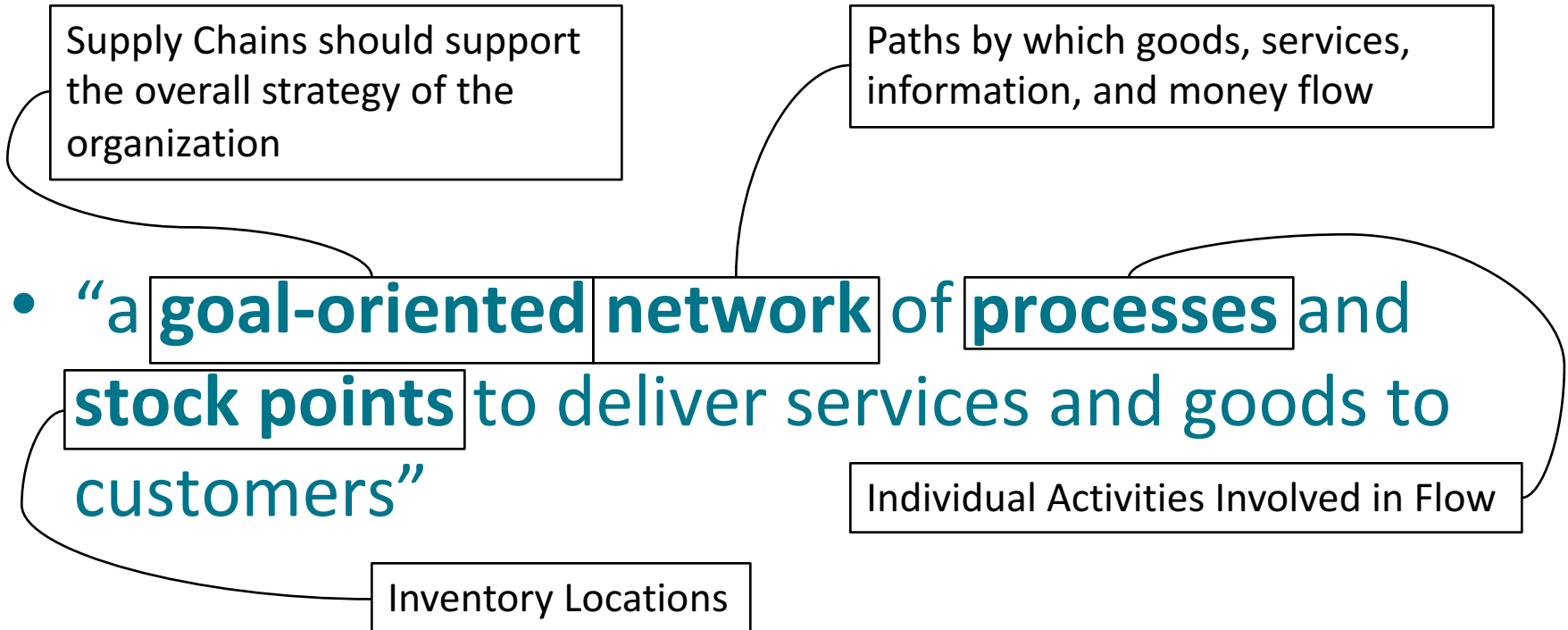


Fundamental Process Analysis

Supply Chains – a Process Perspective



Agenda

- Simple process measurement & Little's Law
- Capacity, utilization and bottlenecks
- Process variability
- Introduction to Queueing Theory
- Extending the model
- Quantifying the impact of variability
- Multiple servers

Measuring a Simple Process

Simple Process



- How can we measure a process?
 - Throughput (TH):
 - ♦ Rate at which items are processed (items/time)
 - Work in Process (WIP):
 - ♦ Number of items in the system (items)
 - Cycle Time (CT):
 - ♦ Time required for a unit to traverse the system (time)

Little's Law

$$WIP = TH * CT$$

$$\text{items} = (\text{items/time}) * \text{time}$$

- Little's Law
 - Over the long-term, the average **work in process** (*WIP*) is equal to the average **throughput** (*TH*) times the average **cycle time** (*CT*).
- Reminders:
 - Holds for long-term averages, not minute to minute or day to day
 - Assumes a stable process, i.e., no trending

Example: Order Confirmation Process



- Inputs & Outputs
 - Inputs: orders from customers
 - Process: Check availability & credit and print shipping documents
 - Output: product ready for delivery
- Example 1
 - Suppose that on average a clerk processes 15 orders per hour and the typical order takes 1.5 hours to be processed from entering the system to finish. What is the average number of orders in process?

$$\begin{aligned}WIP &= TH * CT \\&= 15 \text{ (orders/hour)} * 1.5 \text{ (hours)} \\&= 22.5 \text{ orders in process}\end{aligned}$$

Example: Order Confirmation Process

- Example 2

- Suppose that on average a different clerk processes 12 orders per hour and there are typically 22 orders in the queue or in process. What is the average cycle time?

$$\begin{aligned}CT &= WIP / TH \\&= 22 \text{ (orders)} / 12 \text{ (orders/hour)} \\&= 1.83 \text{ hours}\end{aligned}$$

- Once you know two terms, you can find the third.
- Applies very generally:
 - How much money in outstanding accounts receivable do I have if we bill an average of \$120,000 per week and customers, on average, pay in 4 weeks?

$$\begin{aligned}WIP &= TH * CT \\&= 120 \text{ (k$/week)} * 4 \text{ (weeks)} \\&= \$480,000 \text{ in AR}\end{aligned}$$

Connecting *WIP* to *CT*

$$WIP = TH * CT$$

$$WIP / CT = TH$$

- Which is better?
 - Reduce the work in process (*WIP*) in a process?
 - Decrease the cycle time (*CT*) of a process?
- If throughput stays the same, they move together!
 - Longer *CT* implies greater *WIP*!
 - Decrease in *WIP* means a shorter *CT*!
- Suppose you have a process with $TH = 10$ items/hour, would you rather have
 1. $WIP = 100$ items and $CT = 10$ hours or
 2. $WIP = 10$ items and $CT = 1$ hour?

