} = \frac{\mathrm{STD(X)}}{\mathrm{E(X)}}$$

Quick Quiz
Two stocks: google, mean price 700$, standard deviation 5$
GE mean price 20$, standard deviation $1.
Which stock do you think is more stable?

A Single Server Process



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 $\lambda < \mu$

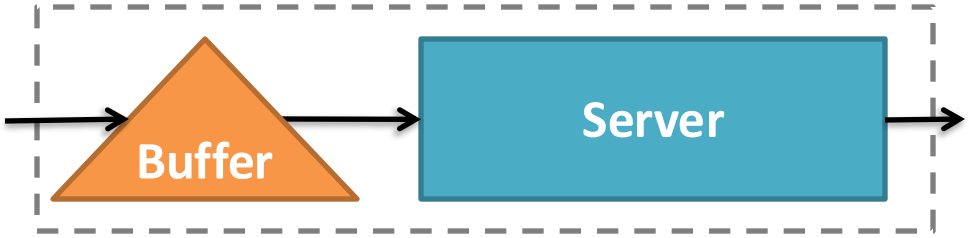
K	Buffer size (for now, let K = ∞)
c	Number of servers in the resource pool (for now, let c=1)

Single-Server Queuing Model

Assumption: $\lambda < \mu$

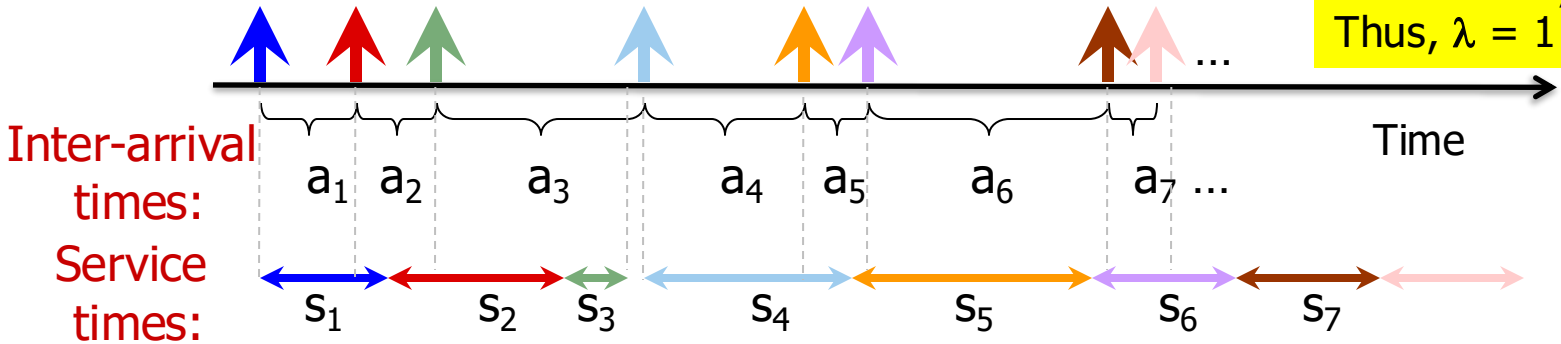
Service rate: μ persons/min
(average capacity rate)

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



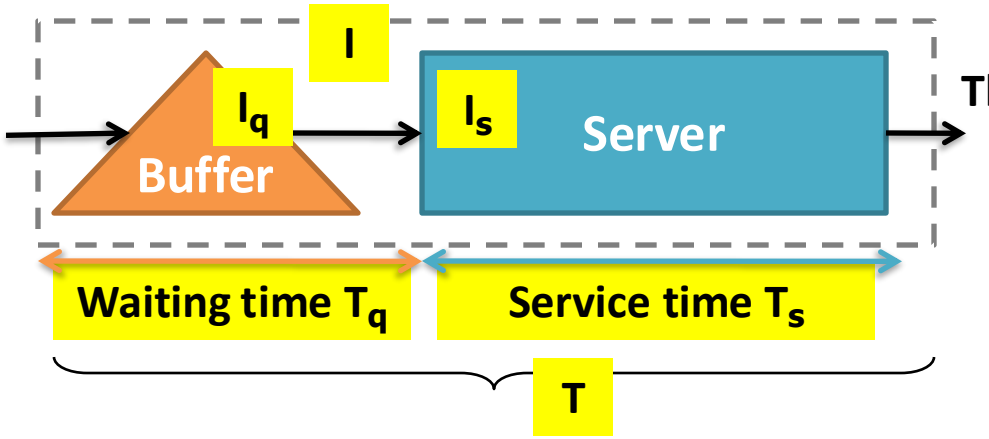
Average throughput rate
λ persons/min

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



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

What are we trying to quantify?



