

MSBA 7004

Operations Analytics

Tutorial 2

Learning objectives

- Review Inventory build-up and Little's Law
- Apply Little's Law to practice problems
- Review queuing theory and PK formula
- Analyze real life practice problems

Warm-up quiz

Q1:[TRUE/FALSE] The long-run average input rate cannot be smaller than the long-run average output rate.

Q2:[TRUE/FALSE] In short run, average input rate can be less than the average output rate.

Q3:[TRUE/FALSE] Reduce cycle time increase capacity rate, so increases productivity.

Q4:[TRUE/FALSE] Reduce flow time increases inventory turnover, so reduces inventory cost.

Q5:[TRUE/FALSE] To process a multi-unit order, the bottleneck resource may depend on the mix of the units.

Q6 :[TRUE/FALSE] If the inter-arrival time has a larger variance, then there will be more people in service on average.

Q7:[TRUE/FALSE] Since a high utilization leads to a high level of congestions, the manager should keep the utilization of the servers as low as possible.

Q8:[TRUE/FALSE] A project is a process that produces standardized commodities.

Q9: [TRUE/FALSE] Risk pooling can help to reduce the impact of variability.

Q1: The long-run average input rate cannot be smaller than the long-run average output rate.

True. In short time, if output rate $>$ input rate, the rest is from the inventory. However, in the long run, the inventory is finite, so the long-run average of output cannot be larger than the long run average of the input rate.

Q2: In short run, average input rate can be less than the average output rate.

True. For short-run, if there's an amount of inventory in the buffer, then the output rate equals to capacity rate no matter whether the input rate is higher or less than output rate.

Q3: Reduce cycle time increase capacity rate, so increases productivity.

True

Q4: Reduce flow time increases inventory turnover, so reduces inventory cost.

True.

Q5: To process a multi-unit order, the bottleneck resource may depend on the mix of the units.

True. The unit load of each resource depends on the product mix, and the bottleneck resource is the one with the maximal unit load.

Q6 :If the inter-arrival time has a larger variance, then there will be more people in service on average.

False. The P-K formula implies more people in the queue, not more people in the service on average. The average number of people in service, I_s , solely depends on the ratio of arrival rate and the service rate of each server.

Q7:Since a high utilization leads to a high level of congestions, the manager should keep the utilization of the servers as low as possible.

False. Although utilization should not be close to one, it does not mean it has to be as low as possible. Low utilization means wastage of the service resource. There is a trade-off between increasing resource utilization and reducing customer's waiting time.

Q8:A project is a process that produces standardized commodities.

False. The output of a project is usually a few or one of a kind.

Q9: Risk pooling can help to reduce the impact of variability.

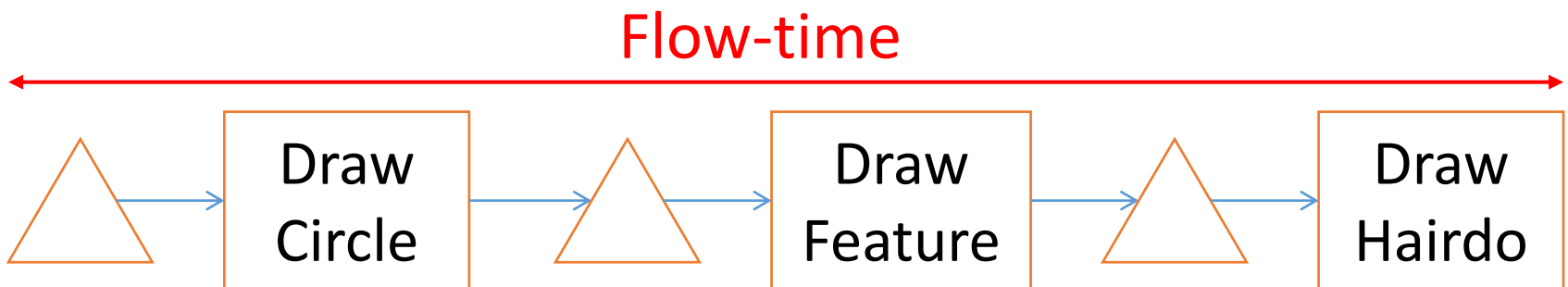
True.

Operational Performance Measures

- **Flow time - T**
- **Throughput / Flow rate – R**
 - Actual output rate
 - Capacity rate: maximum possible output rate
 - $\text{Min}\{\text{Capacity, Arrival (Input) Rate}\}$
 - Utilization = throughput rate/capacity rate
 - Implied utilization = input rate/capacity rate
- **Inventory – I**

Definition: Flow-time

- **Flow-time (Raw process time):** The total time spent by a flow unit within process boundaries
 - Also known as “time in system”
 - Average flow-time: T
 - Units: hours, minutes, seconds



Definition: Theoretical flow time

- **Theoretical flow time:** The minimum amount of time required for a flow unit to go through the process without any waiting or interruptions.
- **Flow time**
 - = \sum Activity time + Waiting(buffer) time
 - = Theoretical flow time + Waiting time

Definition: Throughput/Flow rate

- **Throughput/Flow rate:** The number of flow units that flow through a specific point in the process per unit of time