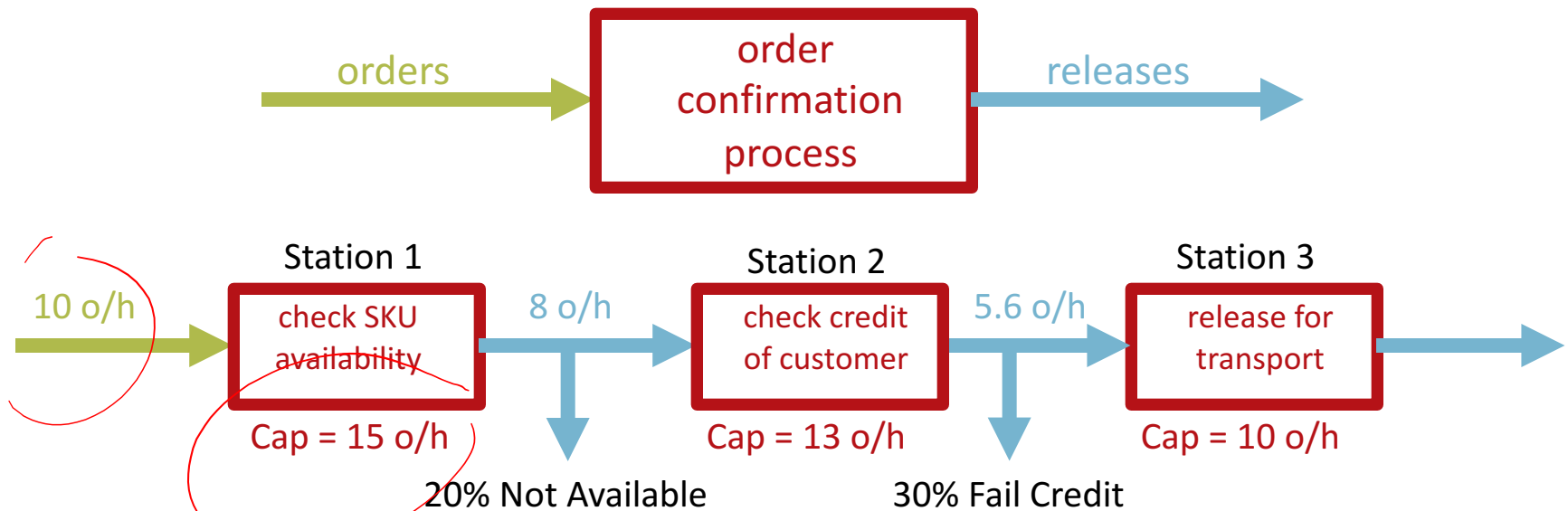
Capacity, Utilization, and Bottlenecks

Capacity and Utilization



- How else can we measure a process?
 - Capacity:
 - ♦ Maximum average rate at which items can flow through the system (units/time)
 - ♦ Equal to Base Capacity (under ideal conditions) minus Detractors (failures, disruptions, rework, maintenance, etc.)
 - Utilization:
 - ♦ Rate of flow into a process divided by its capacity
 - Bottleneck:
 - ♦ The process in a system with the highest utilization

Order Confirmation Process: Bottleneck



Utilization

$$\text{Station 1} = 10/15 = \mathbf{0.67}$$

$$\text{Station 2} = (10 \times 0.8)/13 = \mathbf{0.62}$$

$$\text{Station 3} = (8 \times 0.7)/10 = \mathbf{0.56}$$

Utilization (SKU availability @ 90%)

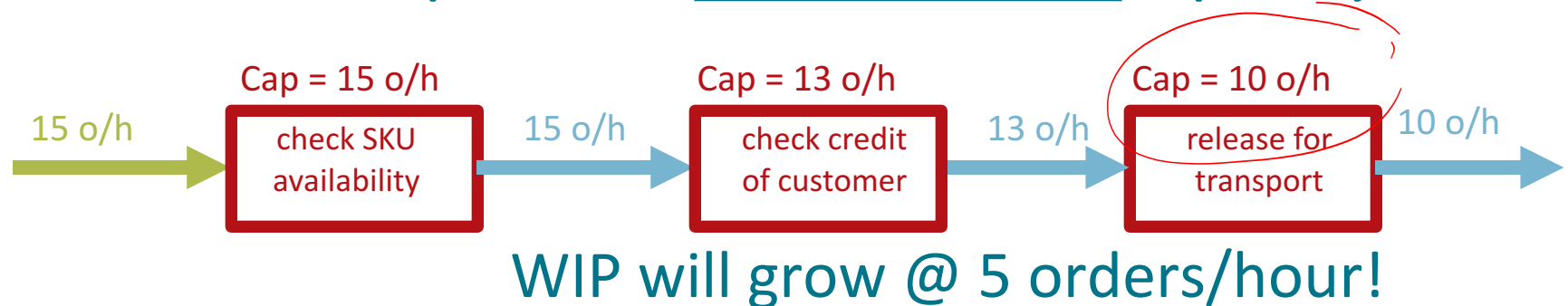
$$\text{Station 1} = 10/15 = \mathbf{0.67}$$

$$\text{Station 2} = (10 \times 0.9)/13 = \mathbf{0.69}$$

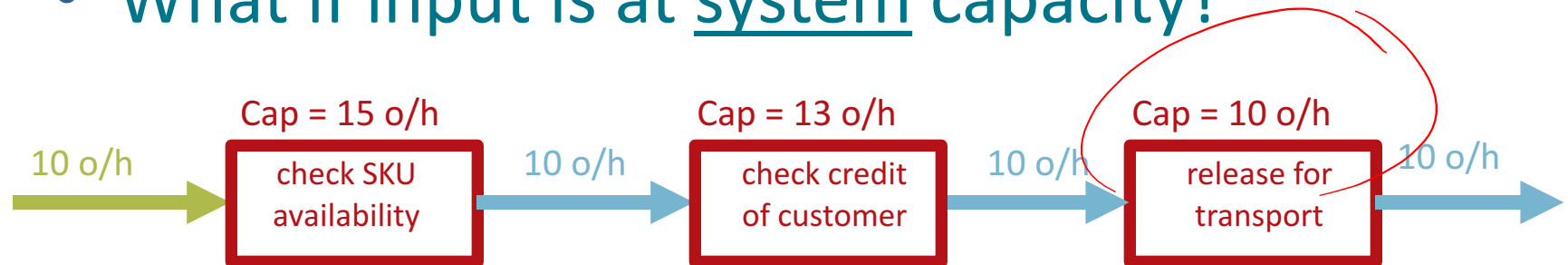
$$\text{Station 3} = (9 \times 0.7)/10 = \mathbf{0.63}$$

Capacity & Utilization Limits

- What if input is at initial station capacity?