Little's Law holds

$$I_q = \lambda T_q$$

$$I_s = \lambda T_s$$

$$I = \lambda T$$

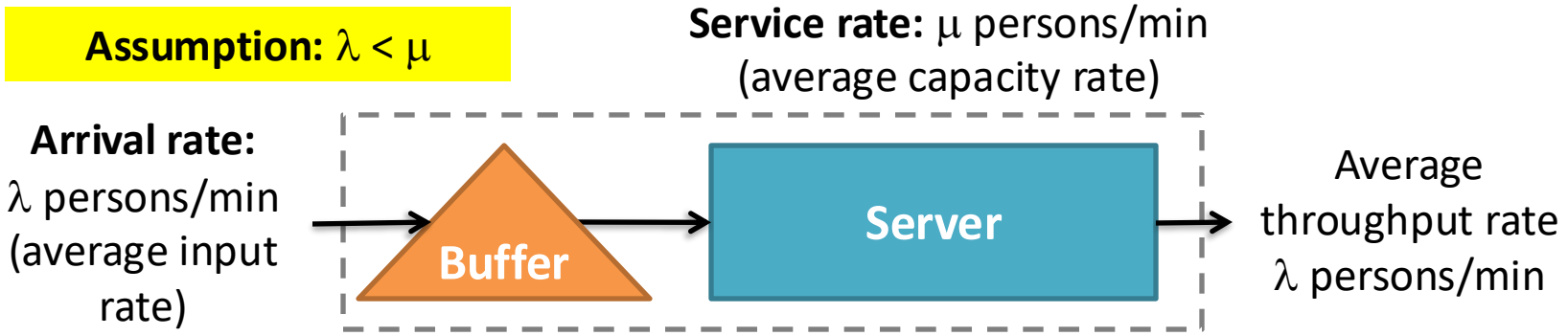
System Characteristics

Utilization	ρ (In a stable system, $\rho = \lambda/\mu \leq 100\%$)
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"



- Average number of persons in the system:

$$I = I_q + I_s$$

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

Pollaczek-Khinchin (PK) Formula (VUT Equation): Single server

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

"=" for special cases

"≈" in general

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

utilization (U), variability (V) and service time (T)

PK Formula and OM Triangle

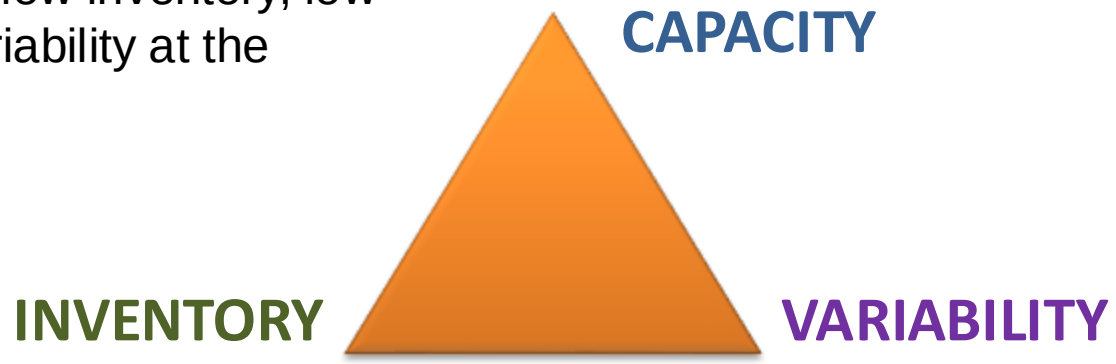
$$I_q \cong \frac{\rho^2}{1-\rho} \times \frac{C_a^2 + C_s^2}{2} = \frac{\lambda}{\mu} \times \frac{\lambda}{\mu - \lambda} \times \frac{C_a^2 + C_s^2}{2}$$

