COMM 204
2018W1 Midterm Review Package

By Dr. Sir Will Zhang, CBE

CUS
commerce
undergraduate
society

Table of Contents

Processes: Introduction

Process: a series of actions or steps taken in order to achieve a particular end. Essentially everything is a process! For this course, we narrow our definition as something that takes *inputs* and turns them into *outputs*. Generally, there are two categories of outputs: products and services. The output of production is a tangible good, the output of a service in intangible.

Production: think of the steps to build a guitar, from wood sourcing, cutting, assembly, stringing, and finishing/painting. Or the process of making a Subway sandwich.

Service: think of a salon, from being greeted, being in queue (if you don't have an appointment), having your hair shampooed, having your hair cut/dyed/burned/whatever, and finally paying and leaving.

These are all fairly standard, linear and common processes.

Note, these are not mutually exclusive. Usually, businesses produce both product and services, and so their business processes involve elements of both production and service. They may also have different processes for both. For example, restaurants produce a product (food), but they also offer service (everything else); the process of waiting tables is obviously distinct from the process of preparing food.

Inventory or "flow unit": some measurable thing that moves through a process (for production: work-in-progress inventory, for a service: a customer, an order, a parcel, a patient). We will discuss this more in "Mapping Processes"

	Product	**Service**
Inventory or flow unit	Work-in-progress/Thing being produced	Can be virtually anything, even you!
Customer contact	Minimal	Extensive
Consumption	Can be delayed	Immediate
Resources	Capital intensive	Labour intensive

Processes: Measure and Assess

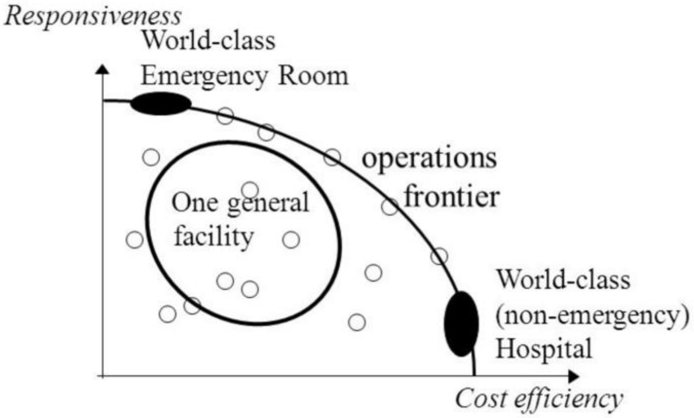
Measuring and comparing processes usually involves some metrics such as

- Cost
- Productivity (outputs/inputs),
- Utilization (inputs/capacity),
- Defect Rate (for production),
- Service Levels (speed/wait times, satisfaction ratings, "friendliness")
- or other judgments.

Efficiency: a broad term that describes minimal waste. Measures of efficiency include productivity (output/input), utilization (inputs/capacity), sales speed (inventory turnover/days of inventory), or other metrics.

Many times, processes are compared with their efficiency, but this is not always the case. For some processes, greater priority may be placed on quality or cost alone!

Choosing a Process: The Operations Frontier



Essentially, 1. plot two dimensions on the axes (usually cost vs. quality) that represent an important trade-off to your business/organization. Each point on the graph represents another business or organization that faces the same trade-off. 2. Along the outermost points (e.g. either greatest quality for any given cost, or lowest cost for any given quality). 3. Draw a line; if you have enough companies on your graph and they are dispersed nicely, you will roughly get a curved "frontier".

In the example above, the trade-off is responsiveness vs. cost-efficiency.

You want to be on the frontier somewhere. If you are "One general facility", for the same cost efficiency, you could have higher responsiveness (or for the same responsiveness, have greater cost efficiency) by improving your processes.

Choosing where you are on the frontier is a ***strategic issue*** for management. Being on the frontier is a ***tactical (execution) issue***. Operational innovation/new technology can ***move the frontier*** outward.

Example of a strategic issue: GM choosing to ignore product quality in favor of cost reductions (which haven't worked) and focusing on profit through their financing division.

Example of tactical issue: GM failing to reduce costs, so they ended up with low quality and high costs. Tactical issues, or execution, is what process/operations deals with primarily.

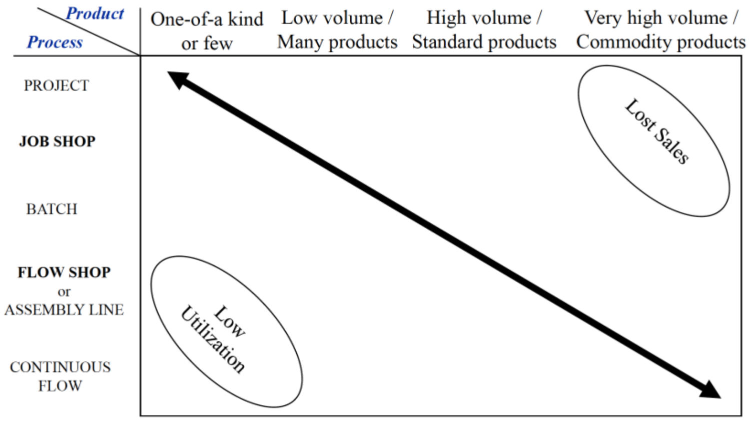
Example of innovation moving the frontier (for cost-quality): Uber? Low-cost carrier airlines (depends how you measure quality, by flight reliability, or amenities/extras such as in-flight entertainment)?

Choosing a Process: The PP-Matrix

There are a few types of processes you must recognize

- ***Project:*** One-time production/service (e.g. a house, any kind of development, special builds). Low volume. Compete on product feature/differentiation. Resource flexibility (low utilization, high cost)
- ***Job shop:*** each customer or product is a "job" that may be at least somewhat different from each other ("many products). Something like a hospital or a (diversified) car repair shop where there are multiple "stations" or tools (e.g. as a hospital is divided into rooms and departments such as radiology, waiting area, admitting, surgical care unit etc. or as a car repair shop would have a hoist, maybe special tools to mount tires, special tools for engine removal etc.) Not all stations/tools will be used for each "job" that goes through the shop!
- ***Batch:*** things are made in batches. Batches may be different from each other.
- ***Flow shop/assembly line*:** individual products are more-or-less mass produced (e.g. iPhones, cars, most things you use in day-to-day life). High volume. Compete on costs. Resource specialization (high utilization)
- ***Continuous flow:*** a flow shop where flow units are not discrete (or are small enough that they may not be considered discrete) e.g. beer or rice. Very high volume.

It's not always clear how to categorize a process. A process can have characteristics of more than one of the above (e.g. Subway has job shop qualities but can be considered an assembly line).