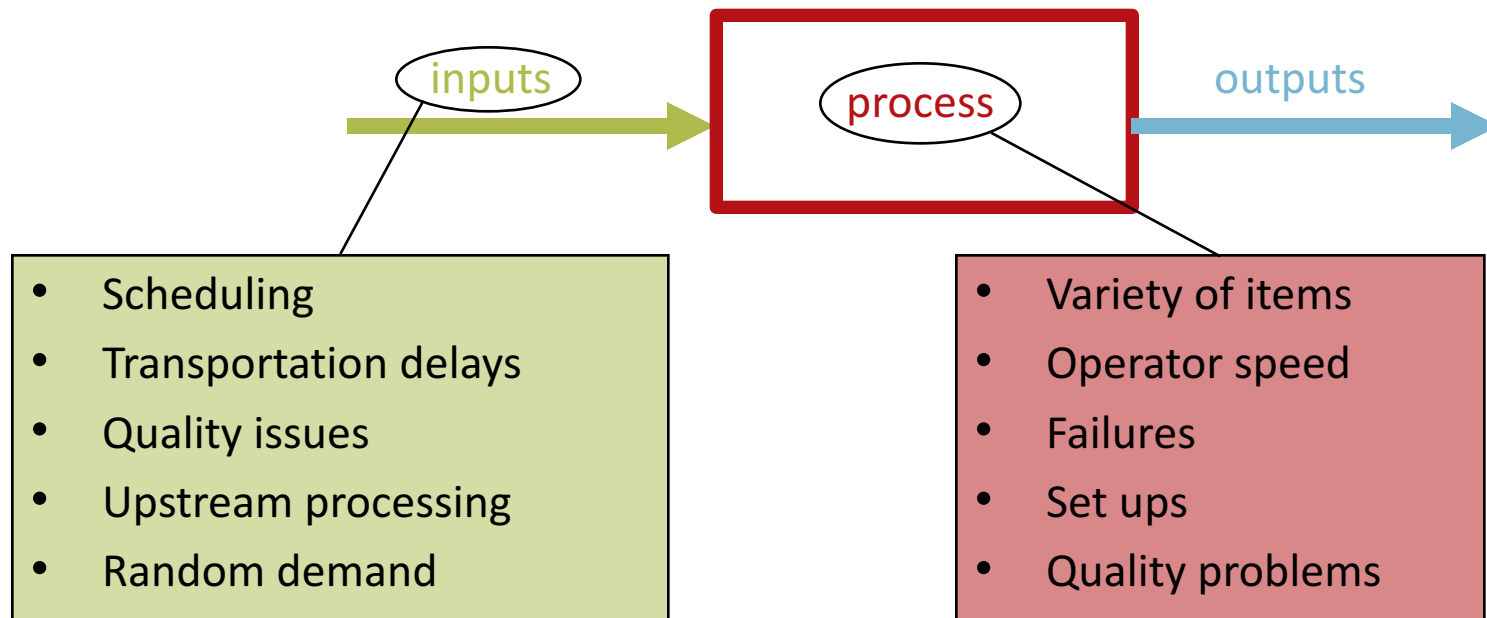
- What if input is at system capacity?



We will see that both *WIP* & *CT* will increase dramatically as utilization approaches 100%!

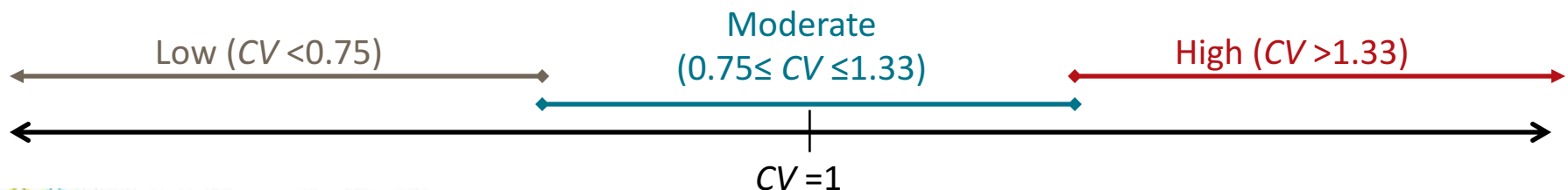
Process Variability

Sources of Process Variability

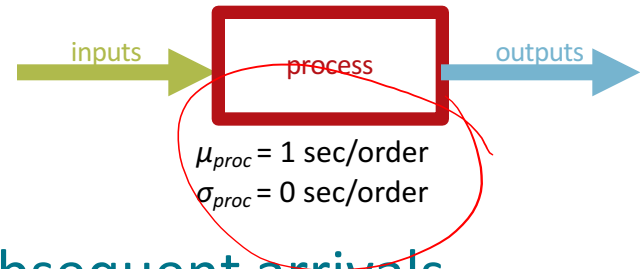


Describing Variability

- Central tendency – typically the mean, μ
- Dispersion around the mean – standard deviation, σ
- Coefficient of Variation – ratio of dispersion to mean, $CV = \sigma/\mu$