Variability

μ= Capacity Rate

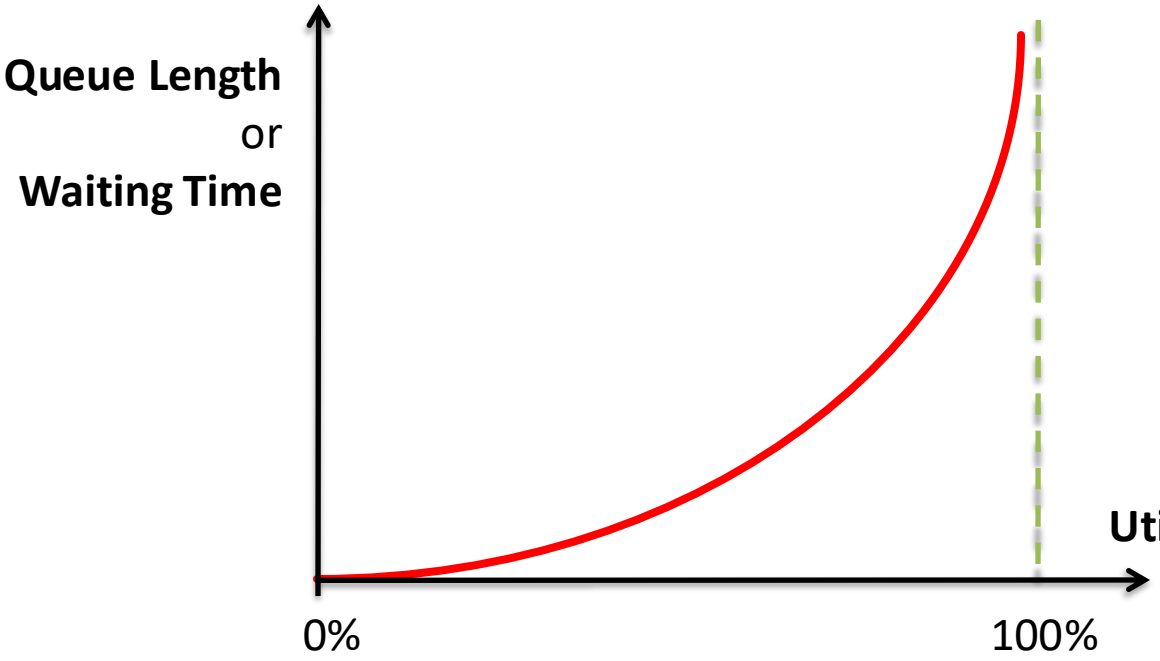
λ = Input Rate

You cannot have low inventory, low capacity, high variability at the same time.



Impact of Utilization ($\rho = \lambda/\mu$)

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**



Managerial implication?