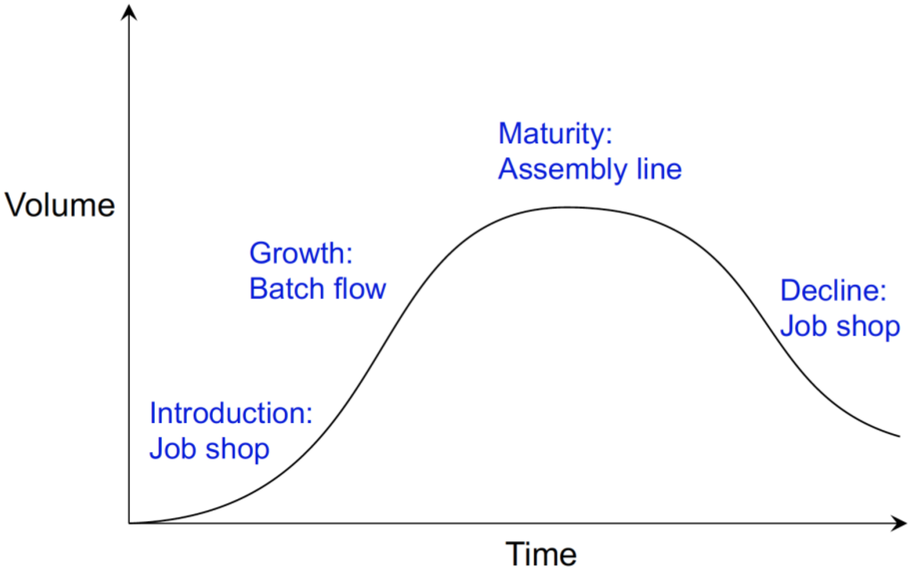

Product
Process
One-of-a kind or few
Low volume / Many products
High volume / Standard products
Very high volume / Commodity products
PROJECT
JOB SHOP
BATCH
FLOW SHOP or ASSEMBLY LINE
CONTINUOUS FLOW
Lost Sales
Low Utilization

On the product-process matrix:
Low utilization = you've spent too much on capital costs! e.g. buying robots and building an assembly line to produce limited edition Louis Vuitton baobaos. Also, potential negative effect on "perceived quality".

Lost sales = Opportunity costs! You make a product with high demand, but you treat each one as a project so you lose sales by not having capacity to meet demand.

In general, be on the line.

Your product's needs may evolve over time (product life cycle). Here's a general idea of how that works.



Mapping Processes

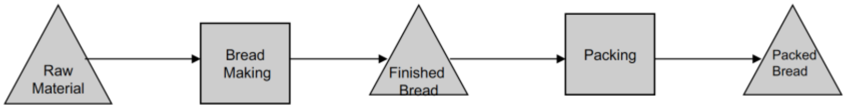
OK before we begin this, you must first know the difference between a rate and a "time". If something I do takes 5secs (my "unit load", defined below), the rate that I can do this is "1/5 units per second" or 1/5(units/second)*60(seconds/minute) = 12/minute or 12(units/minute) x 60(minutes/hour) = 720/hr.

Conversely, if I can do something at a rate of 5/hour, it takes me 1/5 hours (or 12 minutes) to do this. Be comfortable with these conversions, and don't mix the two up! Also, make sure you're always comparing same unit of time (min/hours)

There are many ways to show a process; in general, we will use a process flow diagram. Drawing one out is called "mapping" a process.

First, some definitions:
Flow units (from before): The items that flow through the process (can be heterogenous)
Activities: The transformation steps in the process. Each takes some time to complete
Resources: Things that perform the activities; they have capacities (e.g. machine or worker)
Buffers: Storage units for flow units (queues where flow units wait before another activity can be performed on them. May have finite size.
Flow units are not mapped/drawn on the process flow diagram. Activities are represented with squares, queues are triangles, anything involving a decision is a diamond.

Example: Bread making



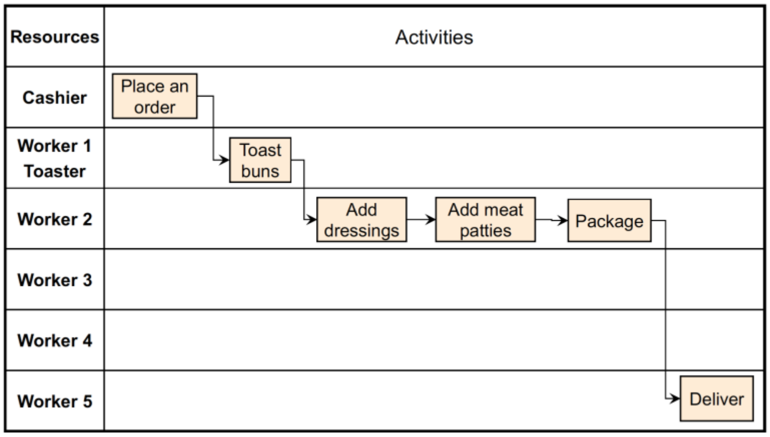
Note for different bread types, 'bread making' and 'packing' may differ for each

To represent different bread types, use a diamond before, say, breadmaking and have it branch out into two+ branches. This can get really complicated and turn into a large tree.

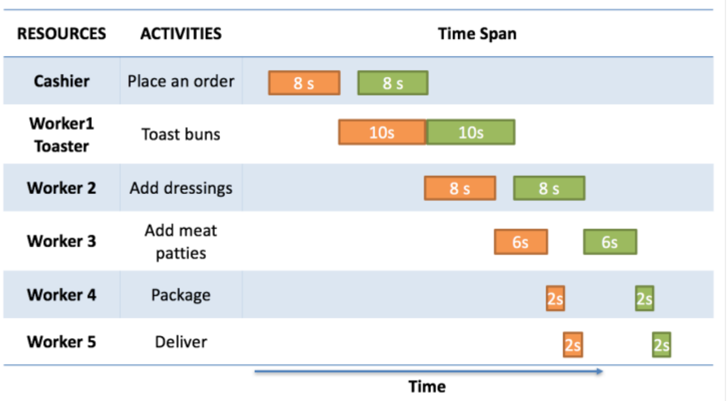
Be reasonable when choosing the level of detail and aggregation. I've actually seen students map a pizza shop process (or something) with about 40 steps, where the last step was "Employee hands order to customer. Avg. flow time 88.5ms. Capacity rate 4328 handing/hour" … Firstly, the capacity is wrong, second, DON'T DO THIS!

Other ways to map:

Swim lane flow chart: resources on left, activities on right



Gantt chart: basically, put time and list the activities on side. Useful for processes that repeat like below (first order/burger is orange; second order/burger is green. You can see the bottleneck is the process with no gap).

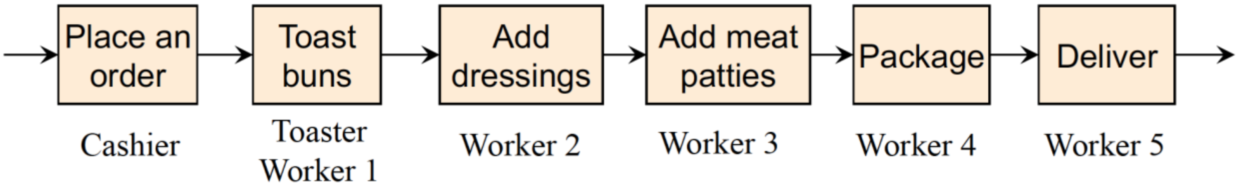


Wait, what's a bottleneck?

More definitions:

Unit load: the amount of time it takes a resource to process each flow unit
Capacity: how much output the process is about to product in a given time (can be per minute, per hour, per day, per year etc.)
Bottleneck: the *activity or resource* that limits the capacity of the whole process. e.g. in the process above, either Worker1 and Toaster is the bottlenecks, or "Toast buns". When we study the capacities of these later, we will see which one specifically is the bottleneck. Usually only one, but can be more than 1 if process is limited by 2 resources with the same slow capacity. Whatever the capacity of the bottleneck is = the capacity of whole process!!!!

For example, this is a process flow diagram.