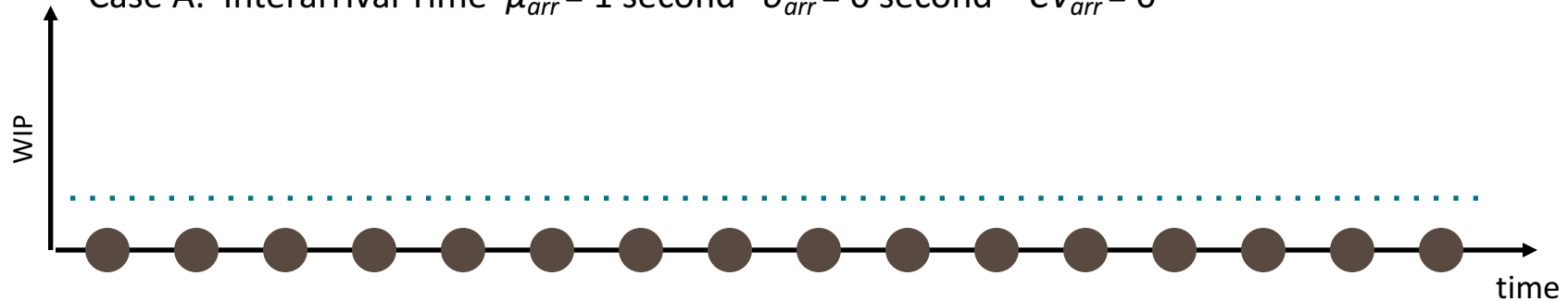


Input Variability

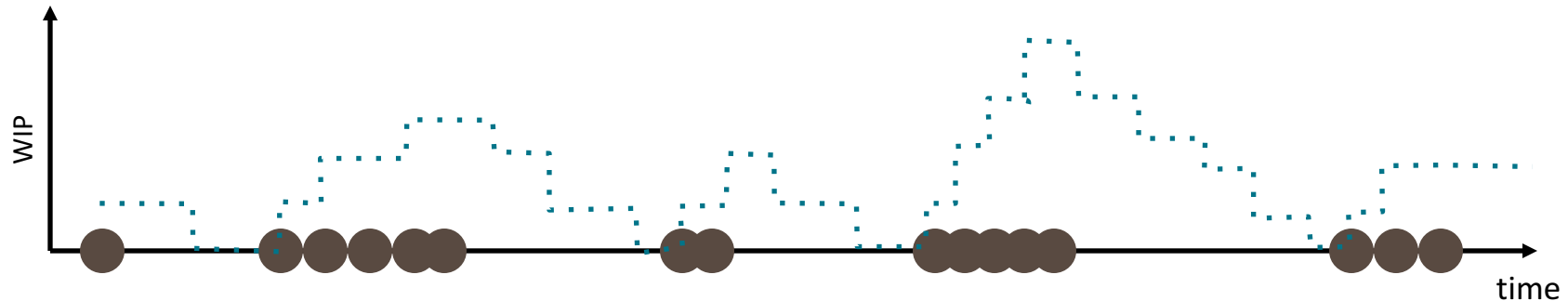


Interarrival Time – time between subsequent arrivals

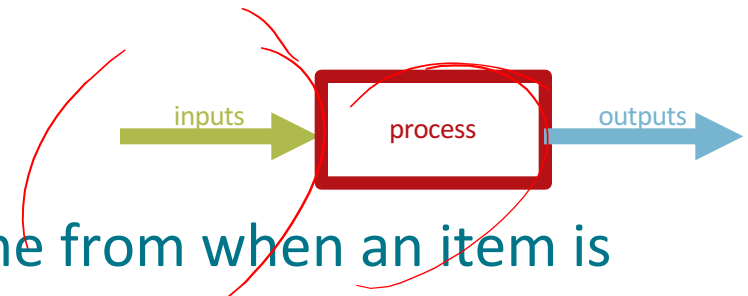
Case A. Interarrival Time $\mu_{arr} = 1 \text{ second}$ $\sigma_{arr} = 0 \text{ second}$ $CV_{arr} = 0$



Case B. Interarrival Time $\mu_{arr} = 1 \text{ second}$ $\sigma_{arr} = 1.3 \text{ second}$ $CV_{arr} = 1.3$



Process Variability



- Effective Process Time – the time from when an item is ready for processing and when it is completed

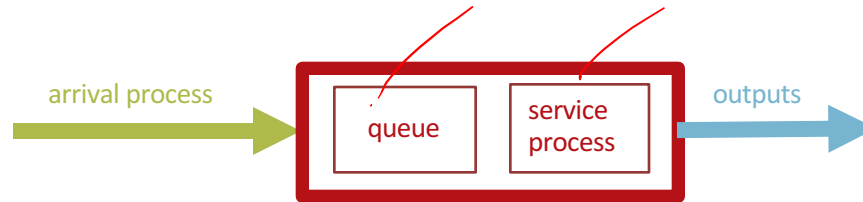
Job	Case A	Case B	Case C
1	21	4	6
2	23	60	4
3	22	4	140
4	24	21	8
5	24	48	5
6	20	6	8
7	22	38	3
8	22	8	4
9	24	54	6
10	26	9	9
11	21	7	120
12	23	8	3
13	18	6	6
14	19	8	7
15	17	4	8
16	16	60	6
Mean	21.4	21.6	21.4
Std. Dev.	2.8	22.1	42.6
CV	0.1	1.0	2.0
Variability	LOW	MOD	HIGH

Increasing variability of the arrival and/or process times increases both *WIP* and *CT* of the system and leads to having items wait or queue.

Introduction to Queueing Theory

Queueing Theory

The study of waiting line phenomena – “Science of Waiting”



Nomenclature

r_a = rate of arrivals (items/time)
 t_a = mean time between arrivals = $1/r_a$ (time/item)
 CV_a = coefficient of variation of interarrivals
 b = buffer size or maximum number of items allowed in system

r_p = rate or capacity of work station (items/time)
 t_p = mean effective process time (time/item) = m/r_p
 CV_p = coefficient of variation of process time
 m = number of parallel servers

Performance Metrics

- t_q = expected waiting time in the queue
- CT = expected time in system = $t_q + t_p$
- WIP = average work in process (items) at station
- WIP_q = average work in process (items) in queue
- u = utilization of the server ($= r_a / r_p$)

Characterizing the Queueing System



- Kendall's notation (X/Y/m/b)

- X = distribution of the interarrival times
- Y = distribution of the effective process times
- m = number of servers at the station
- b = system capacity at any one time

D = Deterministic
M = Exponential
G = General

Deterministic

$$CV = 0$$

M – Exponential (Markovian)

Recall that in a Poisson distribution the interarrival times are distributed exponentially. So $CV \cong 1$.

General –Any distribution with given parameters for mean and standard deviation.

Performance Metrics - M/M/1 Queue

$$WIP = \frac{u}{1-u}$$

$$CT = \frac{WIP}{r_a} = \frac{u}{r_a(1-u)} = \frac{u}{(ur_p)(1-u)} = \frac{t_p}{(1-u)}$$

$$t_q = CT - t_p = \frac{t_p}{(1-u)} - t_p = t_p \left(\frac{1}{1-u} - 1 \right) = \frac{u}{(1-u)} t_p$$

$$WIP_q = r_a t_q = \frac{u}{(1-u)} t_p r_a = \frac{u}{(1-u)} \left(\frac{1}{r_p} \right) (\cancel{r_p} u) = \frac{u^2}{(1-u)}$$

Example – Order Confirmation Processing

- Orders arrive at about 18 per hour and the system can process about 25 orders per hour. There is a single server and assume that both processes are exponentially distributed ($CV=1$).
 1. What is the current utilization?
 2. What is expected number of orders in process?
 3. What is the expected cycle time?
 4. What is the expected time an order waits in the queue?
 5. What is the expected number of orders waiting in the queue?

