Service System Management

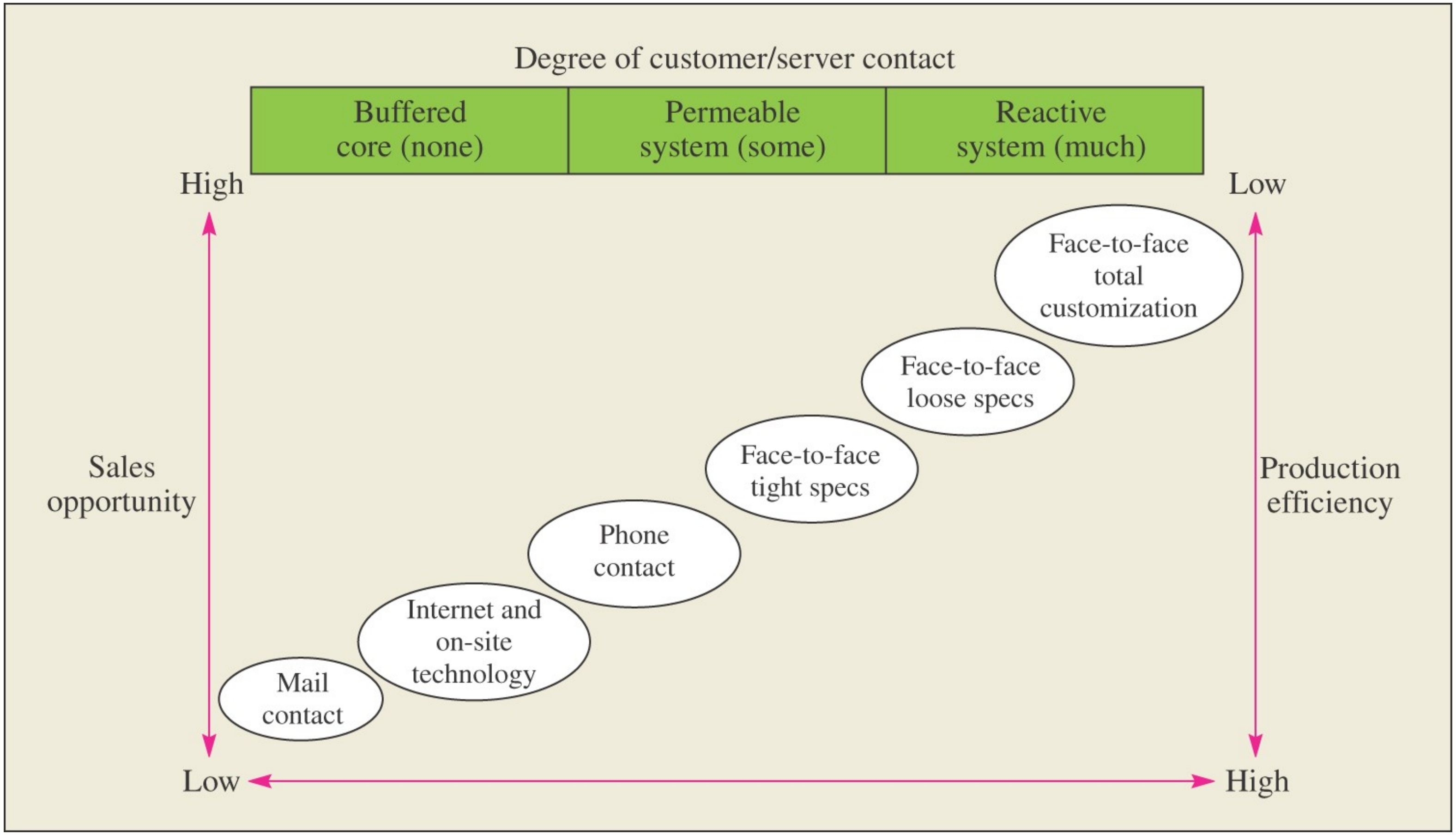
KELLEY SCHOOL OF BUSINESS

INDIANA UNIVERSITY

Nature of Services

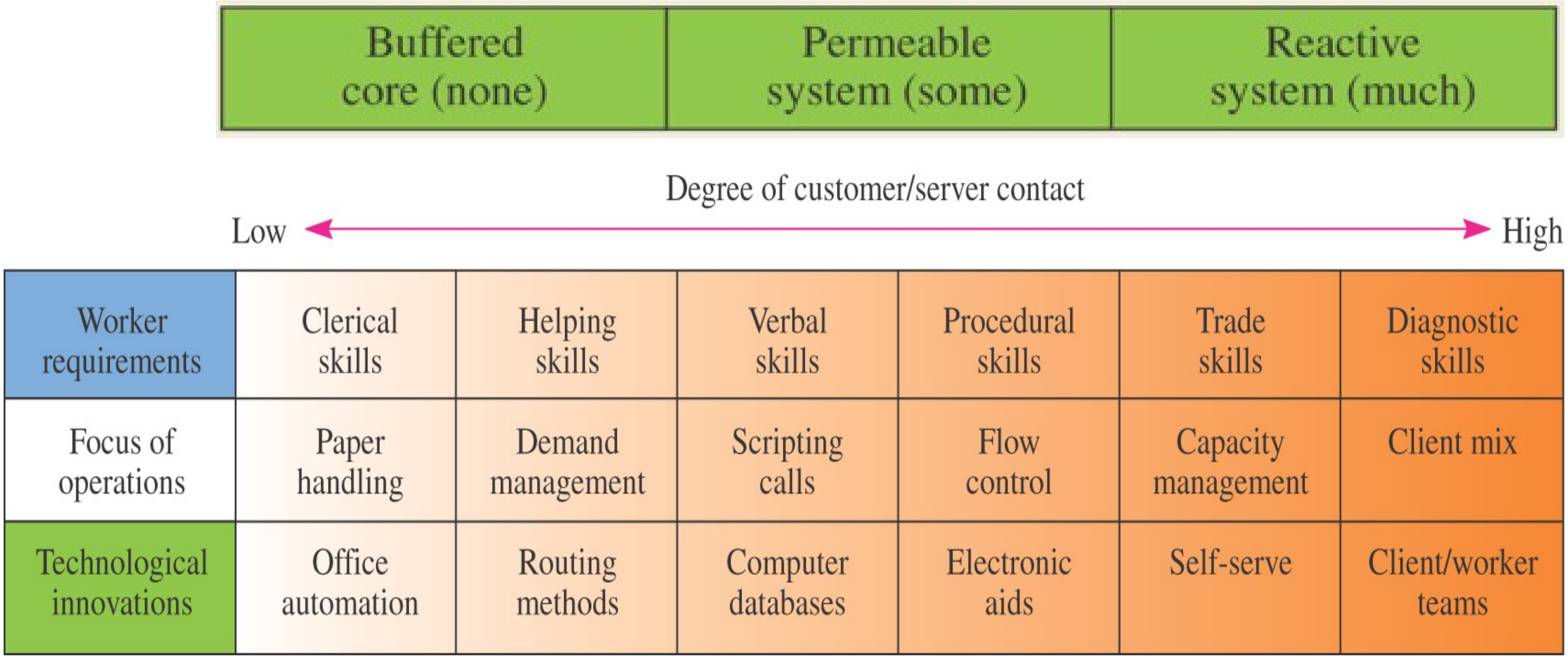
- Produced and consumed (almost) simultaneously; no physical inventory
- Everyone is an expert (quality is “experienced” by people ℗ subjective; no single best approach)
- Idiosyncratic-what works for one may not work for others
- Quality of work is not quality of service (much harder to measure the quality of a service)
- Mix of tangible and intangible attributes
- Strong need to understand marketing and personnel
- Cycles of encounters ℗ “phases” in a “system”