8s	10s	8s	6s	2s	2s
450/hr	360/hr	450/hr	600/hr	1800/hr	1800/hr

Here are the flow times of each activity (8s, 10s etc). Similarly, time for each resource to finish a task is called Unit Load. Capacity of each activity (ideally, we would have capacity for each resource... We will assume from hereon that the capacity of the toaster is 360/hr and the capacity for worker 1 is equal to or greater than that)

Theoretical flow time: the sum of all the flow times. Ignores "waiting" (when a flow unit has gone through one activity but the following resource/activity is not finished with the previous flow unit, e.g. cashier takes an order but the order waits for toaster/worker 1 to finish the previous order.) = 36s

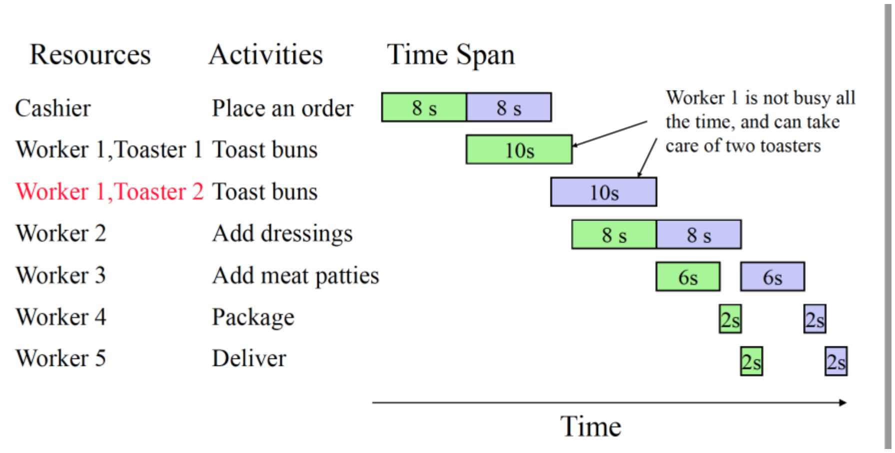
Capacity rate of the whole process = 360/hr (the capacity rate of the bottleneck!)

To increase capacity, you MUST target the bottleneck. Either by increasing resource pool (add another toaster) or targeting the *unit load* or how fast/effective the resource is (buy a better toaster). Targeting non-bottlenecks does nothing, and you'll probably be fired for wasting company resources.

Note that adding another toaster (working in parallel) does not decrease theoretical flow time, but doubles capacity of the resource nonetheless (capacity of both toasters now 720/hr). Is capacity for the entire process doubled???? NO! Something else will be the bottleneck (Cashier and Worker 2).

Increasing capacity of one resources only increases capacity of process up until another resource becomes the bottleneck (i.e. when you increase capacity right to the point where reaches the capacity of another resource).

Two toaster Gantt chart: Order 1 in green, order 2 in purple



Another useful benefit of mapping is it also allows us to identify useless/wasted resources. By the time worker 3 is done adding meat patties on the first order, order 2 won't have its dressing finished yet so he will wait/do nothing until it is (he has no work to do: this is called being "*starved*"). In that time, he could also just package the first order thus worker 4 is pretty redundant. Or, alternatively worker 4 can also deliver.

Look for large gaps between activities for each resource to see if another activity can fit in the gap.

Similarly, in the one toaster case, the "Place an order" process is "*blocked*". Flow unit has nowhere to go because the toaster won't be finished with previous order yet (it is in *buffer*).

For a more complicated example (examinable) we have multiple types of flow units:

Resource	Unit Load (minutes/unit)			
	Product A	Product B	Product C	1A + 2B + 2C
1	2.5	2.5	2.5	12.5
2	1.5	2	2.5	10.5
3	12	0	0	12
4	0	3	3	12
5	3	3	3	15

Assume our process produces 1 of product A, 2 of B and 2 of C. We have 5 resources. If we only produce A, the capacity of our process is 60(min/hour)/12(min/unit) = 5unit/hr. If we produce this mix, add the unit loads of each resource and determine each resource's capacity. It's going to be 60/15=4/hr. Note: to find process capacity, you don't have to find the capacity of each resource, just the one with the longest unit load.

Changing your product mix changes process capacity.

Inventory (Queues - Part 1)

Enough (purely) theoretical stuff for now. Let's think about what would actually happen if these processes were to work/be used. Think of this part separately from previous sections.

Definitions

Throughput rate = the rate at which flow units actually move through the process. It is the lower of the input rate (AKA demand rate) and the capacity rate - if only 3 orders are being made per hour and your burger joint has capacity of 20 orders per hour, your throughput rate is 3/hr. Usually expressed as an average.

Flow time: for this section is the total time it takes something to go through a process, including waiting. NOT THE SAME AS THEORETICAL FLOW TIME. Usually expressed as average.

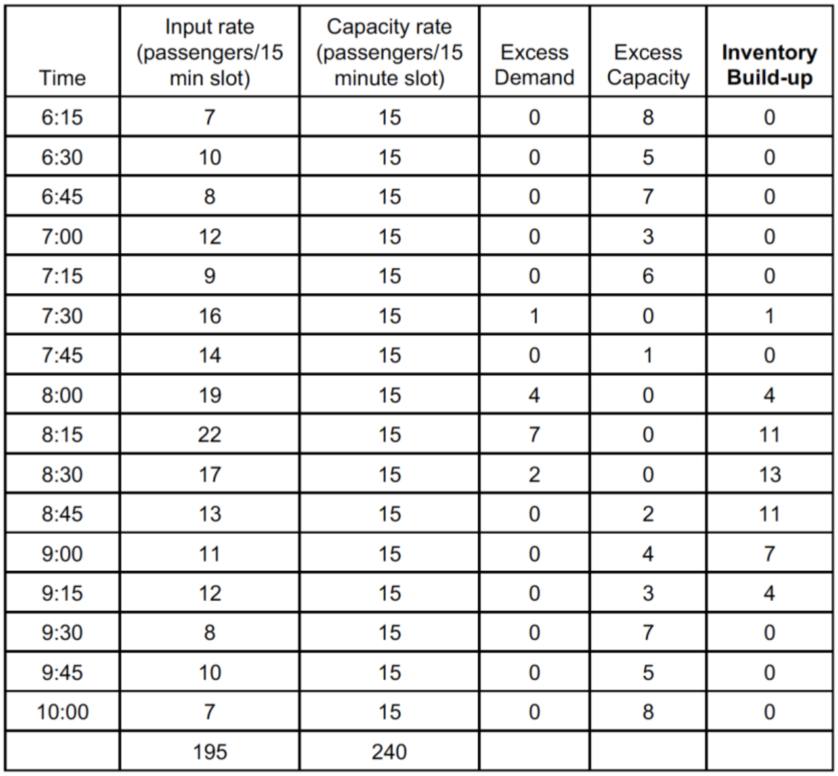
Cycle time: average time between completion of flow units (between flow units exiting the process).

Utilization = (throughput rate)/(capacity) (so essentially, input rate/capacity unless input rate is > capacity, in which case it is just 100%). Can never be greater than 1 (100%).

Implied Utilization: input rate/capacity (assumes organizations can do "overtime" or pump up extra capacity out of nowhere). Kind of useful metric, I guess? If it's too high you should probably think about improving your process capacity at certain times (by targeting the bottleneck! As we learned above).

You want utilization maybe somewhere around 70-80% (usually in a low-variability process) unless you have no variability in demand (deterministic/not stochastic demand) then you can approach 100%. Desired utilization depends on your demand and service time variability.