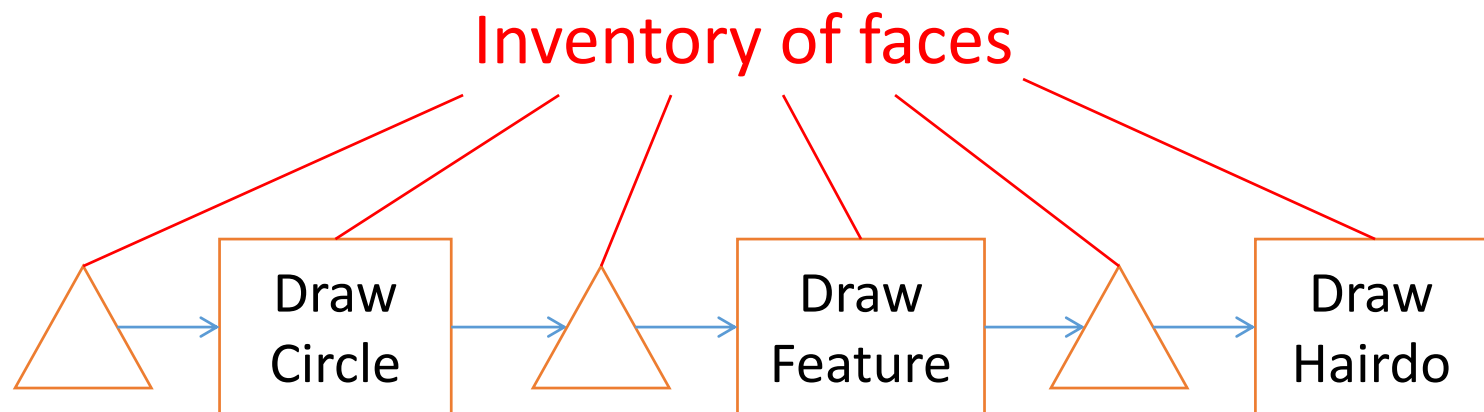
- Units: items/hour or items/day



- How many faces per hour?

Definition: Inventory

- **Inventory:** The total number of flow units present within process boundaries at time t .
 - Denoted as $I(t)$
 - Also known as number in system, work in process (WIP)
 - Units: items, Faces, patients, etc.

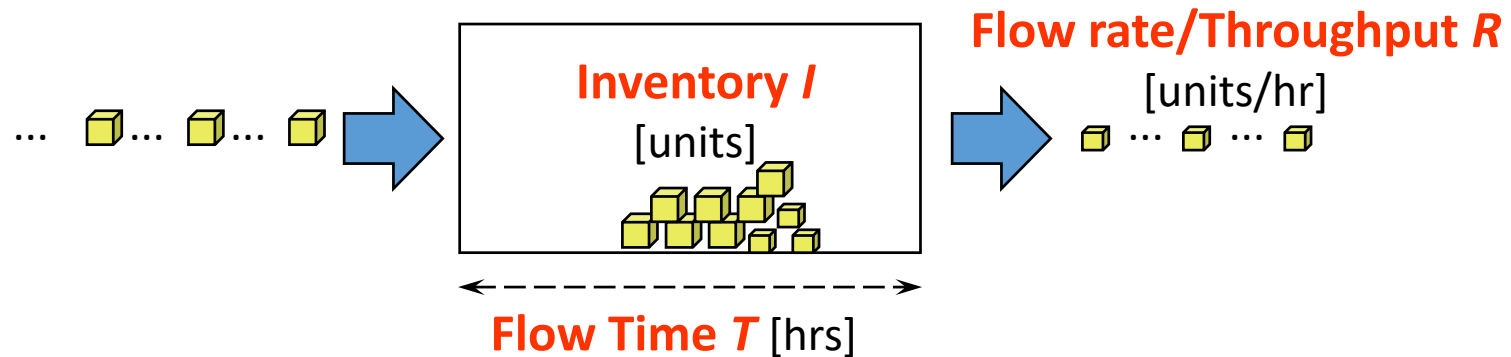


Long-run analysis: Little's Law

Definition in Little's Law

- 3 key operational measures
 - **Flow time – T:** The total time spent by a flow unit within process boundaries
 - **Throughput / flow rate – R:** The number of flow units that flow through a specific point in the process per unit of time
 - Actual output rate
 - $\text{Min}\{\text{Capacity, Input Rate}\}$
 - **Inventory – I:** The total number of flow units present within process boundaries at time t.
- The formula for little's law
 - $I = R * T$

Relating operational measures (flow time T , throughput R & inventory I) with Little's Law



Inventory	=	Throughput x Flow Time
I	=	$R \times T$

Applies to any process if:

- consistent units are used
- long-term averages are used
- consistent flows are measured

$$\begin{aligned}\text{Inventory Turnover} &= \text{Throughput} / \text{Inventory} \\ &= 1 / T\end{aligned}$$

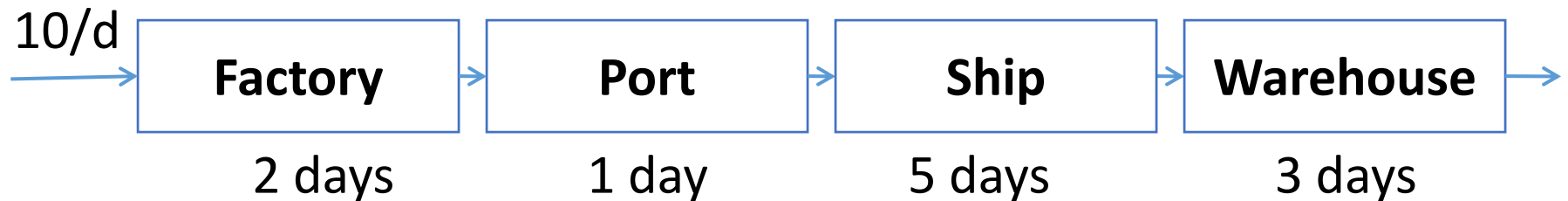
Little's Law: Example 1

A bank handles loans. Management knows the bank receives 40 applications per week on average, which is equivalent to \$20 million. It knows that it completes processing of loans in 25 business days on average. How many loans are typically in process?