Service System Design Matrix



Not impossible to operate off the diagonal, just not often seen and frequently not sustainable

Worker, Process, and Technological Attributes



Buffered core (none)	Permeable system (some)	Reactive system (much)

Degree of customer/server contact

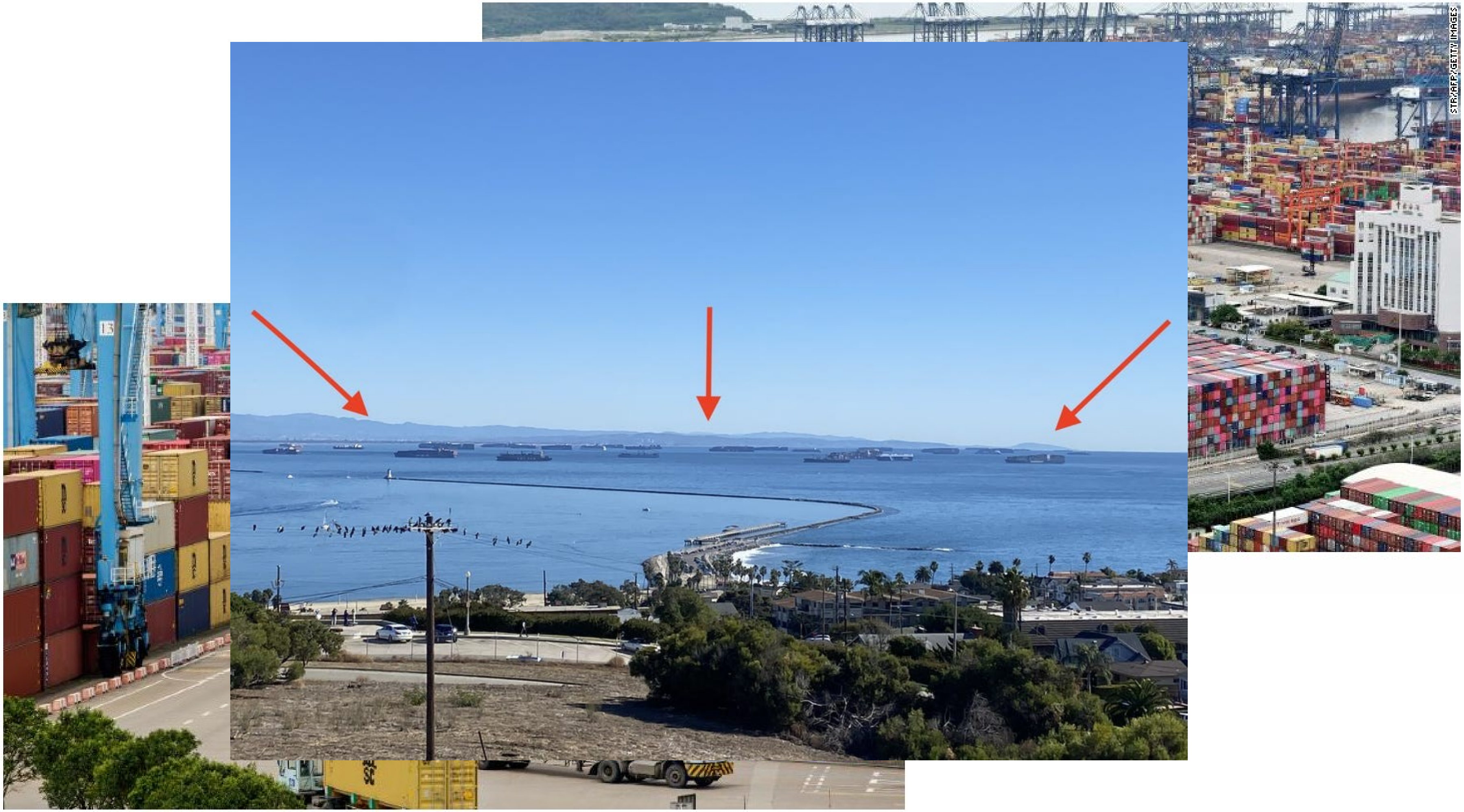
Low ← → High

Worker requirements	Clerical skills	Helping skills	Verbal skills	Procedural skills	Trade skills	Diagnostic skills
Focus of operations	Paper handling	Demand management	Scripting calls	Flow control	Capacity management	Client mix
Technological innovations	Office automation	Routing methods	Computer databases	Electronic aids	Self-serve	Client/worker teams

Characteristics of Workers, Operations, and Innovations Relative to the Degree of Customer/Service Contact

Ever had to wait in line? Any idea why?

What happened during Covid with global supply chains?



