
Chapter 9. Variability and Its Impact on Process Performance: Waiting Time Problems

For consumers, one of the most visible and annoying forms of supply-demand mismatches \Rightarrow **waiting time**

Expected demand rate $>$ expected supply rate

When the implied utilization $<$ 100%

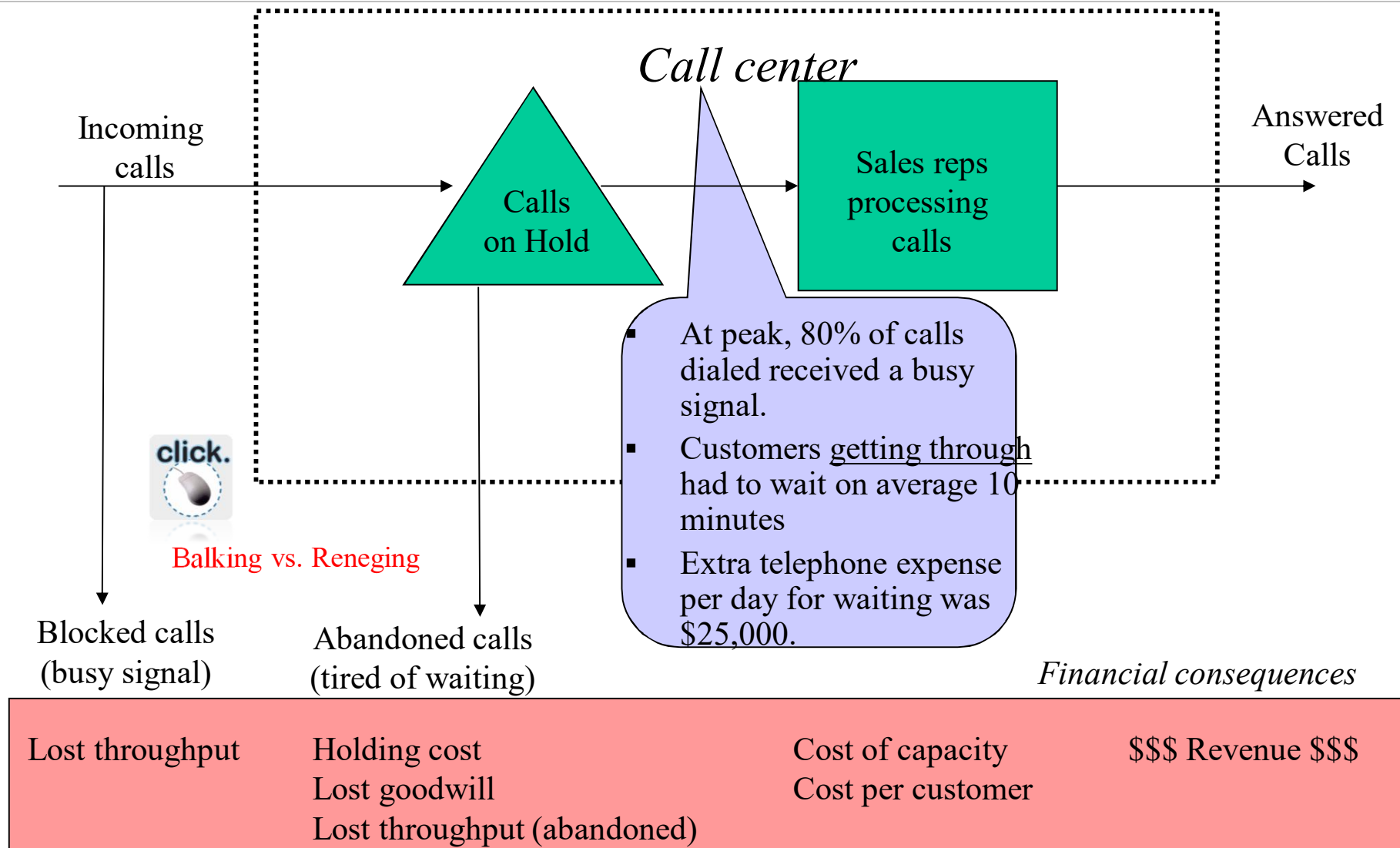
\Rightarrow In the presence of **variability**

-Predict waiting times

- Recommend ways of reducing waiting time \Rightarrow

(choose appropriate capacity level, redesign the service system, reduce variability)

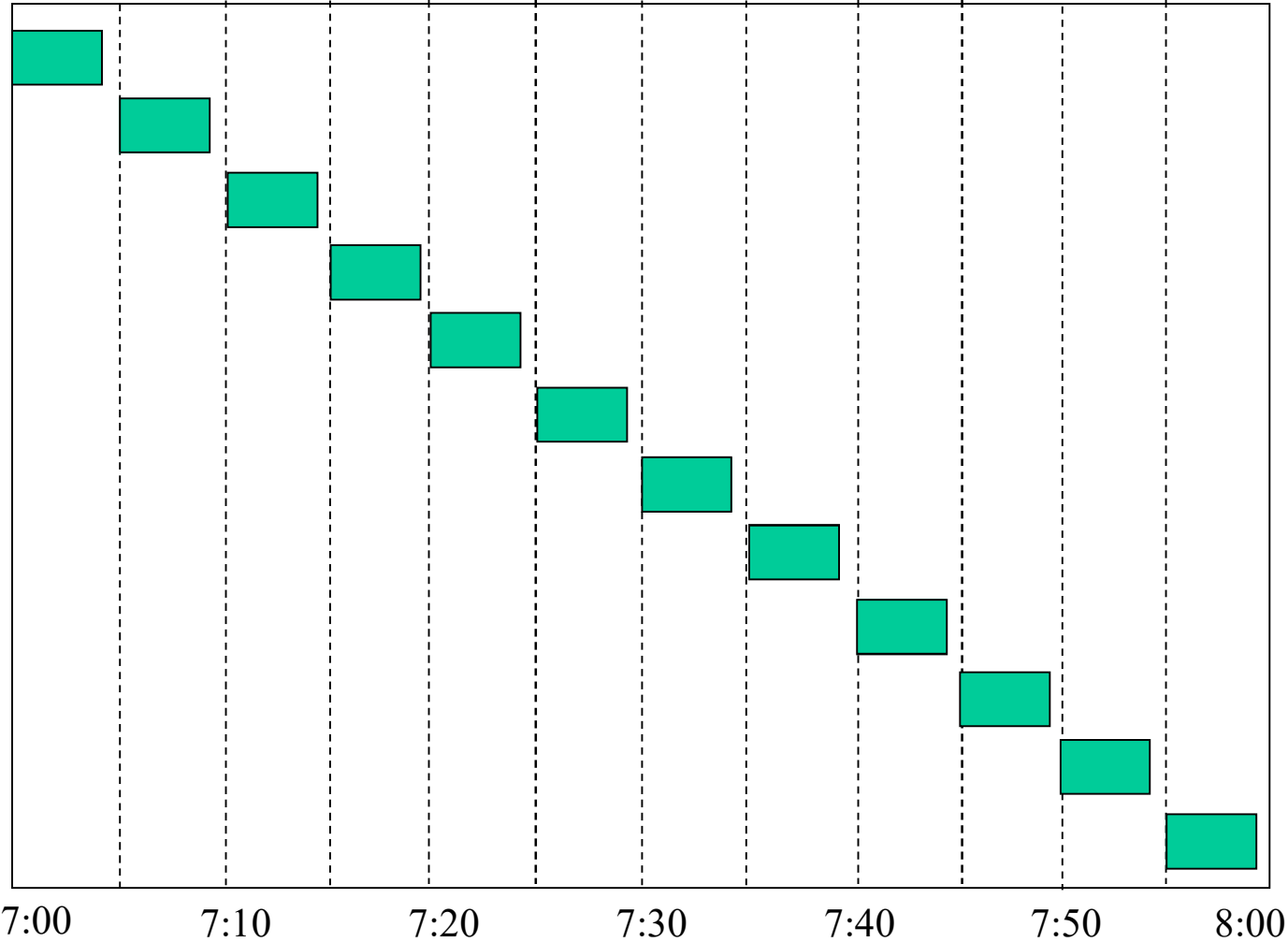
An Example of a Simple Queuing System



A Somewhat Odd Service Process

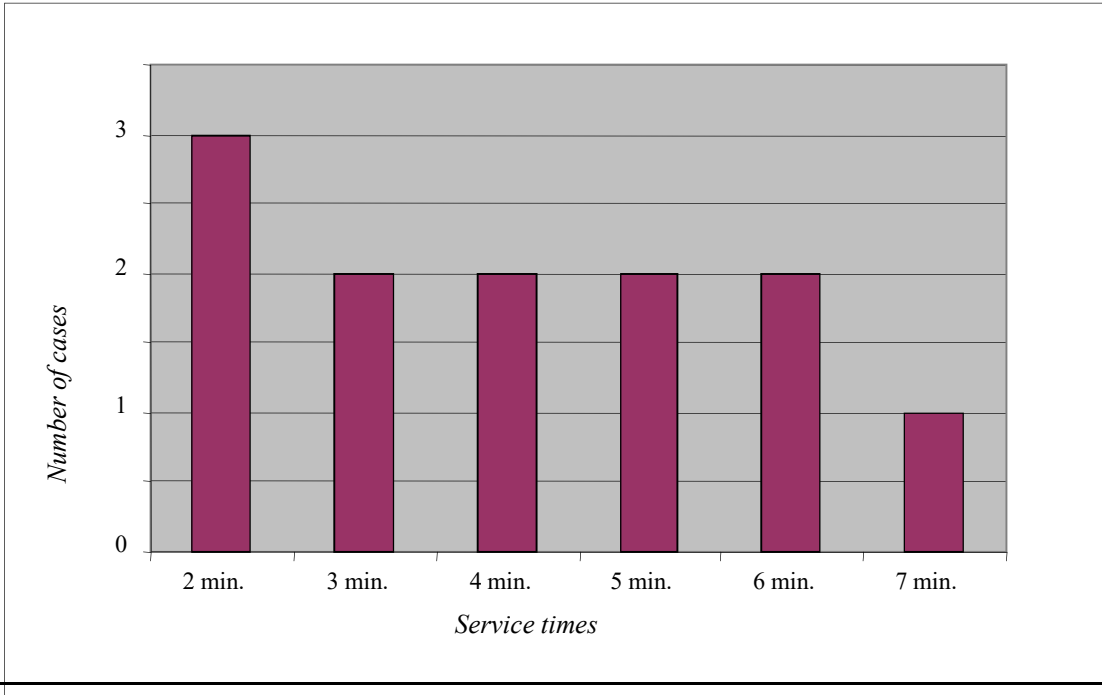
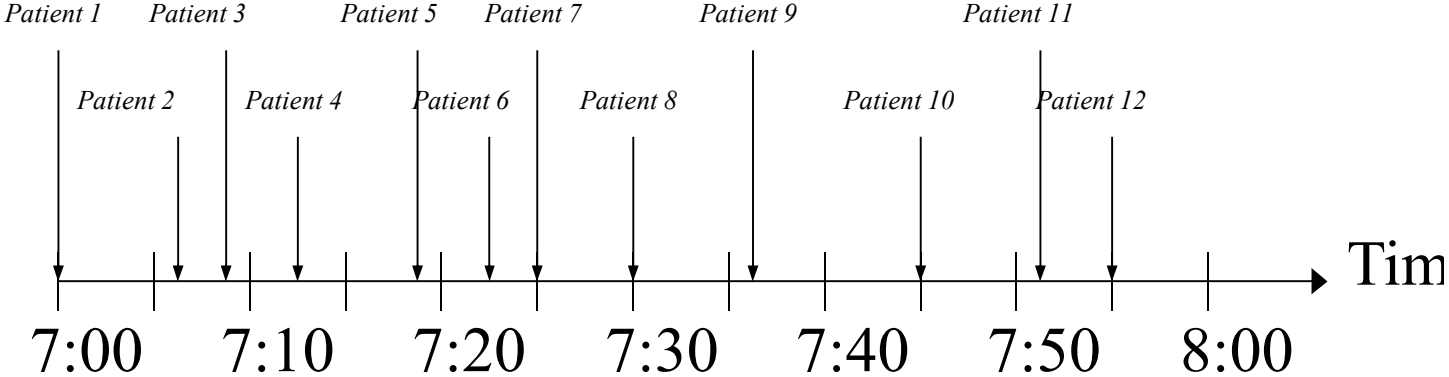
Patient	Arrival Time	Service Time
1	0	4
2	5	4
3	10	4
4	15	4
5	20	4
6	25	4
7	30	4
8	35	4
9	40	4
10	45	4
11	50	4
12	55	4