- Question I
- Flow-units (items) # of applications
- Entering system / exiting system receipt to complete applications
- Throughput (R) $R=40/\text{week}$
- Flow-time (T) $T=25 \text{ business days} = 5 \text{ weeks}$
- Inventory (I)
- $R \times T = I$ $I = R \times T = 40 \times 5 = 200 \text{ applications}$

Little's Law: Example 2

- Walmart imports Product X from an overseas factory. Each order from Walmart goes through several stages before it gets to the store, and it takes time to “flow” each stage



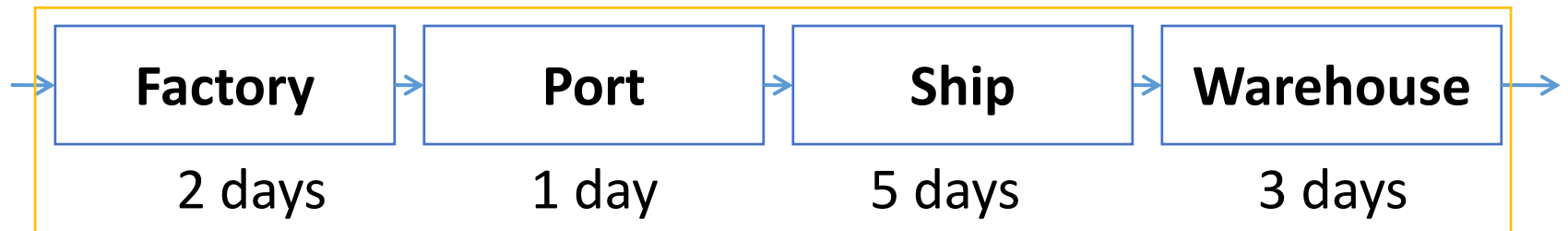
- How much inventory is tied up at the warehouse?
- How much inventory is tied up in the supply chain?

Little's Law can be applied to any process, or any part of it

Little's Law: Example 2

- How to deal with a chained process?
- **Little's Law can be applied to any process, or any part of it**

$$R_p \times T_p = I_p$$
$$I = \sum I_p$$



Little's Law: Exercise 1

Customer Flow: Taco Bell processes on average 1,500 customers per day (15 hours). On average there are 75 customers in the restaurant (waiting to place the order, waiting for the order to arrive, eating *etc.*).

How long does an average customer spend at Taco Bell and what is the average customer turnover?

$$T=I/R=75/(100/\text{hr})=0.75 \text{ hr.}$$

$$1/T=1.33 \text{ times/hr}$$

Little's Law: Exercise 2

Job Flow: The Travelers Insurance Company processes 10,000 claims per year. The average processing time is 3 weeks.

Assuming 50 weeks in a year, what is the average number of claims “in process”.

$$I=RT=10000/50 \text{ claims/week} * 3 \text{ weeks} = 600 \text{ claims}$$

Little's Law: Exercise 3

Question: A general manager at Baxter states that her inventory turns three times a year. She also states that everything that Baxter buys gets processed and leaves the docks within six weeks. (Assume there are 50 weeks per year.)

Are these statements consistent?