Containers are shown at Ningbo-Zhoushan port on August 15, 2021.

Queuing Systems

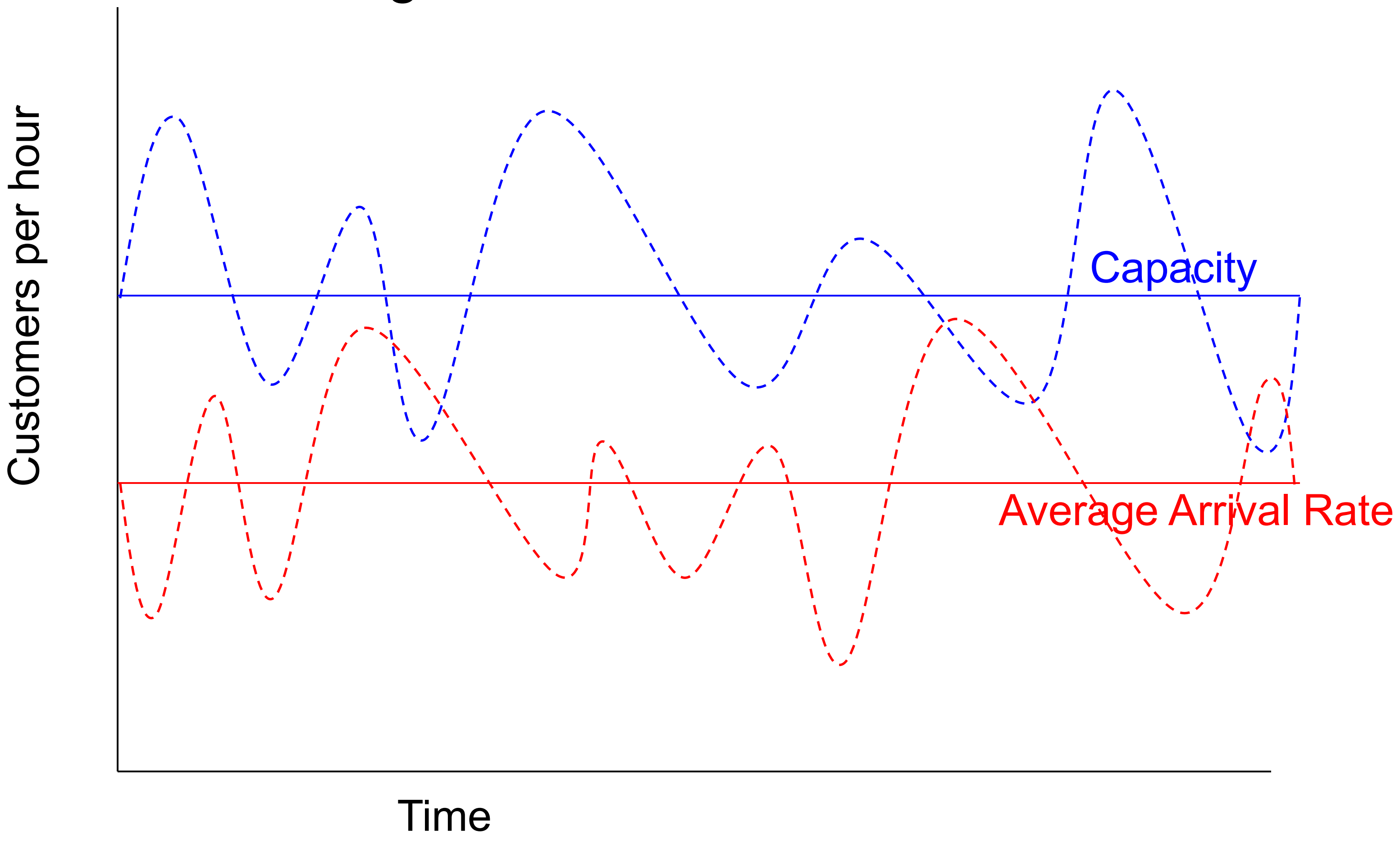
The familiar "waiting in line" situation

Frustrating, annoying

Managing well is key

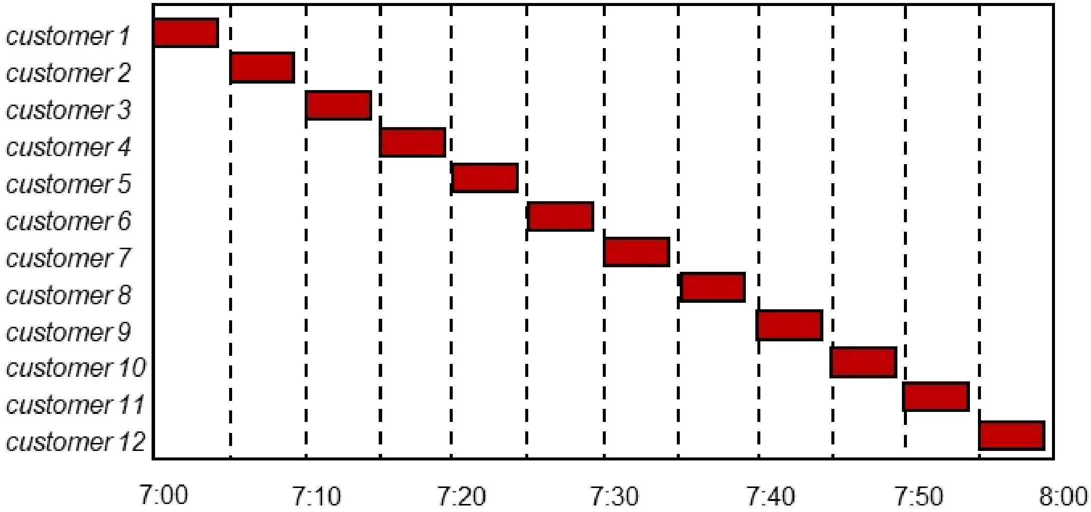
- Objectives - depend on situation
- Balance service with productivity

Question: Would you ever have to wait in a line where the average capacity to serve customers was greater than the average arrival rate of customers?



Why do Queues Form?

