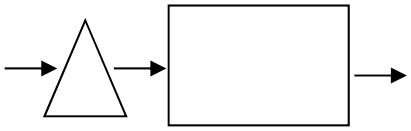
Response Time

Introduction

Example

Physician office



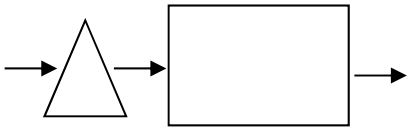
- Patients arrive, on average, every five minutes
- It takes ten minutes to serve a patient
- Patients are willing to wait

$\Rightarrow$ What is the implied utilization of the barber shop?

$\Rightarrow$ How long will patients have to wait?

Example

Physician office



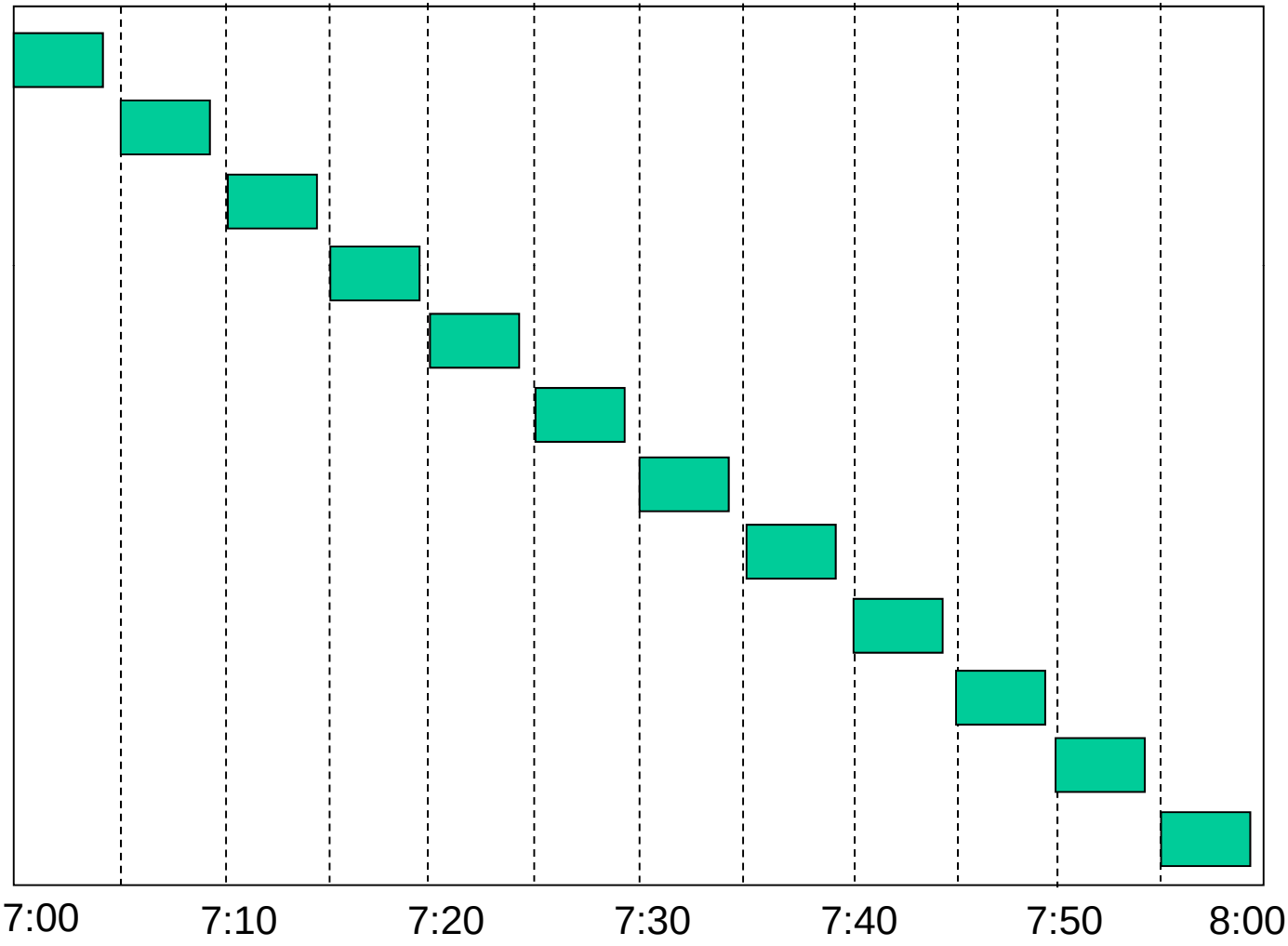
- Patients arrive, on average, every five minutes
- It takes four minutes to serve a patient
- Patients are willing to wait

$\Rightarrow$ What is the utilization of the barber shop?

$\Rightarrow$ How long will patients have to wait?

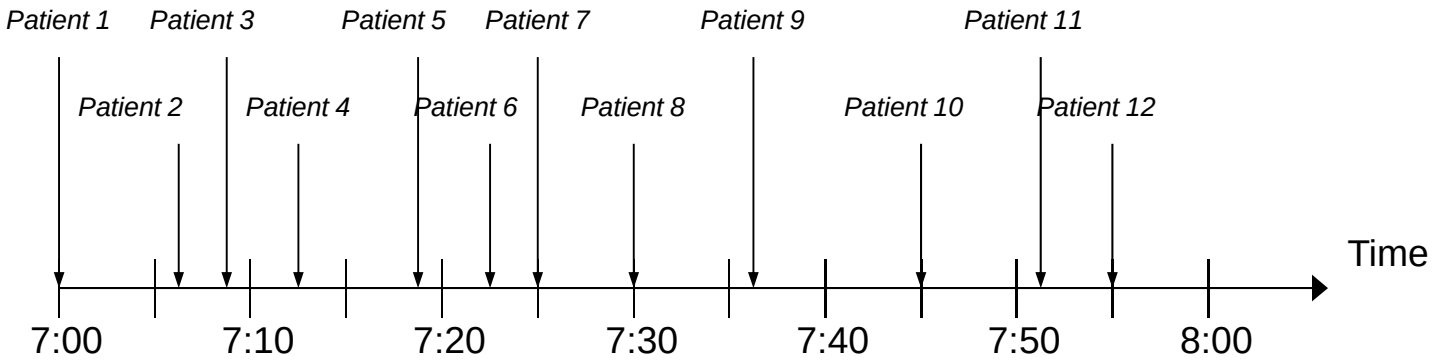
A Somewhat Odd Service Process

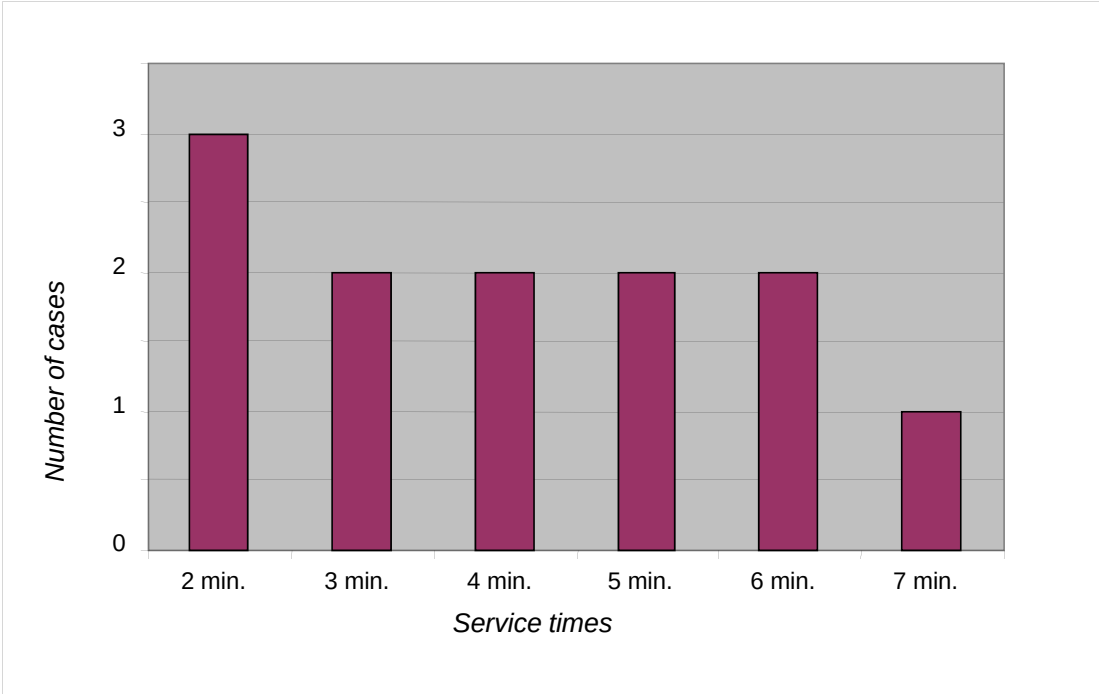
Patient	Arrival Time	Service Time
1	0	4
2	5	4
3	10	4
4	15	4
5	20	4
6	25	4
7	30	4
8	35	4
9	40	4
10	45	4
11	50	4
12	55	4