$$T=6 \text{ weeks} \Rightarrow$$

$$1/T=1/6 \text{ per week} =$$

$$50/6 \text{ per yr} > 3 \text{ per year.}$$

Inconsistent

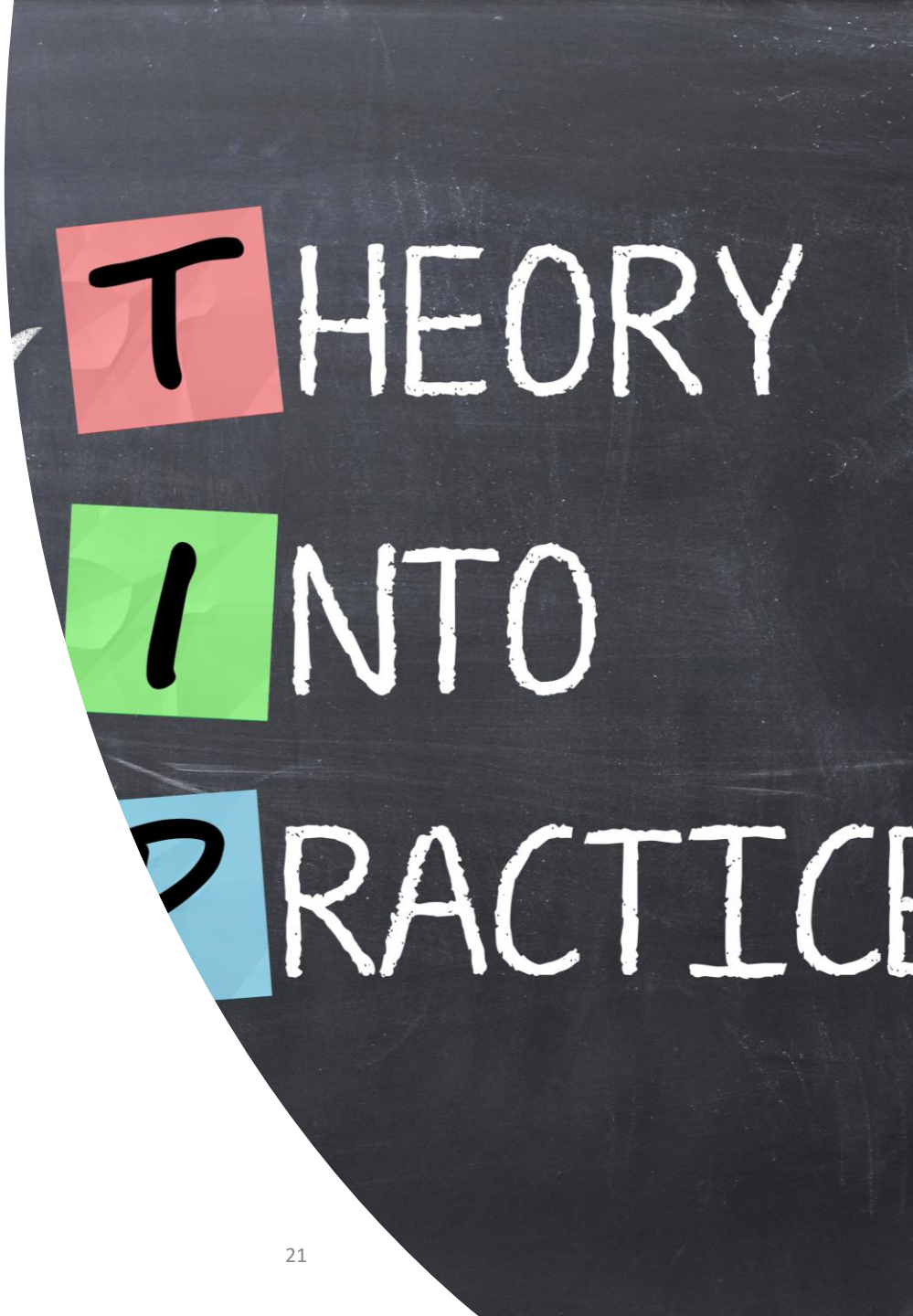
Little's Law: Implications

- 3 Key measures: $I = T * R$
- Given two quantities, one can always compute the third, no matter which part(s) of process.
- As a manager you can only control two. The third is determined by your choices of the first two.

Practice Questions for today

Long-run analysis

- Operational Innovation Can Transform Your Company



THEORY
INTO
PRACTICE

Practice 1:

The *Harvard Business Review* article “Deep Change: How Operational Innovation Can Transform Your Company” contains the following example of operational innovation at Progressive Insurance. Instead of taking between 7 and 10 days to process an insurance claim, the company’s new target was 9 hours.

Assume that the average time to process a claim used to be 8.5 days, and assume that through operational innovation the company has reduced the average time down to 1 day. Given that Progressive processes 10,000 claims/day, and assuming that the company incurs a cost of \$28 per day on each claim that is “in-process,” how much money did Progressive save (per day) because of the reduced claims processing time? (Hint: Every day Progressive incurs a cost of \$28 per claim “in-process.”)

- From the description,
- $R = 10000$ claims/day
- $T_{old} = 8.5$ days
- $I_{old} = R * T_{old} = 85000$ claims

- By reducing the flow time to 1 day, we obtain
- $T_{new} = 1$ day
- $I_{new} = R * T_{new} = 10000$ claims

- Thus,
- $\text{Savings} = (I_{old} - I_{new}) * \$28 = \$ 2.1 \text{ million (per day)}$

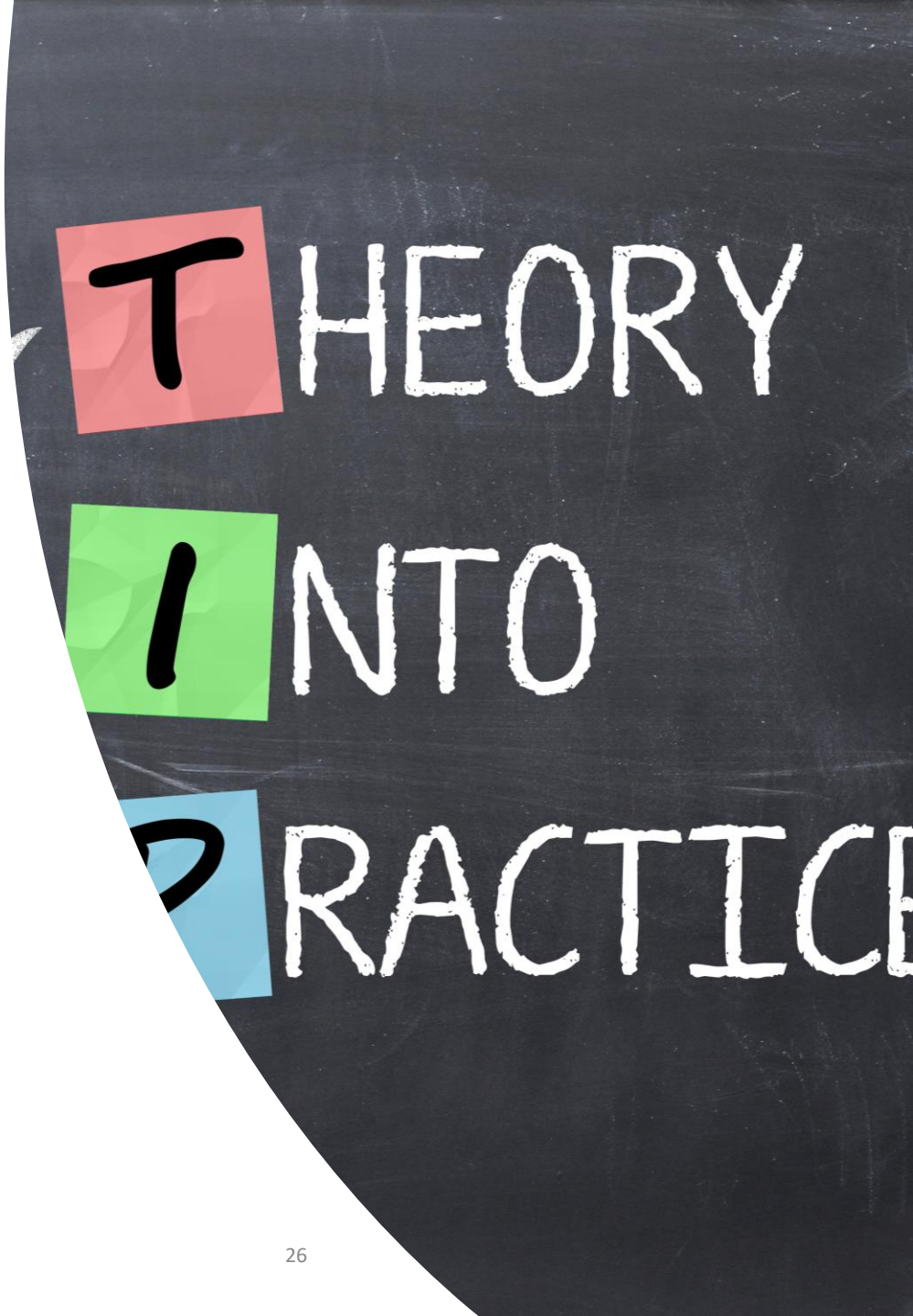
Short-run analysis: inventory build-up diagram