Consider a perfect kiosk:



A customer arrives every 5 minutes

- Demand rate = 12 customers per hour

It takes 4 minutes for a customer to get their tickets

- Capacity rate = 15 customers per hour

Utilization of the "service node" and customer flow time?

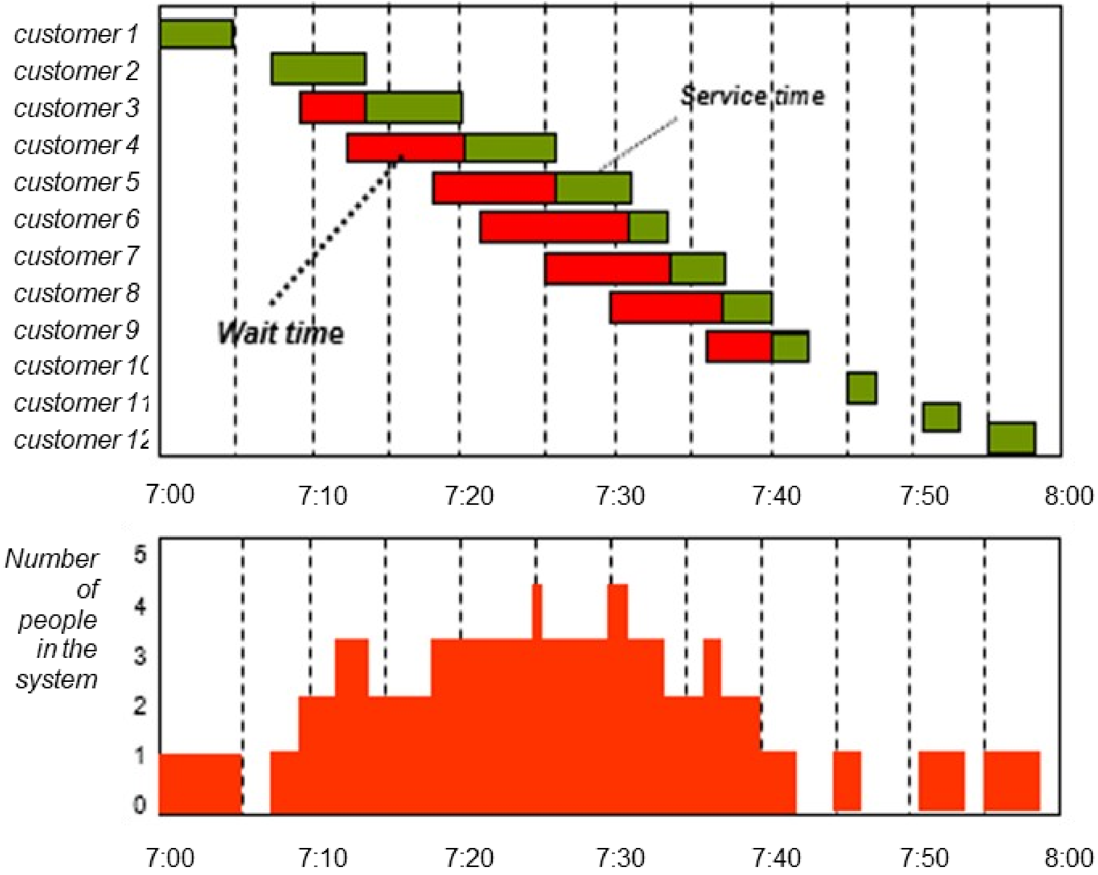
- Utilization = 12/15 = 80%; flow time = **4 min** for each customer

No backlog, no buildup, no waiting… what's wrong with this picture?

Why do Queues Form?

Why do Queues Form?

- **A *real* ticket booth:**
- **Same twelve customers, same hour, same averages, same service node utilization, but…**
- **Average actual customer flow time (est) 8.3 minutes?!**
- **Flow time is more than double due only to variability of demand rates and service times**



Server utilization, interarrival variability, and service time variability drive time in queue

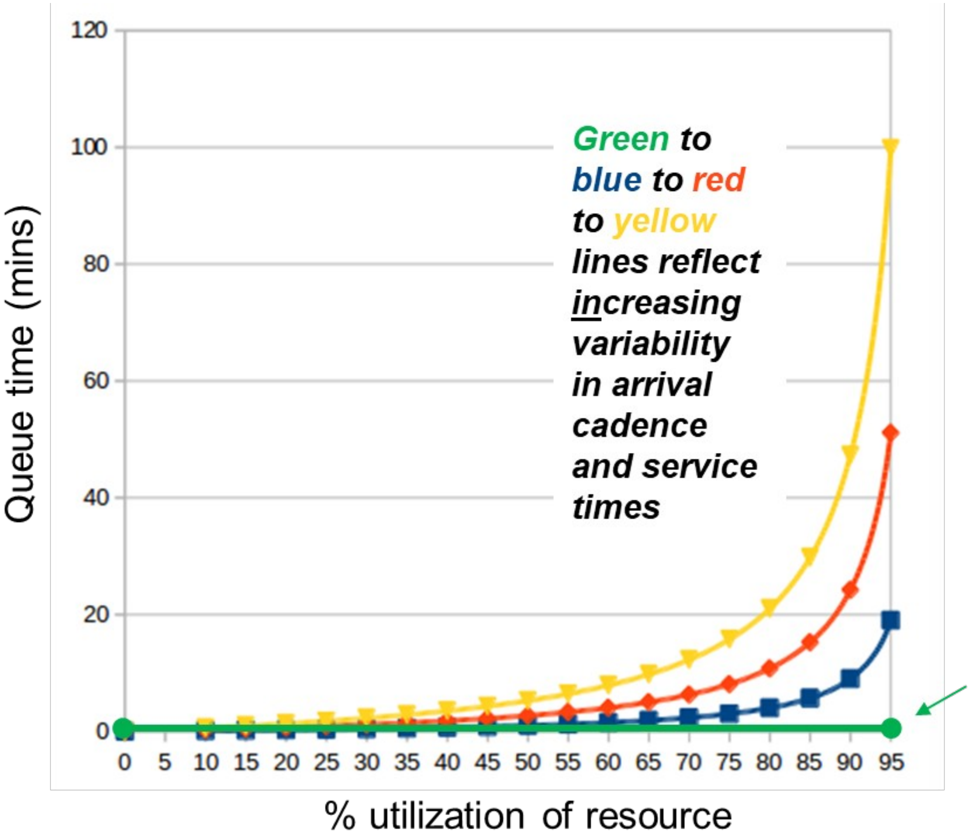
For example, higher touch (more reactive) service system designs would have what impact on service time variability?