Utilization=Flow Rate/Capacity=(12patients/hr)/(15patients/hr) =80%



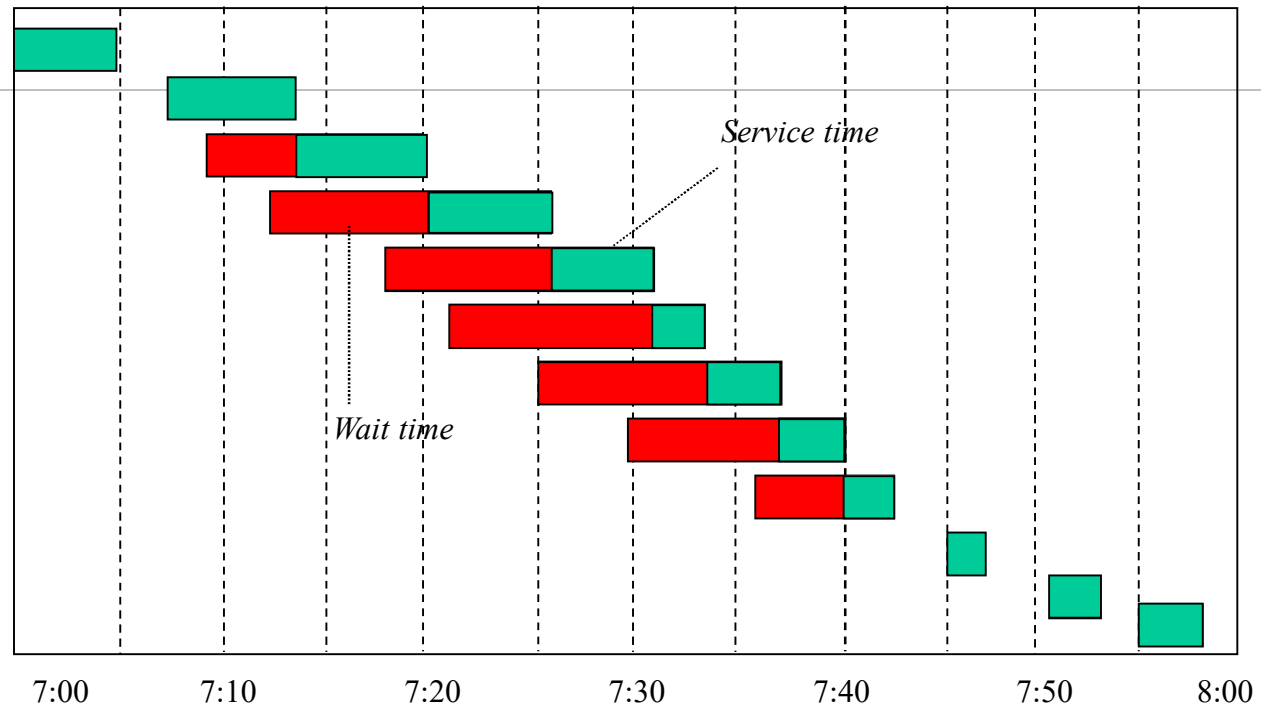
A More Realistic Service Process

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

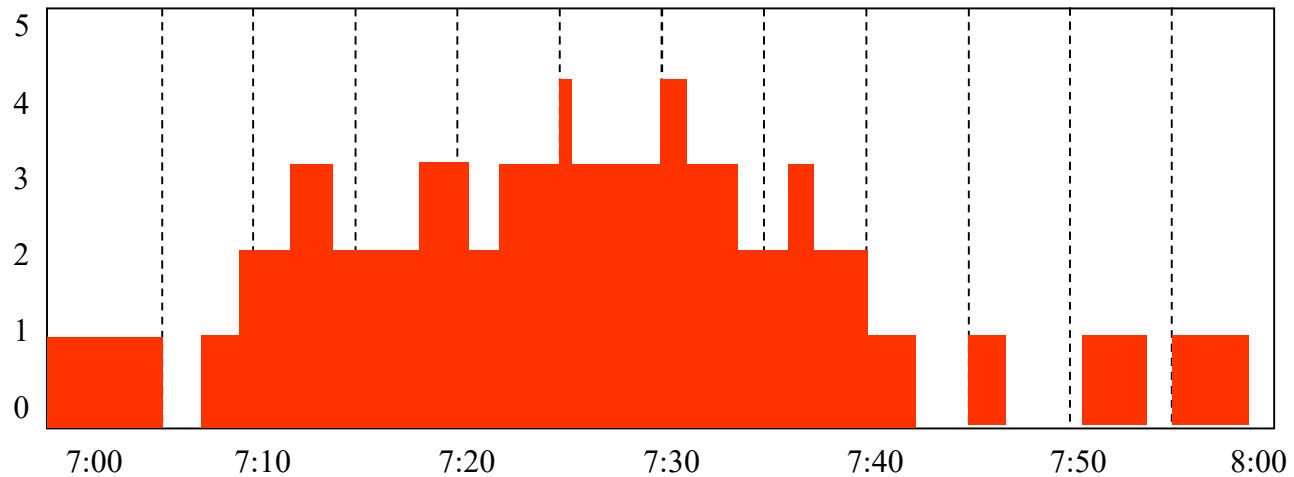
Patient	Arrival Time	Service Time
1	0	5
2	7	6
3	9	7
4	12	6
5	18	5
6	22	2
7	25	4
8	30	3
9	36	4
10	45	2
11	51	2
12	55	3



**Capacity can never run ahead of demand!
(service system)!**

Average Wait Time?

*Inventory
(Patients at lab)*



What generates queues?

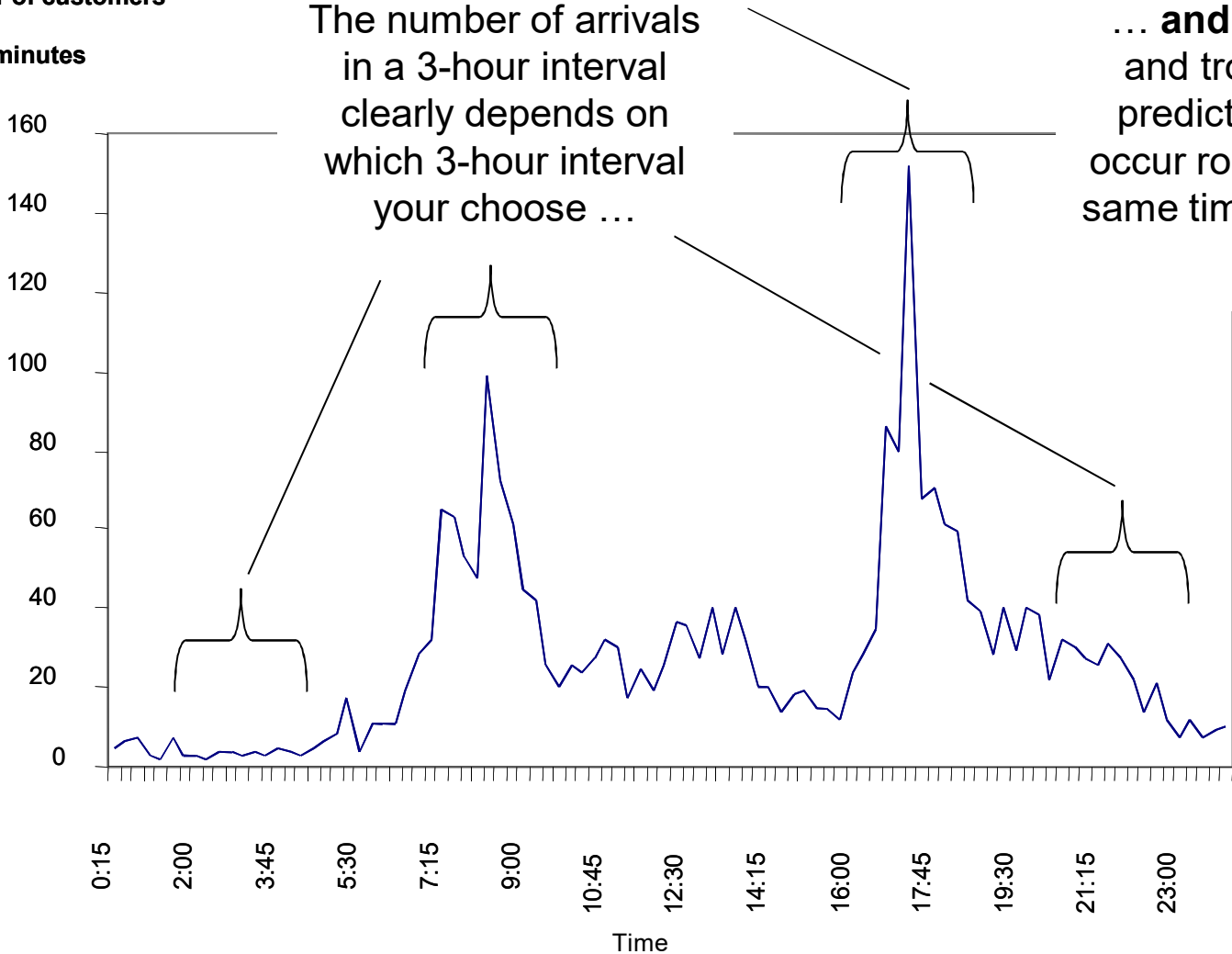
- An arrival rate that predictably exceeds the service rate (i.e., capacity) of a process.
 - e.g., toll booth congestion on the NJ Turnpike during the Thanksgiving Day rush.
- Variation in the arrival and service rates in a process where the average service rate is more than adequate to process the average arrival rate.
 - e.g., calls to a brokerage are unusually high during a particular hour relative to the same hour in other weeks (i.e., calls are high by random chance).

Defining interarrival times and a stationary process

- An **interarrival time** is the amount of time between two arrivals to a process.
- An arrival process is **stationary** over a period of time if the number of arrivals in any subinterval depends only on the length of the interval and not on when the interval starts.
 - For example, if the process is stationary over the course of a day, then the expected number of arrivals within any three hour interval is about the same no matter which three hour window is chosen (or six hour window, or one hour window, etc.).
 - Processes tend to be **nonstationary** (or seasonal) over long time periods (e.g., over a day or several hours) but stationary over short periods of time (say one hour, or 15 minutes).

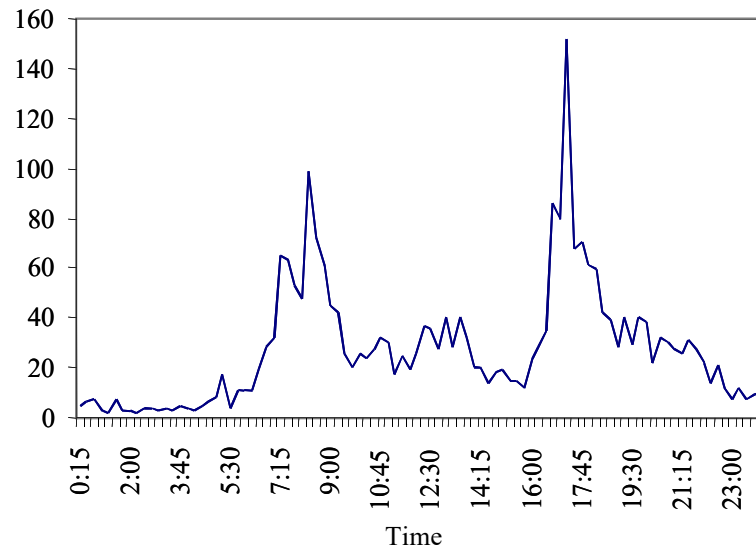
Arrivals to An-ser over the day are nonstationary

Number of customers
Per 15 minutes



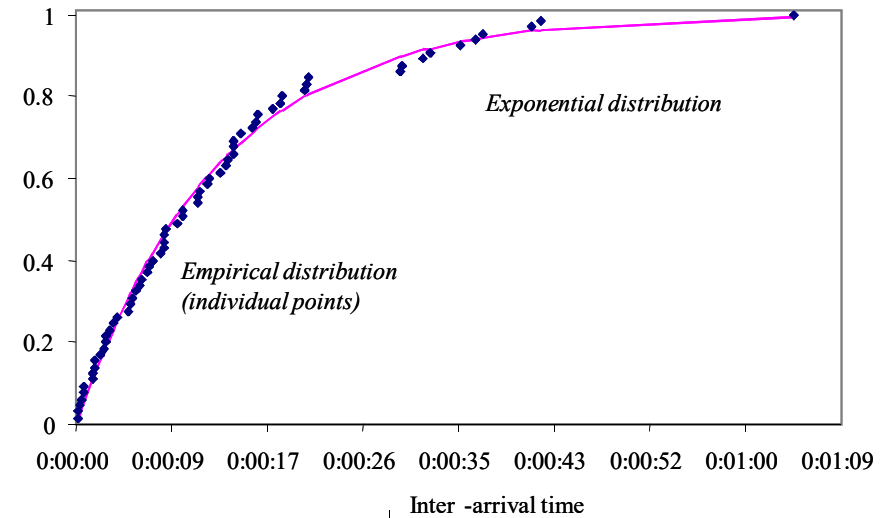
Data in Practical Call Center Setting

Number of customers
Per 15 minutes



- Seasonality vs. variability
- Need to slice-up the data

Distribution Function

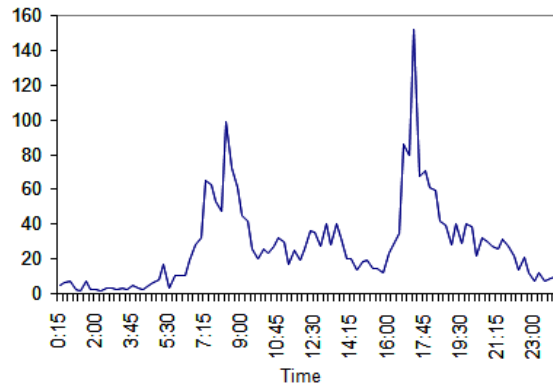


- Within a “slice”, exponential distribution ($CV_a=1$)
- See chapter 6 for various data analysis tools