A More Realistic Service Process

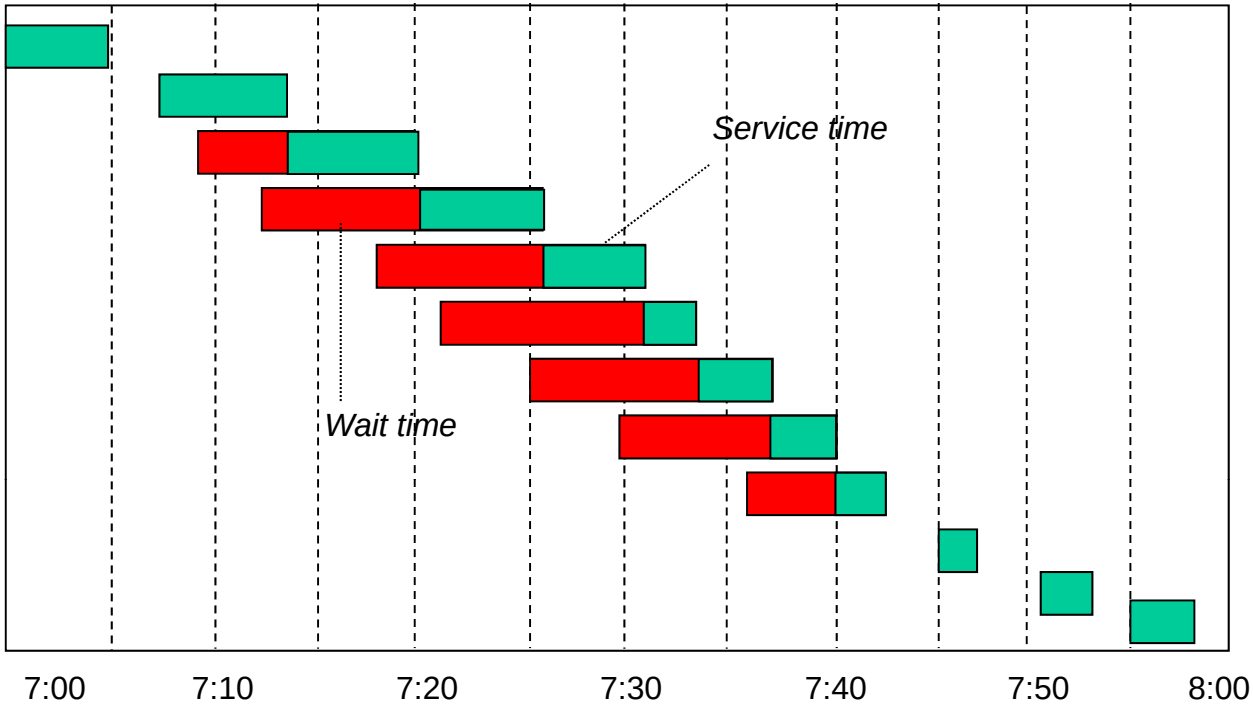
Patient	Arrival Time	Service Time
1	0	5
2	7	6
3	9	7
4	12	6
5	18	5
6	22	2
7	25	4
8	30	3
9	36	4
10	45	2
11	51	2
12	55	3



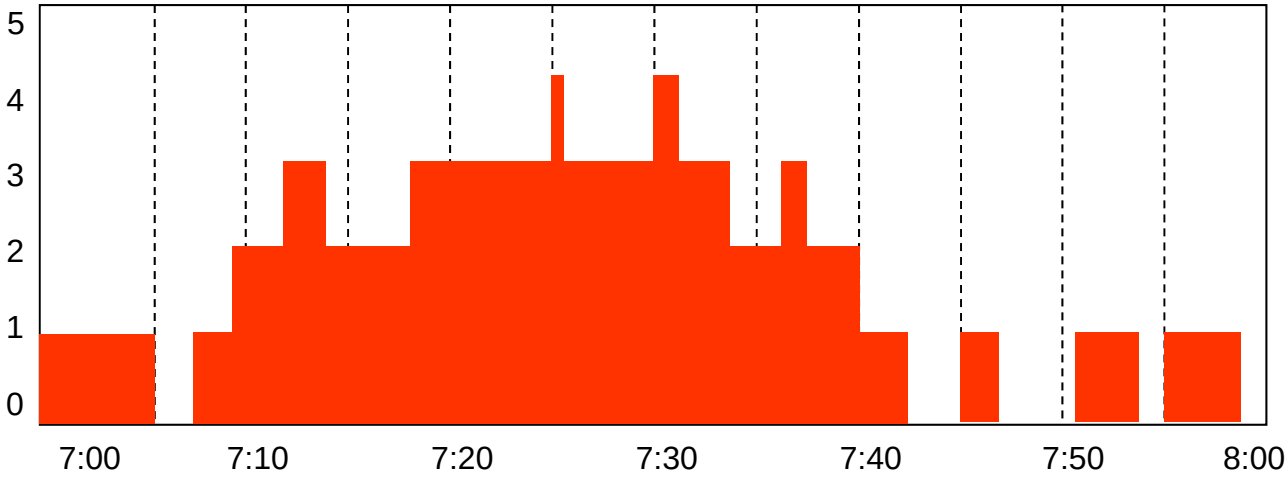


Variability Leads to Waiting Time

Patient	Arrival Time	Service Time
1	0	5
2	7	6
3	9	7
4	12	6
5	18	5
6	22	2
7	25	4
8	30	3
9	36	4
10	45	2
11	51	2
12	55	3



Inventory (Patients at lab)



The Curse of Variability - Summary

Variability hurts flow

With buffers: we see waiting times even though there exists excess capacity

Variability is BAD and it does not average itself out

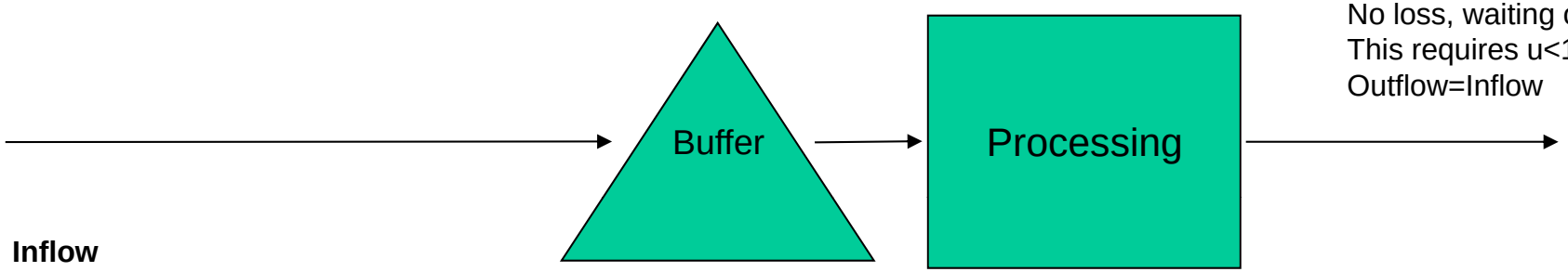
New models are needed to understand these effects

Response Time

Waiting time models: The need for excess capacity

Modeling Variability in Flow

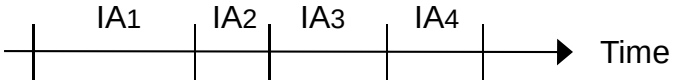
Flow Rate
Minimum{Demand, Capacity} = Demand = 1/a



Outflow
No loss, waiting only
This requires u<100%
Outflow=Inflow

Inflow
Demand process is "random"

Look at the inter-arrival times



a: average inter-arrival time

$$CV_a = \frac{\text{St-Dev(inter-arrival times)}}{\text{Average(inter-arrival times)}}$$

Often Poisson distributed:
- $CV_a = 1$
- Constant hazard rate (no memory)
- Exponential inter-arrivals