- ➢ *Increase*

So at the same level of server utilization, how would you expect queue times for a higher touch, more reactive system to compare with times in a lower touch less reactive system?

- ➢ *Higher*

What might you do to reduce these queue times…?



Green to blue to red to yellow lines reflect increasing variability in arrival cadence and service times

Zero variability would be a flat line

Designs for On-Site Service

Production Line Approach

- Claims processing, class registration, inventory management

Self-Service Approach

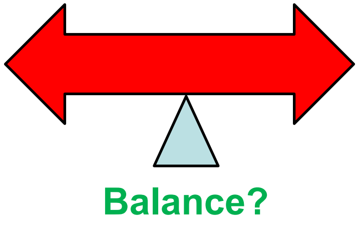
- Salad bar, ATMs, gas stations, any kiosk

Personal Attention Approach

- Ruth's Chris Steakhouse, real estate, auto sales, consulting

Service provider wants…

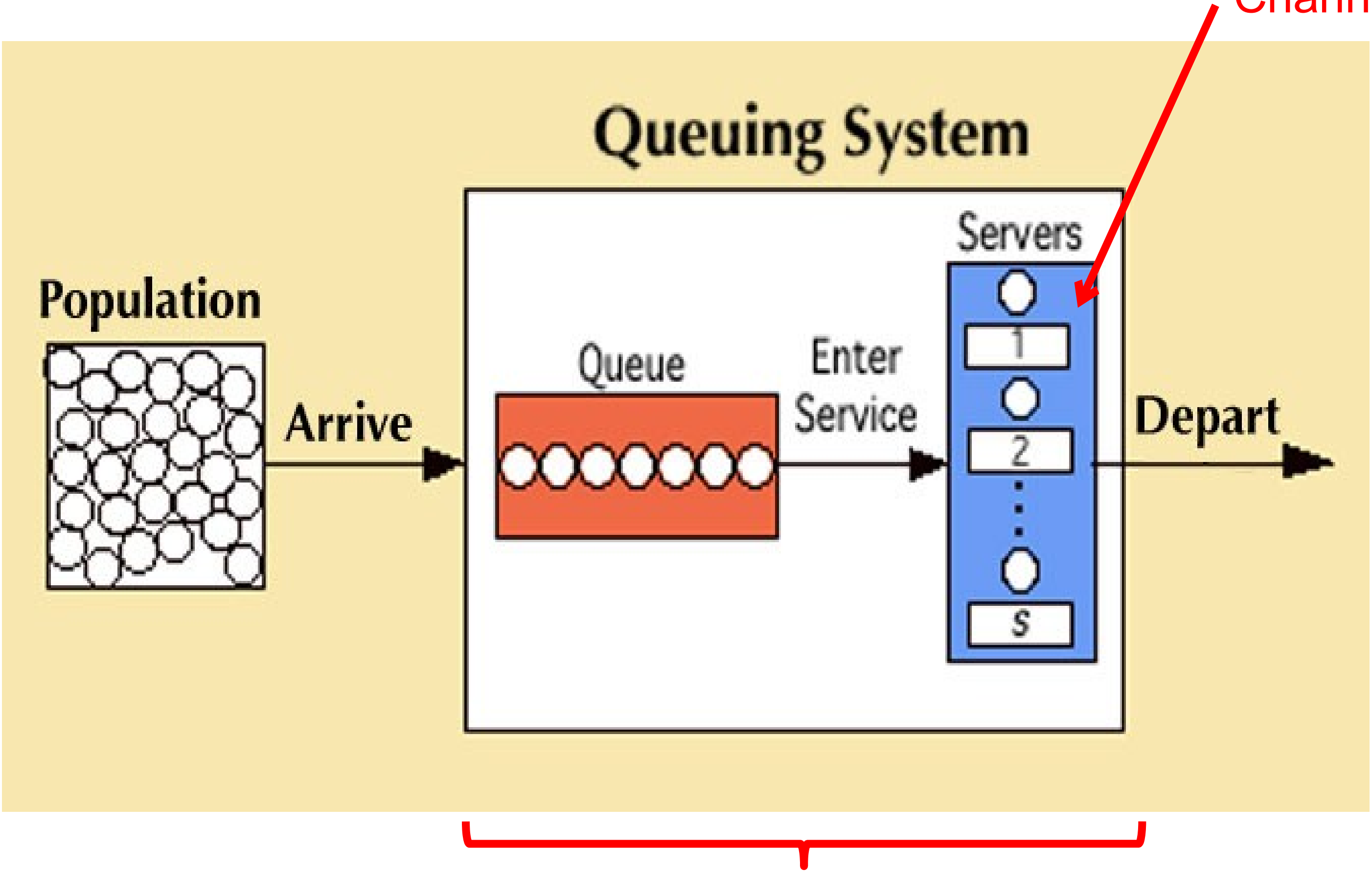
Reduced cost – minimum staffing levels to meet customer expectations