What to Do With Seasonal Data

Measure the true demand data

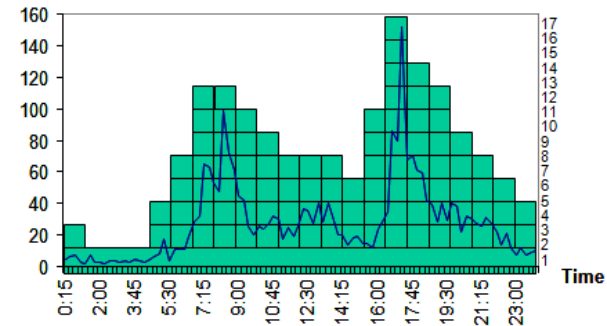
Number of customers
Per 15 minutes



Apply waiting model in each slice

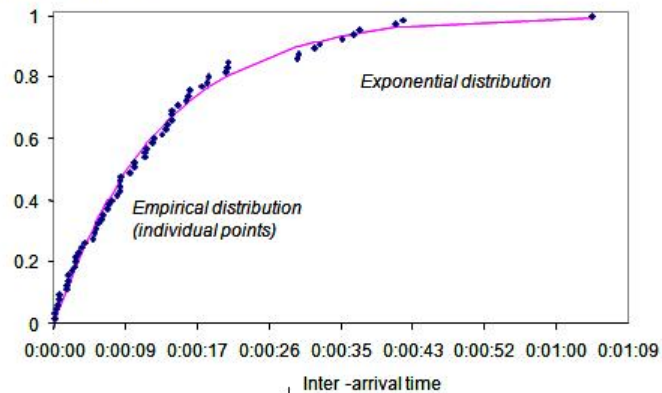
Number of customers
Per 15 minutes

Number of CSRs



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

Distribution Function

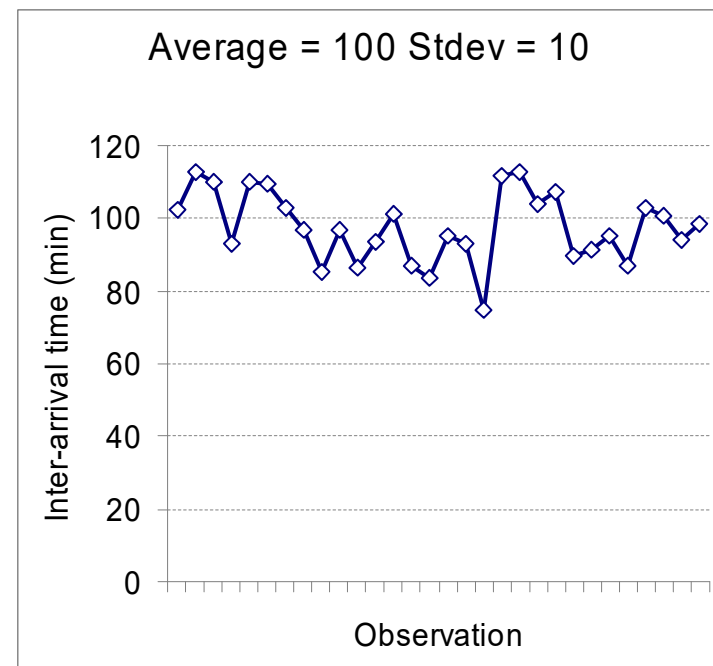
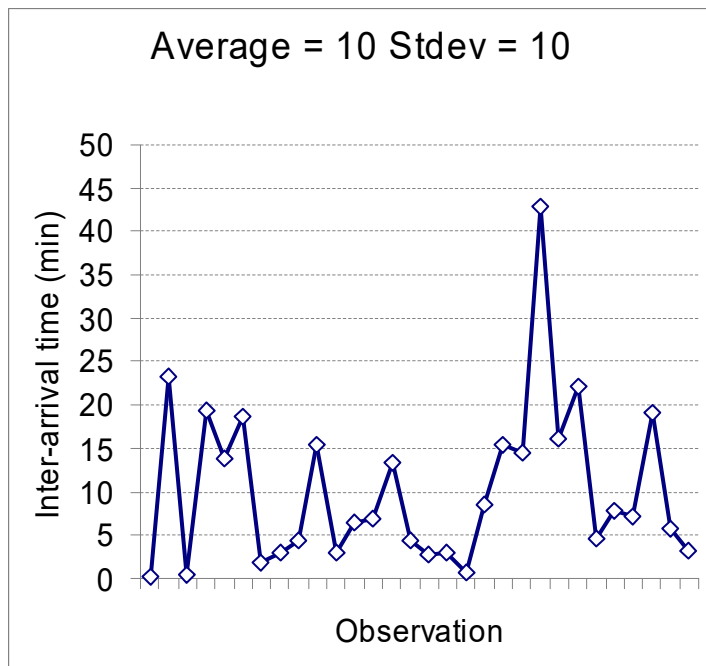


How to describe (or model) interarrival times

- We will use two parameters to describe interarrival times to a process:
 - The **average** interarrival time.
 - The **standard deviation** of the interarrival times.
- What is the standard deviation?
 - Roughly speaking, the **standard deviation** is a measure of how variable the interarrival times are.
 - Two arrival processes can have the same average interarrival time (say 1 minute) but one can have more variation about that average, i.e., a higher standard deviation.

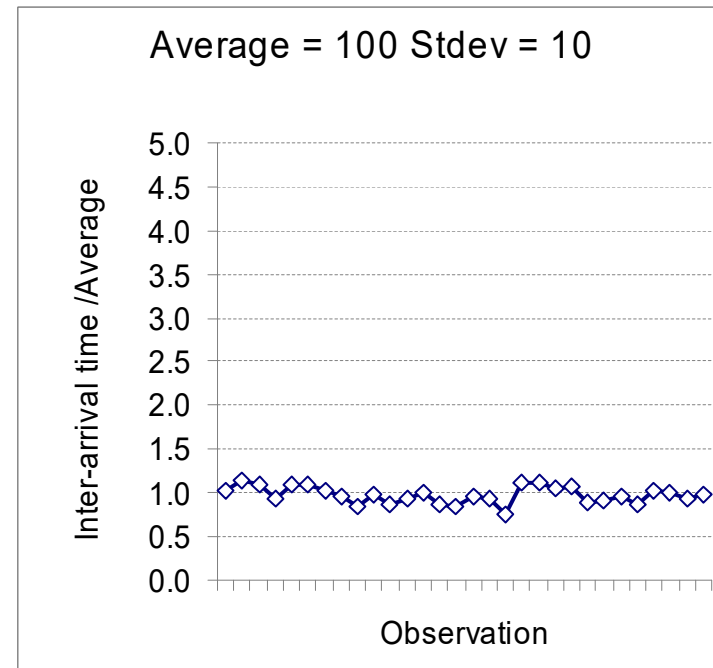
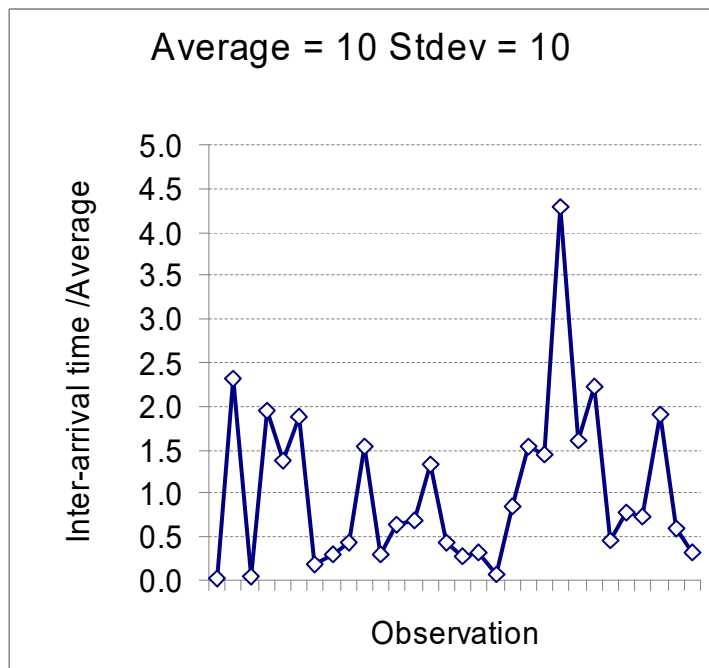
Relative and absolute variability

- The standard deviation is an **absolute** measure of variability.
- Two processes can have the same standard deviation but one can seem much more variable than the other:
 - Below are random samples from two processes that have the same standard deviation. The left one seems more variable.



Relative and absolute variability (continued)

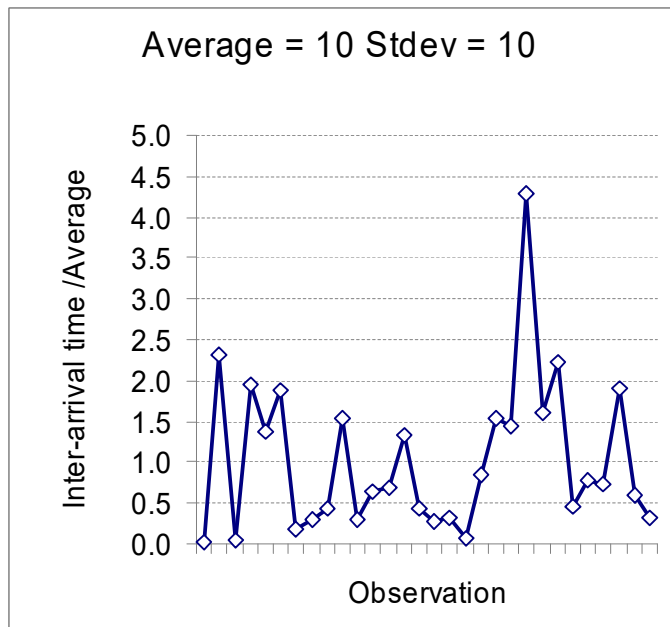
- The previous slide plotted the processes on two different axes.
- Here, the two are plotted **relative** to their average and with the same axes.
- Relative to their average, the one on the left is clearly more variable (sometimes more than 400% above the average or less than 25% of the average).



Coefficient of variation

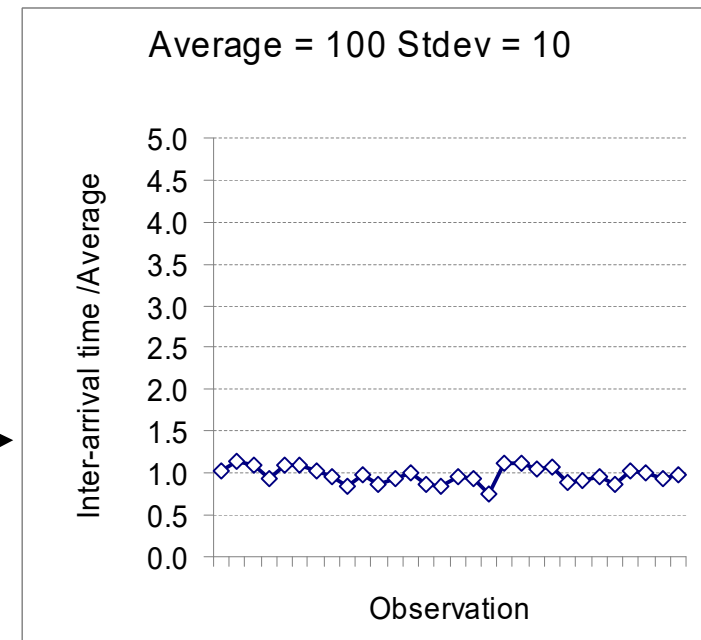
- The coefficient of variation is a measure of the relative variability of a process – it is the standard deviation divided by the average.
- The coefficient of variation of the arrival process:

$$CV_a = \frac{\text{Standard deviation of interarrival time}}{\text{Average interarrival time}}$$