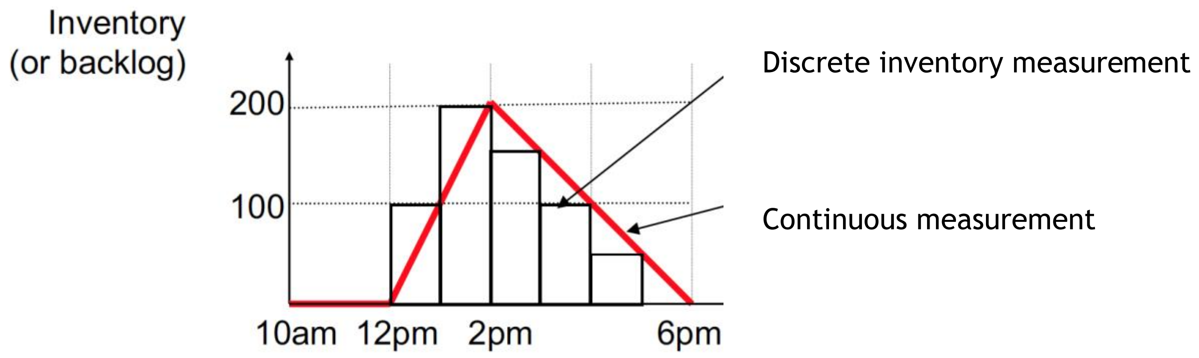
Here's an airport security screening example you've seen in your slides. Figure out how the excess demand/excess capacity/inventory build-up columns are calculated. For a service, inventory usually = queue. Most of our discussion on inventory will be regarding queues.



Time	Input rate (passengers/15 min slot)	Capacity rate (passengers/15 minute slot)	Excess Demand	Excess Capacity	**Inventory Build-up**
6:15	7	15	0	8	0
6:30	10	15	0	5	0
6:45	8	15	0	7	0
7:00	12	15	0	3	0
7:15	9	15	0	6	0
7:30	16	15	1	0	1
7:45	14	15	0	1	0
8:00	19	15	4	0	4
8:15	22	15	7	0	11
8:30	17	15	2	0	13
8:45	13	15	0	2	11
9:00	11	15	0	4	7
9:15	12	15	0	3	4
9:30	8	15	0	7	0
9:45	10	15	0	5	0
10:00	7	15	0	8	0
	195	240			

The above table is *discrete* (there are time periods where we count inventory/queue length instead of it being continuously updated every time someone joins the queue or leaves it)

Inventory diagram: discrete vs. continuous. (Just an illustration; this is not the airport security example above)

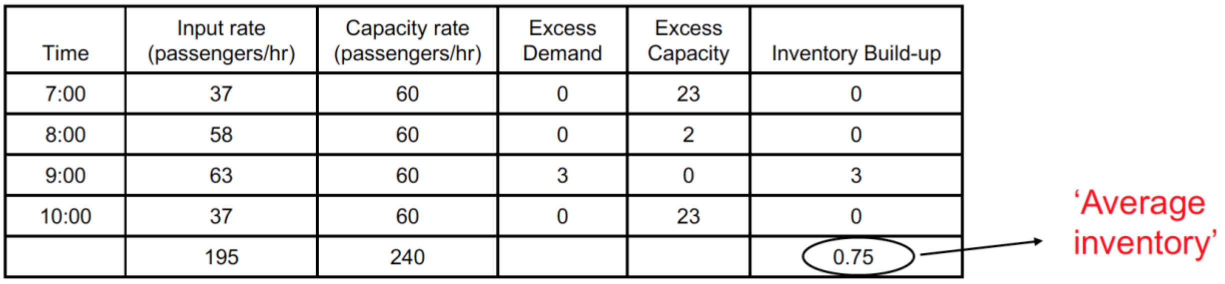


Discrete inventory measurement

Continuous measurement

To calculate average inventory in discrete model, you can just take a simple average. For airport security example, (1+4+11+13+11+7+4)/(16 time periods) = 3.1875.

Note, if we have bigger “blocks” (say 30min or 1hour), average inventory would decrease. Airport security example aggregated by hour instead of 15min blocks:



Time	Input rate (passengers/hr)	Capacity rate (passengers/hr)	Excess Demand	Excess Capacity	Inventory Build-up
7:00	37	60	0	23	0
8:00	58	60	0	2	0
9:00	63	60	3	0	3
10:00	37	60	0	23	0
	195	240			0.75

‘Average inventory’

To compute average inventory in a continuous model, you must do some integration.

$$f_{avg} = \frac{1}{b-a}(\int_a^b f(x)dx$$

Which is a fancy way of denoting: the average of a 2D function (line) between two x-values is the area under the function divided by the range at which the integral is evaluated at, which is yet another fancy way of saying “find the area under the curve and divide by total time elapsed”. See practice question 4 for how to do this exactly.

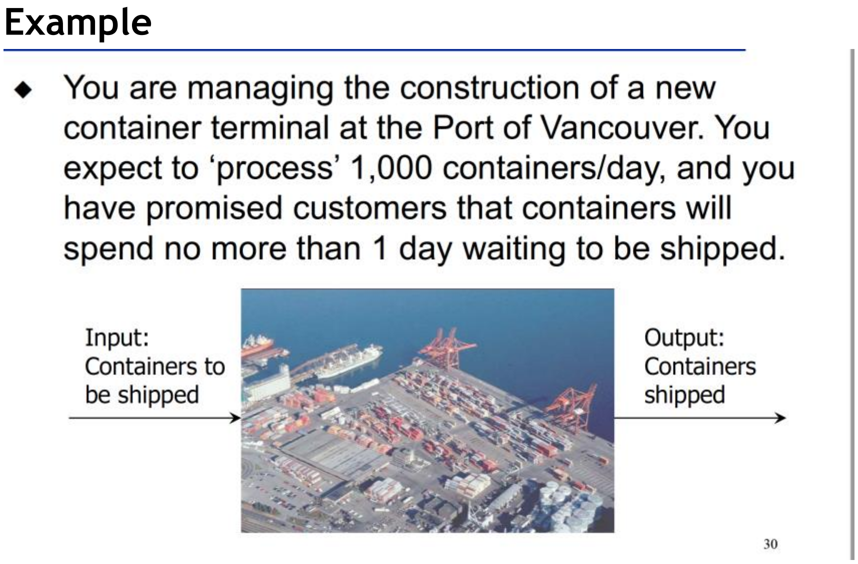
Since the input and output rates may vary over time, both the ‘short-run average’ and ‘long-run average’ rates provide useful information:

- ***Long-run average input rate is assumed to be less than the long run average capacity rate*** This is an important assumption which will carry through for rest of this course. Why? Because if your input rate is greater than capacity, your queue will just grow infinitely… And it’s not even worth discussing. Actually, we could assume less than or equal to, but we won’t.
- And so long-run average throughput rate = Long-run average input rate
- Short-run average input rate can be greater than the short-run average capacity rate but this would lead to long queues.

Little’s Law and Inventory Turnover

I = R x T

(Average) Inventory = (Average) Throughput Rate x (Average) Flow Time
The fact that it is an average is usually implied. If you have a question that gives you two of these things (or something that can be used to find two of these things (e.g. average flow rate instead of flow time), and it’s asking for the third, remember to use this equation. And **please** remember to use same units of time! (Throughput rate 5units/hr cannot multiply by flow time of 8min. The time units won’t cancel out)



Example

- You are managing the construction of a new container terminal at the Port of Vancouver. You expect to 'process' 1,000 containers/day, and you have promised customers that containers will spend no more than 1 day waiting to be shipped.