Customer wants…

Zero waiting – adequate staffing levels to deliver the needed service quickly

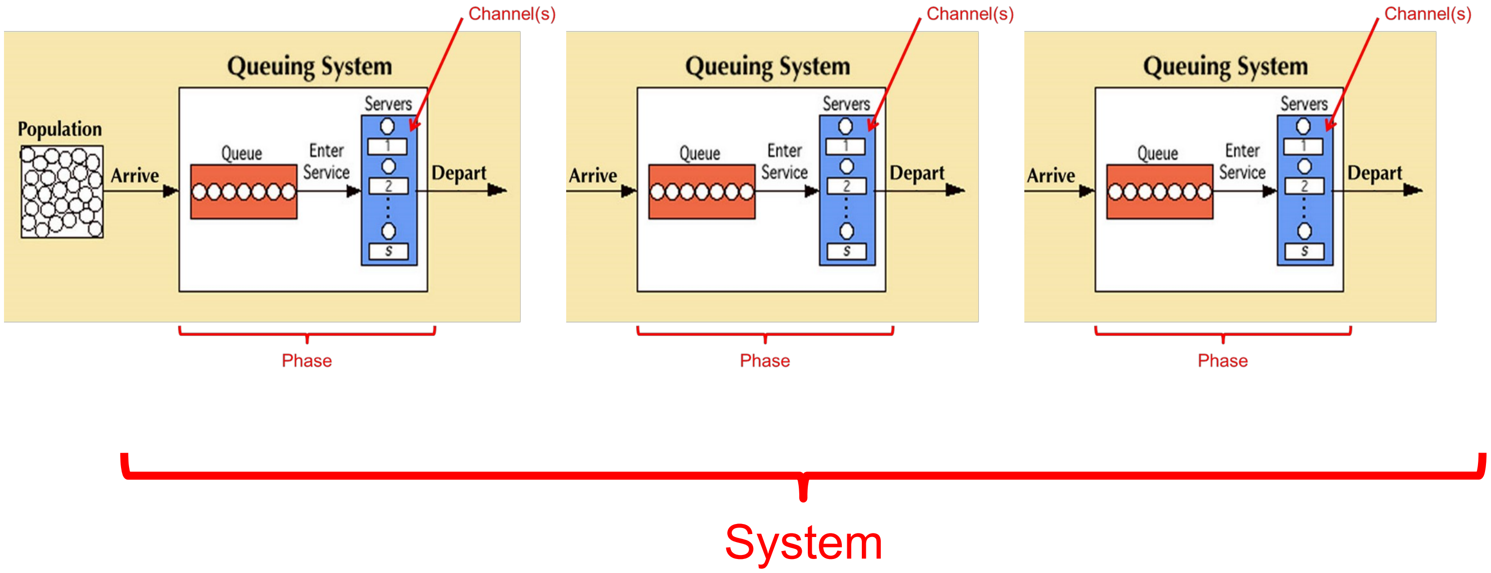
Components of a Queuing System



Channel(s)

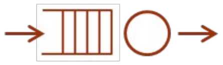
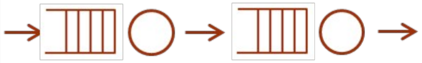
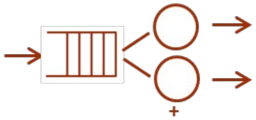
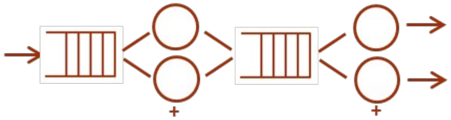
Phase

Components of a Queuing System



In P370, we only worry about single phase, single channel systems (so one phase = the whole system)

Queue Structure and Discipline

Diagram	Description and Examples
	**Single channel, single phase** *(A dentist's office)*
	**Single channel, multi phase** *(A fast food two-window drive through)*
	**Multi channel, single phase** *(Airport ticket counters, grocery self-serve)*
	**Multi channel, multi phase** *(Airport, subway, amusement park)*

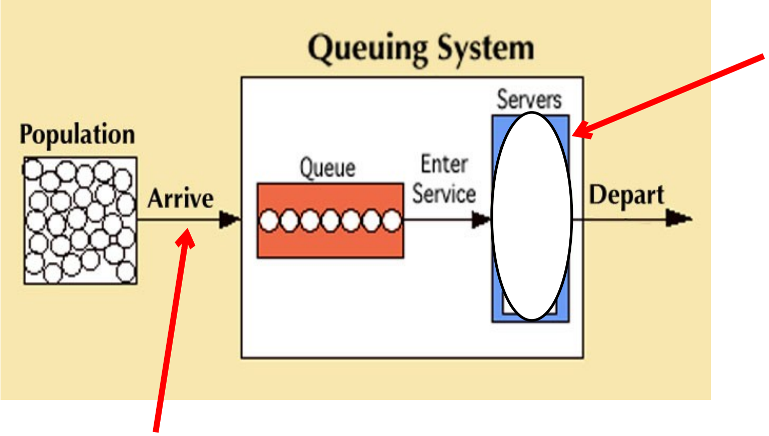
Behaviors

- Balking: walk away, don't join the queue Ⓟ lost revenue
- Reneging: join the queue then walk away Ⓟ bad signal, then lost revenue
- Priorities: FIFO (or FCFS), LIFO, triage, etc. Ⓟ expectation setting is important

Waiting lines impact revenue, reputation, and profitability

Example 1

Suppose we have an airline counter which is single channel, single phase. People arrive exponentially at the rate of 25 per hour and are served exponentially at the rate of 30 per hour.



Average service rate of single server (#/unit time) = **30 ppl/hr**

Average arrival rate of jobs (#/unit time) = **25 ppl/hr**

Would a line ever form?

Components of System

- **Customers**
  - Arrival rate designated by **λ** - e.g. 12 customers per hour - **exponential** distribution
  - Inter-arrival time designated by **1/λ** - e.g. 5 minutes (or of an hour) between each customer arrival
  - Population - finite (small) or infinite (large) ℗ large in P370
- **Servers**
  - Service rate per server designated by **μ** - e.g. serve 15 customers per hour - **exponential** distribution
  - Average service time designated by **1/μ** - e.g. 4 minutes (or of an hour) to serve each customer
  - # of servers is an issue ℗ single server in P370

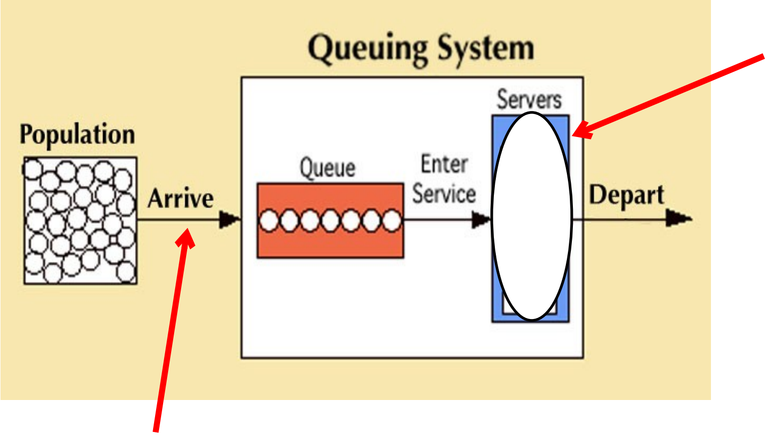
Calculation Trickiness

- “A customer arrives every 30 minutes”: this gives you the inter-arrival time, so if you want to calculate arrival rate **λ**…
- Remember that inter-arrival time (**1/λ**) and arrival rate (**λ**) are *reciprocals*.
- So a customer every 30 minutes = of a customer per minute.
- And of a customer per minute is how many customers per hour?