← $CV_a = 10 / 10 = 1$

$CV_a = 10 / 100$
 $= 0.1$

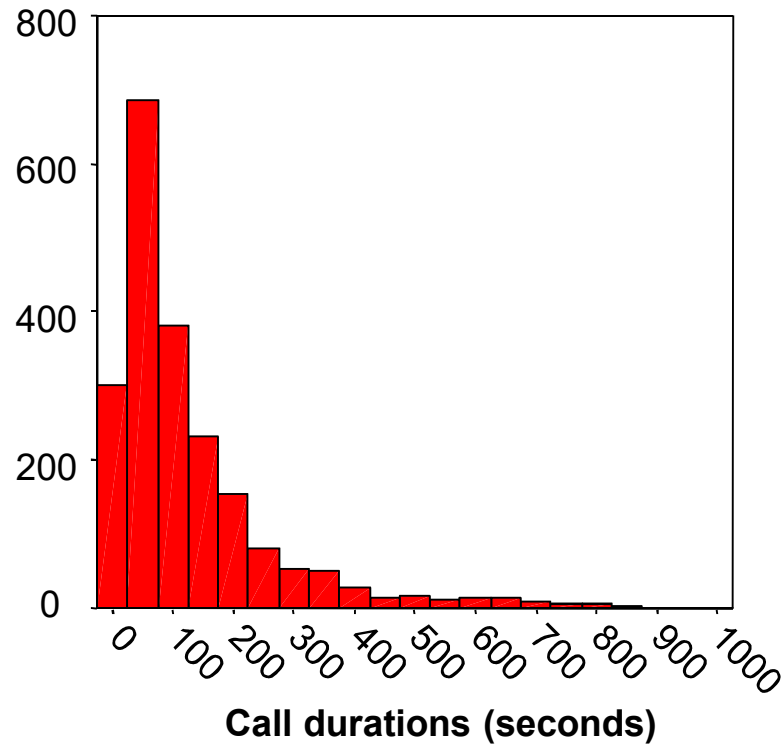


Service time variability

- The coefficient of variation of the service time process:

$$CV_p = \frac{\text{Standard deviation of activity times}}{\text{Average activity time}}$$

Frequency

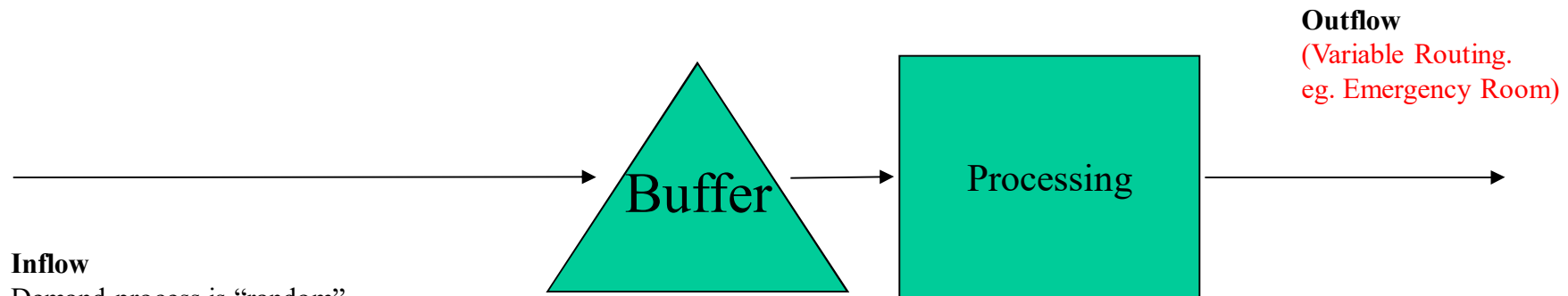


- Standard deviation of activity times = 150 seconds.
- Average call time (i.e., activity time) = 120 seconds.
- $CV_p = 150 / 120 = 1.25$

Modeling Variability in Flow

Flow Rate

$$\text{Minimum}\{\text{Demand, Capacity}\} = \text{Demand} = 1/a$$

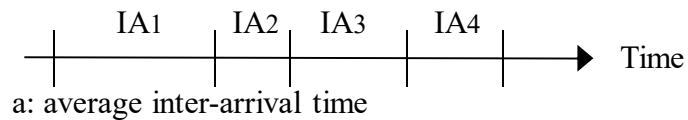


Inflow

Demand process is “random”

(Randomness is the Rule, not the Exception!)

Look at the inter-arrival times



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

Often Poisson distributed:

$$CV_a = 1$$

Constant hazard rate (no memory)

Exponential inter-arrivals

Difference between seasonality and variability

Processing

(Inherent Variation, Machine Breakdown, Operator Absence)

p: average processing time

Same as “activity time” and “service time”

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

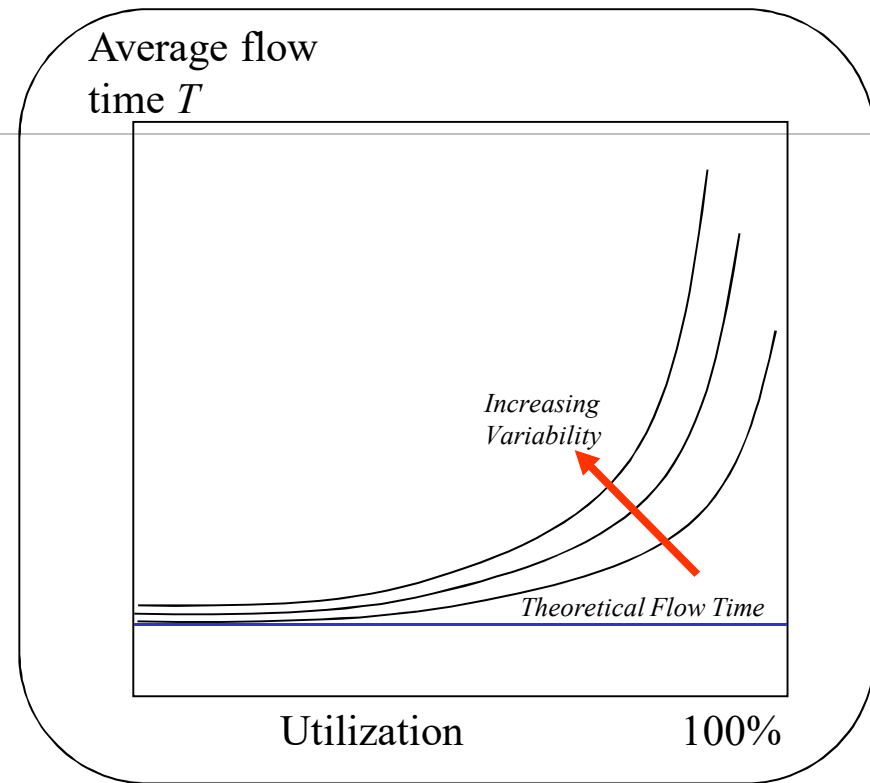
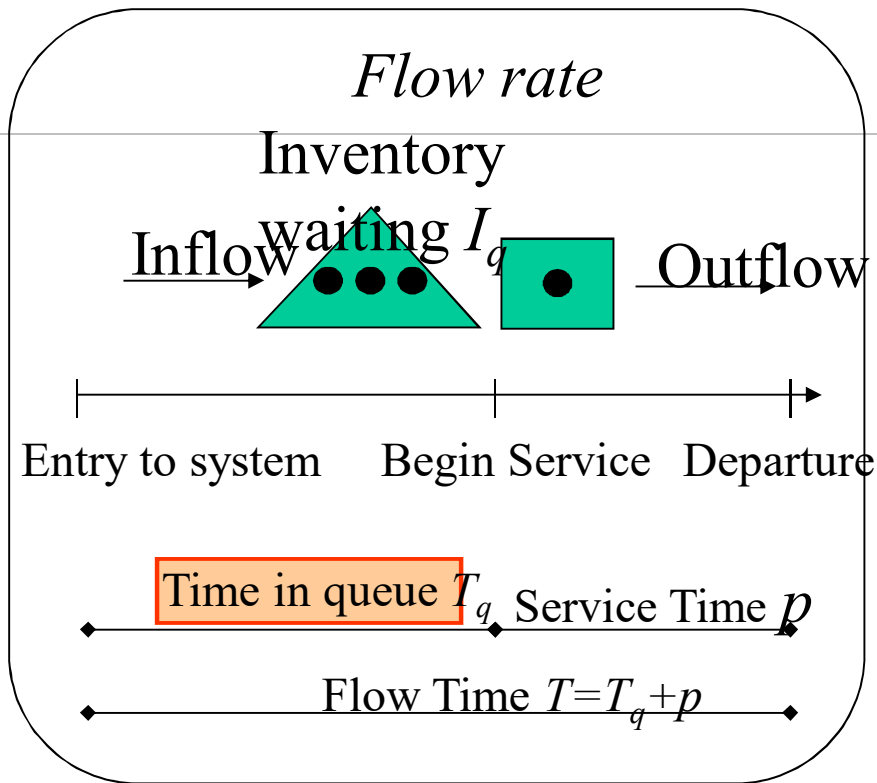
Can have many distributions:

CV_p depends strongly on standardization

Often Beta or LogNormal

CV \Rightarrow measure variability in relative terms

The Waiting Time Formula



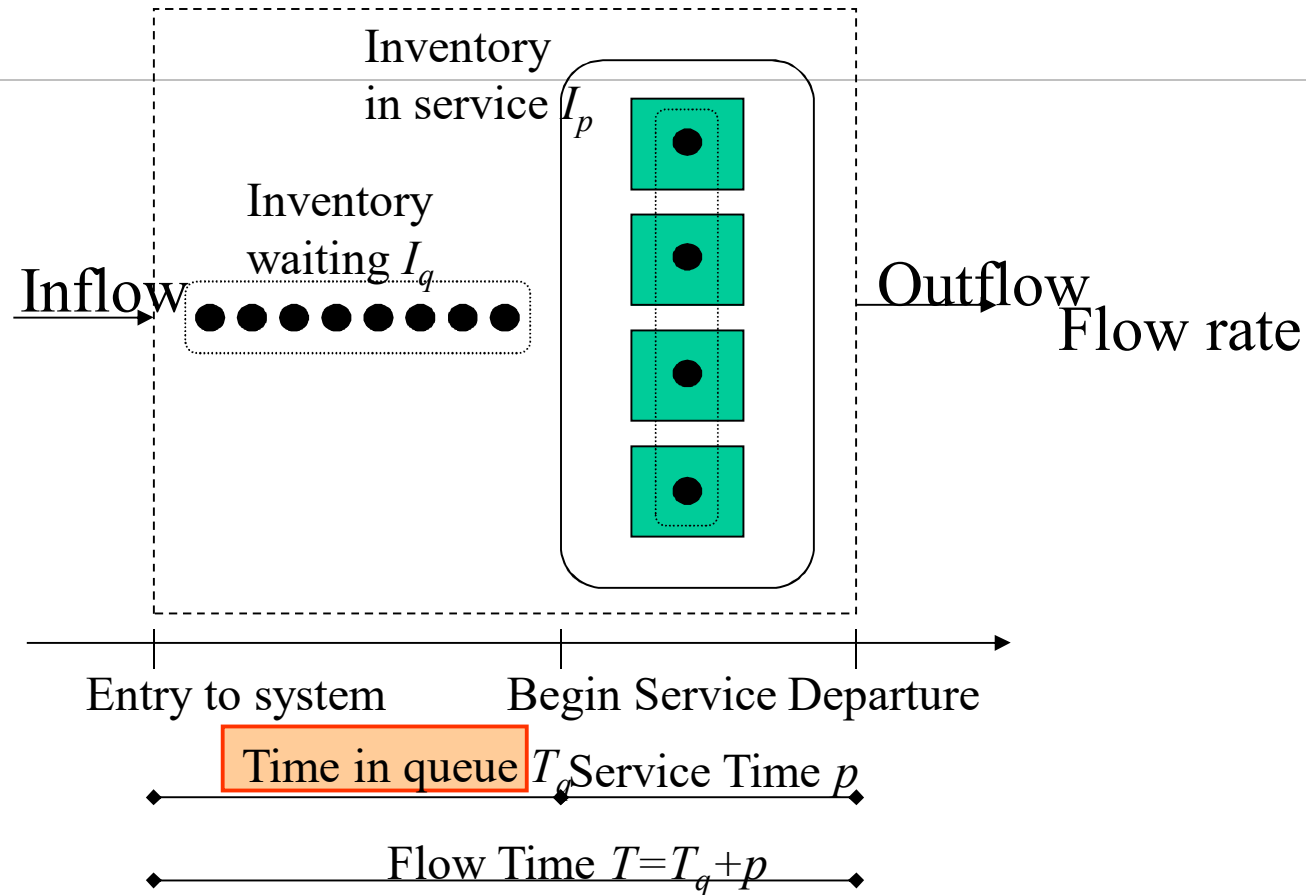
Waiting Time Formula (복잡한 공식은 외울 필요 없음)