Inventory build-up diagram

- Average of process measures matter: both short-run average and long-run average rates provide useful information
- Continuous vs. Discrete
- *# HKG security screening example*
 - Average inventory changes as we select different time intervals
 - Approximation error occurs when you use larger discrete time interval

Practice Questions for today short-run analysis

- HKU Market

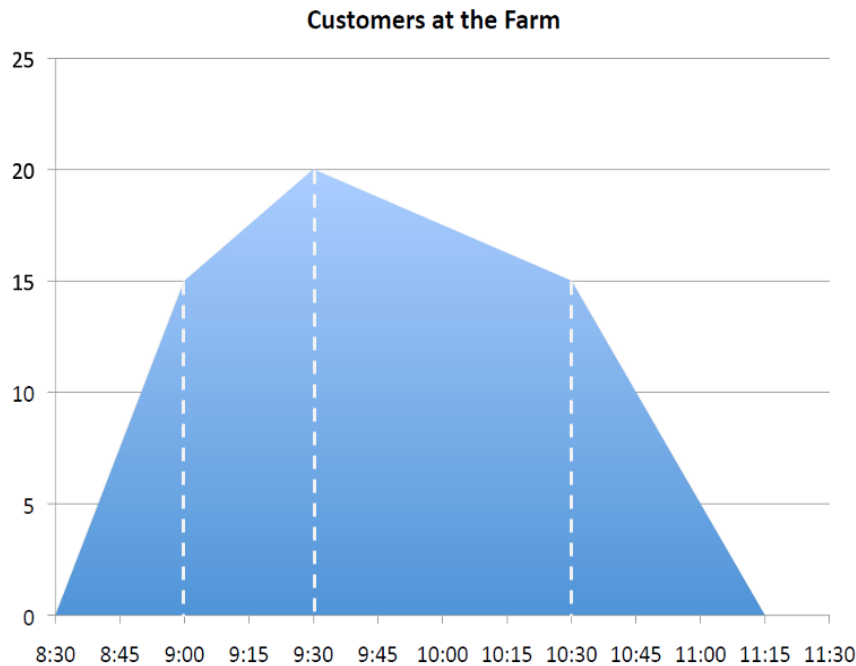


THEORY
INTO
PRACTICE

Practice 2: HKU Market

- The market at the HKU farm every Saturday morning offers fresh organic product that mostly comes from the farm itself.
- The counter opens at 9am. However, customers start lining up at 8:30am. Customers show up at a rate of 30/hr until 9:30am and then at a rate of 15/hr until 10:30am.
- The counter can serve at a rate of 20/hr, and the counter works till all customers served.

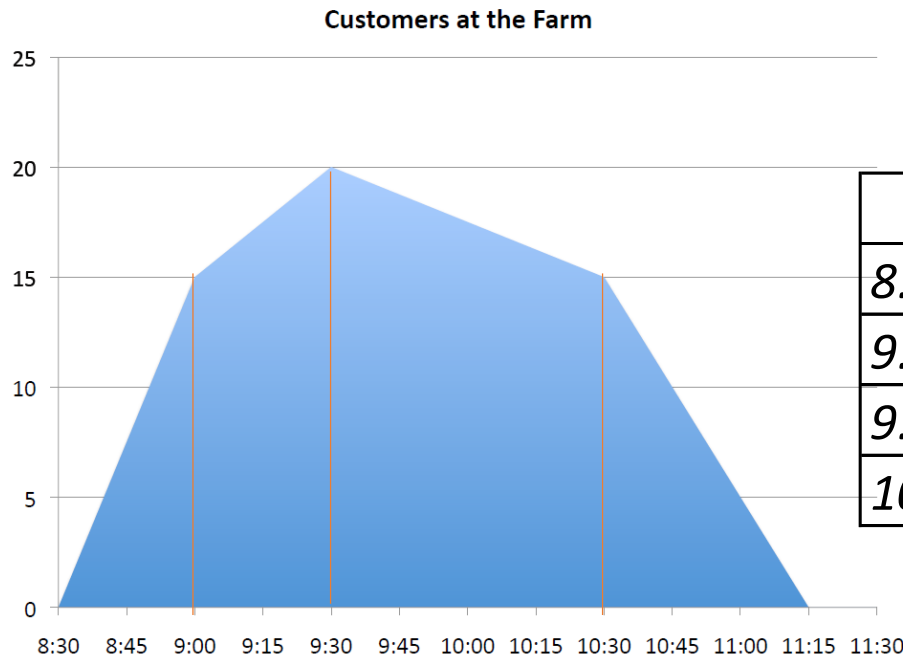
(a) Draw a graph of the number of customers waiting in the line. Start the graph at 8:30am and show the number of waiting customers until the line is empty again. (Use the continuous time setting.)



