

Waiting Times

Chapter 7

Learning Objectives

- ◆ Interarrival and Service Times and their variability
- ◆ Obtaining the average time spent in the queue
- ◆ Pooling of server capacities
- ◆ Priority rules

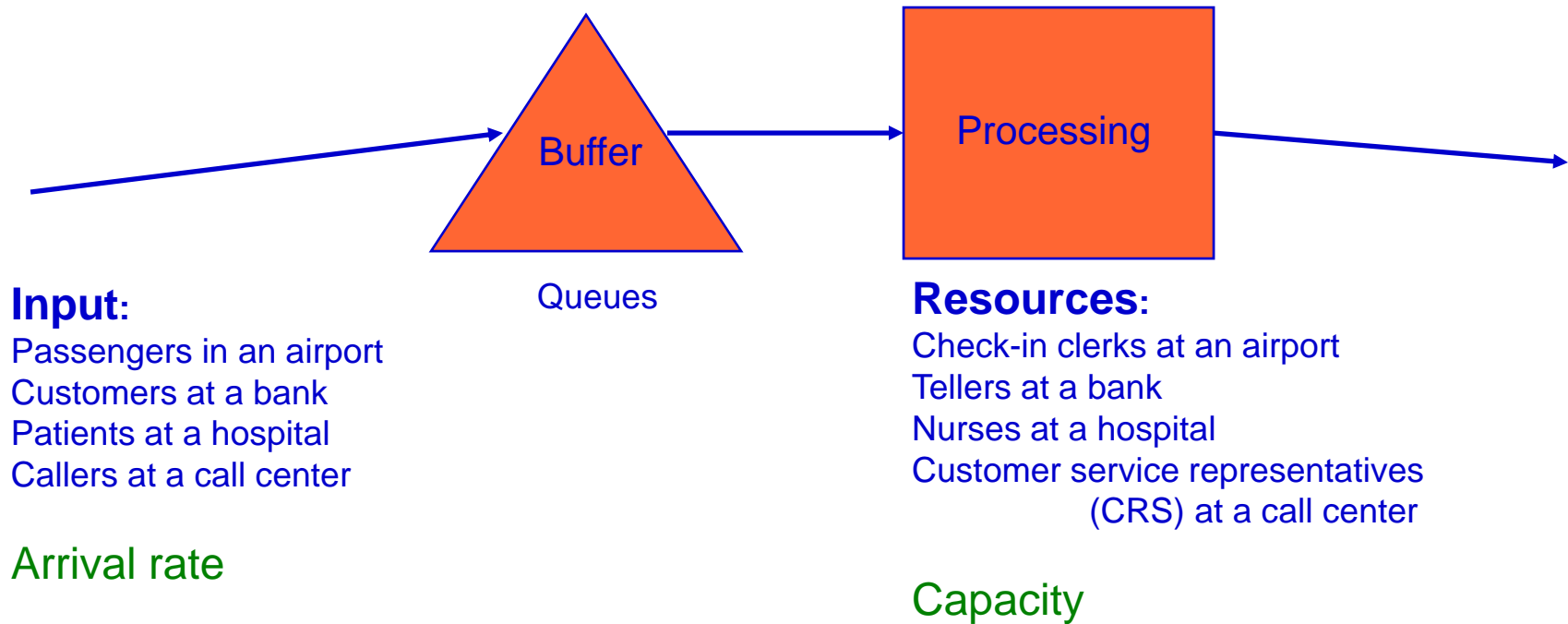
Where are the queues?



Americans spend > 100 M hours/day waiting in a line.
T. Heymann in his book "On an average day"