What do we know?

$$r_a = 18 \text{ items/hour} = 0.30 \text{ items/min}$$

$$t_a = 1/r_a = 0.056 \text{ hrs/item} = 3.33 \text{ min/item}$$

$$CV_a = 1 \text{ (given)}$$

$$r_p = 25 \text{ items/hour} = 0.42 \text{ items/min}$$

$$t_p = 1/r_p = 0.04 \text{ hrs/item} = 2.40 \text{ min/item}$$

$$CV_p = 1 \text{ (given)}$$

$$m = 1 \text{ (given)}$$

$$b = \text{infinity (assumed)}$$

Example – Order Confirmation Processing

1. What is the current utilization?

We know that $u = r_a / r_p = 18/25 = 0.72$

2. What is expected number of orders in process?

For a M/M/1 queue $WIP = u / (1-u) = 0.72/0.28 = 2.57$ orders

3. What is the expected cycle time?

Cycle Time = $t_p / (1-u) = 2.40 / (1 - 0.72) = 8.57$ minutes

4. What is the expected time an order waits in the queue?

$t_q = CT - t_p = 8.57 - 2.40 = 6.17$ minutes

5. What is the expected number of orders waiting in the queue?

$WIP_q = r_a t_q = 0.30(6.17) = 1.85$ orders $= u^2 / (1-u)$

Extending to Other Queues

Estimating t_q



Cycle Time = Waiting Time in Queue + Effective Process Time

$$CT = t_q + t_p$$

t_q = VUT – capturing variability and utilization impacts.

where:

V = the variability (interarrivals and process),

U = the utilization of the station, and

T = the effective processing time (t_p).

Queueing Theory allows us to approximate the cycle time due to the VUT effects for G/G queueing systems.

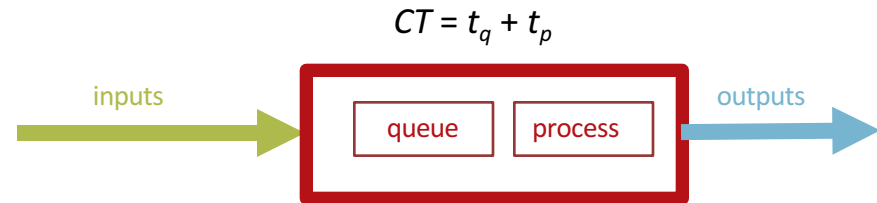
The general case is for a G/G/m/ ∞ queue,

$$t_q \cong \left(\frac{CV_a^2 + CV_p^2}{2} \right) \left(\frac{u^{\sqrt{2(m+1)}-1}}{m(1-u)} \right) t_p$$

VUT Equations

The general case is for a G/G/m/∞ queue,

$$t_q = \left(\frac{CV_a^2 + CV_p^2}{2} \right) \left(\frac{u^{\sqrt{2(m+1)}-1}}{m(1-u)} \right) t_p$$



Where:

CV_a = coefficient of variation of interarrivals
 CV_p = coefficient of variation of process time
 u = utilization of process
 t_p = effective process time
 m = number of parallel servers

The **G/G/1/∞** queue:

For any general distribution for interarrivals and process time and a single server ($m=1$).

$$t_q = \left(\frac{CV_a^2 + CV_p^2}{2} \right) \left(\frac{u}{1-u} \right) t_p$$

The **M/M/1/∞** queue:

For Exponentially distributed (Poisson) interarrivals and process times. This means that $CV_a = CV_p = 1$.

$$t_q = \left(\frac{u}{1-u} \right) t_p$$

The **M/D/1/∞** queue:

For Exponentially distributed (Poisson) interarrivals and Deterministic process times.

This means that $CV_a = 1$ and $CV_p = 0$.

$$t_q = \left(\frac{1}{2} \right) \left(\frac{u}{1-u} \right) t_p$$

The **D/D/1/∞** queue:

For Deterministic interarrivals and process times.

This means that $CV_a = CV_p = 0$.

$$t_q = 0$$

Example – Order Confirmation Processing II

- Orders still arrive at about 18 per hour but actually have a standard deviation of 36. The system can process about 25 orders per hour with a standard deviation of 10. There is a single server.
 1. What is the current utilization?
 2. What is expected number of orders in process?
 3. What is the expected cycle time?
 4. What is the expected time an order waits in the queue?
 5. What is the expected number of orders waiting in the queue?

What do we know?

$$r_a = 18 \text{ items/hour} = 0.30 \text{ items/min}$$

$$t_a = 1/r_a = 0.056 \text{ hrs/item} = 3.33 \text{ min/item}$$

$$CV_a = 36/18 = 2$$

$$r_p = 25 \text{ items/hour} = 0.42 \text{ items/min}$$

$$t_p = 1/r_p = 0.04 \text{ hrs/item} = 2.40 \text{ min/item}$$

$$CV_p = 10/25 = 0.40$$

$$m = 1 \text{ (given)}$$

$$b = \text{infinity (assumed)}$$

Example – Order Confirmation Processing II

1. What is the current utilization?

We know that $u = r_a / r_p = 18/25 = 0.72$

4. What is the expected time an order waits in the queue?

$$t_q = \left(\frac{CV_a^2 + CV_p^2}{2} \right) \left(\frac{u}{1-u} \right) t_p = \left(\frac{2^2 + 0.40^2}{2} \right) \left(\frac{0.72}{1-0.72} \right) (2.40) = 12.84 \text{ min}$$

3. What is the expected cycle time?

$$CT = t_p + t_q = 2.40 + 12.84 = 15.24 \text{ minutes}$$

2. What is expected number of orders in process?

Using Little's Law, $WIP = r_a * \underline{CT} = 0.30(15.24) = 4.57 \text{ orders}$

5. What is the expected number of orders waiting in the queue?

$$WIP_q = r_a \underline{t_q} = 0.30(12.84) = 3.85 \text{ orders}$$

Quantifying the Impact of Variability

Quantifying the Impact of Variability

- Same Example: Orders arrive at 18 per hour with a standard deviation of 36. The system can process about 25 orders per hour with a standard deviation of 10. There is a single server.
 1. What is the current utilization? **72%**
 2. What is expected number of orders in process? **4.57 orders**
 3. What is the expected cycle time? **15.24 minutes**
 4. What is the expected time an order waits in the queue? **12.84 minutes**
 5. What is the expected number of orders waiting in the queue? **3.85 orders**