  * = * 60 cust/hr = = **2 cust/hr**
- So the **λ** for “a customer arrives every 30 minutes” is **customer per minute** or **2 customers per hour** (same thing, different units)
- Same logic applies when the problem says “you can serve a customer every X minutes” and you need to find the service rate **μ**

Example 1

Suppose we have an airline counter which is single channel, single phase. People arrive exponentially at the rate of 25 per hour and are served exponentially at the rate of 30 per hour.



μ = average service rate of single server (#/unit time) = 30 ppl/hr

λ = average arrival rate of jobs (#/unit time) = 25 ppl/hr

Formulas

λ = average arrival rate of jobs (#/unit time)

μ = average service rate of single server (#/unit time)

$\rho = \lambda/\mu$ = utilization rate = P_w (probability of waiting)

$P_0 = 1 - \rho$ = probability no-one is in line

$\overline{n}_s = \lambda/(\mu - \lambda)$ = average number in system

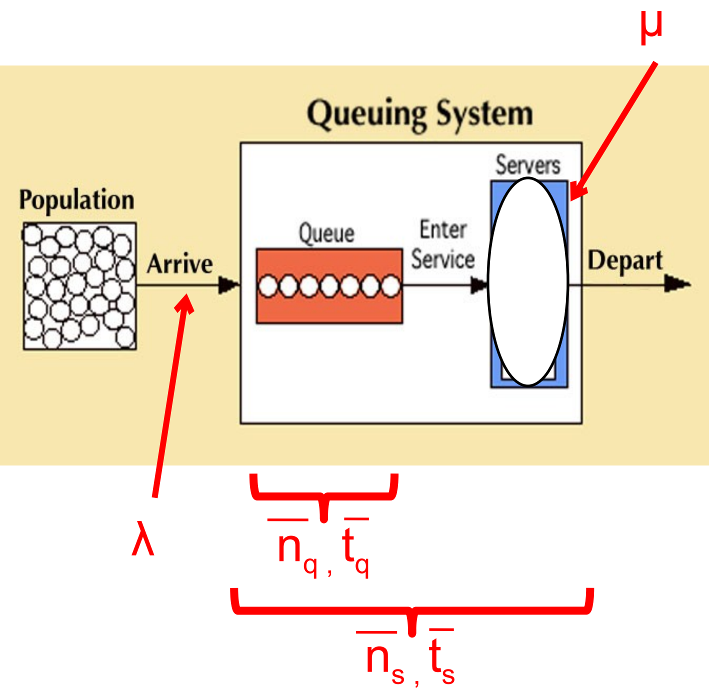
$\overline{n}_q = \overline{n}_s - \rho$ = average number in queue

$\overline{t}_s = \overline{n}_s/\lambda = 1/(\mu - \lambda)$ = average time in system

$\overline{t}_q = \overline{t}_s - 1/\mu$ = average time in queue

Example

Suppose we have an airline counter which is single channel, single phase. People arrive exponentially at the rate of 25 per hour and are served exponentially at the rate of 30 per hour.



$\lambda =$ 25 people/hour

$\mu =$ 30 people/hour

$\rho =$ 25/30 = 0.8333 (83.33% utilized or P_w)

$P_0 = 1 - \rho = 0.1667$ (16.67% chance of no wait)