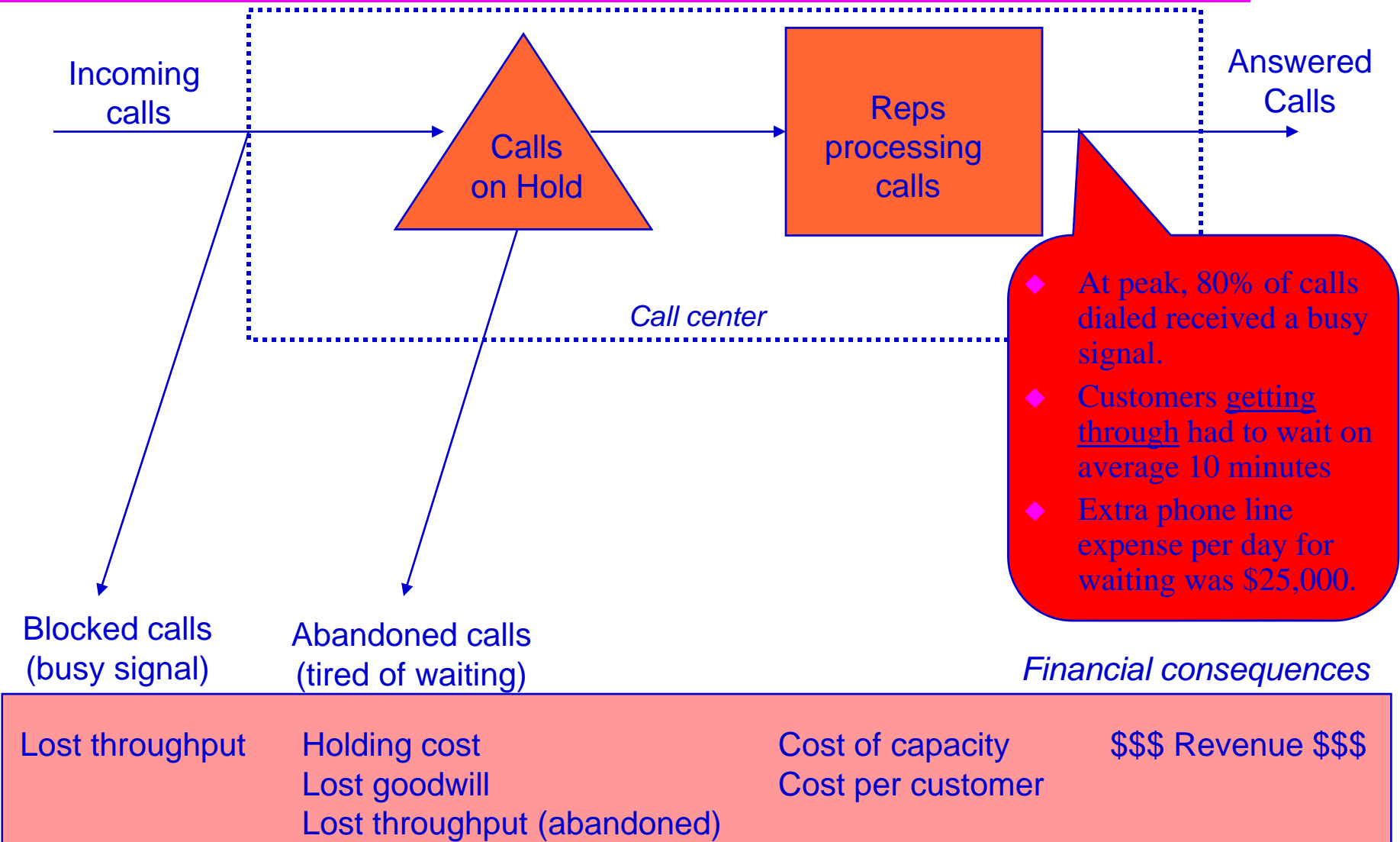


A Queue is made of a server and a queue in front



We are interested in the waiting times in the queue and the queue length.

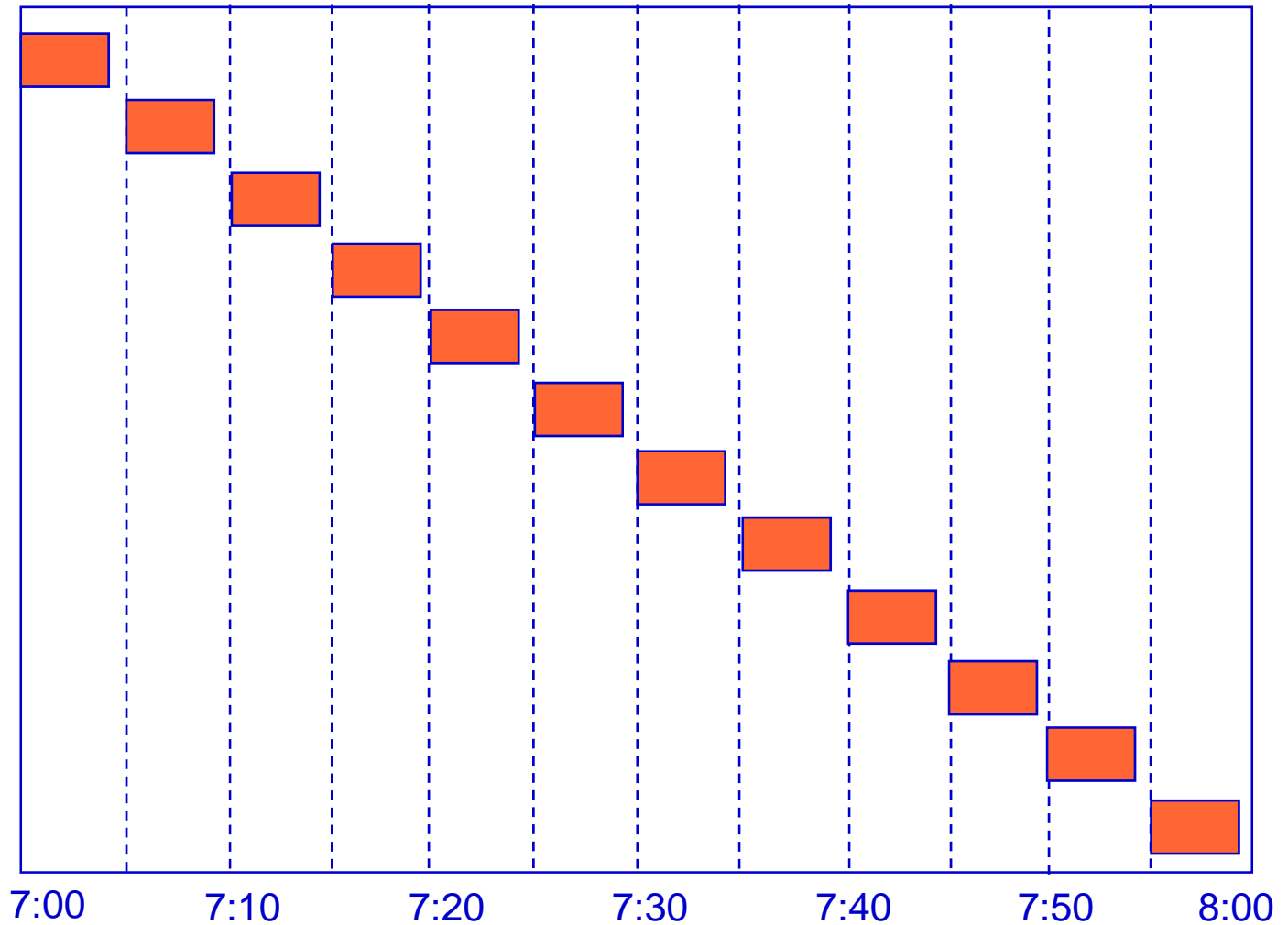
An Example of a Simple Queuing System



A Somewhat Odd Service Process

Constant Arrival Rate (0.2/min) and Service Times (4 min)