We know:

$$r_a = 0.30 \text{ items/min}$$

$$t_a = 3.33 \text{ min/item}$$

$$CV_a = 36/18 = 2$$

$$m = 1 \text{ (given)}$$

$$r_p = 0.42 \text{ items/min}$$

$$t_p = 2.40 \text{ min/item}$$

$$CV_p = 10/25 = 0.40$$

$$b = \text{infinity (assumed)}$$

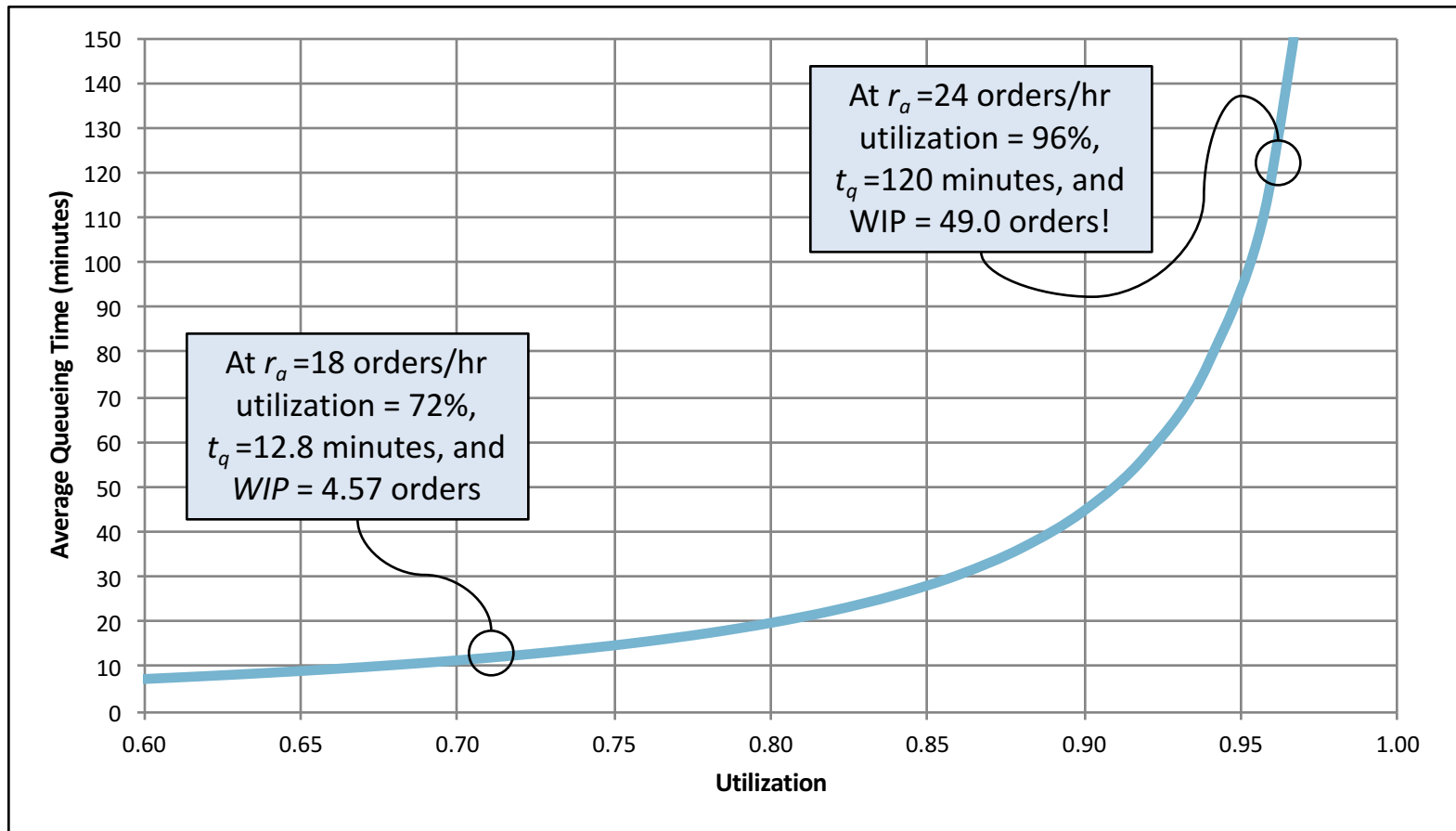
What happens if the arrival rate approaches capacity?

What happens if the variability increases dramatically?

Utilization Impact



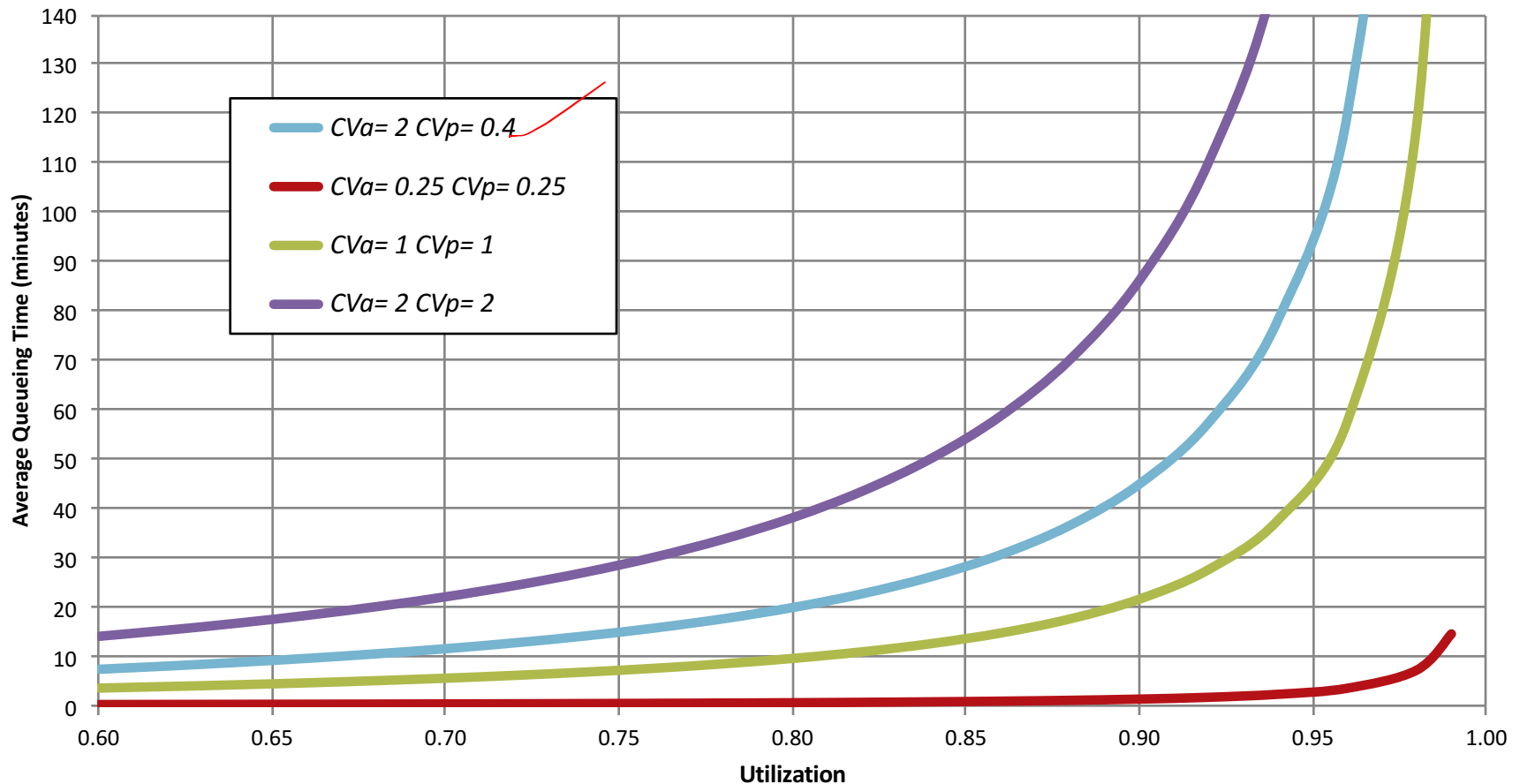
- What happens to average waiting or queueing time if utilization increases?
 - Increases non-linearly along with *WIP*!