	Input rate	Capacity rate	Instantaneous inventory accumulation	Ending Inventory
8:30 – 9:00	30/h		30/h	15
9:00 – 9:30	30/h	20/h	10/h	$15 + 0.5 \cdot (30 - 20) = 20$
9:30 – 10:30	15/h	20/h	-5/h	$20 + (1) \cdot (15 - 20) = 15$
10:30 – 11:15 ($15/20 = 0.75\text{h}$)	0/h	20/h	-20/h	0

(b) Using the inventory build-up diagram(from the first customer arrive till all customers served), calculate the average number of customers in line.

$$\text{Avg. \# of customers} = \frac{\text{Area under the graph}}{\text{Entire time}}$$



	Calculation	Area
8:30-9:00	$0.5 \times 15 / 2$	3.75
9:00-9:30	$0.5 \times (15 + 20) / 2$	8.75
9:30-10:30	$1.0 \times (20 + 15) / 2$	17.5
10:30-11:15	$0.75 \times 15 / 2$	5.625

$$\text{Avg. \# of customers} = \frac{35.625}{2.75} = 12.95$$

(c) What is the average time a person spent in the line?

The total waiting time of all of the customers is the area under the curve, which is 35.625 customers · hours

$$\text{Avg. time each customer spent} = \frac{\text{Area under the graph}}{\text{total num of customers}}$$

The total customers arrived 8:30am-11:15am is 45 customers

$$T = \frac{35.625 \text{ customers} \cdot \text{hours}}{45 \text{ customers}} = 0.792 \text{ hours} = 47.5 \text{ mins}$$

Cause of waiting time

- If inter-arrival time and service time are both deterministic or fixed: **For short-run**, inventory may pile up in buffer because of the mismatch between input rate and capacity rate(e.g. input rate > capacity rate).
- If there's variability in inter-arrival time or service time, although utilization of the process is smaller than 100%, waiting time for service can still be substantial due to variability.

Variability Analysis: basic concepts

- Expectation, Variances and Standard deviation

$$- \bar{X} = E(X) = \sum_{n=0}^{\infty} p(X = n)n \text{ or } \int_{-\infty}^{\infty} f(x)x dx$$

$$- Var(X) = E (X - \bar{X})^2 \quad \text{STD}(X) = \sqrt{Var(X)}$$

- Coefficient of variation (CV)

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

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?

CV of Google = 0.007, CV of GE = 0.05

Variability Analysis: PK formula

Pollaczek-Khinchin (PK) Formula: 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

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

M/M/1 Queue

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 \approx)

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

- Average waiting time in queue

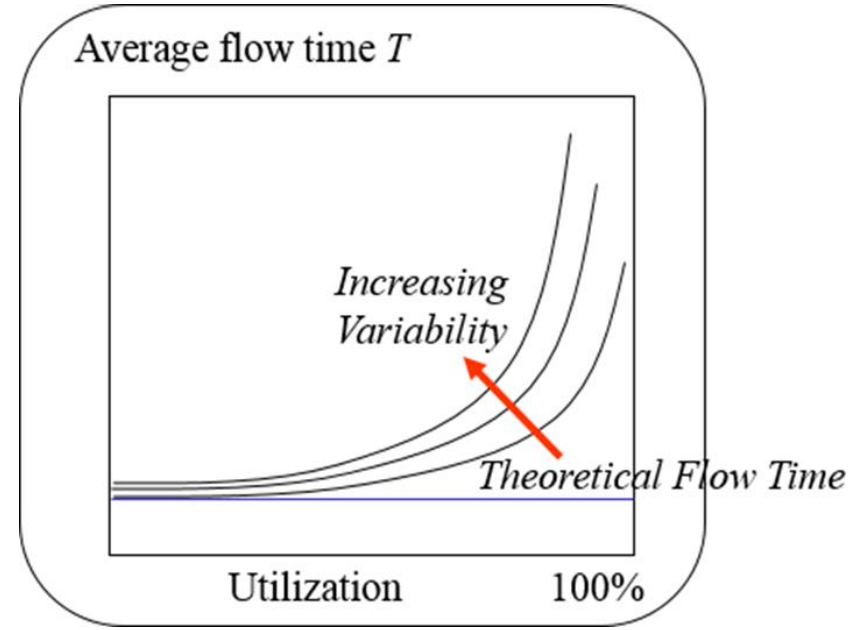
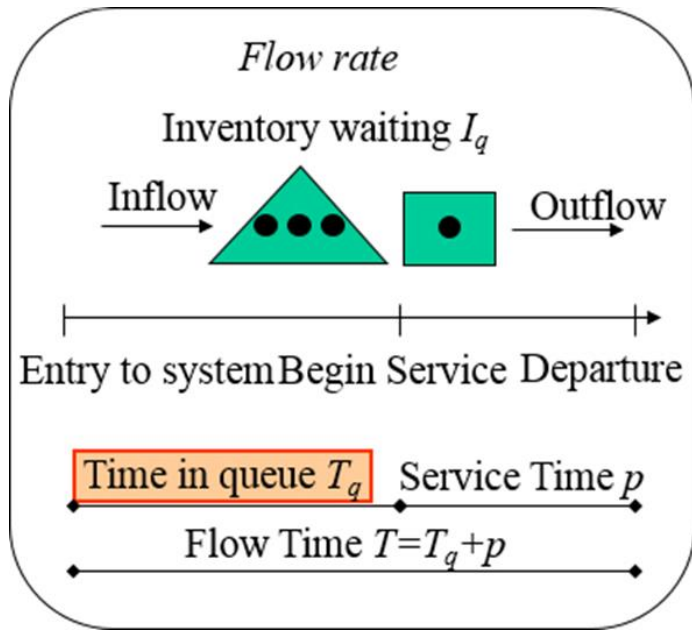
(Little's Law) $T_q = I_q / \lambda$