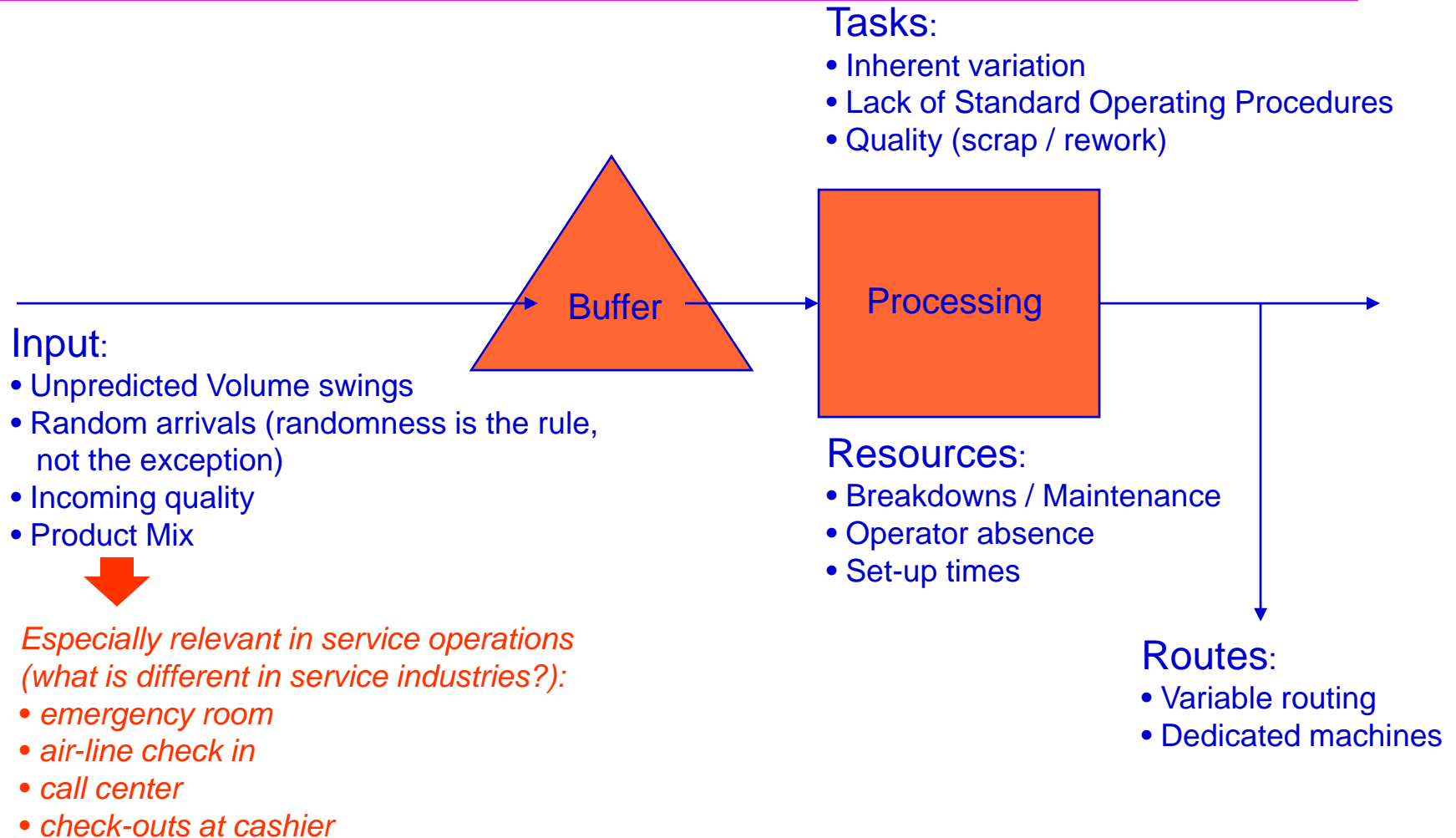
Patient	Arrival Time	Service Time
1	0	4
2	5	4
3	10	4
4	15	4
5	20	4
6	25	4
7	30	4
8	35	4
9	40	4
10	45	4
11	50	4
12	55	4



Where is Variability?

- ◆ There certainly is significant (actually infinite) amount waiting when the arrival rate is greater than the service rate
 - Equivalently, the processing capacity is less than the flow rate
- ◆ More interestingly, variability can cause long waiting times. Variability in
 - Arrival process
 - Processing times
 - Availability of resources
 - Routing of flow units; Recall the Resume Validation Example.