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

└─── Service time factor
 └─── Utilization factor
 └─── Variability factor



Waiting Time Formula for Multiple, Parallel Resources

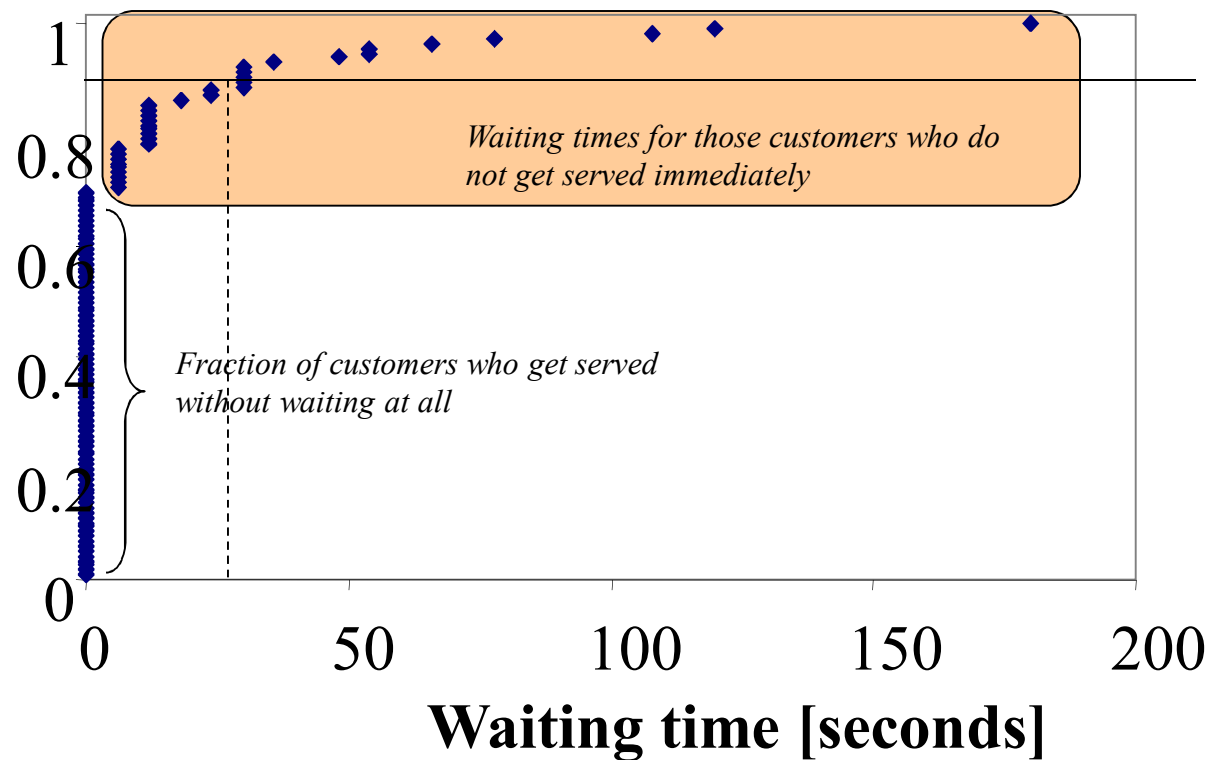


Waiting Time Formula for Multiple (m) Servers

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

Service Levels in Waiting Systems

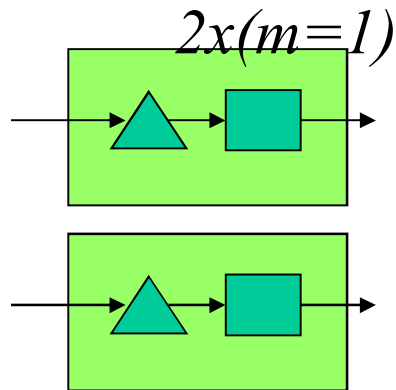
Fraction of customers who have to wait x seconds or less



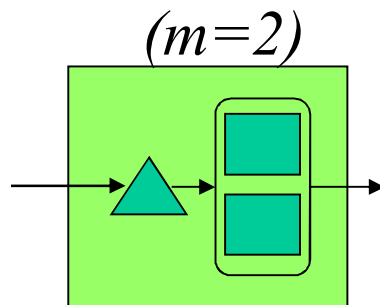
- Target Wait Time (TWT)
- Service Level = Probability{Waiting Time \leq TWT}
- Example: Deutsche Bundesbahn Call Center
 - Year 2020: 30% of calls answered within 20 seconds
 - Target: 80% of calls answered within 20 seconds

Managerial Responses to Variability: Pooling

Independent Resources



Pooled Resources



$$a=4\text{min}, p=3\text{min} \Rightarrow a=2\text{min}, p=3/2\text{min}$$

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

$$= 3 * \left(\frac{0.75}{1 - 0.75} \right) * \left(\frac{1 + 1}{2} \right) = 9 \text{ min}$$

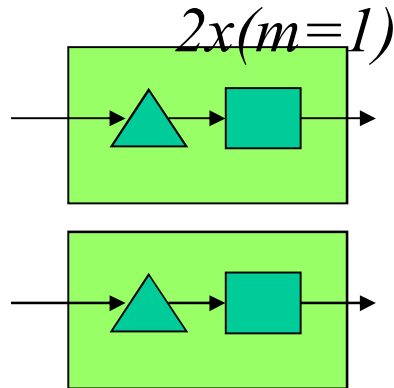
$$Tq = \left(\frac{\text{Activity time}}{m} \right) * \left(\frac{\text{utilization}^{\sqrt{2(m+1)-1}}}{1 - \text{utilization}} \right) * \left(\frac{CV_a^2 + CV_p^2}{2} \right)$$

$$= \left(\frac{3}{2} \right) * \left(\frac{0.75^{\sqrt{2(2+1)-1}}}{1 - 0.75} \right) * \left(\frac{1 + 1}{2} \right) = 3.95 \text{ min}$$

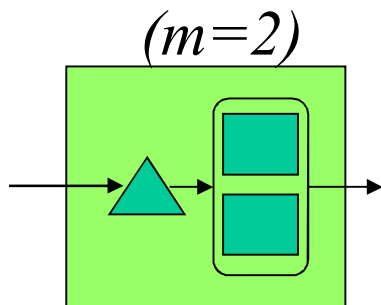
Can serve the same number of customers using the same processing time, but in only half the waiting time!

Managerial Responses to Variability: Pooling

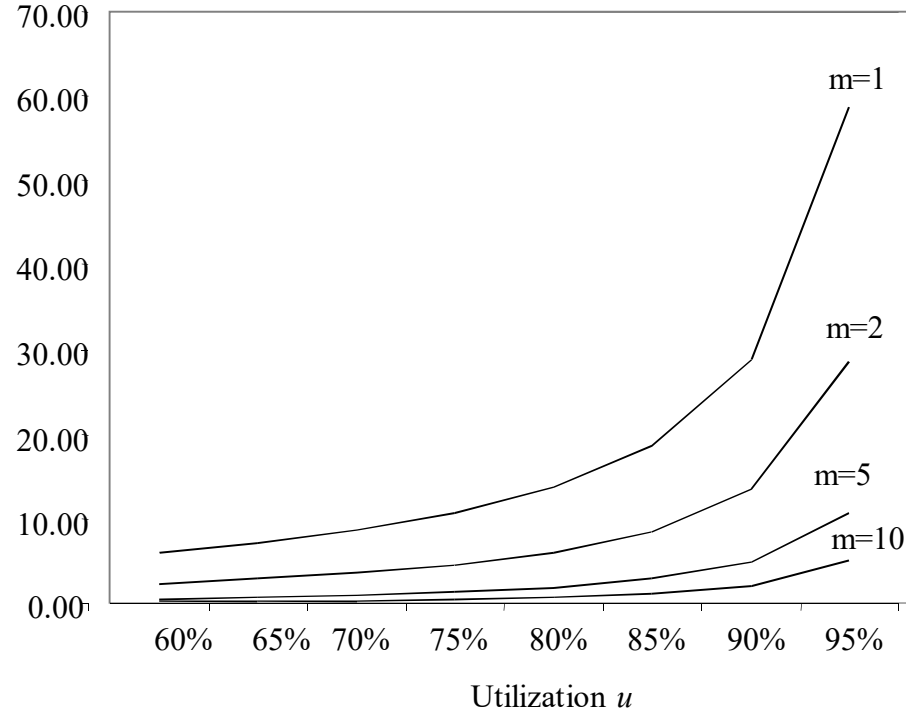
Independent Resources



Pooled Resources



Waiting
Time T_q



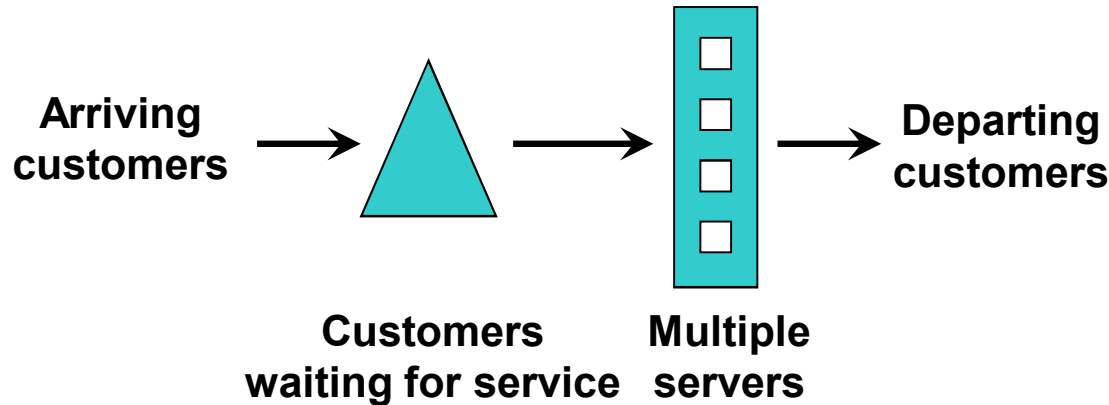
Implications:

+ balanced utilization

Pooling prevents the case that one resource is idle while the other faces a backlog of work.

⇒ 서비스시스템의 번호표 배포

A multi-server queue

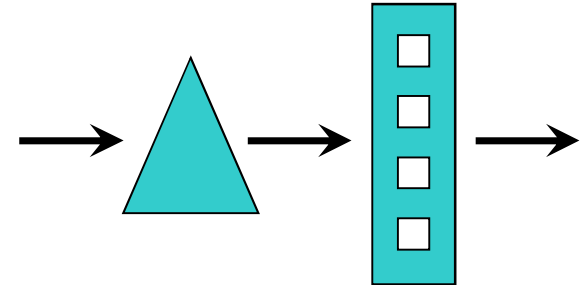


- Assumptions:
 - All servers are equally skilled, i.e., they all take p time to process each customer.
 - Each customer is served by only one server.
 - Customers wait until their service is completed.
 - There is sufficient capacity to serve all demand.
- a = average interarrival time
- Flow rate = $1 / a$
- p = average activity time
- Capacity of each server = $1 / p$
- m = number of servers
- Capacity = $m \times 1 / p$
- Utilization = Flow rate / Capacity
 - = $(1 / a) / (m \times 1 / p)$
 - = $p / (a \times m)$

A multi-server queue – some data and analysis

- Suppose:

- $a = 35$ seconds per customer
- Flow rate = $1 / 35$ customers per second
- $p = 120$ seconds per customer
- Capacity of each server = $1 / 120$ customers per second
- $m =$ number of servers = 4