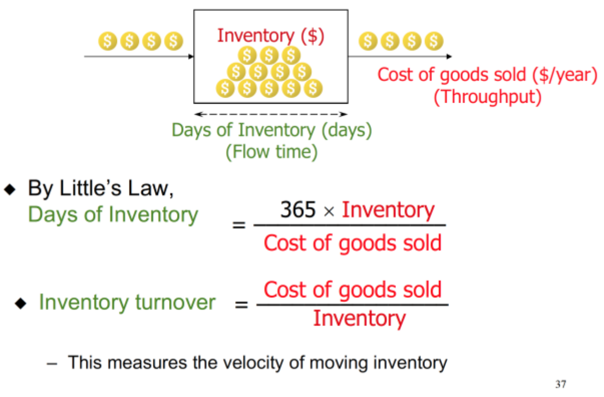
a) On average, your container storage yard can hold 500 containers. Is your yard big enough?
I = R x T
I = 1000 (containers/day) x 1day → Average inventory is 1000 containers. Answer: No lol.
b) Suppose the yard is expanded to hold 2000 containers. Since container traffic is growing rapidly, you are soon processing 2000 containers/day. You are asked to make improvement to the terminal to handle 4000 containers/day. But there is no more room to expand the yard. What changes can you make in order to process 4000 containers/day?

Well, two ways to "handle" this: increase inventory space (which we can't do) or decrease flow time. if you can't target I (by increasing inventory space available), and you can't change R (or it's undesirable to decrease demand), then target T: ship stuff faster (decrease flow time).

Detour to some accounting stuff that has nothing to do with anything else in this course:
Inventory Turnover = some accounting metric that estimates how many times you've essentially moved your entire inventory in a given period of time (usually 1 year).

Here, we are discussing actual inventory (product) held, not queues.
Inventory Turnover (or turns)
= (Cost of goods sold) / (Average inventory investment)
= ($ value of cost of output) / ($ value of average inventory)
= R / I
= 1/T
Where average inventory = (beginning inventory + ending inventory)/2 over the time period being evaluated.

More formulas:



- By Little's Law,
  Days of Inventory $= \dfrac{365 \times \text{Inventory}}{\text{Cost of goods sold}}$
- Inventory turnover $= \dfrac{\text{Cost of goods sold}}{\text{Inventory}}$
  - This measures the velocity of moving inventory

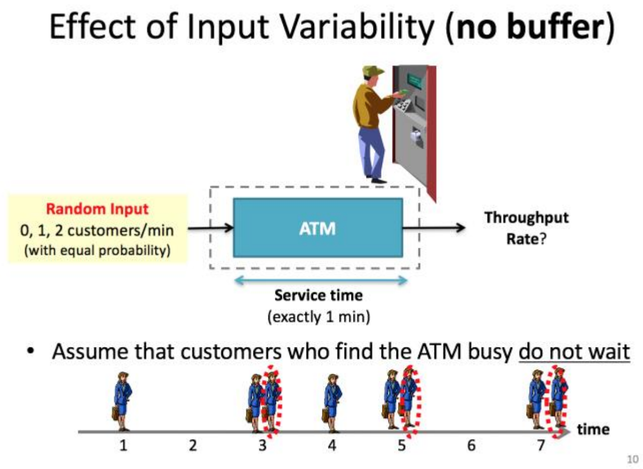
37

Variability

Sometimes demand is higher than other times (as we saw in inventory). Capacity or service time can also be variable (if each flow unit requires different treatment like haircuts take different times, or if your workers have unreliable efficiency and sometimes are fast, other times slow).

Predictable Variability: defined as the name implies. Can control/adjust for this (increase prices to alter demand).

Unpredictable/stochastic variability: defined as the name implies. Due to lack of information. Can prevent this by buying/acquiring information.



Effect of Input Variability (no buffer)

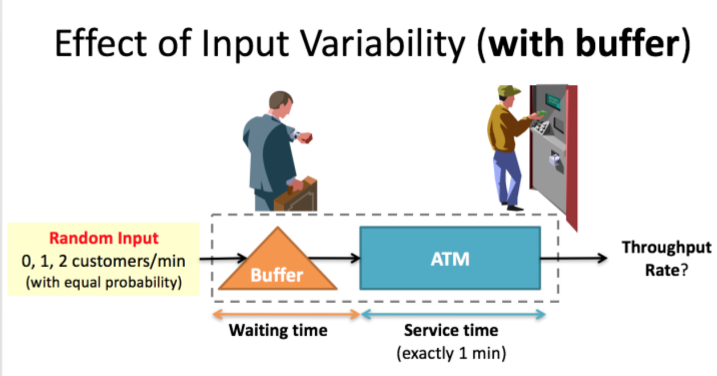
- Assume that customers who find the ATM busy do not wait

10

With no variability, there would be no queues. Because service time would be 1min, and customers come 1 min at a time (inter-arrival time = exactly 1min). Customer 2 would arrive exactly when the ATM is finished with customer 1.

With variability, because customers do not come in this pattern, sometimes there may be times where 2 customers come, and 1 gets “lost” (they see the ATM is busy and walk away). Note, average input rate is still 1/min = (0 x (1/3) + 1 x (1/3) + 2 x (1/3)).

Essentially, you need a buffer (queue) if you face variability or you will get lower throughput (recall throughput = (input rate)/(capacity rate) of process. With variability and no buffer, input rate is effectively reduced because people leave), resulting in lost revenue.

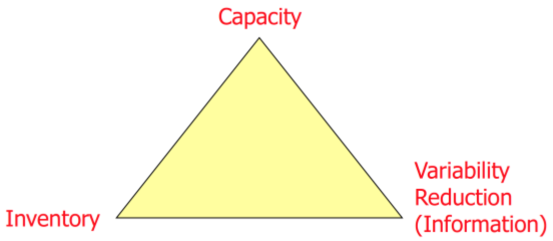

Effect of Input Variability (**with buffer**)

When you add a buffer (queue) and assume it will never “fill up” and nobody leaves it, you don’t lose demand anymore.

But, suppose you have a buffer, compared to no-variability, variability still leads to an increase in the average inventory in the system (which includes the lineup) and an increase in the average flow time (including time in queue).

The OM (Operations Management) Triangle

OM triangle displays a tradeoff. You can’t have low inventory, low capacity, low information at the same time. “A senior executive at Dell described inventory as ‘the physical embodiment of bad information’”



Quantitative Analysis: Let's define some terms:

- λ (lambda) = long-run average input rate (in customers/minute or customers/hour etc.)
- $1/\lambda$ = (average) customer inter-arrival time (one customer every ___ minutes, hours etc.)
- 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 $\lambda < \mu$. Safety capacity = $\mu - \lambda$ (i.e. how much extra capacity we have to account for variability). We assume stability in all our systems that we perform analysis on.
- K = buffer capacity (for now let K = infinity)
- c = number of servers in the resource pool (for now let c = 1)
- Utilization of a process = rho (that "*p*"-looking thing) = λ/μ = input rate/capacity rate
- CV = Coefficient of variation (there is C_s and C_a). For service, it is standard deviation of service **times** over average service **time**. Similar for inter-arrival.

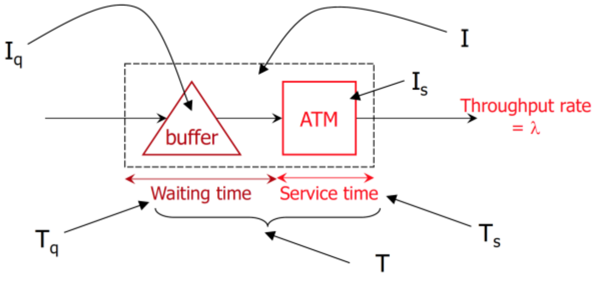
Note: we are focusing on long-run averages, ignoring the predictable variability that may be occurring in the short run. In reality, we should be concerned with both types of variability.