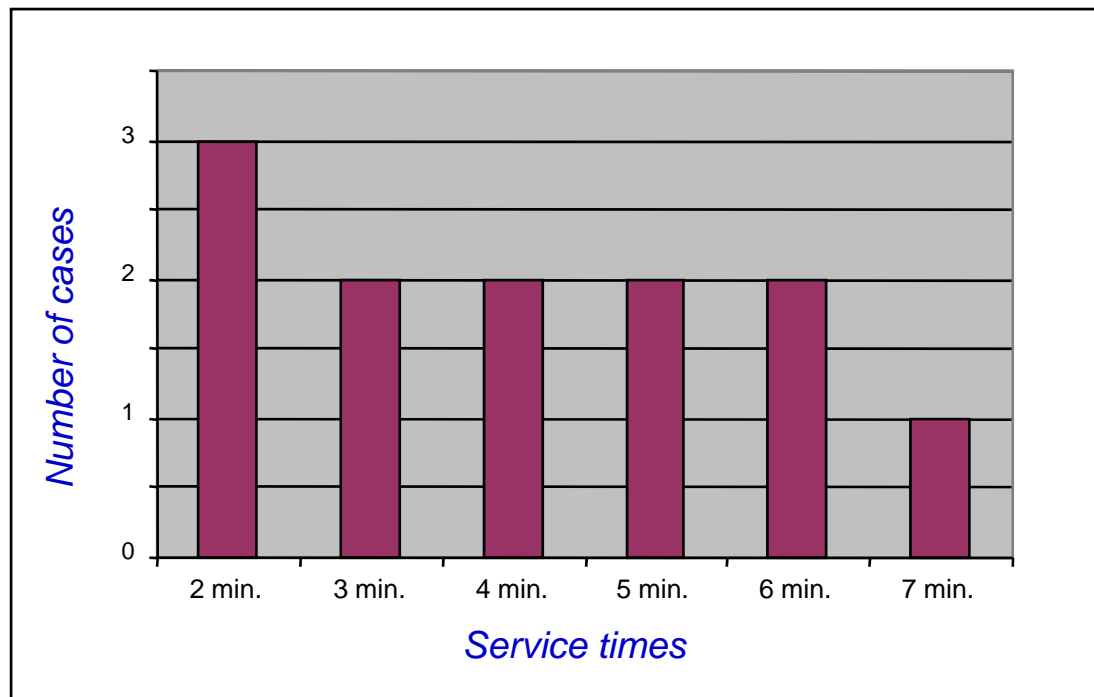
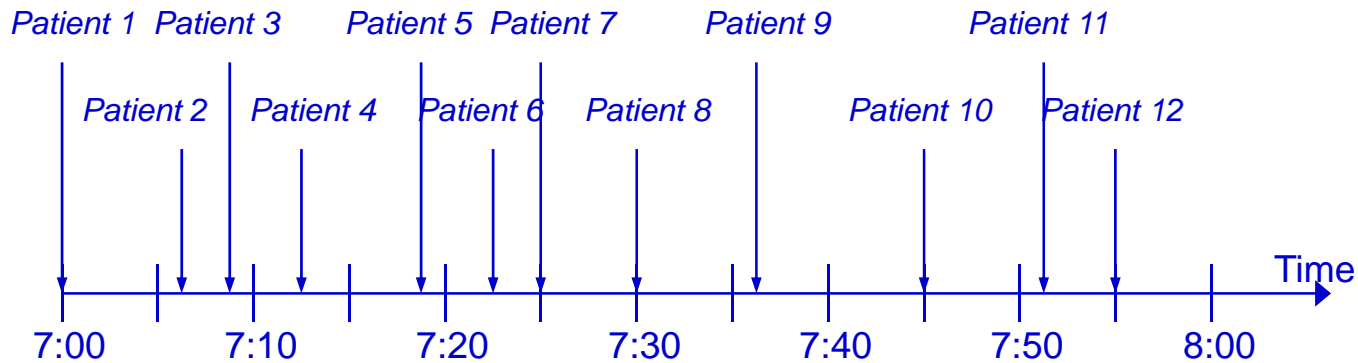
Variability: Where does it come from? Examples



A More Realistic Service Process

Random Arrival Rate and Service Times

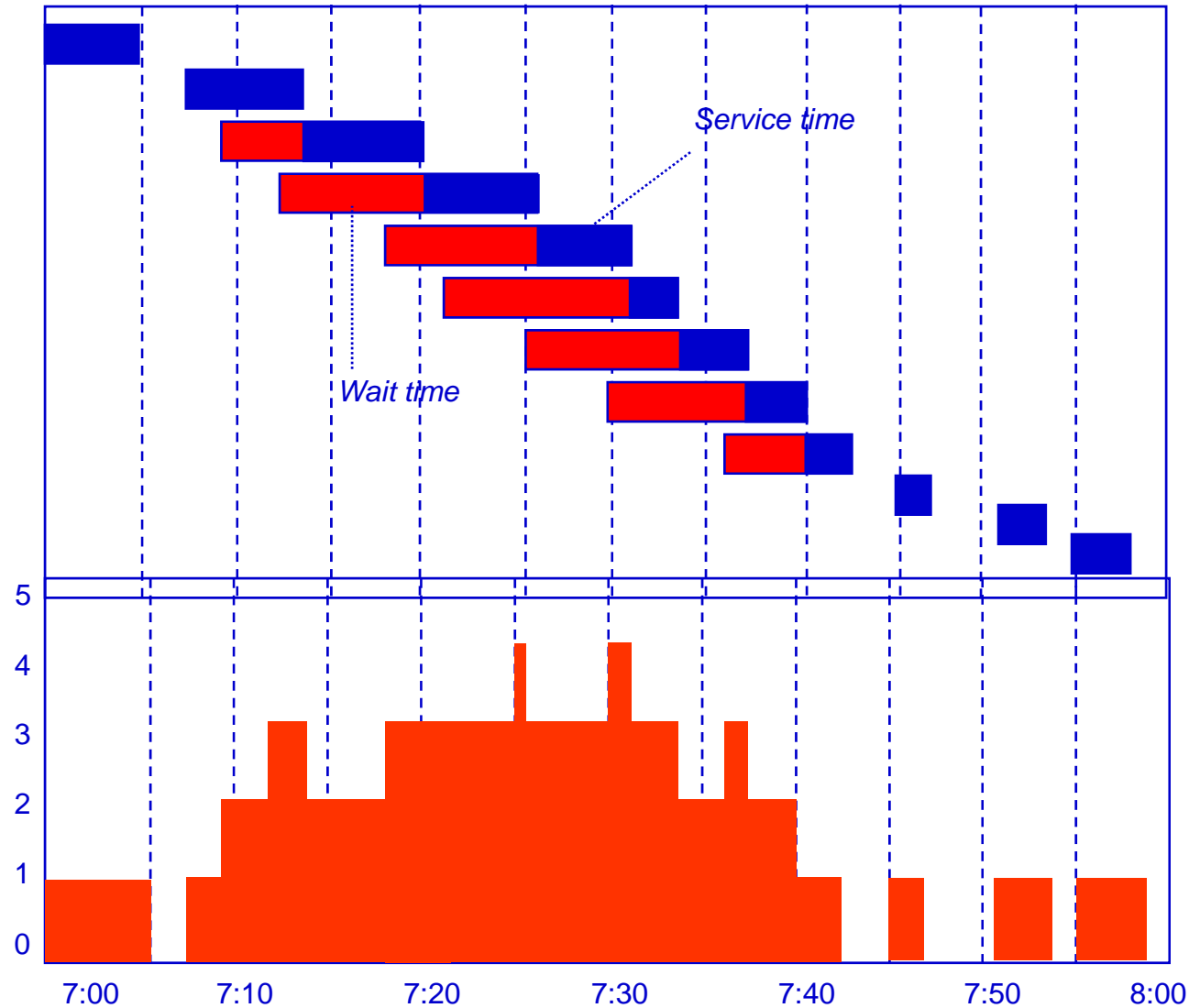
Patient	Arrival Time	Service Time
1	0	5
2	7	6
3	9	7
4	12	6
5	18	5
6	22	2
7	25	4
8	30	3
9	36	4
10	45	2
11	51	2
12	55	3