$$\begin{aligned} I &= I_q + I_s = \frac{\lambda^2}{\mu(\mu-\lambda)} + \frac{\lambda}{\mu} \\ &= \frac{\lambda}{\mu-\lambda} \end{aligned}$$

$$\begin{aligned} T &= T_q + T_s = I_q / \lambda + I_s / \lambda \\ &= \frac{\lambda}{\mu(\mu-\lambda)} + \frac{1}{\mu} = \frac{1}{\mu-\lambda} \end{aligned}$$

* “M” comes from the *memoryless* property of exponential distribution

The Waiting time Formula for single resource



$$CV_a = \frac{\text{St Dev}(\text{interarrival times})}{\text{Average}(\text{interarrival times})}$$

$$CV_p = \frac{\text{St Dev}(\text{processing times})}{\text{Average}(\text{processing times})}$$

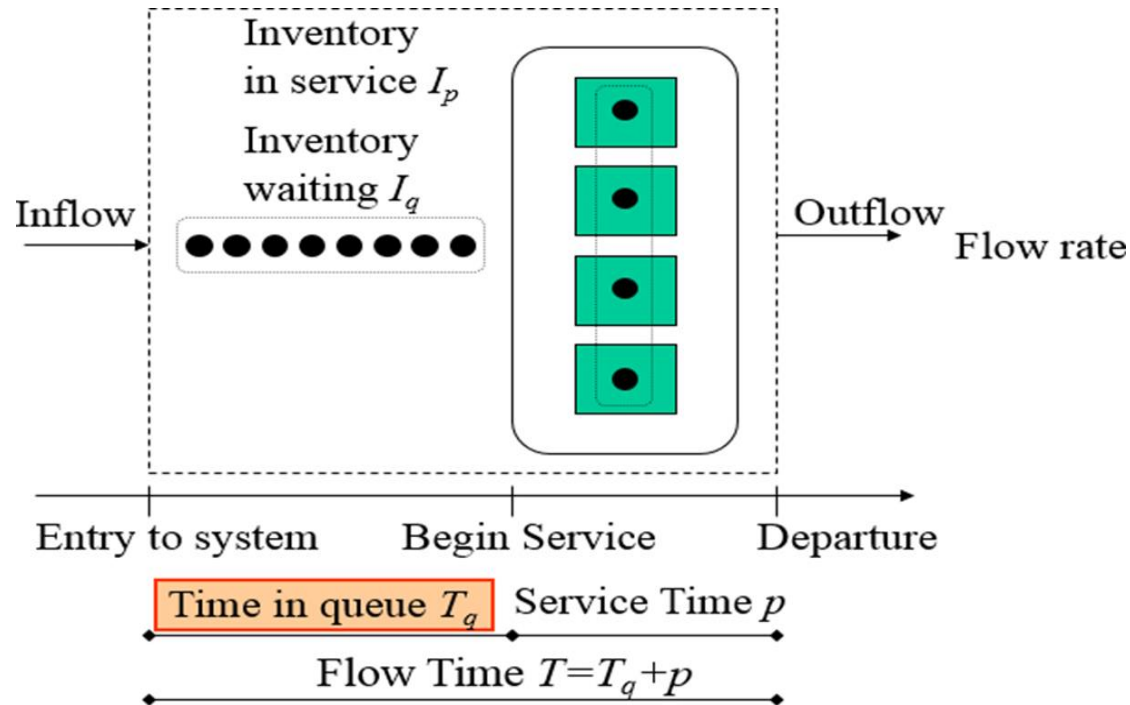
$$\text{Time in queue} = \text{Activity Time} * \left(\frac{\text{utilization}}{1 - \text{utilization}} \right) * \left(\frac{CV_a^2 + CV_p^2}{2} \right)$$

Diagram illustrating the components of the Time in queue formula:

- Activity Time
- Utilization factor: $\left(\frac{\text{utilization}}{1 - \text{utilization}} \right)$
- Variability factor: $\left(\frac{CV_a^2 + CV_p^2}{2} \right)$
- Service time factor