For the remainder of our queueing analysis, we break down our process into two parts: buffer/queue (subscript q) and service (subscript s),

$T = T_q + T_s$. where T = total average flow time (including time in queue and service) T_q is the average queue time, T_s is average service time (not including queue).

$I = I_q + I_s$. where I = total average customers in system which includes both those in queue and being served, I_q is the average customers in queue or average length of queue, I_s = average number of customers being served at any time.

Again, we always assume $\lambda < \mu$, or else all this analysis is completely pointless because the queue will go on forever.



- Little's Law holds ($I_q = \lambda T_q$, $I_s = \lambda T_s$, and $I = \lambda T$)

Make sure you are comfortable with rearranging Little's law formulas ($T_q = I_q/\lambda$, for example)

Variability: P-K Formula

The following formula is used to calculate average inventory in a queue (average queue length). When we have average queue length, we can use it in conjunction with Little's law to compute average queue time and other useful stuff.

For example, if you compute I_q, you can use Little's law ($T_q = I_q/\lambda$) to find T_q, you can add that to average service time (T_s) which is usually given (or can be backed out by doing $1/\mu$ (average service rate), to get total average flow time = T. From here, you can apply Little's law again ($I = T/\lambda$) to get total inventory aka average number of customers in the process (queue and service). Subtract I_q and you will get I_s (average number of customers being processed at any time).

Is there an easier way to get I_s?

Directly apply Little's law: $I_s = T_s \times \lambda$ where $T_s = 1/\mu$ (again, you want to make sure T_s and μ are in the same units of time).

This works because lambda (input rate) is assumed to be our throughput rate (the lower of input and capacity rates), and so lambda = throughput rate of the entire process = lambda of the service activity.

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

"=" for special cases "≈" in general

I_q = average queue length (excl. the one in service)

ρ = (long run) average utilization

= average throughput / average capacity $= \lambda / \mu$

C_a = 'coefficient of variation' of inter-arrival time $= \sigma\{a\}/E\{a\}$

C_s = 'coefficient of variation' of service times $= \sigma\{s\}/E\{s\}$

Can be reformulated to

$$Iq = \frac{\lambda}{\mu} \times \frac{\lambda}{\mu - \lambda} \times \frac{C_a^2 + C_s^2}{2}$$

which is equivalent. Personal preference on which to use.

The P-K formula, as above, is useful under the following 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 until served
- First-in-first-out
- All units arrive independently of each other (the probability of customer arriving doesn’t depend on whether another customer is arriving; in what kinds of situations would this assumption not hold?)
- Inter-arrival times and service times follow a “general” distribution. Called a G/G/1 queue First G represents that inter-arrival times follow a general distribution, second represents that service times follows a general distribution. 1 = number of servers.

G/G/1 SIMPLE EXAMPLE:

Customers arrive at rate 4 per hour, 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 and the average time that a flow unit spends in the queue?

First, define our information (in averages)
λ = arrival rate = 4 customers arrive/hour
1/λ = inter-arrival time = ¼ hours between customer arrivals (keep same units, could also convert all to minutes)

1/μ = service time = 1 customer served every 1/6 hours (i.e. 10 minutes)
μ = 6 served/hour

Stdev(Inter-arrival times) = 5min = 5/60 = 1/12 hours
Stdev(Service times) = 3min = 3/60 = 1/20 hours

rho (*p*) = λ/μ = 4/6 = 2/3 = 66.67% < 1 (Good! Or else we can’t do anything)

C_a = (1/12)/(1/4) = (1/3) = 0.3333 (or 5min/15min)
C_s = (1/20)/(1/6) = (3/10) = 0.3 (or 3min/10min)

$$I_q \cong \frac{\frac{2^2}{3}}{1-(\frac{2}{3})} * \frac{\left(\frac{1}{3}\right)^2+\left(\frac{3}{10}\right)^2}{2} \cong 0.13407$$

On average, there are 0.13407 people in line. (There are times where the line is empty, so avg. can and often will be < 1)

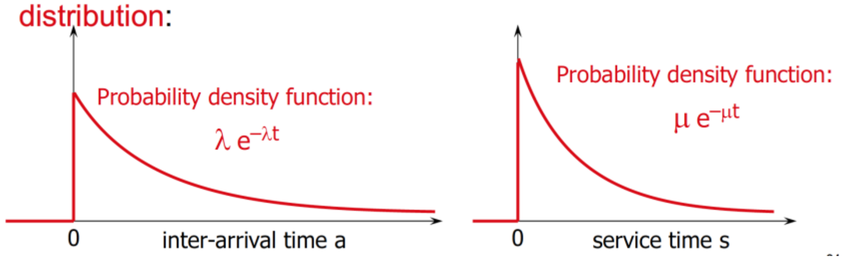
Assuming my I_q is correct, $T_q = I_q$ / λ = 0.13407/4 = 0.03342 hours or 2.01105 minutes avg. queue time.

A quick detour on probability stuff and M/M/1:

Inter-arrival times and service times need not follow a “general” distribution! If you don’t know what kind of distribution it is, a good assumption is using an exponential distribution for both (M/M/1) where M = times are exponentially distributed.

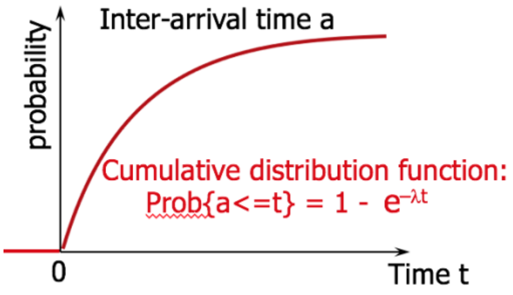
You don’t need to worry too much about what how the probability functions work. But here are the general shapes of an exponential distribution function (exponentially-distributed times).

- A most commonly used distribution is the exponential distribution:



Essentially, shorter times are much more likely than long times.

(Cumulative probability function is just the area under this curve, or the integral of these curves. Don’t worry too much about those either).



For an exponential distribution, CV for arrival and service times = 1. So the second half of the same P-K formula (Cs^2+Ca^2)/2 = 1. Much easier! Also, the formula is exact =, not approximate ≃

So for M/M/1, $I_q = \frac{\rho^2}{1 - \rho}$ which is equivalent to $\frac{\lambda}{\mu} \times \frac{\lambda}{\mu - \lambda}$

EXAMPLE:
Customers arrive at rate 4/hour, and mean service time is 10 minutes (both interarrival time and service time are exponentially distributed)
What is the average size of the queue? What is the average time that a flow unit spends in the queue?

Answer: same process as before in G/G/1, just don't multiply by the second term (as again, Cs and Ca are 1 so (Cs+Ca)/2 = 1)

$$I_q = \frac{\frac{2^2}{3}}{1 - \left(\frac{2}{3}\right)} = 1.3333$$

T_q = 1.3333/4

What is the variance of the service time?

Well, C_s = 1
C_s = standard deviation of service times/mean service time

1 = (standard deviation)/10min

Standard deviation = 10min! Essentially exponential distribution assumes mean and standard deviation are equal! Variance = 10min^2 = 100min^2.

What is the likelihood that the inter-arrival time is at most 20 minutes?

Use the cumulative distribution function: Prob{a<=t} = 1 - $e^{-lambda*t}$ where t = time of 20min and a is something being under that time.