Difference between seasonality and variability

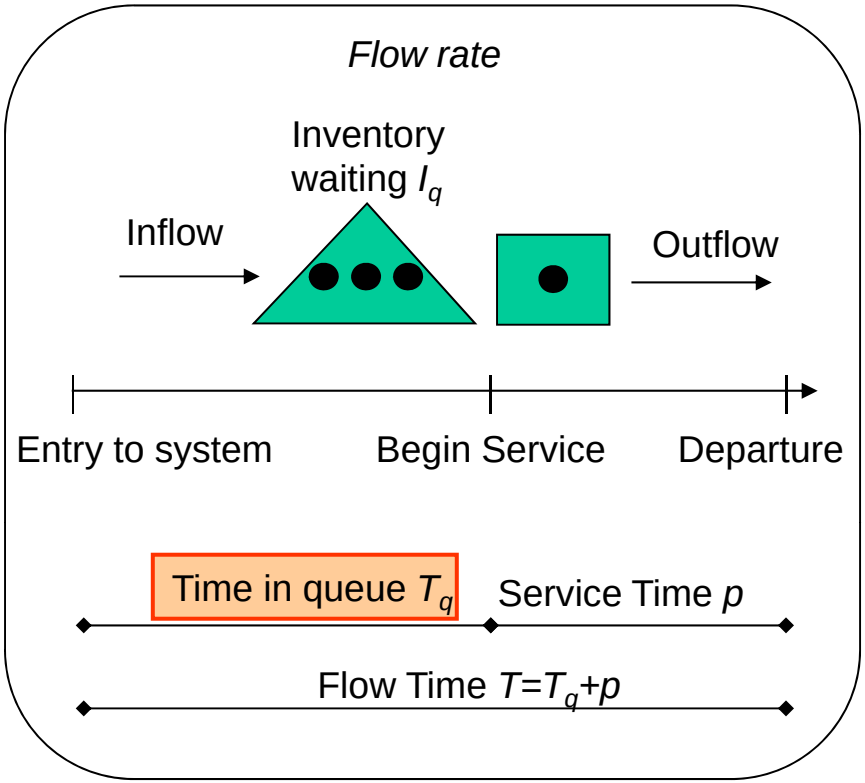
Processing
p: average processing time

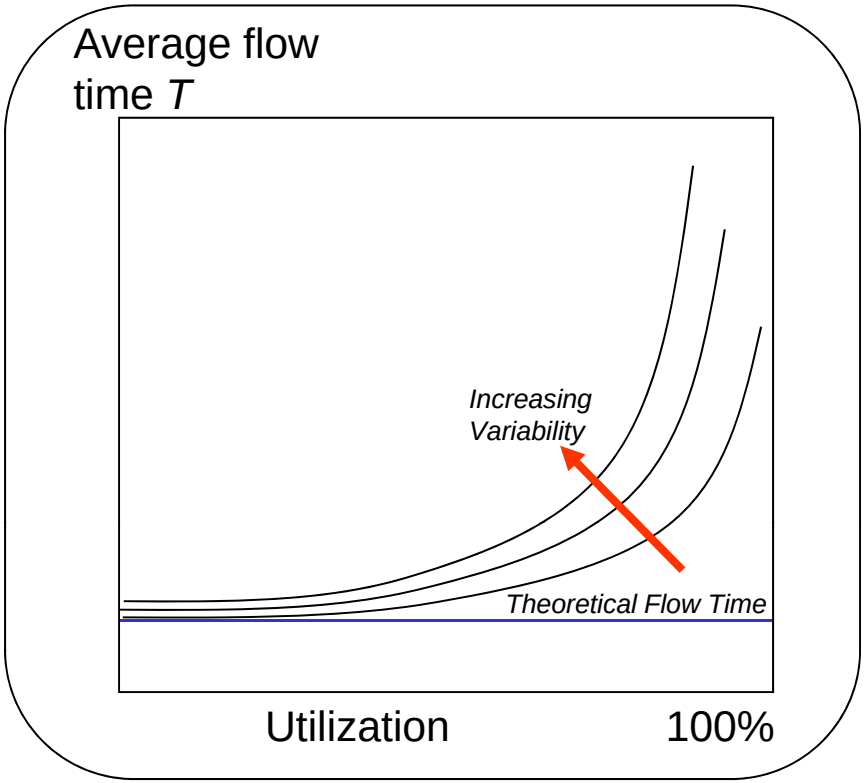
Same as "activity time" and "service time"

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

Can have many distributions:
- CV_p depends strongly on standardization
- Often Beta or LogNormal

The Waiting Time Formula





Waiting Time Formula

$$\text{Time in queue} = \underbrace{\text{Activity Time}}_{\text{Service time factor}} * \underbrace{\left(\frac{\text{utilization}}{1-\text{utilization}}\right)}_{\text{Utilization factor}} * \underbrace{\left(\frac{CV_a^2 + CV_p^2}{2}\right)}_{\text{Variability factor}}$$

Example: Walk-in Doc

Newt Philly needs to get some medical advise. He knows that his Doc, Francoise, has a patient arrive every 30 minutes (with a standard deviation of 30 minutes). A typical consultation lasts 15 minutes (with a standard deviation of 15 minutes). The Doc has an open-access policy and does not offer appointments.

If Newt walks into Francois's practice at 10am, when can he expect to leave the practice again?

Summary

Even though the utilization of a process might be less than 100%, it might still require long customer wait time

Variability is the root cause for this effect

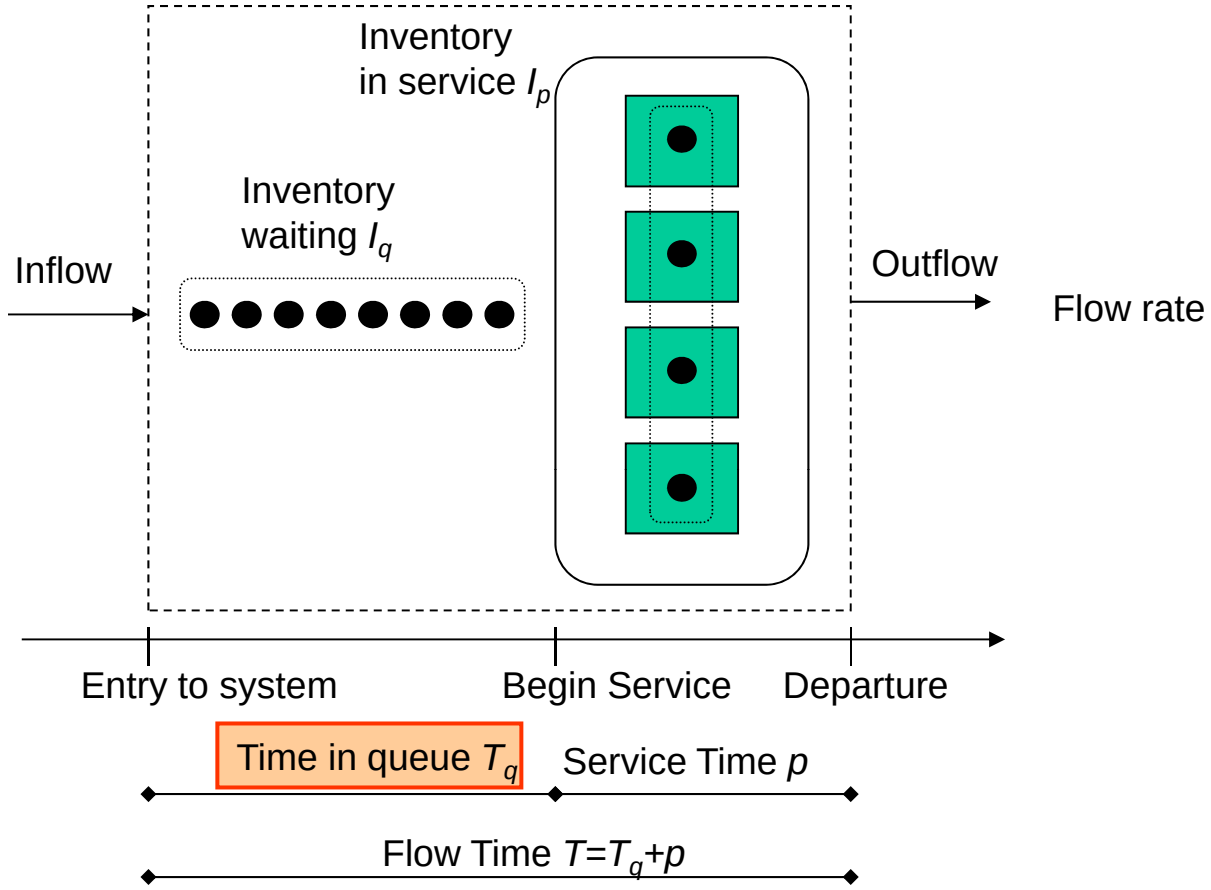
As utilization approaches 100%, you will see a very steep increase in the wait time