Variability Leads to Waiting Time

Average Arrival Rate (0.2/min) and Service Times (4 min)

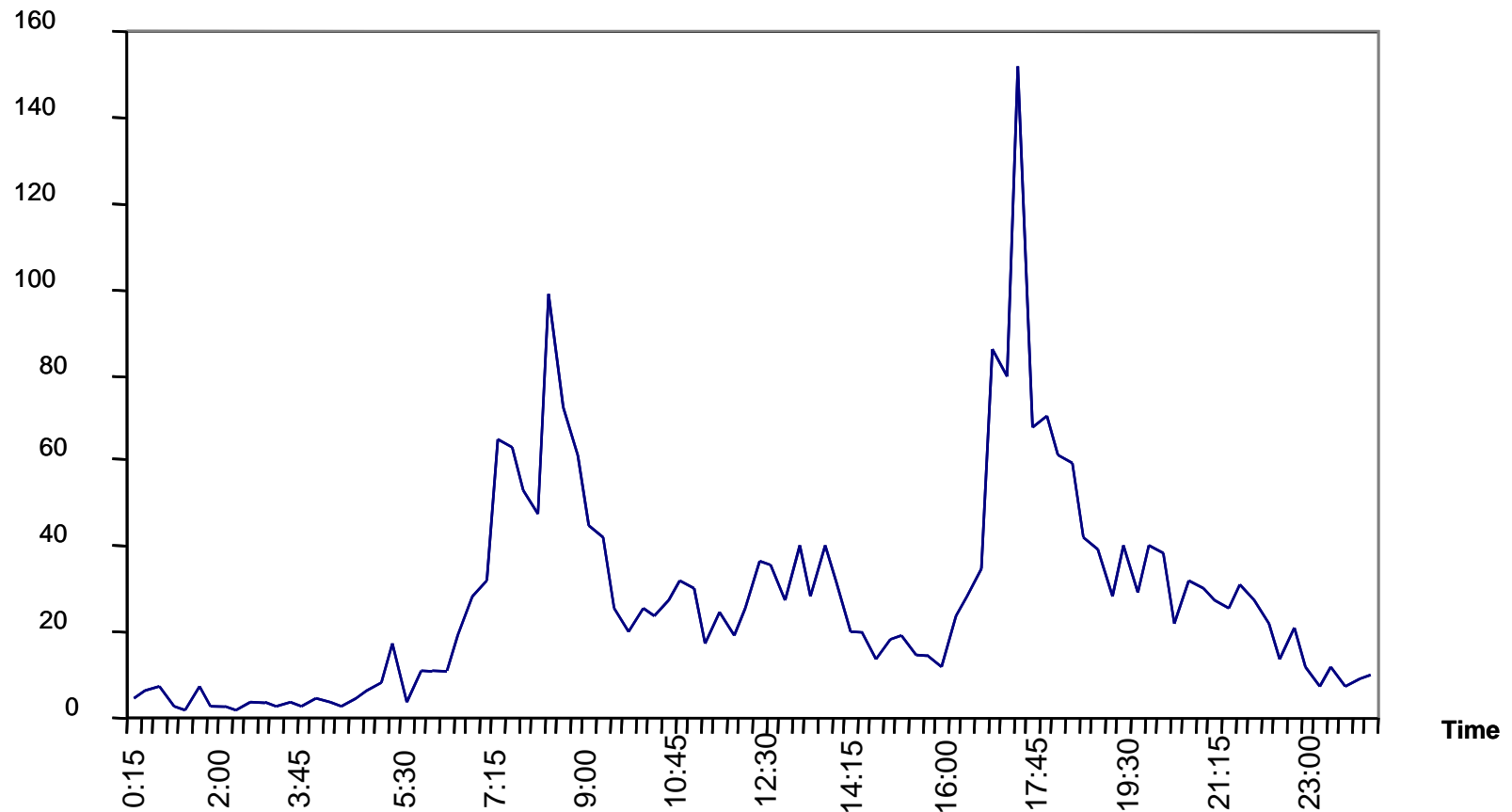
Patient	Arrival Time	Service Time
1	0	5
2	7	6
3	9	7
4	12	6
5	18	5
6	22	2
7	25	4
8	30	3
9	36	4
10	45	2
11	51	2
12	55	3



Inventory
(Patients at lab)