Note, we have t in minutes, lambda currently is in customers per hour (4/hour). Must convert either t to hours or lambda to customers/minute. We will do the former.

20min = 1/3 hour.
Prob{a<=t} = 1 - $e^{-4*(1/3)}$ = 0.7364 or 73.64% probability that any inter-arrival time is under 20min. (Probably don't worry about this example too much).

Another side note: if your times follow an exponential distribution, your rates (1/times) will follow what's called a Poisson distribution.

Last queue type: M/D/1 where D = deterministic (your service times do not change i.e. standard deviation of service times is 0). Say your server is a robot.

The P-K formula gives for an M/D/1 queue:

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

Can you derive this on your own, using what you know about Cs and Ca?)

Yes! We know the Ca = 1 as it is in M/M/1 (because stdev=mean), and Cs = Stdev service time /mean service time= 0 because stdev of service times = 0! So (1+0)/2 = ½

So really, all you need to know is the original P-K formula, everything other type of queue that follows (with 1 server) is an easier/simpler version of that.

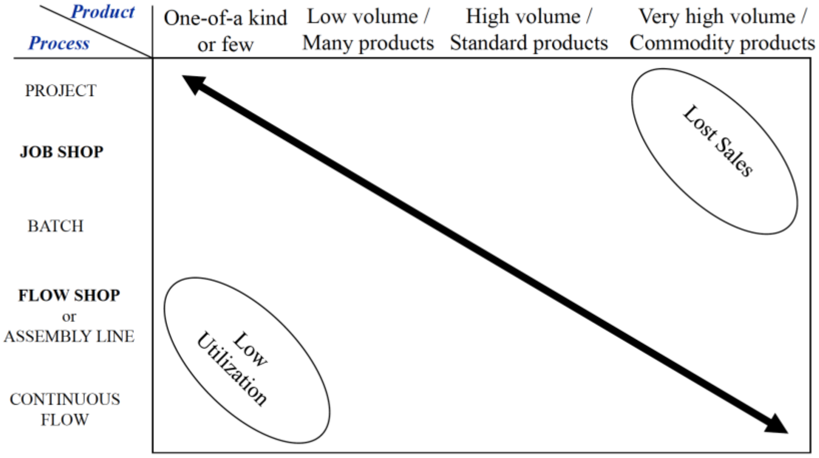
I don't believe M/M/C or project management are on midterm so I didn't cover them here.

Practice Questions

1. Fill in the blanks
The capacity of a process is always equal to the ____________ of the ____________.

____________________ occurs when an activity stops in a process because it has nothing to do. Conversely, an activity is _______________ when the next process is unready to take its output.

2. On the PP-matrix below, what kind of costs does low utilization represent? ___________ costs. What kinds of costs does lost sales represent? _______________ costs.


Product
Process
One-of-a kind or few
Low volume / Many products
High volume / Standard products
Very high volume / Commodity products
PROJECT
JOB SHOP
BATCH
FLOW SHOP or ASSEMBLY LINE
CONTINUOUS FLOW
Lost Sales
Low Utilization

3. Match the following:

A) tactical issues,	1. Being on the frontier
B) strategic issues,	2. Choosing a place on the frontier
C) operation innovation	3. Shifting the frontier

4. Suppose a store opens at 8AM. Customers show up at the rate of 30 per hour until 1PM, and then at the rate of 45 per hour until 3PM. The store closes at 3PM regardless of the number of customers waiting in line, and the unsatisfied customers are sent away.

Suppose that every customer who shows up at the store joins the line and waits until satisfied or sent away. The store can serve customers at the rate of 50 per hour between 8AM and 9AM, at the rate of 10 per hour between 9AM and 12 Noon, and then at the rate of 40 per hour between 12 Noon and 3PM. Use the continuous-time model in your calculation. (a) How many customers do you expect to see in the line at 11:30 AM? How many customers are sent away at the end of the day? (b) Calculate the average number of waiting customers between 8AM and 3PM. Then compute the average amount of time a customer spends on the line.

5. Suppose an airport runway can handle 5 flights/hour and the average flow time is 15 minutes. What is the average number of planes on the ground?

6. True/False: If a process has no buffer and variable input, compared to the no buffer case,

The process would have lower resource utilization (T/F)
The process would have longer average flow time (T/F)
The process would have a lower inventory/work-in-progress (T/F)
The process would have a higher throughput rate (T/F)

7. In a bank with 1 (super-efficient, robotic) teller, the average inter-arrival time of customers is 10 minutes, and this time is exponentially distributed. The service time is 5 minutes.

Compute:

Cs and Ca

Average queue length

Average queue time

8. WIP inventory level in job shop is (higher/lower) than in a flow shop. Equipment specialization is (higher/lower) in a job shop than a flow shop.

9. Suppose we have the following process for production of Hershey's cookies n crème.

Tasks	Resources	Unit Load
Add cookies	Worker A	5min/unit
Add crème	Workers B and C in parallel	12min/unit per worker
Package	Worker D	2min/unit
Inspect	Worker E	1min/unit

a) Draw the process map for making Hershey's cookies n crème (on separate paper please). Label the capacities of the individual resources and activities.

b) What is the theoretical flow time? What is the capacity of this entire process? What is the bottleneck?

c) Your SFU friend claims that if you add another worker to the cookies process, you can increase capacity by 50%. In very simple and basic English (because SFU), explain to him why he's wrong.

d) He claims to have misspoken, and meant to suggest adding another worker to the crème process, prove that he belongs in SFU (explain why he's still wrong).

e) Now determined to prove he's just as smart as you (lol), he claims you can reduce some costs as you have way too many workers right now. Assuming you don't do anything, and we still have the original process, determine which processes are blocked or starved, and determine if he's correct. If he is, who could you fire?

10. a) HSBC bank launders money for terrorists and drug cartels. They also seem to think they have a problem with long queues. Suppose on average, 10 customers arrive every hour, and average service rate is 12 customers per hour (that is, average service time is 5 minutes). Suppose also that there is no variability at all in either arrival times or service times. Based on this information, would there be a queue? Would it be long or would it be short?

b) Suppose there is variability now. Standard deviation of inter-arrival times is 2min and standard deviation of service time is 1min. Both times follow a general distribution. What kind of queue system is this? What is the average length of queue, what is the average waiting time in line?

c) Further, what is the average number of people in the bank (including waiting and in queue), and what is the average amount of time each customer spends in the bank (including waiting and in queue)

d) Let's say the bank customers interarrival times and the banks' service times both follow an exponential distribution. Without doing any calculations, would average queue length be longer or shorter than the case in part b)?

e) Let's say the bank customers interarrival times are as the same as in part b) but the service times are now deterministic, again, with no calculations, would queues be longer or shorter than in part b)?

f) Assume both inter-arrival times and service times (means and stdevs) are as in b). Give 3 suggestions on how could the bank could reduce queue lengths and queue time, relating each back to the operations management triangle.

Solutions

1. The capacity of a process is always equal to the capacity of the bottleneck.

Starving occurs when an activity stops in a process because it has nothing to do. Conversely, an activity is blocked when the next process is unready to take its output.

2. (High) capital costs, opportunity costs

3. Tactical = being on frontier, strategic = choosing a spot on frontier, innovation = moving the frontier

4.