If you want fast service, you will have to hold excess capacity

Response Time

More on Waiting time models / Staffing to Demand

Waiting Time Formula for Multiple, Parallel Resources



Waiting Time Formula for Multiple (m) Servers

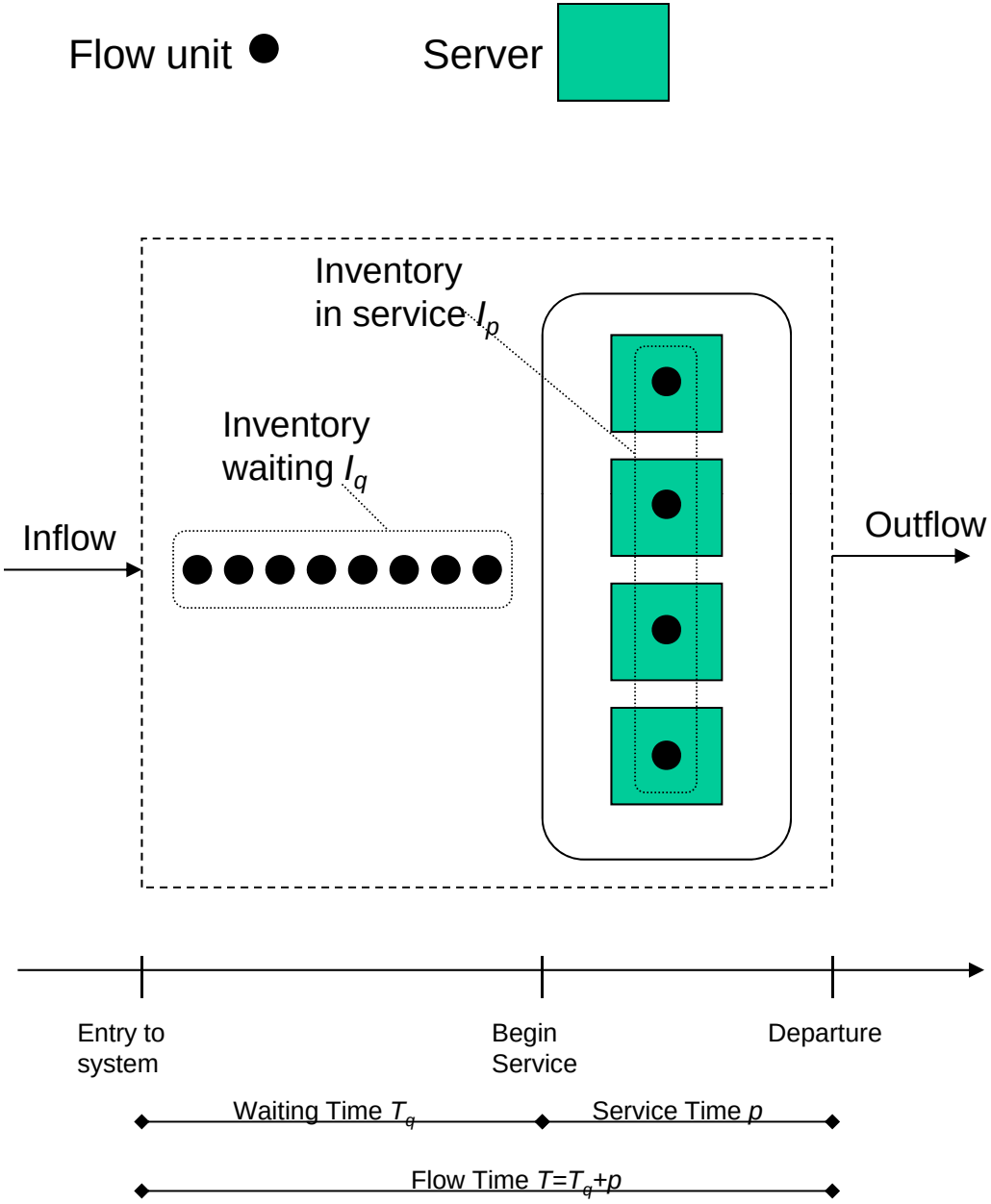
$$Time\ in\ queue = \left(\frac{Activity\ time}{m}\right) * \left(\frac{utilization^{\sqrt{2(m+1)}-1}}{1-utilization}\right) * \left(\frac{CV_a^2 + CV_p^2}{2}\right)$$

Example: Online retailer

Customers send emails to a help desk of an online retailer every 2 minutes, on average, and the standard deviation of the inter-arrival time is also 2 minutes. The online retailer has three employees answering emails. It takes on average 4 minutes to write a response email. The standard deviation of the service times is 2 minutes.

Estimate the average customer wait before being served.

Summary of Queuing Analysis



Utilization (Note: make sure <1)

$$u = \frac{1/a}{m * 1/p} = \frac{p}{am}$$

Time related measures

$$T_q \approx \left(\frac{p}{m}\right) * \left(\frac{u^{\sqrt{2(m+1)}-1}}{1-u}\right) * \left(\frac{CV_a^2 + CV_p^2}{2}\right)$$

$$T = T_q + p$$

Inventory related measures (Flow rate=*1/a*)

$$I_q = \frac{1}{a} * T_q$$

$$I_p = u * m$$

$$I = I_p + I_q$$

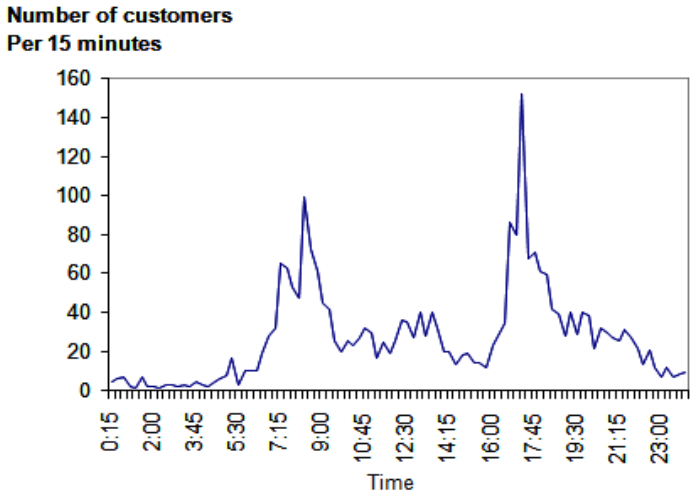
Staffing Decision

Customers send emails to a help desk of an online retailer every 2 minutes, on average, and the standard deviation of the inter-arrival time is also 2 minutes. The online retailer has three employees answering emails. It takes on average 4 minutes to write a response email. The standard deviation of the service times is 2 minutes.

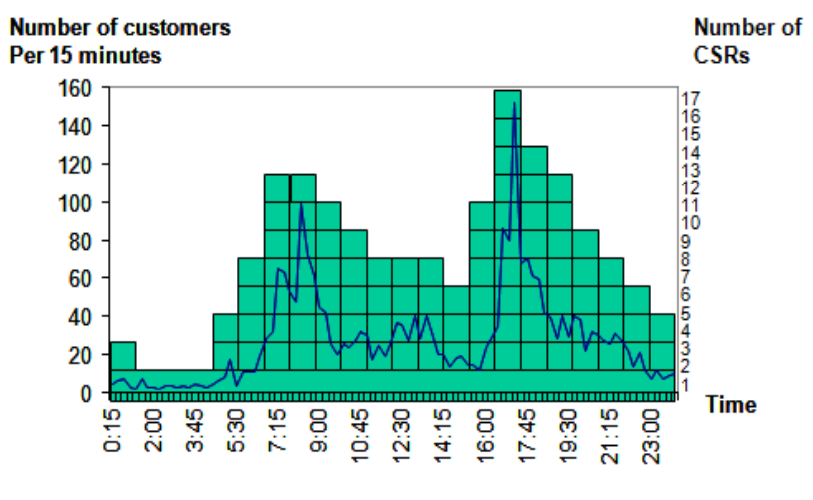
How many employees would we have to add to get the average wait time reduced to x minutes?

What to Do With Seasonal Data

Measure the true demand data



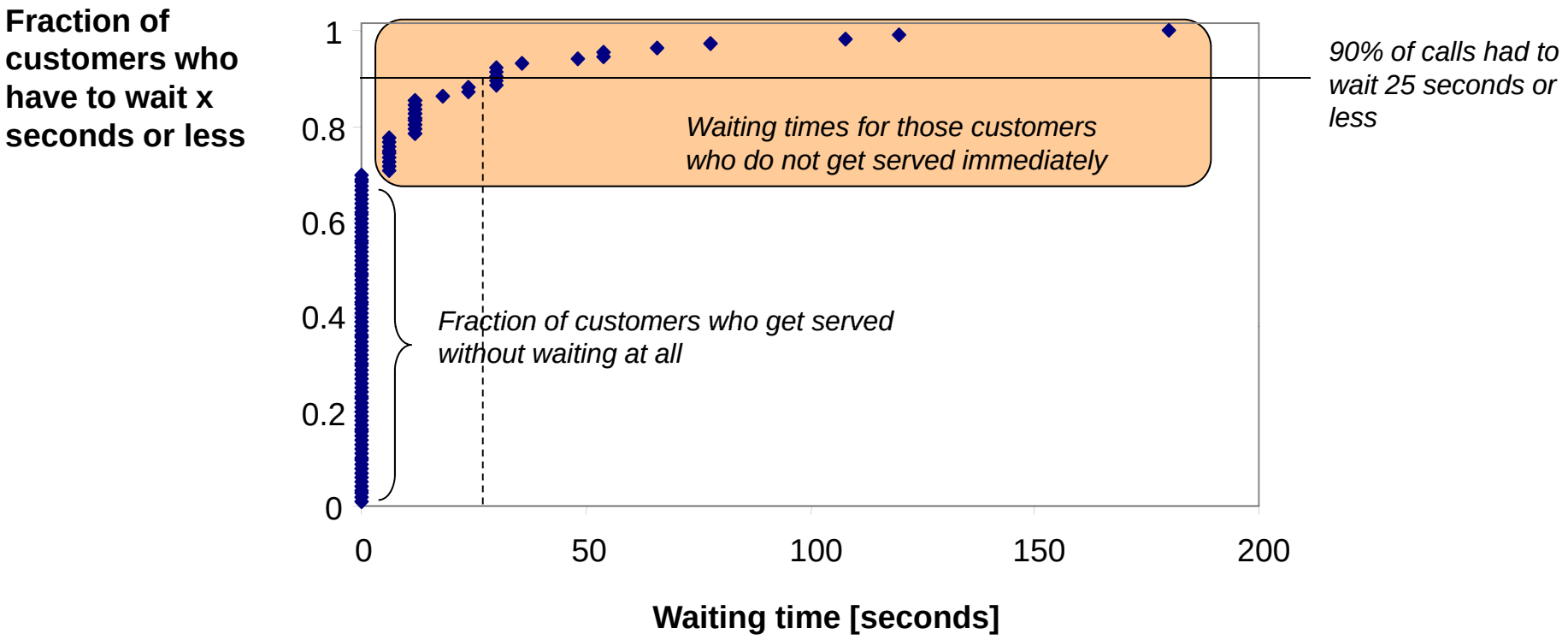
Apply waiting model in each slice



Slice the data by the hour (30min, 15min)