Waiting Time Formula for Multiple, Parallel Resources

- Waiting Time Formula for Multiple (m) Servers



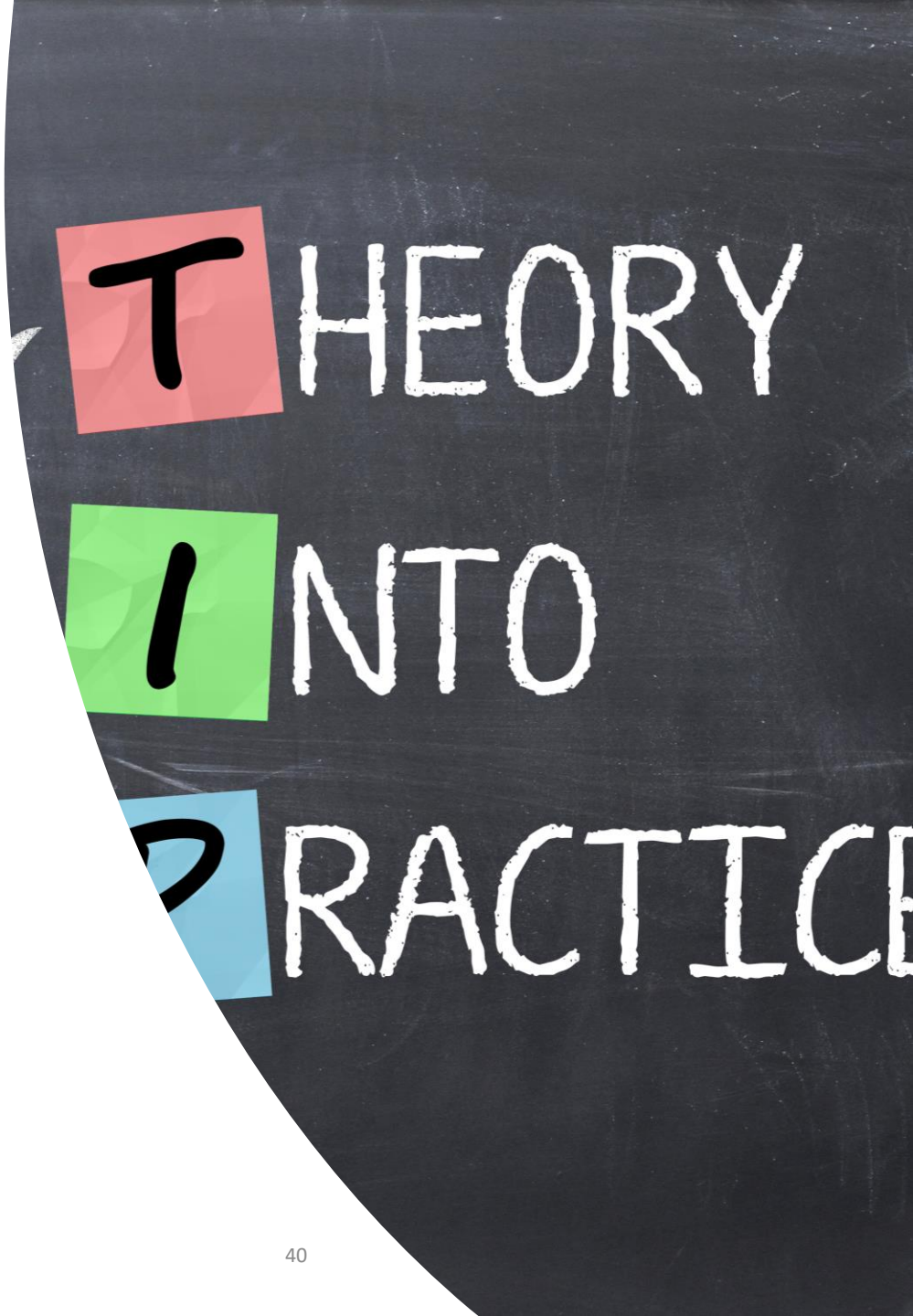
$$\text{Time in queue} = \left(\frac{\text{Activity time}}{m} \right) \times \left(\frac{\text{utilization}^{\sqrt{2(m+1)}-1}}{1 - \text{utilization}} \right) \times \left(\frac{CV_a^2 + CV_p^2}{2} \right)$$

Summary

- Even when a process is demand constrained (utilization is less than 100%), waiting time for service can be substantial due to variability in the arrival and/or service process.
- Waiting times tend to increase dramatically as the utilization of a process approaches 100%.
- Pooling multiple queues can reduce the time-in-queue with the same amount of labor (or use less labor to achieve the same time-in-queue).

Practice Questions for today variability analysis

- Queuing



THEORY
INTO
PRACTICE

Practice 3

A catalog mail-order retailer, has one customer service representative (CSR) to take orders at an 800 telephone number. If the CSR is busy, the next caller is put on hold. For simplicity, assume that any number of incoming calls can be put on hold and nobody hangs up in frustration over a long wait. Suppose that, on average, one call comes every 5 minutes and that it takes the CSR an average of 4 minutes to take an order. Both interarrival and activity times are exponentially distributed. The CSR is paid \$20 per hour, and the telephone company charges \$5 per hour for the 800 line. The company estimates that each minute a customer is kept on hold costs it \$2 in customer dissatisfaction and loss of future business.

a) The average time that a customer will be on hold

$$\mu = 15/\text{hour}$$

$$\lambda = 12/\text{hour}$$

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

$$\begin{aligned} T_q &= \frac{I_q}{\lambda} = \frac{\lambda}{\mu(\mu-\lambda)} = \frac{12}{15(15-12)} \text{ hours} \\ &= 16 \text{ minutes} \end{aligned}$$

b) The average number of customers in system

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

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

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

$$I = \frac{\lambda}{\mu-\lambda} = 4 \text{ customers}$$

c) The total hourly cost of service and waiting

$$I_q = \frac{\rho^2}{1-\rho} = \frac{\lambda^2}{\mu(\mu-\lambda)} = \frac{12^2}{15 * 3} = 3.2customers$$

$$I_q \cdot H = 3.2customers * \$2/customer/minute = \$384/hour$$

$$\text{Total Cost} = 384 + 20 + 5 = \$409/hour$$

Key Take-aways

- Short run analysis vs. Long run analysis
- Three key operational measures: flow time, inventory, and throughput
 - Little's Law connects them
- Application for Inventory build-up diagram, Little's Law and queuing model.