8AM - 1PM (5Hrs): 30/Hr
1PM - 3PM (2Hrs): 45/Hr

Store can serve 50/Hr 8AM-9AM
10/Hr 9AM-12PM
40/Hr 12PM-3PM

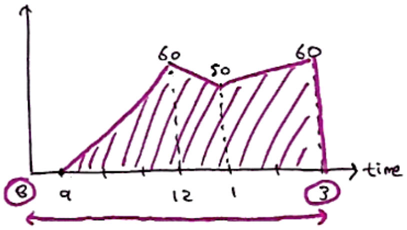
a) I = RT Inventory Build-Up diagram

8-9 AM: No inventory (input < capacity)
9-12PM: (30 * 3) - (10 * 3) = 60
12-1PM: 60 + 30 - 40 = 50
1-3PM: 50 + (45) 2 - 40(2) = 60

There would be 60 * 5/6 = 50 ppl waiting at 11:30 AM. 60 customers are sent away at 3 PM.

Since every customer showed up & left the system,

b) Compute the area under the curve



(60 * 3 * 1/2) + (60+50)/2 + (60+50)*2/2
= 90 + 55 + 110 = 255
255 / 7 = 36.43 Avg # of waiting customers

Total # of customers showed up 30 * 5 = 150
45 * 2 = 90) 240

Throughput rate = 240/7 = 34.29 customers/Hr

Avg Flow Time (t)
= I / R (Inv / throughput)
= 36.43/34.29
= 1.06 Hr ≈ 64 mins

5. Little's law → I = R x T = 5 x 15 = 75

6. *The process would have lower resource utilization (T/F)*

Effectively, without a buffer, some input is “turned away” (people look at the queue and leave). This reduced input rate. Utilization = input/capacity. These “inputs” don’t even enter the process!

*The process would have longer average flow time (T/**F**)*

If we include the buffer as part of the process, then having a buffer obviously means people who spend time there (instead of leaving) would have a longer average flow time. Even if you exclude it (which is wrong, we should include it), then the average flow time is still the same.

*The process would have a lower inventory/work-in-progress (**T**/F)*

Inputs are being “turned away”, so at any time in the process, the process is less full (average utilization is lower). Further, if you include the buffer as part of the process, people in the buffer are considered inventory. Relate longer average flow time to inventory using Little’s law, considering flow time stays the same.

*The process would have a higher throughput rate (T/**F**)*

Input rate is lower, as explained above. Throughput = min(input rate, capacity rate).

7. In a bank with 1 (super-efficient, robotic) teller, the average inter-arrival time of customers is 10 minutes, and this time is exponentially distributed. The service time is 5 minutes.

Compute:

Cs = 0 (standard deviation is 0, deterministic)

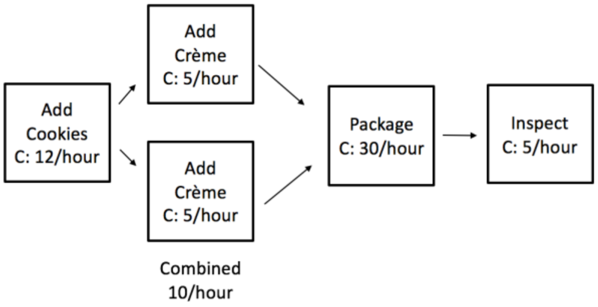
Ca = 1 (standard deviation = mean in exponential distribution)

Average queue length

Input rate (lambda) = 6/hour (from 10min). Service rate = 12/hour. Utilization (rho) = 50%
M/D/1 queue: rho^2/(1-rho) * (1/2) = 0.25
Average queue time
Iq = Tq*Lamda → Iq/Lambda = Tq → 0.25/(6)=0.0417 hours

8. WIP inventory level in job shop is (higher/**lower**) than in a flow shop. Equipment specialization is (higher/**lower**) in a job shop than a flow shop.

9. a)



b) Theoretical flow time = 5+12+2+1 = 20min. Capacity = 10/hour. Bottlenecked by crème process.

c) You must "target" the bottleneck to make any difference in process capacity.

d) If you add another worker, the capacity of the crème activity becomes 15/hour. This represents a 50% increase in the activity capacity. However, the bottleneck of the entire process shifts and by now is the cookies activity (capacity of 12/hour), so the whole process now has a capacity of 12/hour (limited by bottleneck of cookies).

e) He is actually correct. Due to bottleneck of the crème process, the cookies process is blocked, the packaging process is starved. Worker D has nothing to do most of the time. If you plot this out on a gantt chart, you will find that he actually has time to do Inspection as well. You could fire the inspector (or you could give the inspector the packaging job, doesn't really matter who you fire.)

10. a) There is no queue if there is no variability. Why? Because we always assume average input rate is below average capacity rate (or else it doesn't really work, you'll have an infinitely growing lineup). And if people all arrive consistently at a rate that is below capacity, then the process would be able to service everyone. Think of a machine that services 1 customer per minute, and 1 customer arrives every minute, it will service the first customer, be done, and right when it's finished, the second customer would have just arrived… and repeat. Nobody would ever need to wait.

The problem of queue only arises because sometimes maybe 3 people arrive at once (where the queue begins) and sometimes nobody comes.

b) G/G/1 queue. Use P-K Formula

We know
Average service time 5min
Mu = 1/5 per minute or 12/hour
lambda = 10/hour
Average inter-arrival time = 1/lambda = (1/10 hour) = 6min
Stdev of inter-arrival time = 2min so Ca = 2/6
Stdev of service time = 1min so Cs = 1/5

So *p* (rho) = lambda/mu = 10/12 = 5/6

$$I_q \cong \frac{\left(\frac{5}{6}\right)^2}{\left(1-\frac{5}{6}\right)} * \frac{\left(\frac{2}{6}\right)^2 + \left(\frac{1}{5}\right)^2}{2} \cong 0.31482$$

$$T_q = \frac{I_q}{\lambda} = \frac{0.31482}{10} = 0.031482\ hours\ or\ 1.89\ minutes$$

c) T = T_q + T_s = 1.89min + 5min = 6.89min (average total time spent by each customer)
I = I_q + I_s where I_s = T_s x lambda = (5min/(60min/hour))*10/hour = 0.8333 → 0.31482+0.8333 = 1.14815 (average number of people in bank at any time).

(You may arrive at these answers through other calculations, but this is probably the fastest.)

Sanity checks
Our variability is relatively low (Cs and Ca are small), so it makes sense that the average queue length is short and average wait time is less than 2 minutes. I_s should be < 1 because our service only serves 1 customer at once, but there is some time where it is idle (due to variability, there are times nobody comes), so average should be less than 1. None of the numbers are outlandishly large or small.

d) Longer. Exponential distribution assumes mean and standard deviation are equal and so CV of both arrival and service = 1. The standard deviation of the process in part b) is lower than the mean (I wrote this example with low variability). Which means both Cs and Ca (which are stdev/mean) < 1, so it follows that (Cs^2+Ca^2)/2 for our G/G/1 queue is smaller than that in the M/M/1. There is less variability here than if we assume M/M/1. Note, this does not apply to EVERY G/G/1 queue. If we had said that our standard deviations were for example 15min each, then the M/M/1 queue would have a shorter average queue length and time. Essentially, the point of this question is to show that more variability increases queue length (inventory), as M/M/1 would have more variability than our G/G/1 example.

e) Shorter. Deterministic service times = no more service time variability. Less variability = lower inventory (shorter queues). Or, in the P-K formula, Cs = 0 instead of some positive number.

f) Get a faster worker or add another worker (increase capacity)
Get more information on demand or train your workers to be more consistent (reduce variability)

Note, simply getting more information about demand isn't going to reduce your queue lengths. You have to actually use the information to adjust demand (raise prices at certain times etc.) essentially, you are reducing input variability by manipulating demand to match your capacity, or adjust capacity in response to demand.