Level the demand
Assume demand is “stationary” within a slice

Service Levels in Waiting Systems



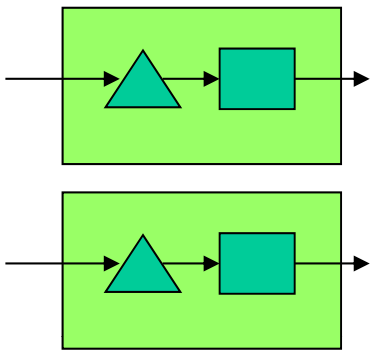
- Target Wait Time (TWT)
- Service Level = Probability{Waiting Time≤TWT}
- Example: Big Call Center
  - starting point / diagnostic: 30% of calls answered within 20 seconds
  - target: 80% of calls answered within 20 seconds

Response Time

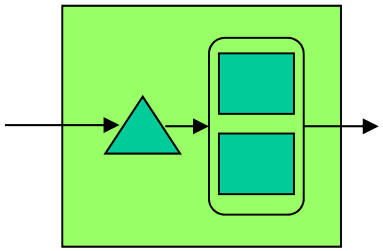
Capacity Pooling

Managerial Responses to Variability: Pooling

Independent Resources
2x(m=1)



Pooled Resources
(m=2)



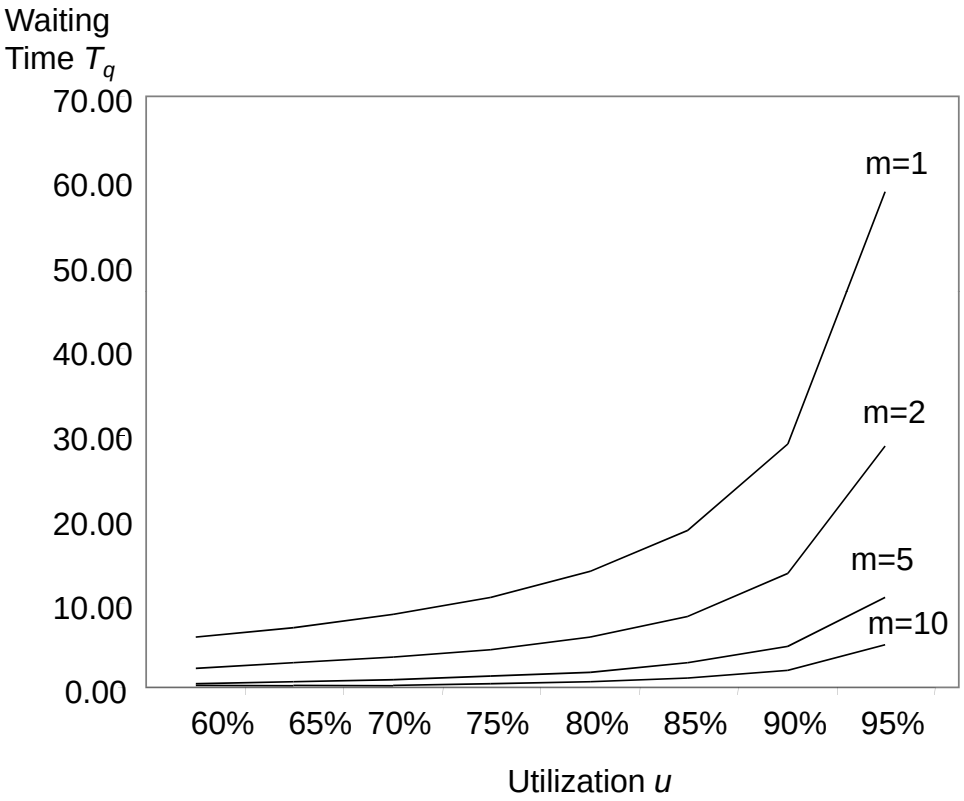
Example:
Processing time=4 minutes
Inter-arrival time=5 minutes (at each server)
m=1, Cva=CVp=1

⇒Tq =

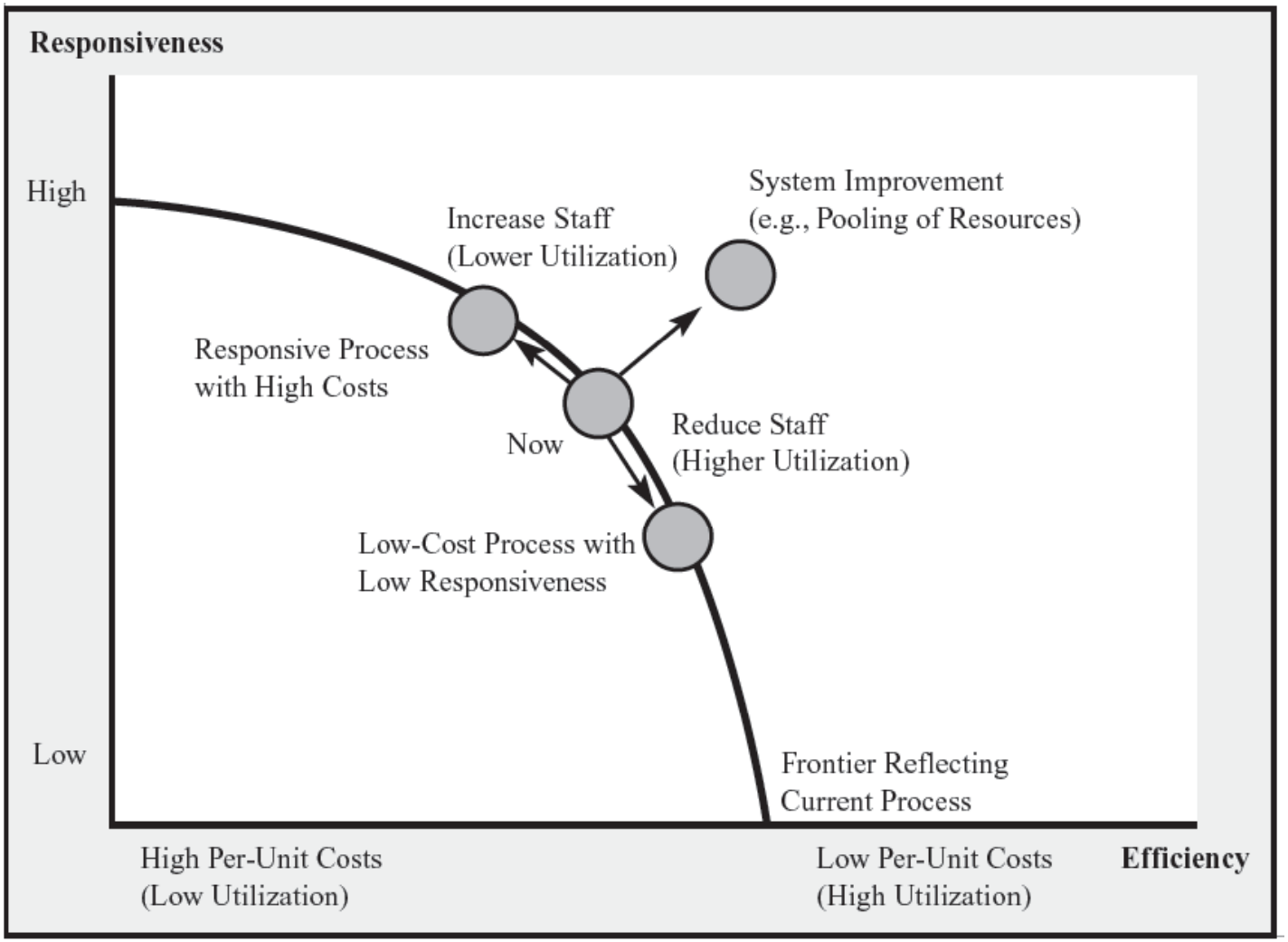
Processing time=4 minutes
Inter-arrival time=2.5 minutes
m=2, Cva=CVp=1

⇒Tq =

Managerial Responses to Variability: Pooling



Pooling: Shifting the Efficient Frontier



Summary

What is a good wait time?

Fire truck or IRS?