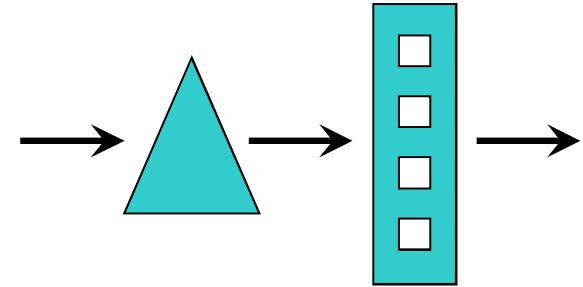


- Then

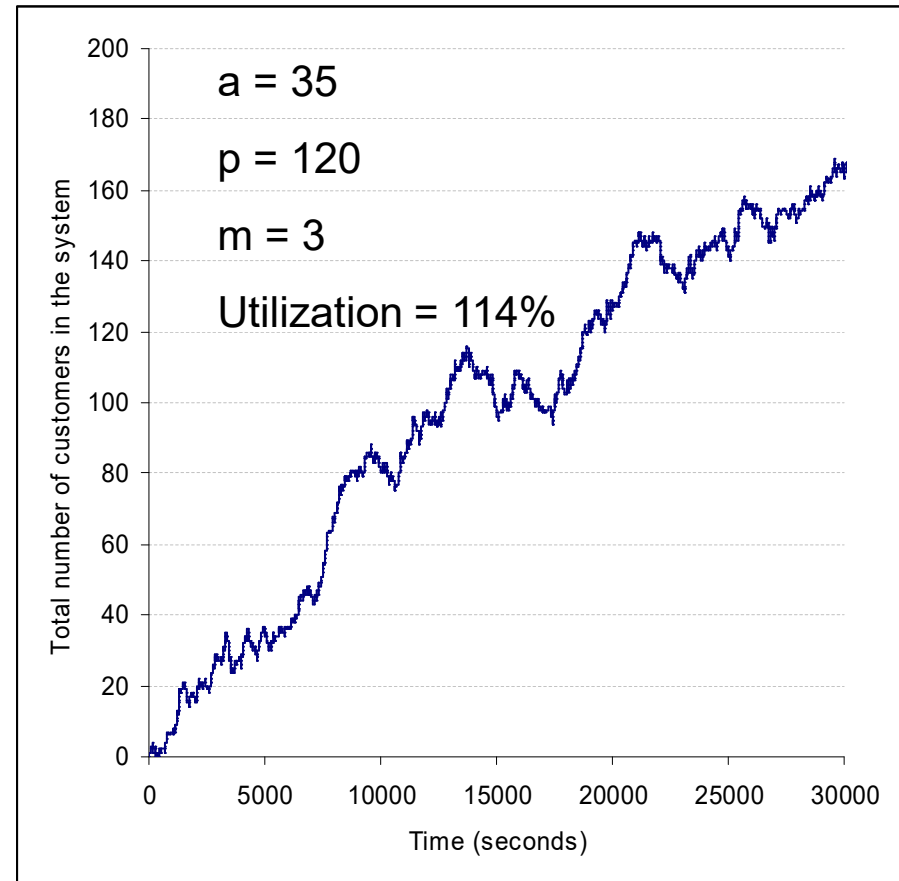
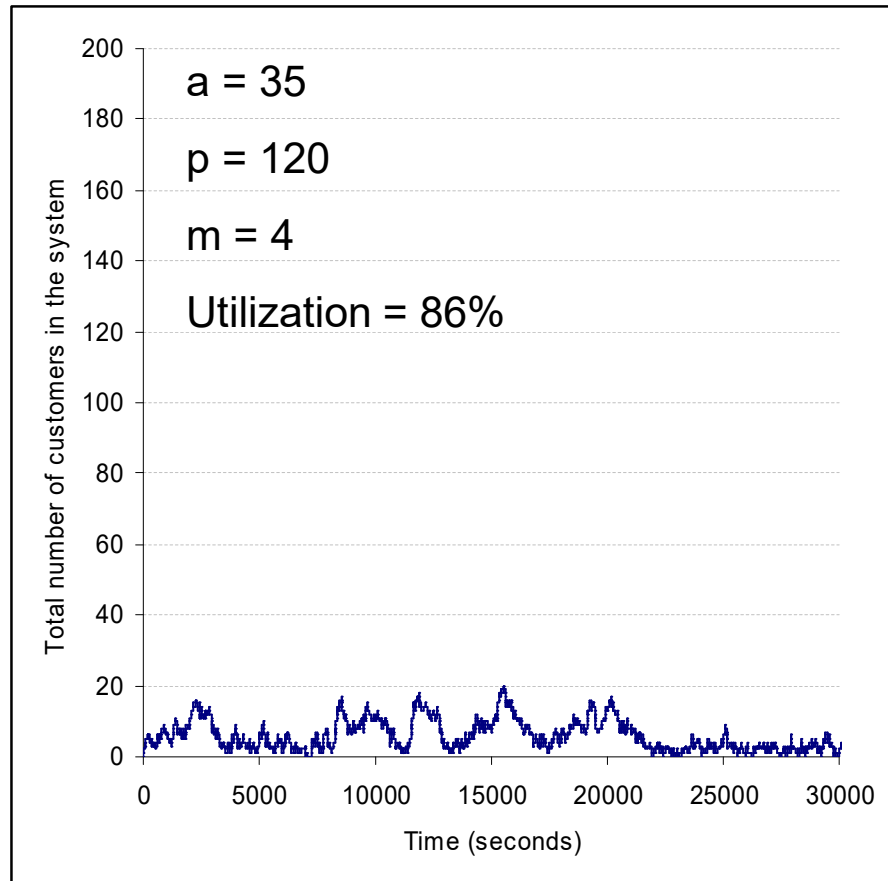
- Capacity = $m \times 1 / p = 4 \times 1 / 120 = 1 / 30$ customers per second
- Utilization = Flow rate / Capacity
 - = $p / (a \times m)$
 - = $120 / (35 \times 4)$
 - = 85.7%

The implications of utilization

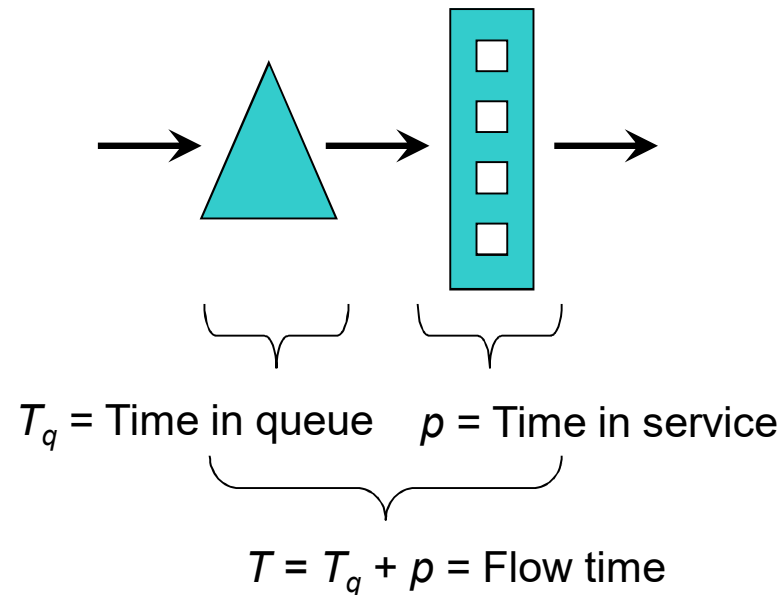
- A 85.7% utilization means:
 - At any given moment, there is a 85.7% chance a server is busy serving a customer and a $1 - 0.857 = 14.3\%$ chance the server is idle.
 - At any given moment, on average 85.7% of the servers are busy serving customers and $1 - 0.857 = 14.3\%$ are idle.
- If utilization $< 100\%$ then:
 - The system is **stable** which means that the size of the queue does not keep growing over time. (Capacity is sufficient to serve all demand.)
- If utilization $\geq 100\%$ then:
 - The system is **unstable**, which means that the size of the queue will continue to grow over time.



Stable and unstable queues



Time spent in the two stages of the system



Recall:

$p = \text{Activity time}$

$m = \text{number of servers}$

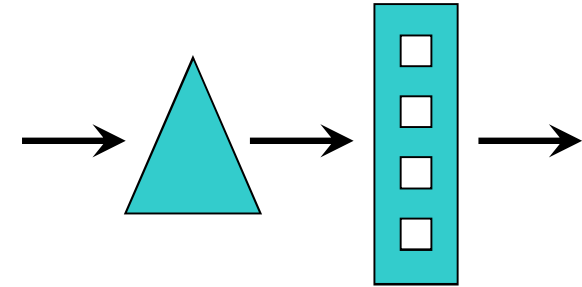
$$\text{Time in queue} = \left(\frac{\text{Activity time}}{m} \right) \times \left(\frac{\text{Utilization}^{\sqrt{2(m+1)}-1}}{1 - \text{Utilization}} \right) \times \left(\frac{CV_a^2 + CV_p^2}{2} \right)$$

- The above Time in queue equation works only for a stable system, i.e., a system with utilization less than 100%

What determines time in queue in a stable system?

The **capacity factor**:

Average processing time of the system = (p/m) .
Demand does not influence this factor.



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

The **utilization factor**:

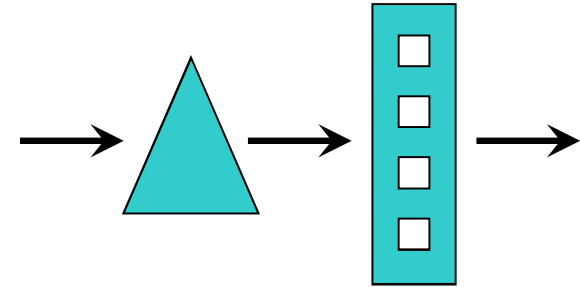
Average demand influences this factor because utilization is the ratio of demand to capacity.

The **variability factor**:

This is how variability influences time in queue – the more variability (holding average demand and capacity constant) the more time in queue.

Evaluating Time in queue

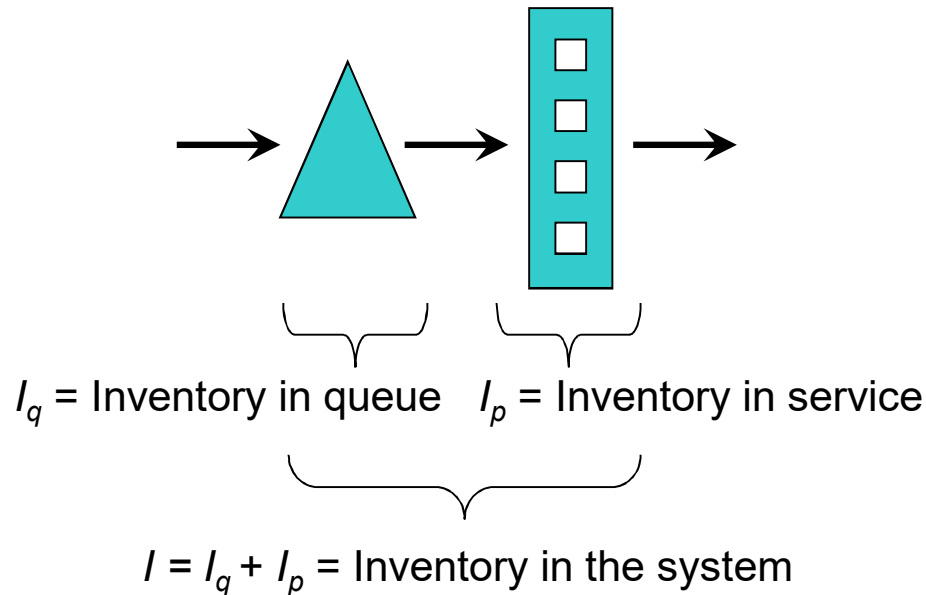
- Suppose:
 - $a = 35$ seconds, $p = 120$ seconds, $m = 4$
 - Utilization = $p / (a \times m) = 85.7\%$
 - $CV_a = 1, CV_p = 1$



$$\begin{aligned} \text{Time in queue} &= \left(\frac{\text{Activity time}}{m} \right) \times \left(\frac{\text{Utilization}^{\sqrt{2(m+1)-1}}}{1 - \text{Utilization}} \right) \times \left(\frac{CV_a^2 + CV_p^2}{2} \right) \\ &= \left(\frac{120}{4} \right) \times \left(\frac{0.857^{\sqrt{2(4+1)-1}}}{1 - 0.857} \right) \times \left(\frac{1^2 + 1^2}{2} \right) = 150 \end{aligned}$$

- So Flow time = $150 + 120 = 270$ seconds.
 - In other words, the average customer will spend 270 seconds in the system (waiting for service plus time in service)

How many customers are in the system?

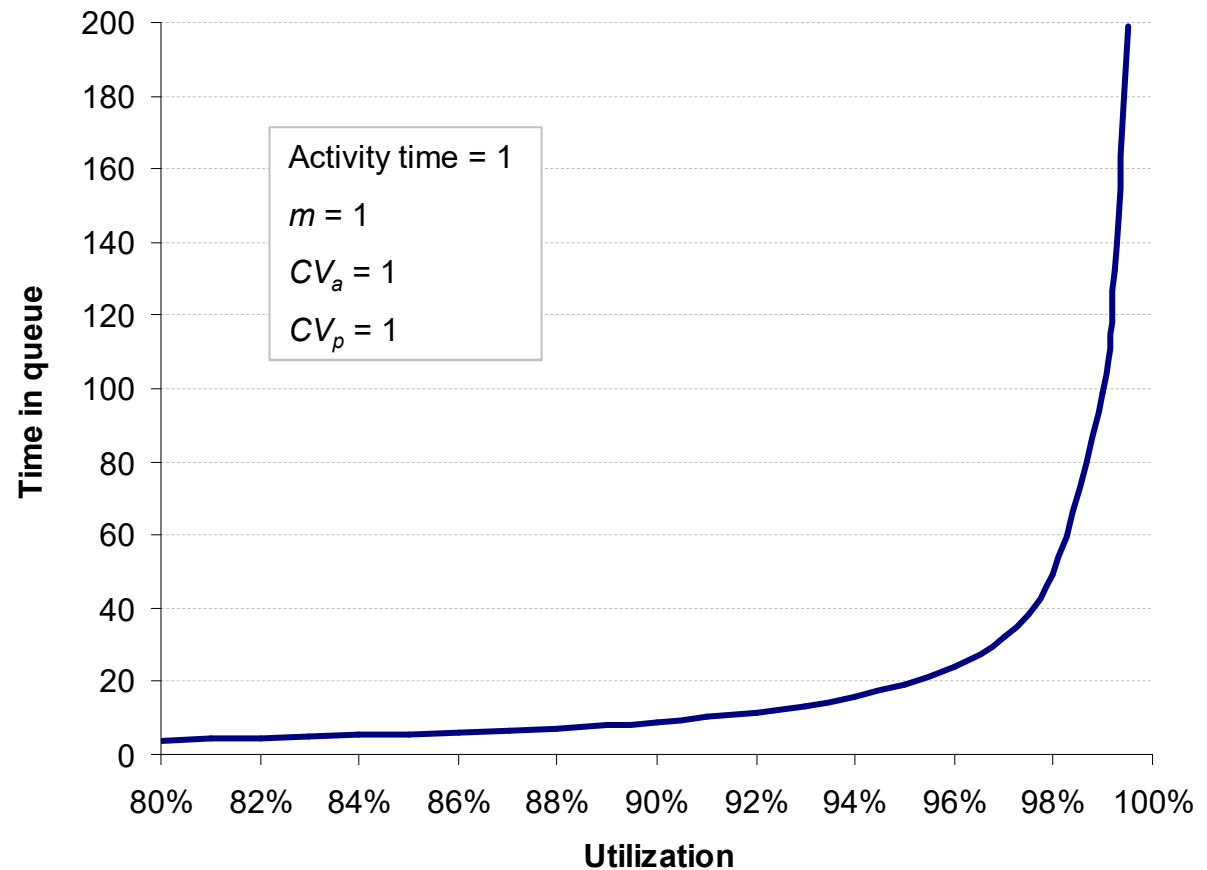


- Use Little's Law, $I = R \times T$
- $R = \text{Flow Rate} = (1/a)$
 - The flow rate through the system equals the demand rate because we are demand constrained (utilization is less than 100%)
- $I_q = (1/a) \times T_q = T_q / a$
- $I_p = (1/a) \times p = p / a$
 - Note, time in service does not depend on the number of servers because when a customer is in service they are processed by only one server no matter how many servers are in the system.