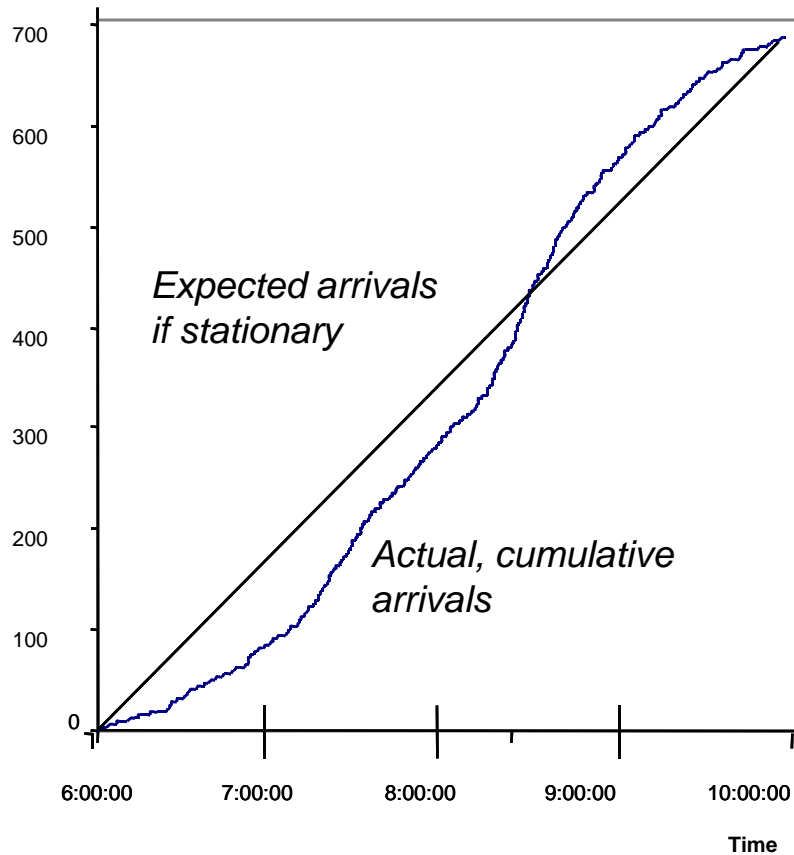
Is the incoming call rate stationary?

Number of customers
Per 15 minutes

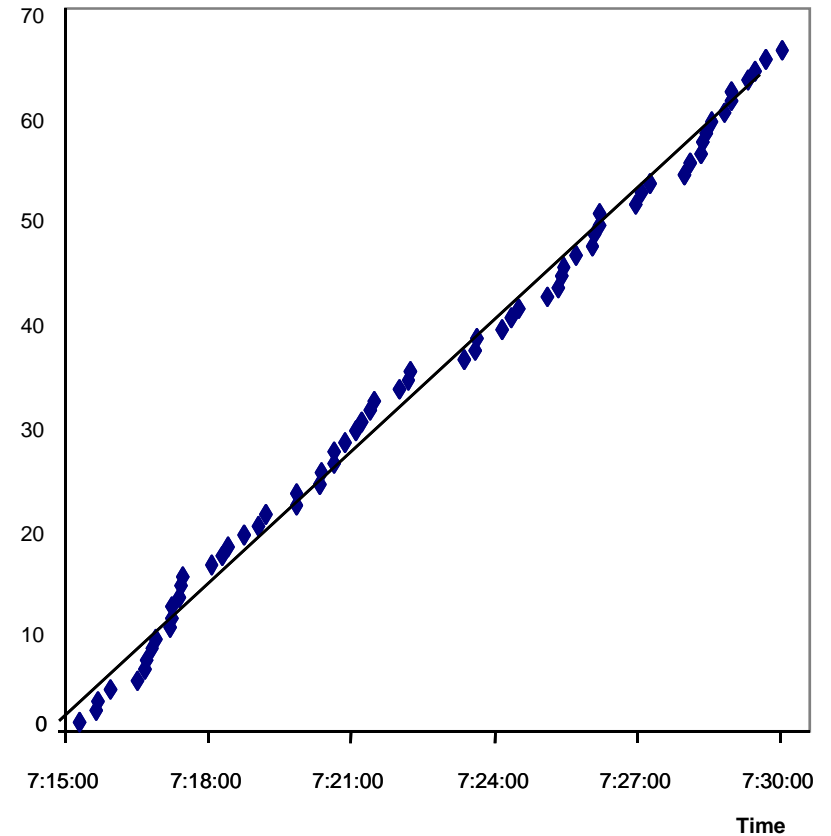


How to test for stationary?

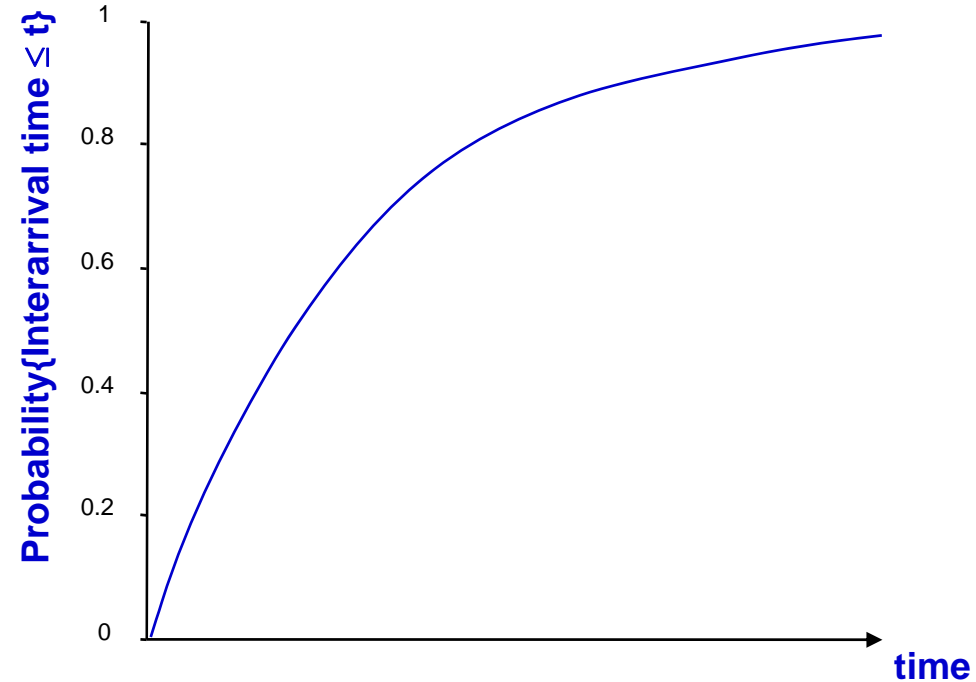
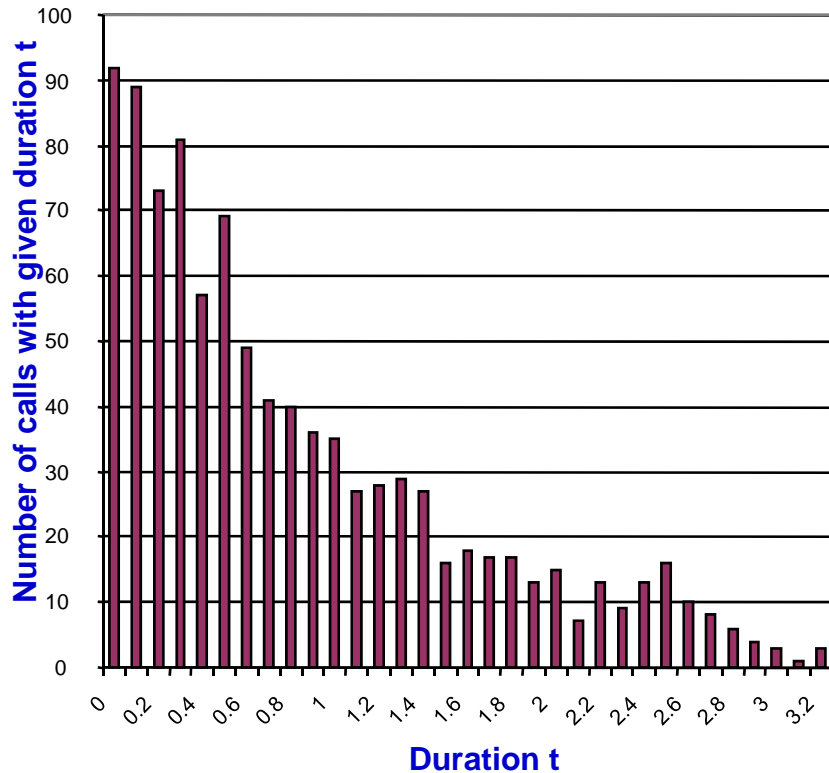
Cumulative
Number of
Customers