Variability Impact



- What happens to average waiting or queueing time if variability increases?



Variability Impact



- Heat Map of waiting time in queue (t_q) for changes in variability of arrivals and processing time. Assumes a 72% utilization.

Cv _{arr} \Cv _{proc}	-	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00	1.10	1.20	1.30	1.40	1.50	1.60	1.70	1.80	1.90	2.00
-	-	0.0	0.1	0.3	0.5	0.8	1.1	1.5	2.0	2.5	3.1	3.7	4.4	5.2	6.0	6.9	7.9	8.9	10.0	11.1	12.3
0.10	0.0	0.1	0.2	0.3	0.5	0.8	1.1	1.5	2.0	2.5	3.1	3.8	4.5	5.2	6.1	7.0	7.9	8.9	10.0	11.2	12.4
0.20	0.1	0.2	0.2	0.4	0.6	0.9	1.2	1.6	2.1	2.6	3.2	3.9	4.6	5.3	6.2	7.1	8.0	9.0	10.1	11.3	12.5
0.30	0.3	0.3	0.4	0.6	0.8	1.0	1.4	1.8	2.3	2.8	3.4	4.0	4.7	5.5	6.3	7.2	8.2	9.2	10.3	11.4	12.6
0.40	0.5	0.5	0.6	0.8	1.0	1.3	1.6	2.0	2.5	3.0	3.6	4.2	4.9	5.7	6.5	7.4	8.4	9.4	10.5	11.6	12.8
0.50	0.8	0.8	0.9	1.0	1.3	1.5	1.9	2.3	2.7	3.3	3.9	4.5	5.2	6.0	6.8	7.7	8.7	9.7	10.8	11.9	13.1
0.60	1.1	1.1	1.2	1.4	1.6	1.9	2.2	2.6	3.1	3.6	4.2	4.8	5.6	6.3	7.2	8.1	9.0	10.0	11.1	12.3	13.5
0.70	1.5	1.5	1.6	1.8	2.0	2.3	2.6	3.0	3.5	4.0	4.6	5.2	6.0	6.7	7.6	8.5	9.4	10.4	11.5	12.7	13.9
0.80	2.0	2.0	2.1	2.3	2.5	2.7	3.1	3.5	3.9	4.5	5.1	5.7	6.4	7.2	8.0	8.9	9.9	10.9	12.0	13.1	14.3
0.90	2.5	2.5	2.6	2.8	3.0	3.3	3.6	4.0	4.5	5.0	5.6	6.2	6.9	7.7	8.5	9.4	10.4	11.4	12.5	13.6	14.8
1.00	3.1	3.1	3.2	3.4	3.6	3.9	4.2	4.6	5.1	5.6	6.2	6.8	7.5	8.3	9.1	10.0	11.0	12.0	13.1	14.2	15.4
1.10	3.7	3.8	3.9	4.0	4.2	4.5	4.8	5.2	5.7	6.2	6.8	7.5	8.2	8.9	9.8	10.7	11.6	12.7	13.7	14.9	16.1
1.20	4.4	4.5	4.6	4.7	4.9	5.2	5.6	6.0	6.4	6.9	7.5	8.2	8.9	9.7	10.5	11.4	12.3	13.4	14.4	15.6	16.8
1.30	5.2	5.2	5.3	5.5	5.7	6.0	6.3	6.7	7.2	7.7	8.3	8.9	9.7	10.4	11.3	12.2	13.1	14.1	15.2	16.4	17.6
1.40	6.0	6.1	6.2	6.3	6.5	6.8	7.2	7.6	8.0	8.5	9.1	9.8	10.5	11.3	12.1	13.0	13.9	15.0	16.0	17.2	18.4
1.50	6.9	7.0	7.1	7.2	7.4	7.7	8.1	8.5	8.9	9.4	10.0	10.7	11.4	12.2	13.0	13.9	14.8	15.9	16.9	18.1	19.3
1.60	7.9	7.9	8.0	8.2	8.4	8.7	9.0	9.4	9.9	10.4	11.0	11.6	12.3	13.1	13.9	14.8	15.8	16.8	17.9	19.0	20.2
1.70	8.9	8.9	9.0	9.2	9.4	9.7	10.0	10.4	10.9	11.4	12.0	12.7	13.4	14.1	15.0	15.9	16.8	17.8	18.9	20.1	21.3
1.80	10.0	10.0	10.1	10.3	10.5	10.8	11.1	11.5	12.0	12.5	13.1	13.7	14.4	15.2	16.0	16.9	17.9	18.9	20.0	21.1	22.3
1.90	11.1	11.2	11.3	11.4	11.6	11.9	12.3	12.7	13.1	13.6	14.2	14.9	15.6	16.4	17.2	18.1	19.0	20.1	21.1	22.3	23.5
2.00	12.3	12.4	12.5	12.6	12.8	13.1	13.5	13.9	14.3	14.8	15.4	16.1	16.8	17.6	18.4	19.3	20.2	21.3	22.3	23.5	24.7