Utilization and system performance

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

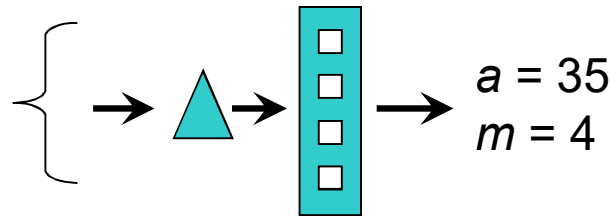
- Time in queue increases dramatically as the utilization approaches 100%



Which system is more effective?

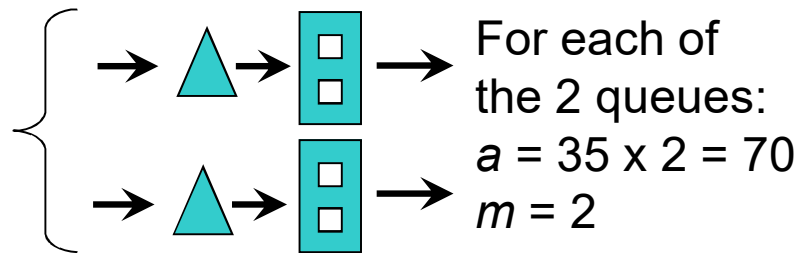
Pooled system:

One queue, four servers



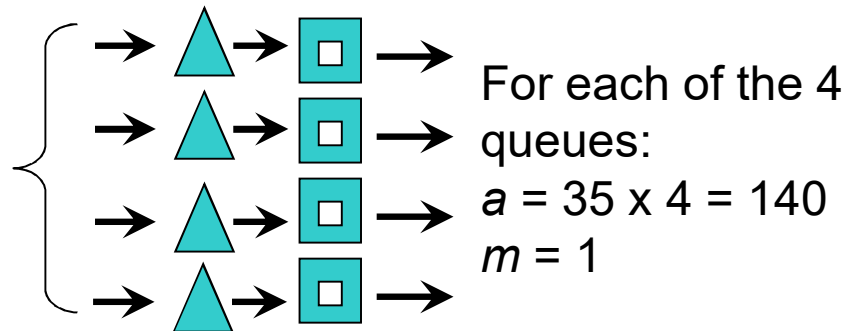
Partially pooled system:

Two queues, two servers with each queue.



Separate queue system:

Four queues, one server with each queue.



Across these three types of systems:

Variability is the same:
 $CV_a = 1, CV_p = 1$

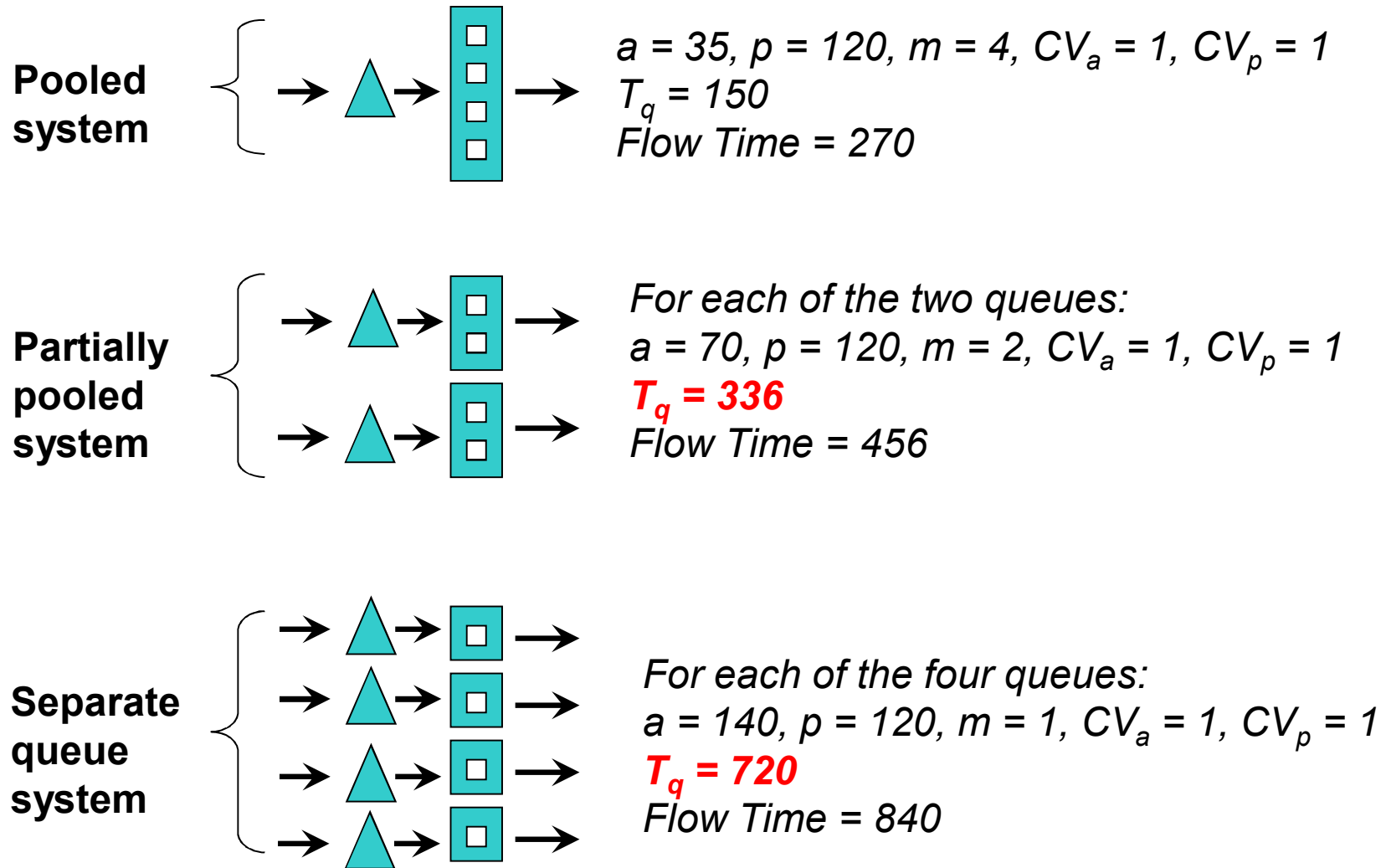
Total demand is the same: 1/35 customers per second.

Activity time is the same: $p = 120$

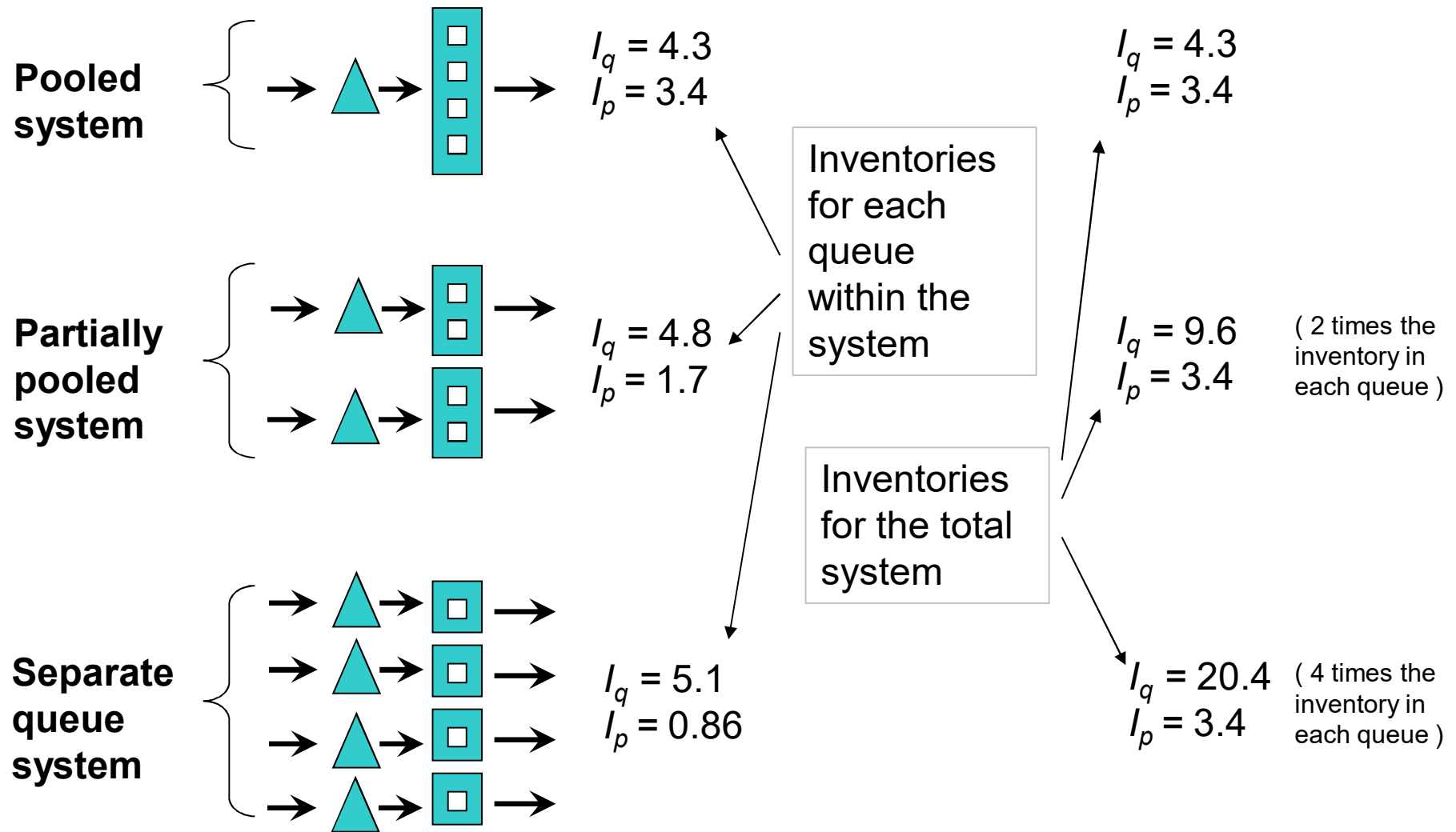
Utilization is the same: $p / a \times m = 85.7\%$

The probability a server is busy is the same = 0.857

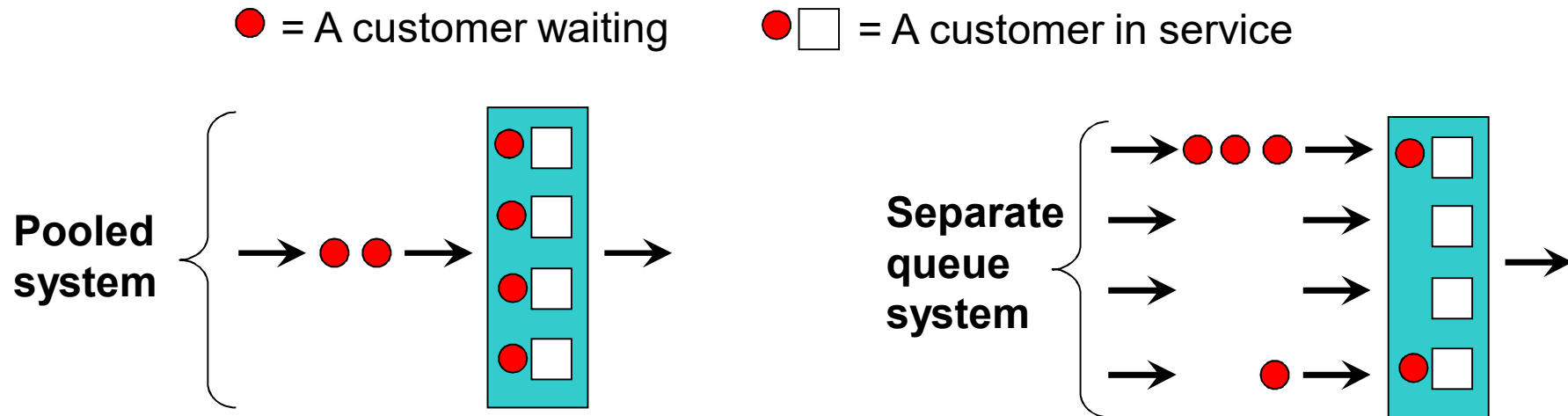
Pooling can reduce Time in queue considerably



Pooling reduces the number of customers waiting but not the number in service



Why does pooling work?