Cumulative
Number of
Customers



Exponential distribution for Interarrival times



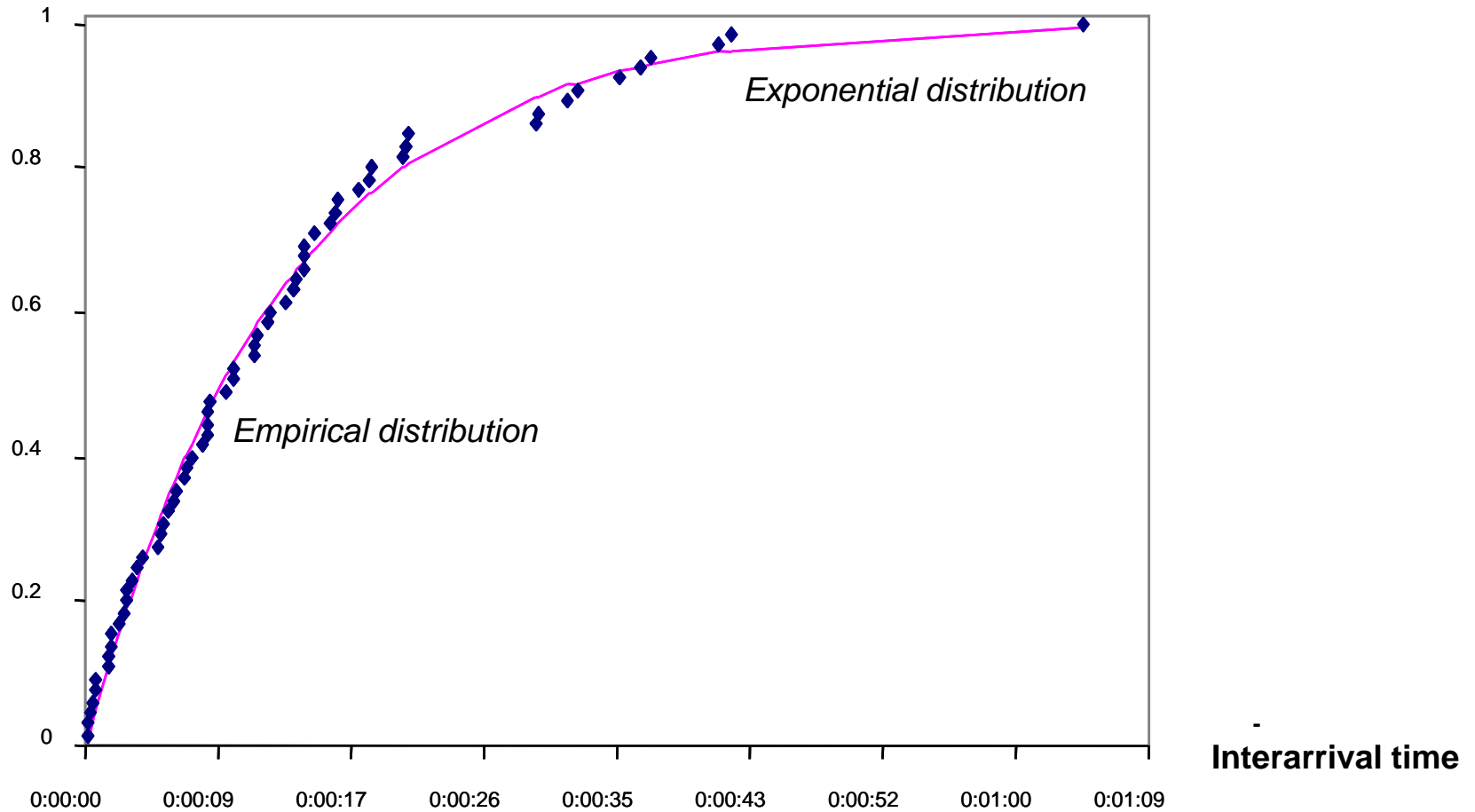
$$\text{Prob} \{ IA \leq t \} = 1 - \exp(-t/a)$$