Limitations of Pooling

Assumes flexibility

Increases complexity of work-flow

Can increase the variability of service time

Interrupts the relationship with the customer / one-face-to-the-customer

Group clinics

Electricity grid / smart grid

Flexible production plants

The Three Enemies of Operations

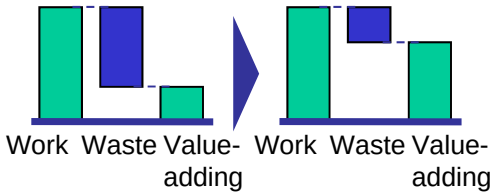
Additional costs due to variability in demand and activity times

Is associated with longer wait times and / or customer loss

Requires process to hold excess capacity (idle time)

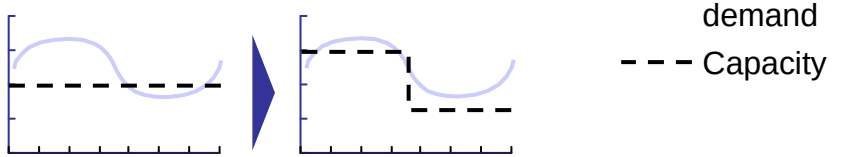
Use of resources beyond what is needed to meet customer requirements

- Not adding value to the product, but adding cost
- Reducing the performance of the production system
- 7 different types of waste



Additional costs incurred because of supply demand mismatches

- Waiting customers or
- Waiting (idle capacity)



Response Time

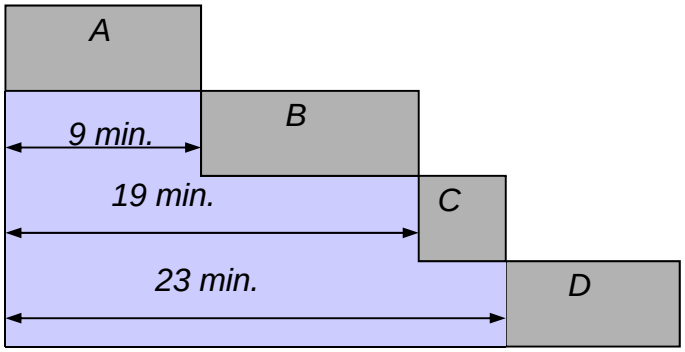
Scheduling / Access

Managerial Responses to Variability: Priority Rules in Waiting Time Systems

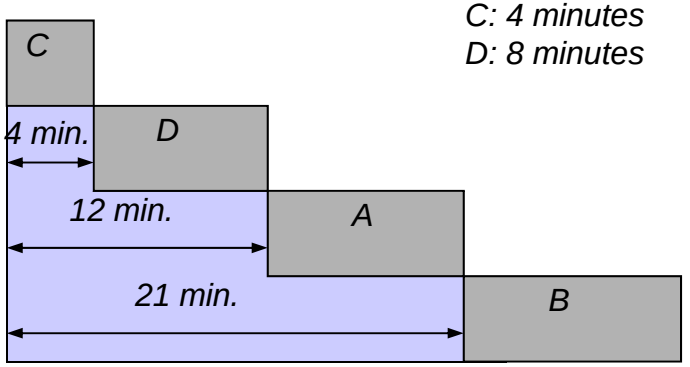
- Flow units are sequenced in the waiting area (triage step)
- Provides an opportunity for us to move some units forwards and some backwards
- First-Come-First-Serve
  - easy to implement
  - perceived fairness
  - lowest variance of waiting time
- Sequence based on importance
  - emergency cases
  - identifying profitable flow units

Managerial Responses to Variability: Priority Rules in Waiting Time Systems

Service times:
A: 9 minutes
B: 10 minutes
C: 4 minutes
D: 8 minutes



Total wait time: 9+19+23=51min



Total wait time: 4+13+21=38 min

- Shortest Processing Time Rule
  - Minimizes average waiting time
  - Problem of having “true” processing times

Appointments

- Open Access
- Appointment systems

Response Time

Redesign the Service Process

Reasons for Long Response Times (And Potential Improvement Strategies)

- Insufficient capacity on a permanent basis
  => Understand what keeps the capacity low

- Demand fluctuation and temporal capacity shortfalls
  Unpredictable wait times => Extra capacity / Reduce variability in demand
  Predictable wait times => Staff to demand / Takt time

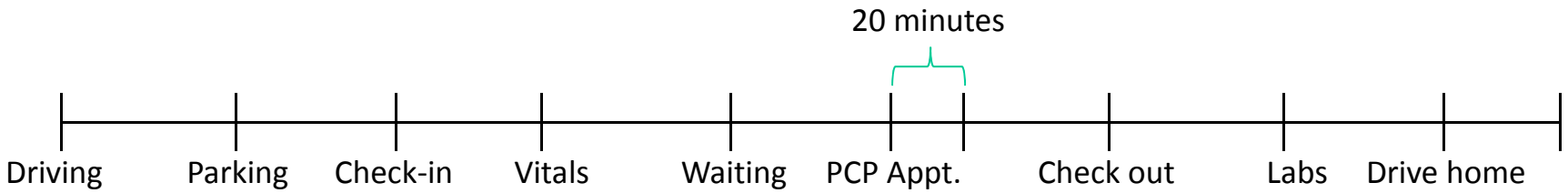
- Long wait times because of low priority
  => Align priorities with customer value

- Many steps in the process / poor internal process flow (often driven by handoffs and rework loops)
  => Redesign the service process

http://www.minyanville.com/businessmarkets/articles/drive-thrus-emissions-fast-food-mcdonalds/5/12/2010/id/28261

The Customer's Perspective

How much time does a patient spend on a primary care encounter?



Two types of wasted time:

- Auxiliary activities required to get to value add activities (result of process location / lay-out)
- Wait time (result of bottlenecks / insufficient capacity)

$$\textit{Flow Time Efficiency (or \%VAT)} = \frac{\textit{Total value add time of a unit}}{\textit{Total time a unit is in the process}}$$

Process Mapping / Service Blue Prints

Customer actions	Walk into the branch / talk to agent	Customer supplies more data	Customer supplies more data	Sign contracts
Line of interaction				
Onstage actions	Collect basic information	Request for more data	Request for more data	Explain final document
Line of visibility				
Backstage actions	Pre Approval process; set up workflow / account responsibility	Pre Approval process; set up workflow / account responsibility		
Line of internal interaction				
Support processes			Run formal credit scoring model	

Source: Yves Pigneur

Process Mapping / Service Blue Prints
How to Redesign a Service Process

Move work off the stage
- Example: online check-in at an airport

Reduce customer actions / rely on support processes
- Example: checking in at a doctor's office

Instead of optimizing the capacity of a resource, try to eliminate the step altogether
- Example: Hertz Gold – Check-in offers no value; go directly to the car

Avoid fragmentation of work due to specialization / narrow job responsibilities
- Example: Loan processing / hospital ward

If customers are likely to leave the process because of long wait times, have the wait occur later in the process / re-sequence the activities
- Example: Starbucks – Pay early, then wait for the coffee

Have the waiting occur outside of a line
- Example: Restaurants in a shopping malls using buzzers
- Example: Appointment

Communicate the wait time with the customer (set expectations)
- Example: Disney

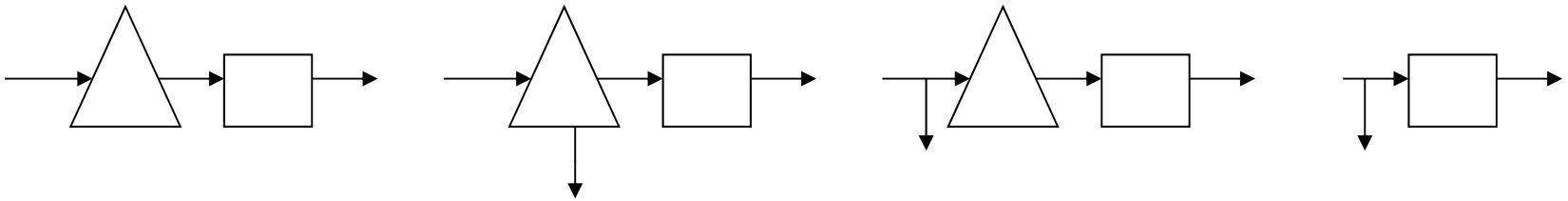
Response Time

Loss Models

Different Models of Variability

Waiting problems
Utilization has to be less than 100%
Impact of variability is on Flow Time

Loss problems
Demand can be bigger than capacity
Impact of variability is on Flow Rate



Pure waiting problem, all customers are perfectly patient.