$\bar{n}_s = \lambda/(\mu - \lambda) = 25/(30-25) = 5$ people in system

$\bar{n}_q = \bar{n}_s - \rho = 5 - 0.8333 = 4.1667$ people in line

$\bar{t}_s = \bar{n}_s/\lambda = 5/25 = 0.20$ hrs = 12 min in system

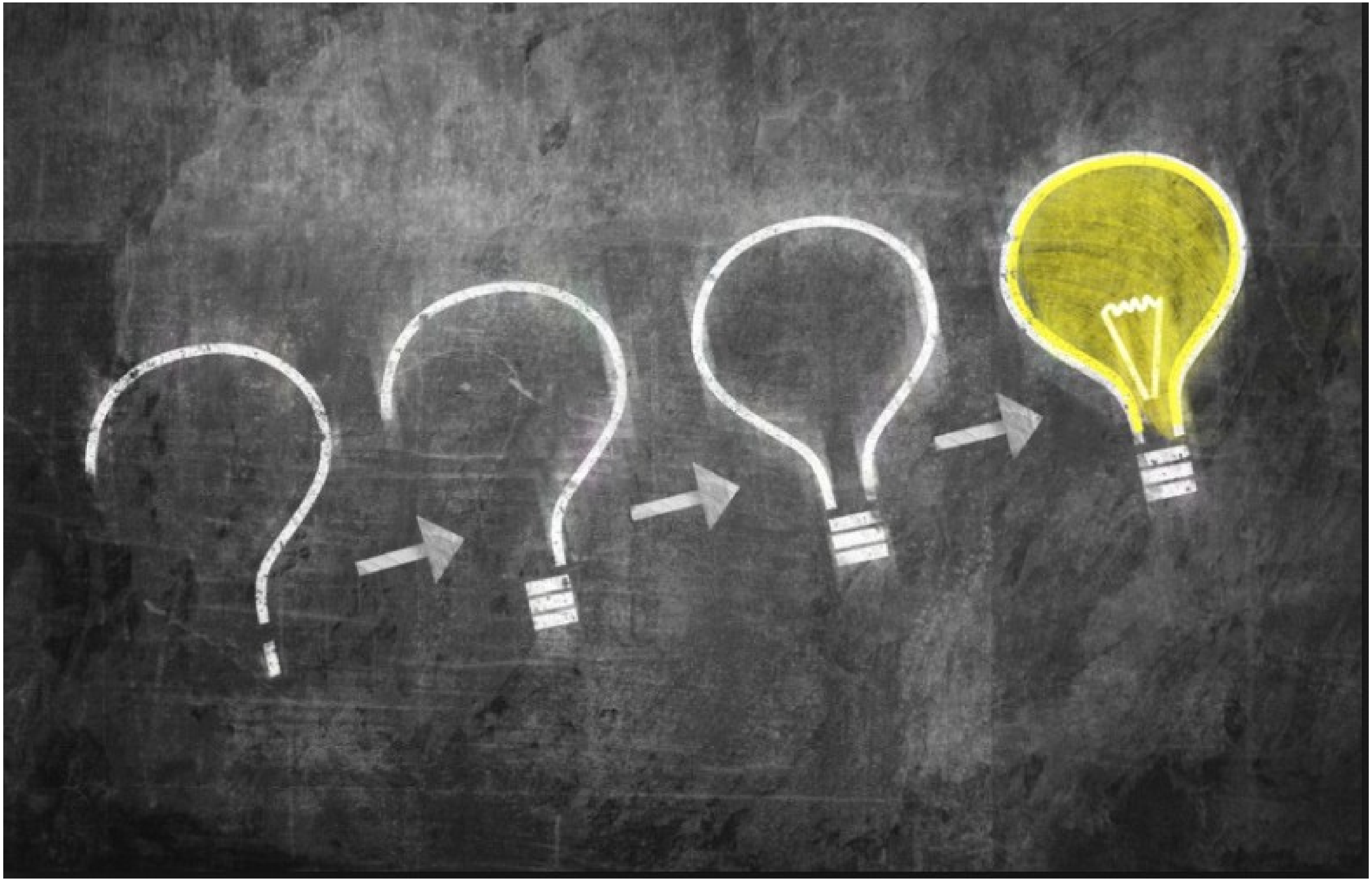
$\bar{t}_q = \bar{t}_s - 1/\mu = 0.20 - 1/30 = 0.1667$ hrs = 10 min in line

Queue Psychology

- Unoccupied time vs. occupied time
- Pre-process wait vs. in-process wait
- Uncertain waits vs. certain waits
- Unexplained waits vs. explained waits
- Unfair waits vs. equitable waits
- Willingness to wait related to value
- Solo waits vs. group waits
- The front-end and the back-end of the encounter are not created equal
- Segment the pleasure; combine the pain
- Let the customer control the process

Today's play list…

- I Walk the Line (Johnny Cash)
- Tired of Waiting (The Kinks)
- The Waiting (Tom Petty)
- Sitting, Waiting, Wishing (Jack Johnson)
- Jump in the Line (Harry Belafonte)

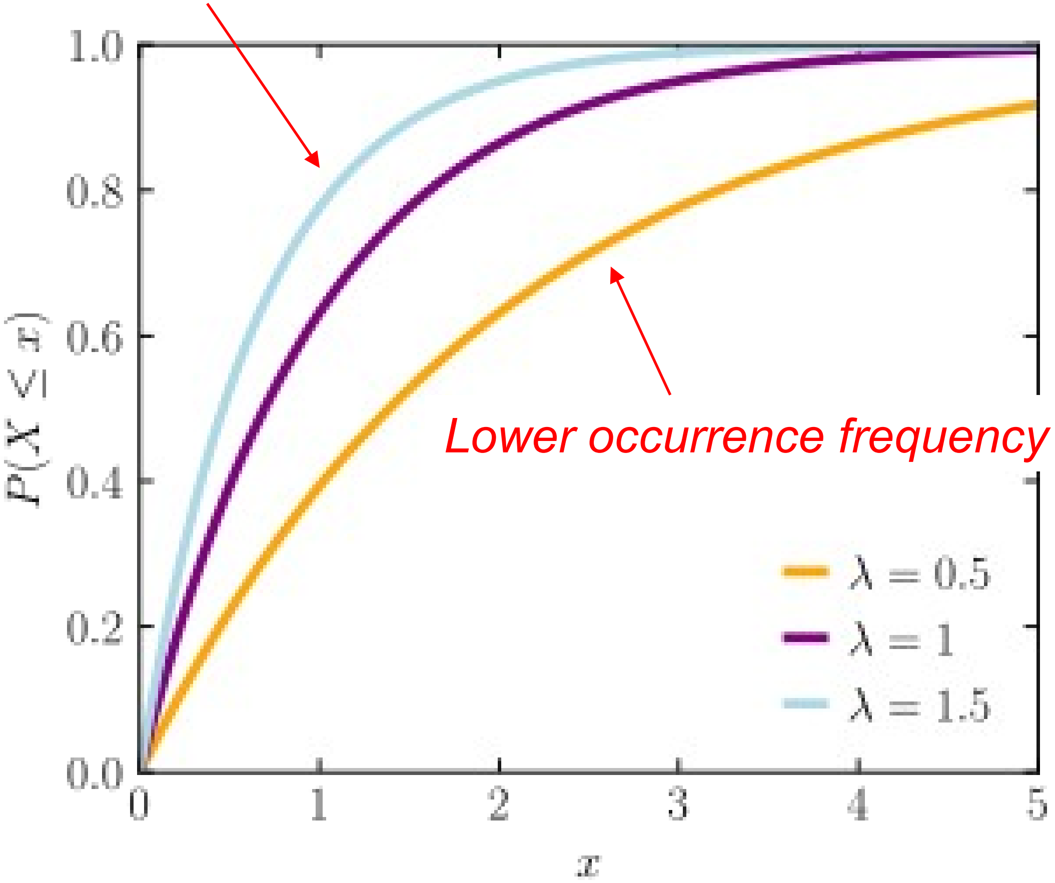


Reminders

- Week 4 quiz and industry article **due midnight Sunday**
- Prep for next lecture: Capacity Management

Terms

- **“Line” = “Queue”**
- **Exponential – e.g., on average something happens λ times per minute, so the probability of the thing occurring approaches 100% as time passes (more quickly with a larger λ)**



Higher occurrence frequency

Lower occurrence frequency

Probability of occurrence

Time elapsed