Utilization

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

- 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

- Maximizing utilization is a good idea in a process with no variability

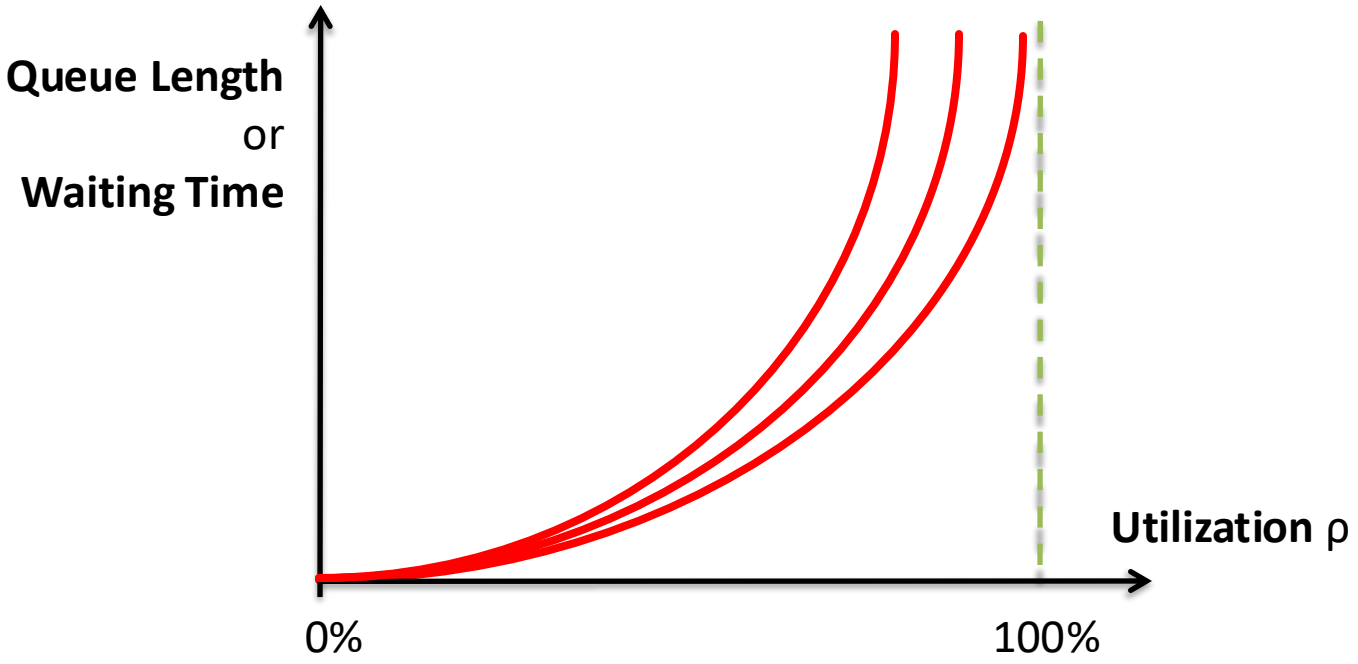
With Variability

- 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**



Ways to reduce waiting

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

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

Queuing Theory

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

The single server 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 (No units “balk” at the length of the queue)
Any unit entering the system stays in the queue till served
First-in-first-out (FIFO)
All units arrive independently of each other

Queuing Notation: 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_s

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

G/G/1 Example

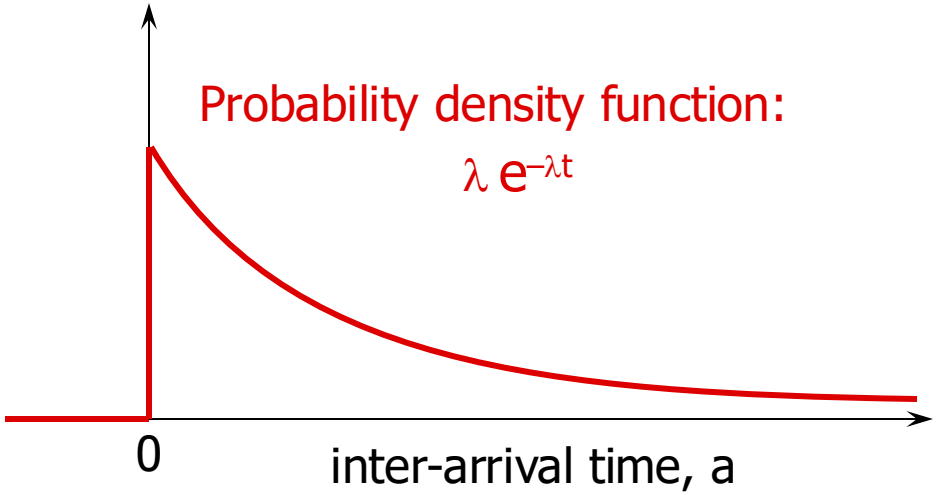
- Customers arrive at rate 4/hour, and mean service time is 10 minutes
- Assume that standard deviation of inter-arrival times equals 5 minutes, and the standard deviation of service time equals 3 minutes
- What is the average size of the queue? What is the average time that a flow units spends in the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

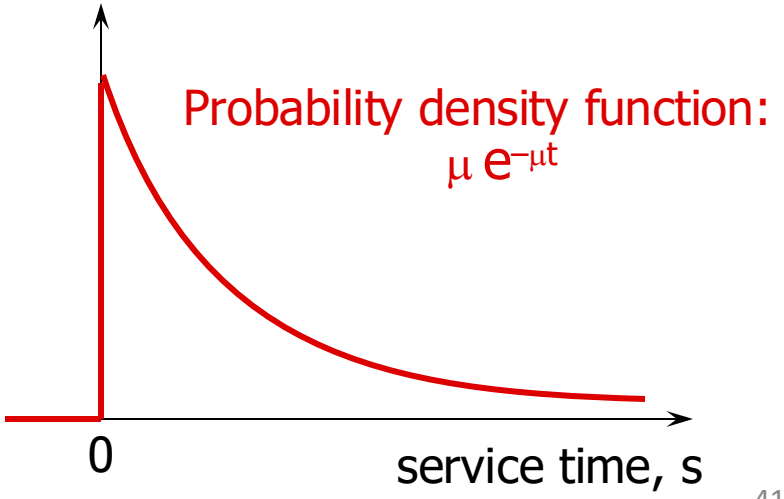
$$I_q \cong \frac{\rho^2}{1-\rho} \times \frac{C_a^2 + C_s^2}{2} =$$

$$T_q =$$

What if we don't have data about the process?