All customers enter the process, some leave due to their impatience

Customers do not enter the process once buffer has reached a certain limit

Customers are lost once all servers are busy

Same if customers are patient

Same if buffer size=0

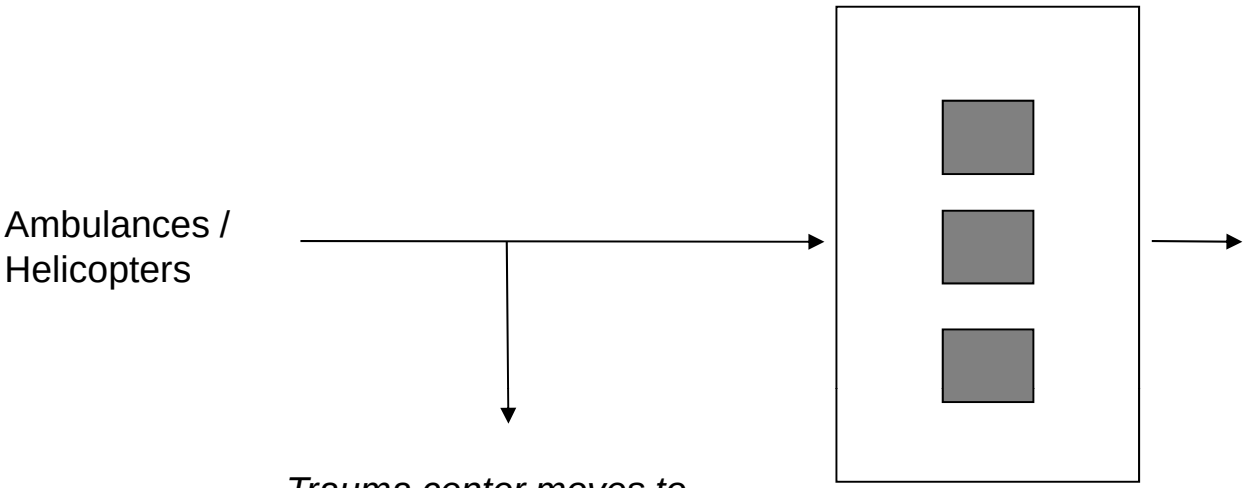
Same if buffer size is extremely large

Variability is always bad – you pay through lower flow rate and/or longer flow time

Buffer or suffer: if you are willing to tolerate waiting, you don't have to give up on flow rate

Analyzing Loss Systems

Resources
3 trauma bays (*m=3*)



Trauma center moves to diversion status once all servers are busy

incoming patients are directed to other locations

Demand Process

One trauma case comes in every 3 hours

(*a*=3 hours)

a is the interarrival time

Exponential interarrival times

Service Process

Patient stays in trauma bay for an average of 2 hours

(*p*=2 hours)

p is the service time

Can have any distribution

What is P_m, the probability that all m resources are utilized?

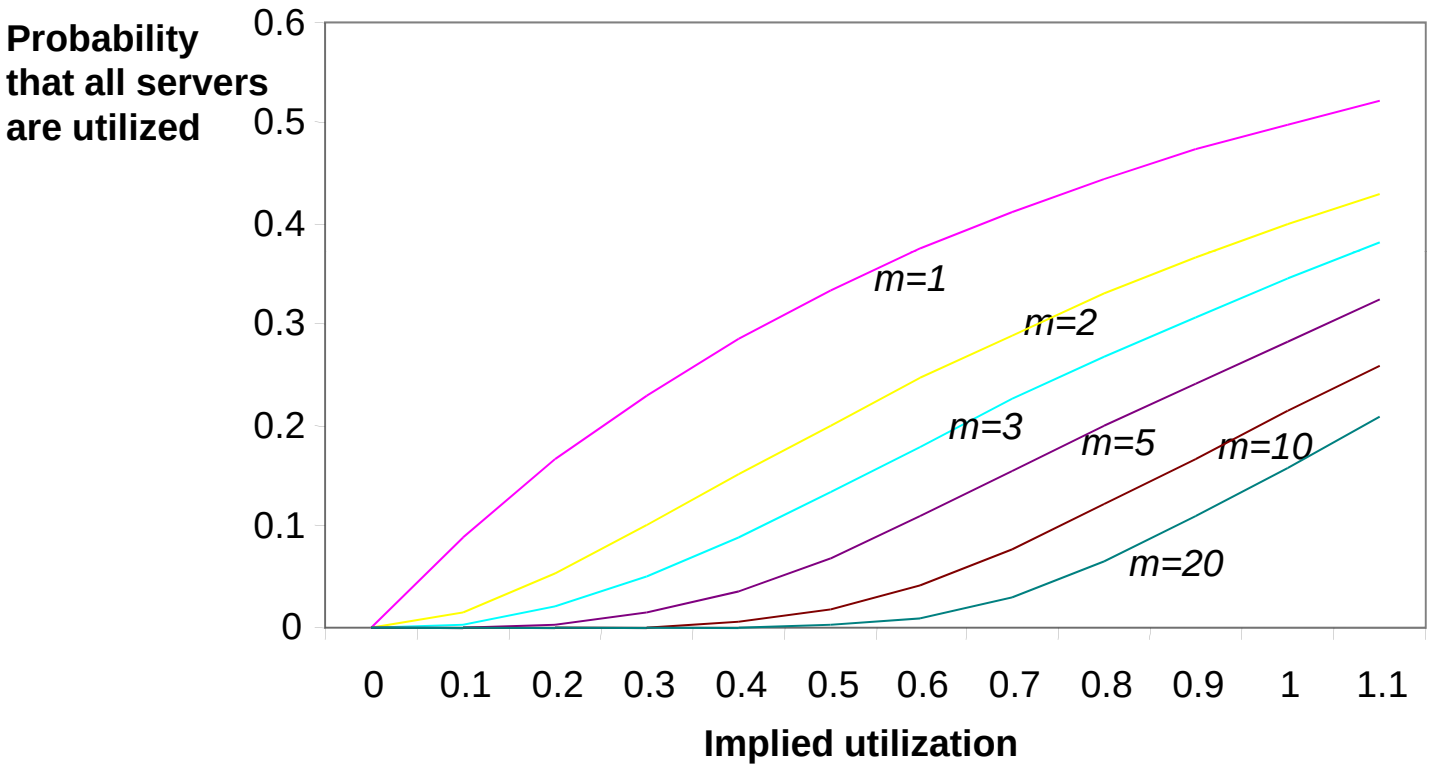
Analyzing Loss Systems: Finding $P_m(r)$

- Define $r = p / a$
- Example: r= 2 hours/ 3 hours
  r=0.67
- Recall m=3
- Use Erlang Loss Table
- Find that $P_3(0.67)=0.0255$

r = p / a	*m* = 1	2	3	4	5
0.10	0.0909	0.0045	0.0002	0.0000	0.0000
0.20	0.1667	0.0164	0.0011	0.0001	0.0000
0.25	0.2000	0.0244	0.0020	0.0001	0.0000
0.30	0.2308	0.0335	0.0033	0.0003	0.0000
0.33	0.2500	0.0400	0.0044	0.0004	0.0000
0.40	0.2857	0.0541	0.0072	0.0007	0.0001
0.50	0.3333	0.0769	0.0127	0.0016	0.0002
0.60	0.3750	0.1011	0.0198	0.0030	0.0004
0.67	0.4000	0.1176	0.0255	0.0042	0.0006
0.70	0.4118	0.1260	0.0286	0.0050	0.0007
0.75	0.4286	0.1385	0.0335	0.0062	0.0009
0.80	0.4444	0.1509	0.0387	0.0077	0.0012
0.90	0.4737	0.1757	0.0501	0.0111	0.0020
1.00	0.5000	0.2000	0.0625	0.0154	0.0031

Given $P_m(r)$ we can compute:
- Time per day that system has to deny access
- Flow units lost = 1/a * $P_m(r)$

Implied utilization vs probability of having all servers utilized: Pooling Revisited