Multiple Servers

Example – Order Confirmation Processing III

- What happens if arrival rate doubles to 36 orders per hour?
Assume the CV_a does not change and each server can process about 25 orders per hour with a standard deviation of 10.

$$r_a = 0.60 \text{ items/min}$$

$$t_a = 1.67 \text{ min/item}$$

$$CV_a = 2$$

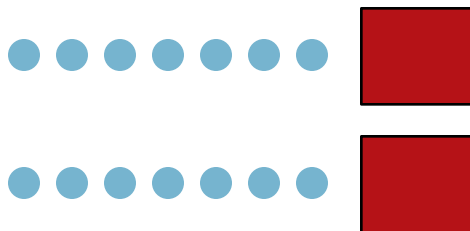
$$r_p = 0.42 \text{ items/min}$$

$$t_p = 2.40 \text{ min/item}$$

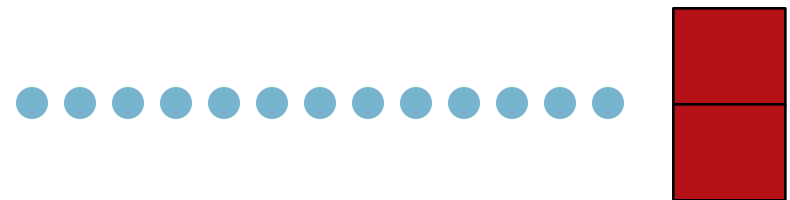
$$CV_p = 0.40$$

- We need to add an additional server – but how?
 - Case 1. Create separate queues with individually dedicated servers.
 - Case 2. Create a single queue for two parallel servers.

Case 1. Multiple Lines Dedicated Servers



Case 2. Single Line Multiple Parallel Servers



Multiple Servers: Case 1 – separate queues

- Create two separate queues $r_a = 0.60$ items/min
 - The arrival rate is simply cut in half – half to each queue
 - Same as earlier example (II)

Dividing arrival rate equally among the two servers gives $r_a = 0.60/2 = 0.30$

Thus, utilization of each server = $0.30/0.42 = 0.72$

$$t_q = \left(\frac{CV_a^2 + CV_p^2}{2} \right) \left(\frac{u}{1-u} \right) t_p = \left(\frac{2^2 + 0.40^2}{2} \right) \left(\frac{0.72}{1-0.72} \right) (2.40) = 12.84 \text{ min}$$

Cycle Time = $t_p + t_q = 2.40 + 12.84 = 15.24$ minutes

Using Little's Law, $WIP = r_a * CT = 0.30(15.24) = 4.57$ orders in each queue = 9.14 total

$WIP_q = r_a t_q = 0.30(12.84) = 3.85$ orders in each queue = 7.70 total

Multiple Servers: Case 2 – parallel servers

What is the current utilization?

We know that utilization will change with 2 servers.

Recall that $u = r_a / r_p = r_a t_p / m = (0.60)(2.40) / 2 = 0.72$

What is the expected time an order waits in the queue?

$$t_q = \left(\frac{CV_a^2 + CV_p^2}{2} \right) \left(\frac{u^{\sqrt{2(m+1)}-1}}{m(1-u)} \right) t_p = \left(\frac{2^2 + 0.40^2}{2} \right) \left(\frac{0.72^{\sqrt{2(2+1)}-1}}{2(0.28)} \right) 2.40 = 5.54 \text{ min}$$

What is the expected cycle time?

$$CT = t_p + t_q = 2.40 + 5.54 = 7.94 \text{ minutes}$$

What is expected number of orders in process?

$$\text{Using Little's Law, } WIP = r_a * CT = 0.60(7.94) = 4.76 \text{ orders}$$

What is the expected number of orders waiting in the queue?

$$WIP_q = r_a t_q = 0.60(5.54) = 3.32 \text{ orders}$$

Comparing Cases

	Case 1 (2 queues)	Case 2 (single queue)
Utilization	0.72	0.72
Cycle Time (CT)	15.24 min	7.94 min
Expected Time in Queue (t_q)	12.84 min	5.54 min
WIP	9.14 orders	4.76 orders
WIP in queue (WIP_q)	7.70 orders	3.32 orders

$$t_q = \left(\frac{CV_a^2 + CV_p^2}{2} \right) \left(\frac{u^{\sqrt{2(m+1)}-1}}{m(1-u)} \right) t_p$$

Case 1: $t_q = (2.08)(2.57)(2.40) = 12.84 \text{ min}$

Case 2: $t_q = (2.08)(1.11)(2.40) = 5.54 \text{ min}$

- Parallel servers will outperform dedicated servers when utilization and variability are the same – due to variability pooling.
- A delay in processing one order has a larger impact with dedicated lines.

Key Points from the Lesson

Key Points (1/2)

- Process Measurements

- Throughput (TH) - Rate at which items are processed (items/time)
- Work in Process (WIP) - Number of items in the system (items)
- Cycle Time (CT) - Time required for a unit to traverse the system (time)
- Capacity - Maximum average rate at which items can flow through the system (units/time)
- Utilization - Rate of flow into a process divided by its capacity
- Bottleneck - The process in a system with the highest utilization

- Little's Law $WIP = CT * TH$

- Over the long-term, the average **work in process** (WIP) is equal to the average **throughput** (TH) times the average **cycle time** (CT).

Key Points (2/2)

- Impact of Variability (arrivals & process)
 - Cycle Time $CT = \text{time in queue} + \text{time in process} = t_q + t_p$
 - Introduced Queueing Theory – Science of Waiting
 - ♦ M/M/1 Queues – exact equations for exponential arrivals/process
 - ♦ G/G/m queues – approximations for more general cases
 - VUT Equations – allow examination of impact of variability

$$t_q = \left(\frac{CV_a^2 + CV_p^2}{2} \right) \left(\frac{u^{\sqrt{2(m+1)}-1}}{m(1-u)} \right) t_p$$

- *WIP* and *CT* increase exponentially as:
 - ♦ Utilization approaches capacity
 - ♦ Variability in either arrivals or processing increases
- Parallel machines/servers outperform dedicated ones due to pooling

Questions, Comments, Suggestions? Use the Discussion!



"Mac – wishing he did not have to queue!"
courtesy of Danaka Porter, Calgary Humane Society
<http://www.calgaryhumane.ca/>



MIT Center for
Transportation & Logistics

caplice@mit.edu
ctl.mit.edu