- Suppose you are starting a service business. You haven't yet seen the actual customer 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
- The most commonly used distribution is the **exponential distribution**





Why use these assumptions?

- In many situations, the exponential distribution assumption is a good approximation for what really happens
  - Ambulance arrivals to an emergency department
  - Service time in call centers
- Easy to analyze because coefficient of variation (CV) is 1 for exponential distributions

$$\text{CV} = \frac{\text{Standard Deviation}}{\text{Mean}} = \frac{1/\lambda}{1/\lambda}$$

- 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

M/M/1 queue

The first **"M"*** indicates that the **inter-arrival times** are **exponentially** distributed (Arrivals follow Poisson Process)

The second **"M"** indicates that the **service times** are **exponentially** distributed

The **"1"** refers to the fact that there is a **single server**

- 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) $T_q = I_q / \lambda$

$I = I_q + I_s = ???$

$T = T_q + T_s = ???$

* "M" comes from the *memoryless* property of exponential distribution

Simple Example

- Customers arrive at rate 4/hour, and mean service time is 10 minutes
- We do not have variability information for both inter-arrival and service time. Need to approximated with exponential distribution
- What is the average size of the queue? What is the average time that a flow units spends in the queue?

$$\lambda = 4$$

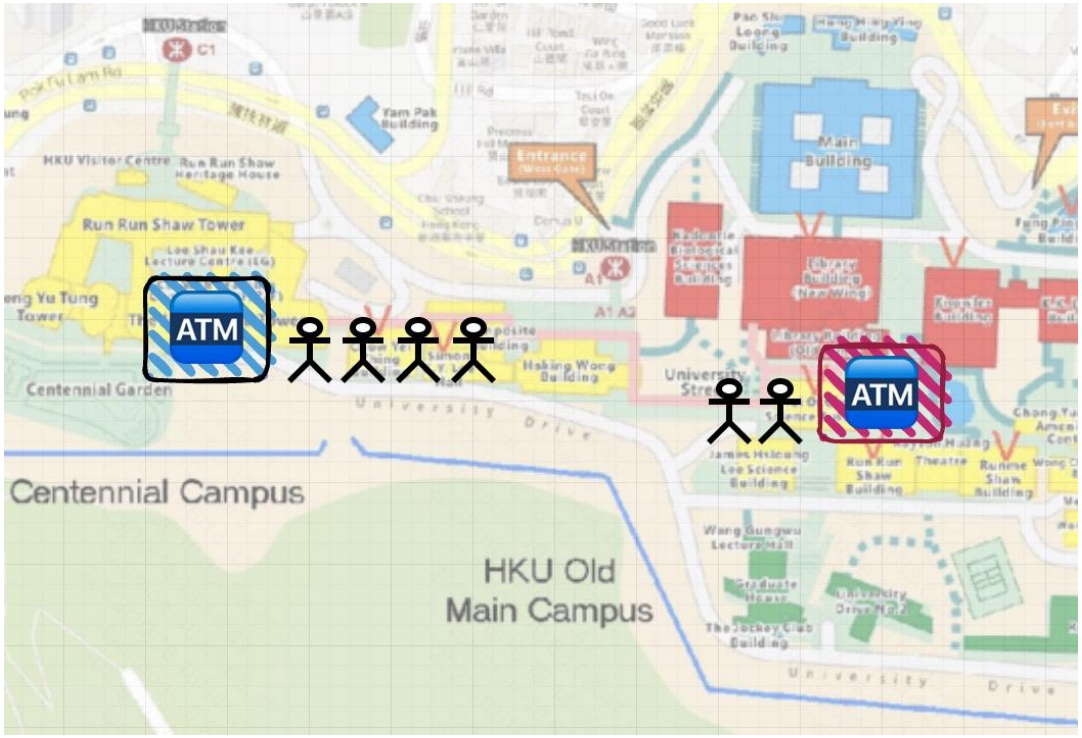
$$E[s] = 1/6 \text{ hour}$$

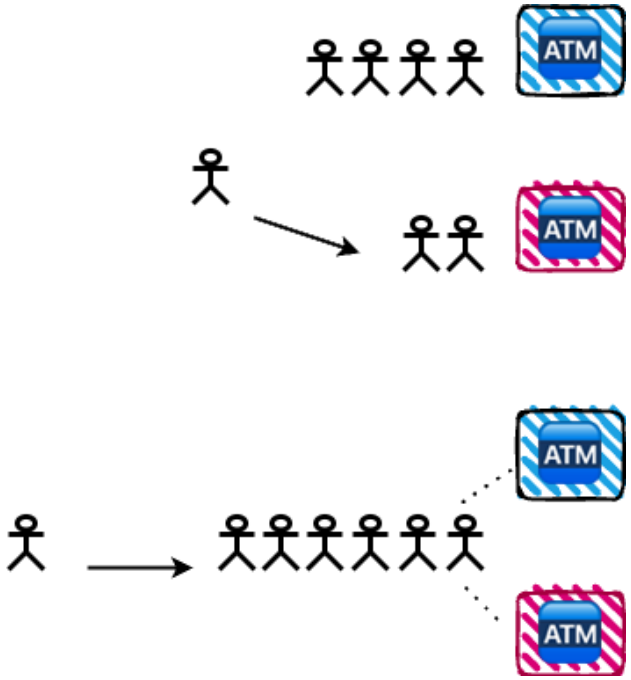
$$\rho = \lambda/\mu =$$

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

$$T_q = \frac{I_q}{\lambda}$$

Pooling or not pooling





Inputs: Variability and Utilization

Service time

Mean = s min

(Avg. service rate per server $= 1/s = \mu$)

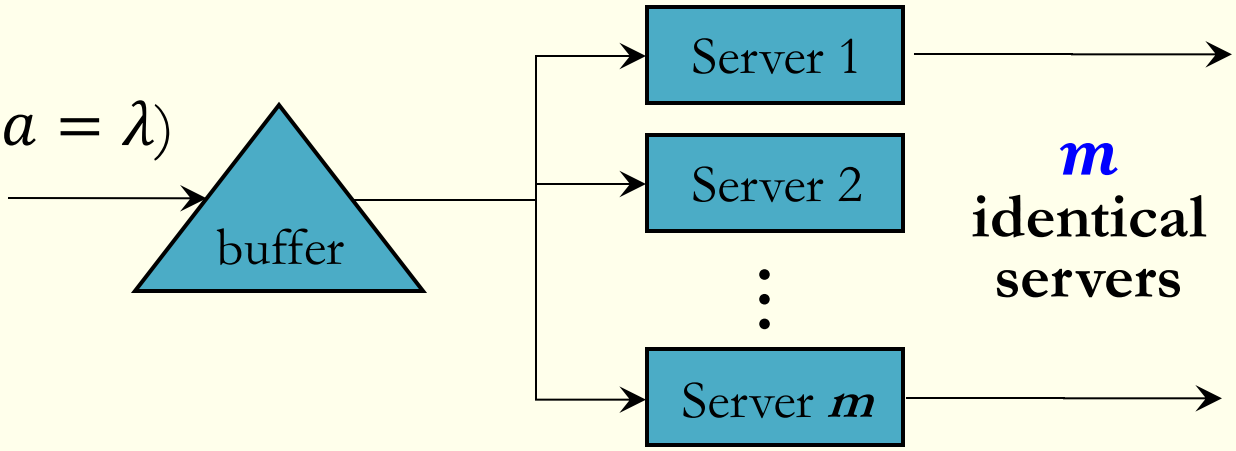
Stdev = σ_s min

Inter-arrival time

Mean = a min

(Avg. arrival rate $= 1/a = \lambda$)

Stdev = σ_a min



$C_a = \sigma_a / a$

$C_s = \sigma_s / s$

$u = \rho$ = avg. system utilization

= avg. arrival rate / avg. aggregate service rate

$= \frac{1/a}{m \cdot (1/s)} = s/(m\,a)$

Key Equation: VUT

Avg. Wait Time = Variability × Utilization × Service Time

$$T_q \cong \left(\frac{{C_a}^2 + {C_s}^2}{2}\right) \times \left(\frac{u^{\sqrt{2(m+1)}-1}}{m(1-u)}\right) \times s$$

Variability effect

Utilization effect

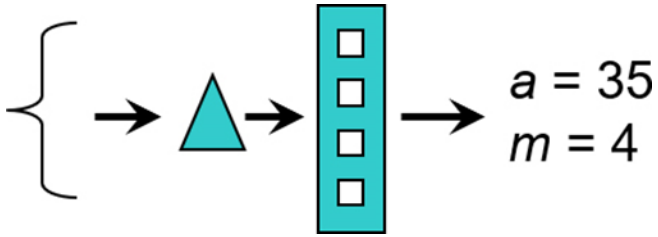
Time scale

When $m = 1$ (single server), VUT Equation reduces to:

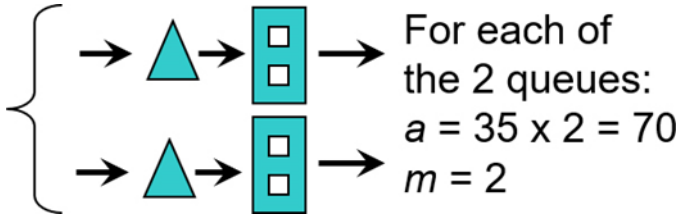
$$T_q \cong \left(\frac{{C_a}^2 + {C_s}^2}{2}\right) \times \left(\frac{u}{1-u}\right) \times s$$

Which System Is More Effective?

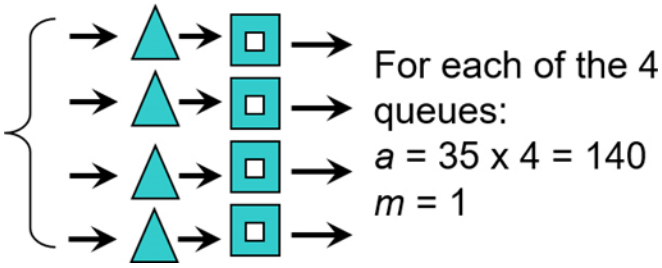
- **Pooled system:**
- One queue, four servers



- **Partially pooled system:**
- Two queues, two servers with each queue.



- **Separate queue system:**
- Four queues, one server with each queue.



Unit: seconds.

- Across these three types of systems:
- Variability is the same:

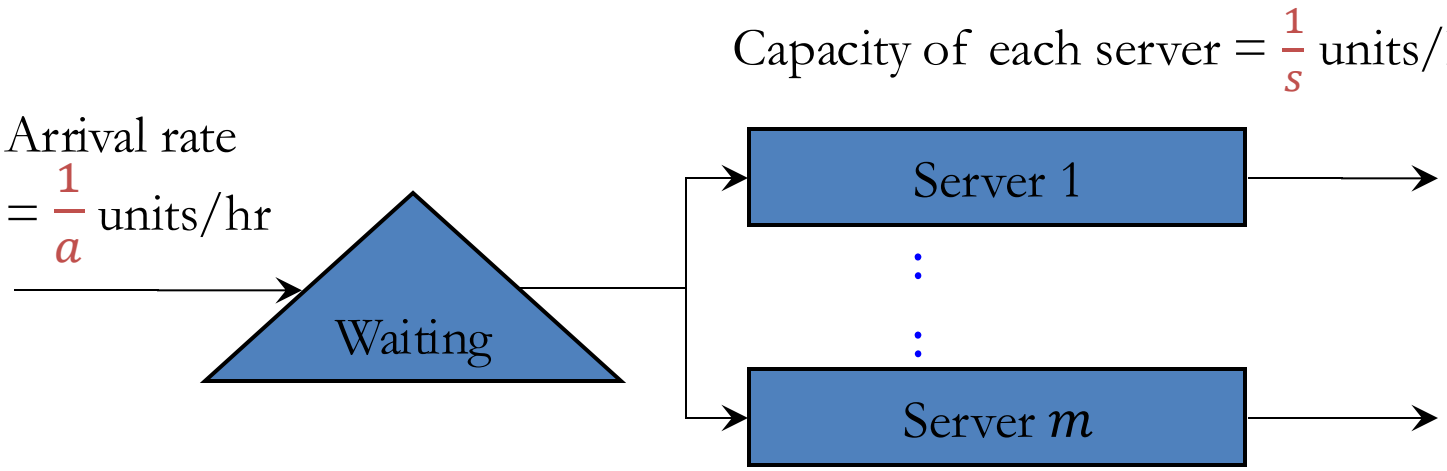
$$C_a = 1, C_s = 1$$

- Total demand is the same:

$1/35$ customers per second.

- Activity time is the same: s = 120.
- Utilization is the same: $\frac{s}{a \cdot m} = 85.7\%$
- The probability a server is busy is the same = 0.857.

How Many are Waiting in Line on Average?