Erlang Loss Table

	m									
	1	2	3	4	5	6	7	8	9	10
0.10	0.0909	0.0045	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.20	0.1667	0.0164	0.0011	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.25	0.2000	0.0244	0.0020	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.30	0.2308	0.0335	0.0033	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.33	0.2500	0.0400	0.0044	0.0004	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.40	0.2857	0.0541	0.0072	0.0007	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
0.50	0.3333	0.0769	0.0127	0.0016	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000
0.60	0.3750	0.1011	0.0198	0.0030	0.0004	0.0000	0.0000	0.0000	0.0000	0.0000
0.67	0.4000	0.1176	0.0255	0.0042	0.0006	0.0001	0.0000	0.0000	0.0000	0.0000
0.70	0.4118	0.1260	0.0286	0.0050	0.0007	0.0001	0.0000	0.0000	0.0000	0.0000
0.75	0.4286	0.1385	0.0335	0.0062	0.0009	0.0001	0.0000	0.0000	0.0000	0.0000
0.80	0.4444	0.1509	0.0387	0.0077	0.0012	0.0002	0.0000	0.0000	0.0000	0.0000
0.90	0.4737	0.1757	0.0501	0.0111	0.0020	0.0003	0.0000	0.0000	0.0000	0.0000
1.00	0.5000	0.2000	0.0625	0.0154	0.0031	0.0005	0.0001	0.0000	0.0000	0.0000
1.10	0.5238	0.2237	0.0758	0.0204	0.0045	0.0008	0.0001	0.0000	0.0000	0.0000
1.20	0.5455	0.2466	0.0898	0.0262	0.0063	0.0012	0.0002	0.0000	0.0000	0.0000
1.25	0.5556	0.2577	0.0970	0.0294	0.0073	0.0015	0.0003	0.0000	0.0000	0.0000
1.30	0.5652	0.2687	0.1043	0.0328	0.0085	0.0018	0.0003	0.0001	0.0000	0.0000
1.33	0.5714	0.2759	0.1092	0.0351	0.0093	0.0021	0.0004	0.0001	0.0000	0.0000
1.40	0.5833	0.2899	0.1192	0.0400	0.0111	0.0026	0.0005	0.0001	0.0000	0.0000
1.50	0.6000	0.3103	0.1343	0.0480	0.0142	0.0035	0.0008	0.0001	0.0000	0.0000
1.60	0.6154	0.3299	0.1496	0.0565	0.0177	0.0047	0.0011	0.0002	0.0000	0.0000
1.67	0.6250	0.3425	0.1598	0.0624	0.0204	0.0056	0.0013	0.0003	0.0001	0.0000
1.70	0.6296	0.3486	0.1650	0.0655	0.0218	0.0061	0.0015	0.0003	0.0001	0.0000
1.75	0.6364	0.3577	0.1726	0.0702	0.0240	0.0069	0.0017	0.0004	0.0001	0.0000
r = p/a 1.80	0.6429	0.3665	0.1803	0.0750	0.0263	0.0078	0.0020	0.0005	0.0001	0.0000
1.90	0.6552	0.3836	0.1955	0.0850	0.0313	0.0098	0.0027	0.0006	0.0001	0.0000
2.00	0.6667	0.4000	0.2105	0.0952	0.0367	0.0121	0.0034	0.0009	0.0002	0.0000
2.10	0.6774	0.4156	0.2254	0.1058	0.0425	0.0147	0.0044	0.0011	0.0003	0.0001
2.20	0.6875	0.4306	0.2400	0.1166	0.0488	0.0176	0.0055	0.0015	0.0004	0.0001
2.25	0.6923	0.4378	0.2472	0.1221	0.0521	0.0192	0.0061	0.0017	0.0004	0.0001
2.30	0.6970	0.4449	0.2543	0.1276	0.0554	0.0208	0.0068	0.0019	0.0005	0.0001
2.33	0.7000	0.4495	0.2591	0.1313	0.0577	0.0220	0.0073	0.0021	0.0005	0.0001
2.40	0.7059	0.4586	0.2684	0.1387	0.0624	0.0244	0.0083	0.0025	0.0007	0.0002
2.50	0.7143	0.4717	0.2822	0.1499	0.0697	0.0282	0.0100	0.0031	0.0009	0.0002
2.60	0.7222	0.4842	0.2956	0.1612	0.0773	0.0324	0.0119	0.0039	0.0011	0.0003
2.67	0.7273	0.4923	0.3044	0.1687	0.0825	0.0354	0.0133	0.0044	0.0013	0.0003
2.70	0.7297	0.4963	0.3087	0.1725	0.0852	0.0369	0.0140	0.0047	0.0014	0.0004
2.75	0.7333	0.5021	0.3152	0.1781	0.0892	0.0393	0.0152	0.0052	0.0016	0.0004
2.80	0.7368	0.5078	0.3215	0.1837	0.0933	0.0417	0.0164	0.0057	0.0018	0.0005
2.90	0.7436	0.5188	0.3340	0.1949	0.1016	0.0468	0.0190	0.0068	0.0022	0.0006
3.00	0.7500	0.5294	0.3462	0.2061	0.1101	0.0522	0.0219	0.0081	0.0027	0.0008
3.10	0.7561	0.5396	0.3580	0.2172	0.1187	0.0578	0.0249	0.0096	0.0033	0.0010
3.20	0.7619	0.5494	0.3695	0.2281	0.1274	0.0636	0.0283	0.0112	0.0040	0.0013
3.25	0.7647	0.5541	0.3751	0.2336	0.1318	0.0666	0.0300	0.0120	0.0043	0.0014
3.30	0.7674	0.5587	0.3807	0.2390	0.1362	0.0697	0.0318	0.0130	0.0047	0.0016
3.33	0.7692	0.5618	0.3843	0.2426	0.1392	0.0718	0.0331	0.0136	0.0050	0.0017
3.40	0.7727	0.5678	0.3915	0.2497	0.1452	0.0760	0.0356	0.0149	0.0056	0.0019
3.50	0.7778	0.5765	0.4021	0.2603	0.1541	0.0825	0.0396	0.0170	0.0066	0.0023
3.60	0.7826	0.5848	0.4124	0.2707	0.1631	0.0891	0.0438	0.0193	0.0077	0.0028
3.67	0.7857	0.5902	0.4191	0.2775	0.1691	0.0937	0.0468	0.0210	0.0085	0.0031
3.70	0.7872	0.5929	0.4224	0.2809	0.1721	0.0960	0.0483	0.0218	0.0089	0.0033
3.75	0.7895	0.5968	0.4273	0.2860	0.1766	0.0994	0.0506	0.0232	0.0096	0.0036
3.80	0.7917	0.6007	0.4321	0.2910	0.1811	0.1029	0.0529	0.0245	0.0102	0.0039
3.90	0.7959	0.6082	0.4415	0.3009	0.1901	0.1100	0.0577	0.0274	0.0117	0.0046
4.00	0.8000	0.6154	0.4507	0.3107	0.1991	0.1172	0.0627	0.0304	0.0133	0.0053

Erlang Loss Table

Probability{all *m* servers busy}=

$$P_m(r) = \frac{\frac{r^m}{m!}}{1 + \frac{r^1}{1!} + \frac{r^2}{2!} + \ldots + \frac{r^m}{m!}}$$

Response Time

Review

(My-law.com)
My-law.com is a recent start-up trying to cater to customers in search of legal services online. Unlike traditional law firms, My-law.com allows for extensive interaction between lawyers and their customers via telephone and the Internet. This process is used in the upfront part of the customer interaction, largely consisting of answering some basic customer questions prior to entering a formal relationship. In order to allow customers to interact with the firm's lawyers, customers are encouraged to send e-mails to my-lawyer@My-law.com. From there, the incoming e-mails are distributed to the lawyer who is currently "on call." Given the broad skills of the lawyers, each lawyer can respond to each incoming request.

E-mails arrive from 8 A.M. to 6 P.M. at a rate of 10 e-mails per hour (coefficient of variation for the arrivals is 1). At each moment in time, there is exactly one lawyer "on call," that is, sitting at his or her desk waiting for incoming e-mails. It takes the lawyer, on average, 5 minutes to write the response e-mail. The standard deviation of this is 4 minutes.

a. What is the average time a customer has to wait for the response to his/her e-mail, ignoring any transmission times? *Note:* This includes the time it takes the lawyer to start writing the e-mail *and* the actual writing time.

b. How many e-mails will a lawyer have received at the end of a 10-hour day?

c. When not responding to e-mails, the lawyer on call is encouraged to actively pursue cases that potentially could lead to large settlements. How much time on a 10-hour day can a My-law.com lawyer dedicate to this activity

Jim's Computer

Jim wants to find someone to fix his computer. PC Fixers (PF) is a local service that offers such computer repairs. A new customer walks into PF every 10 minutes (with a standard deviation of 10 minutes). PF has a staff of 5 computer technicians. Service times average around 40 minutes (with a standard deviation of 40 minutes).

JC1. If Jim walks into PF, how long must he wait in line before he can see a technician? (Only include the waiting time, not any service time)

JC2. How many customers will, on average, be waiting for their computer to be fixed?

Real Compute

RealCompute offers real-time computing services. The company owns 4 supercomputers that can be accessed through the internet. Their customers send jobs that arrive on average every 4 minutes (inter-arrival times are exponentially distributed and, thus, the standard deviation of the inter-arrival times is 4 minutes).

Each job takes on average 10 minutes of one of the supercomputers (during this time, the computer cannot perform any other work). Customers pay $20 for the execution of each job. Given the time-sensitive nature of the calculations, if no supercomputer is available, the job is redirected to a supercomputer of a partner company called OnComp, which charges $40 per job to Real Compute (OnComp always has supercomputer capacity available).

RC1. What is the probability with which an incoming job can be executed by one of the supercomputers owned by RealCompute?

RC2. How much does RealCompute pay on average to OnComp (in $s per hour)?

Contractor

A contractor building houses and doing renovation work has currently six projects planned for the season. Below are the items, and the estimated times to complete them:

New construction at Springfield	- 60 days
Bathroom remodeling at Herne	- 10 days
Training time for solar roof installation	- 2 days
Update web-site	- 6 days
Renovation of deck at Haverford	- 8 days
New kitchen at Rosemont	- 20 days

Suppose the contractor starts immediately with the first project, no other projects get added to this list, and the contractor sequences them so as to minimize the average time the project waits before it gets started. What will the contractor be doing in 30 days from the start date of the first project?

Call Center

Consider a call center that has a constant staffing level. Because of increased demand in the morning, the call center has a very high utilization in the morning and a very low utilization in the afternoon. Which of the following will decrease the average waiting time in the call center?

(a) Add more servers
(b) Decrease the service time coefficient of variation
(c) Decrease the average service time
(d) Level the demand between the morning hours and the afternoon hours
(e) All of the above