- Both systems currently have 6 customers.
- In the pooled system, all servers are busy and only 2 customers are waiting.
- In the separate queue system, 2 servers are idle and 4 customers are waiting.
- The separate queue system is inefficient because there can be customers waiting and idle servers at the same time.

Some quick service restaurants pool drive through ordering

- The person saying “Can I take your order?” may be hundreds (or even thousands) of miles away:
 - Pooling the order taking process can improve time-in-queue while requiring less labor.
 - It has been shown that queue length at the drive through influences demand – people don’t stop if the queue is long.
 - However, this system incurs additional communication and software costs.

Limitations to pooling

- Pooling may require workers to have a broader set of skills, which may require more training and higher wages:
 - Imagine a call center that took orders for McDonalds and Wendys ... now the order takers need to be experts in two menus.
 - Suppose cardiac surgeons need to be skilled at kidney transplants as well.

- Pooling may disrupt the customer – server relationship:
 - Patients like to see the same physician.

- Pooling may increase the time-in-queue for one customer class at the expense of another:
 - Removing priority security screening for first-class passengers may decrease the average time-in-queue for all passengers but will likely increase it for first-class passengers.

Why the long queues for the women's restroom?

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

p = Activity time

$$u = \text{Utilization} = \frac{p}{a \times m}$$



Women wait for a public restroom at Yankee Stadium in 2014

Potty-parity Laws

- Women's Restroom Equity Bill:
 - Passed NY City Counsel May 2005.
 - Requires women's restrooms to have twice the flushing capacity of men's restrooms (at least a 2:1 ratio of women's toilets to men's toilets & urinals)
 - In the context of Time in queue equation, this law requires m to increase for women, which decreases the capacity factor (p/m), and decreases utilization.

- Other solutions:
 - Unisex restrooms (pools capacity, at least the toilets) ⇒ **Not in Korea**
 - Flexible partitions that can change the size of each restroom.

$$Time\ in\ queue = \left(\frac{p}{m}\right) \times \left(\frac{u^{\sqrt{2(m+1)}-1}}{1-u}\right) \times \left(\frac{CV_a^2 + CV_p^2}{2}\right)$$

$p = Activity\ time$

$$u = Utilization = \frac{p}{a \times m}$$

Managerial Responses to Variability: Priority Rules in Waiting Time Systems

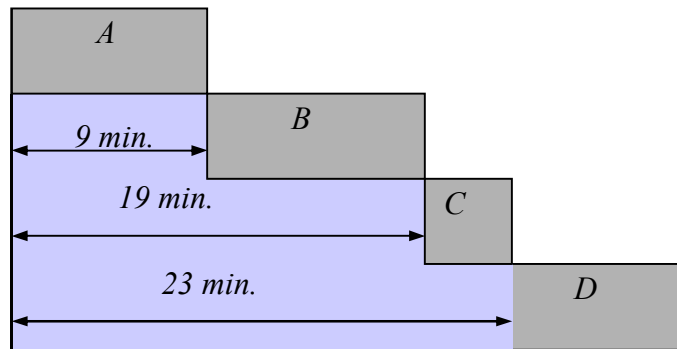
Service times:

A: 9 minutes

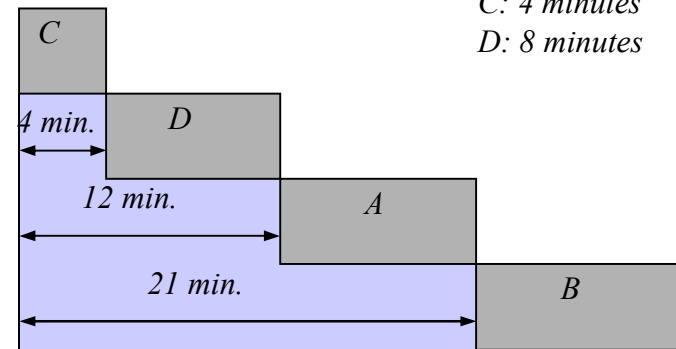
B: 10 minutes

C: 4 minutes

D: 8 minutes



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



Total wait time: $4+12+21=37$ min

- Flow units are sequenced in the waiting area (triage step)
- **Shortest Processing Time (SPT) Rule**
 - Minimizes average waiting time
 - Problem of having “true” processing times (Can I ask a quick question?)
- **First-Come, First-Served (FCFS)**
 - easy to implement
 - perceived fairness
- Sequence based on importance
 - emergency cases
 - identifying profitable flow units

Reducing Variability

Variability is the enemy of all operations.

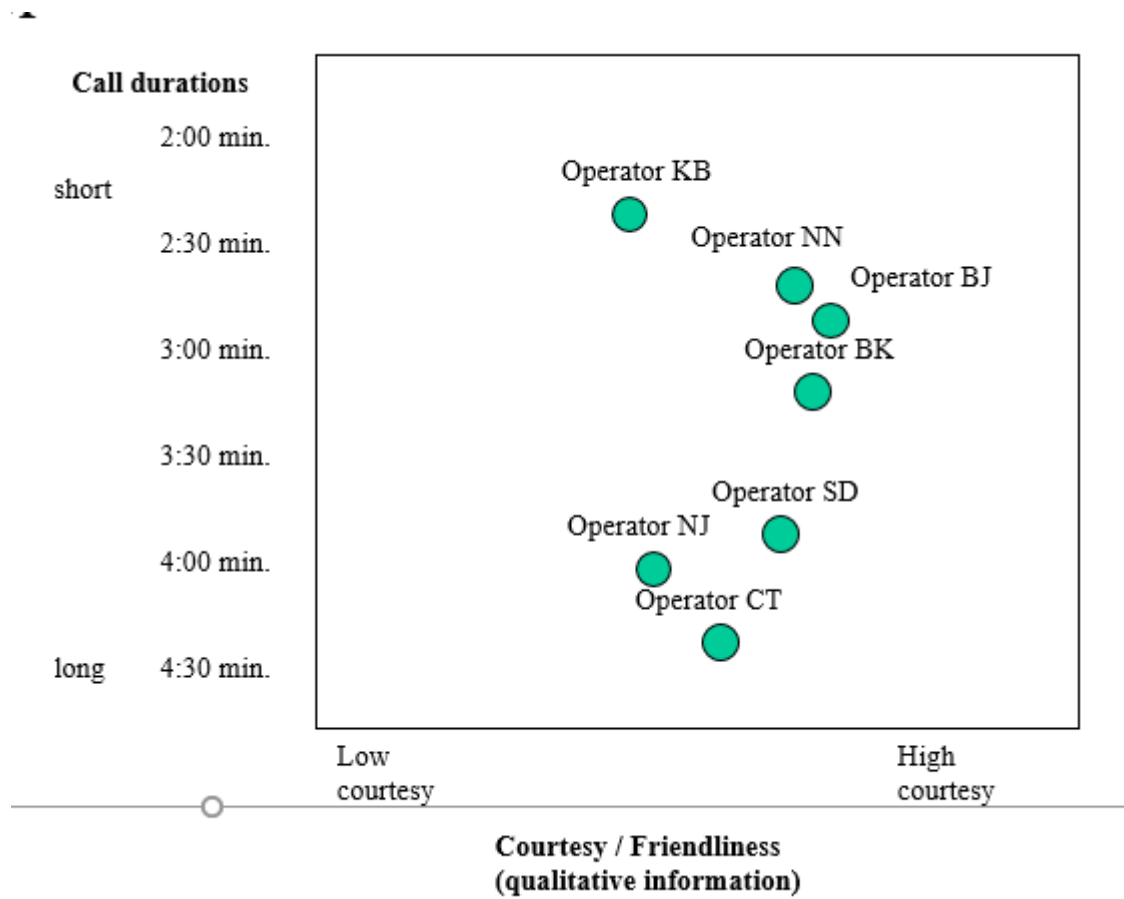
♣ Ways to reduce arrival variability

Appointment systems

- Do not eliminate arrival variability
- If the doctor has the right to be late, why not the patient?
- What portion of the available capacity should be reserved in advance? (Ch. 18 Revenue Management)
- Do not reduce the variability of the true underlying demand
→ providing incentives for customers to avoid peak hours
(Demand Management)

Reducing Variability

- ♣ Ways to reduce processing time variability
 - Find a balance between operational efficiency (e.g. call duration) and the quality of service (perceived courtesy)
 - Investment in training and technology
(eg. Operators received real-time instructions (scripting))



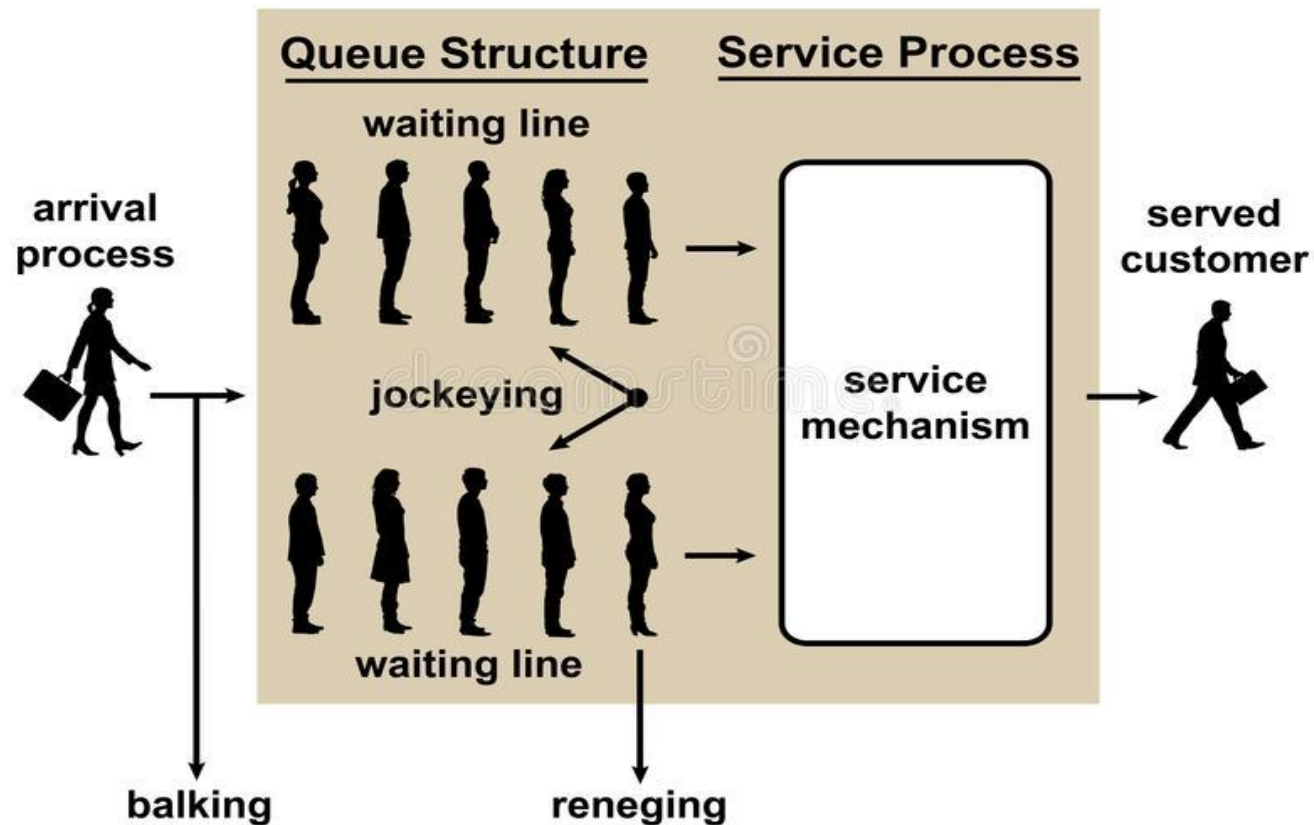
Managing Waiting Systems: Points to Remember

- Variability is the norm, not the exception
- Variability leads to waiting times although utilization $< 100\%$
- Use the Waiting Time Formula to
 - get a qualitative feeling of the system
 - analyze specific recommendations / scenarios
- Managerial response to variability:
 - understand where it comes from and eliminate what you can
 - accommodate the rest by holding excess capacity
- Difference between variability and seasonality
 - seasonality is addressed by staffing to (expected) demand

Summary

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

QUEUEING SYSTEM



-
- (Remark) The utilization factor is nonlinear.
 - Utilization=0.8 $\rightarrow 0.8/(1-0.8)=4$
Utilization=0.9 $\rightarrow 0.9/(1-0.9)=9$
Utilization=0.95 $\rightarrow 0.95/(1-0.95)=19$
Utilization=0.99 $\rightarrow 0.99/(1-0.99)=99$