Avg. number waiting		Avg. number being served		Avg. number in system
I_q	+	I_s	=	I_{sys}

Avg. waiting time		Avg. service time		Avg. time in system
T_q	+	s	=	T_{sys}

Little's Law: $I_q = \frac{T_q}{a}$ $I_s = \frac{s}{a}$ $I_{sys} = \frac{T_{sys}}{a}$

Managerial Insights

By giving a precise quantitative description of what drives waiting, the VUT equation helps sharpen our intuition into the impact of utilization and variability on congestion.

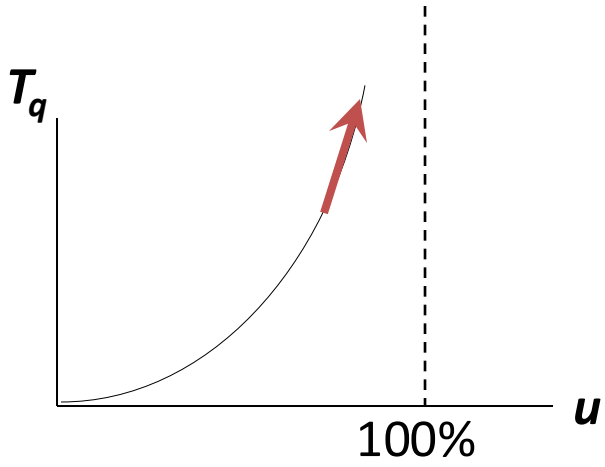
$$T_q \cong \left(\frac{{C_a}^2 + {C_s}^2}{2}\right) \times \left(\frac{u^{\sqrt{2(m+1)}-1}}{m(1-u)}\right) \times s$$

Ways to reduce waiting?

Effect of Utilization

See the math:

$$T_q \cong \left(\frac{{C_a}^2 + {C_s}^2}{2}\right) \times \left(\frac{u^{\sqrt{2(m+1)}-1}}{m(1-u)}\right) \times s$$



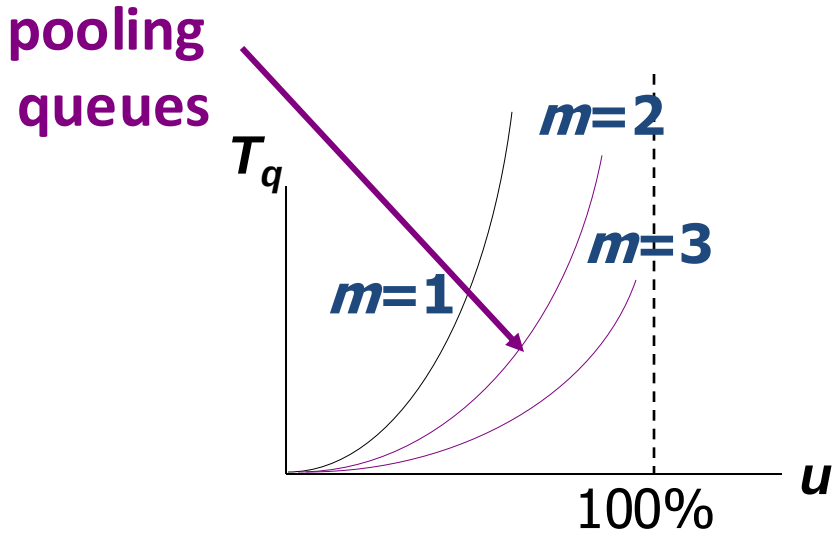
Intuition:

- High ***u*** makes system slow to recover from periods of higher-than-average demand or service times

Managerial implications:

- Must maintain capacity in excess of average demand if cannot tolerate long waits

Pooling Resource



See the math:

$$T_q \cong \left(\frac{{C_a}^2 + {C_s}^2}{2}\right) \times \left(\frac{u^{\sqrt{2(m+1)}-1}}{m(1-u)}\right) \times s$$

Intuition:

- Pooling avoids one server idle while customers wait in other queue
- Postpone server choice

Managerial implications:

- Pooling decreases variability
- Pooling resources spreads risks

Pooling or not pooling

- Is pooling always better?

How to deal with variability?

Reduce Variability

About Input (Demand)
Better Forecasting
Better Scheduling

About Process
Reduce Process Variability
Better Quality

Manage Variability

Choose appropriate "Buffer"

Build adequate *inventory*
and/or
Build adequate *capacity*

Reduce impact of variability by "risk pooling"

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 variability
- 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