$E(IA) = a$, expected time between arrivals

$$\text{StDev}(IA) = a$$

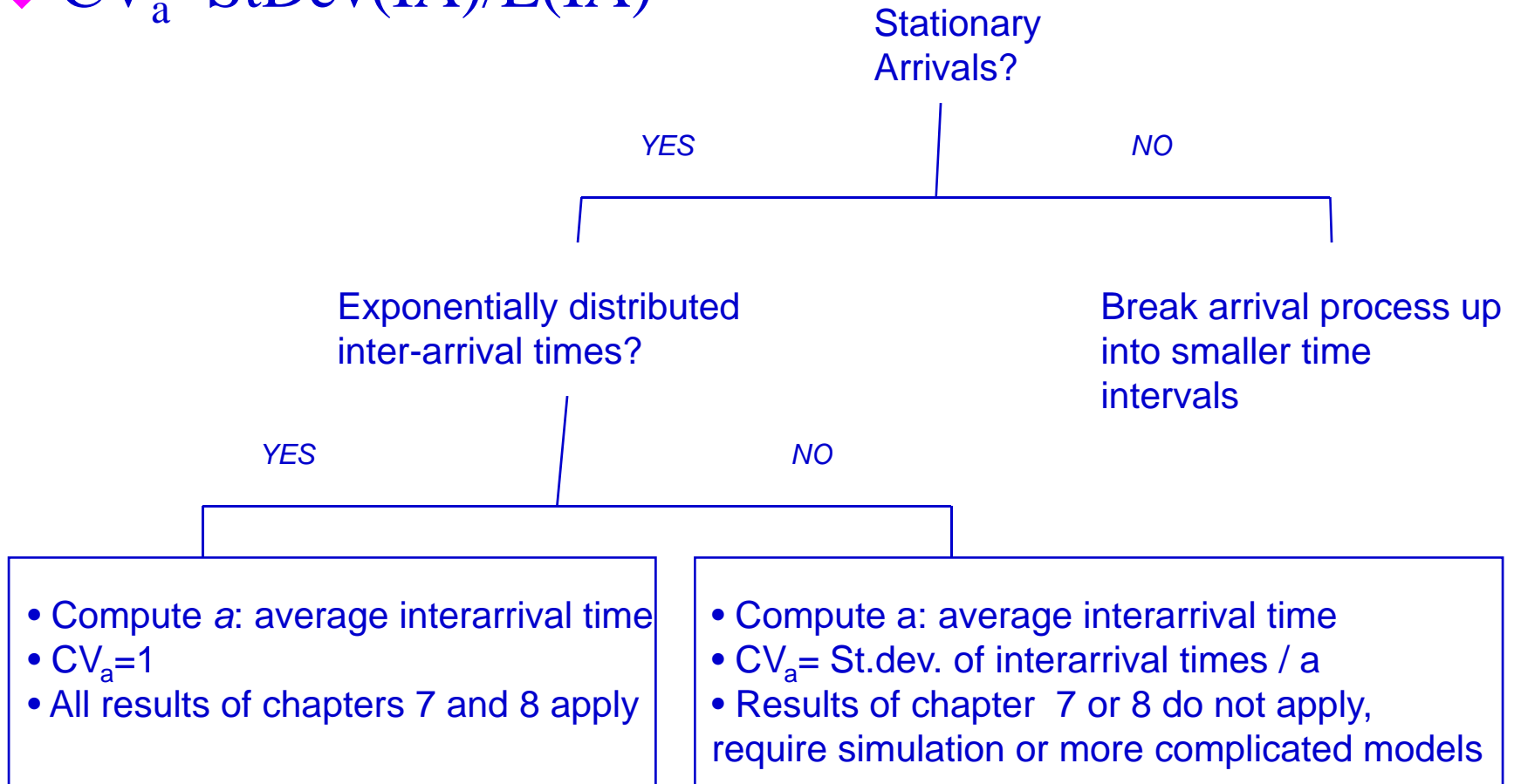
Comparing empirical and theoretical distributions

Distribution Function



Analyzing the Arrival Process

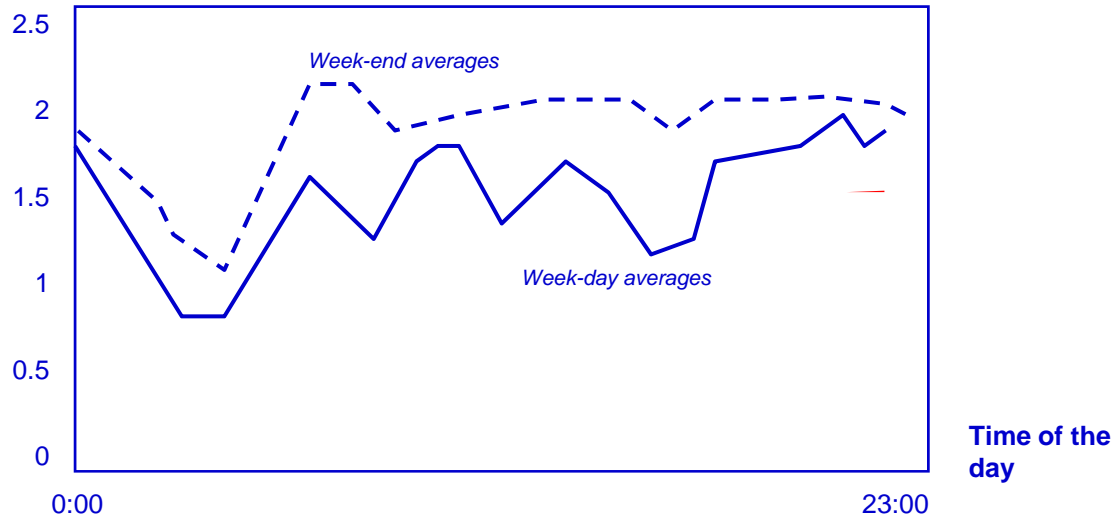
◆ $CV_a = \text{StDev}(\text{IA}) / E(\text{IA})$



Analyzing the Service Times

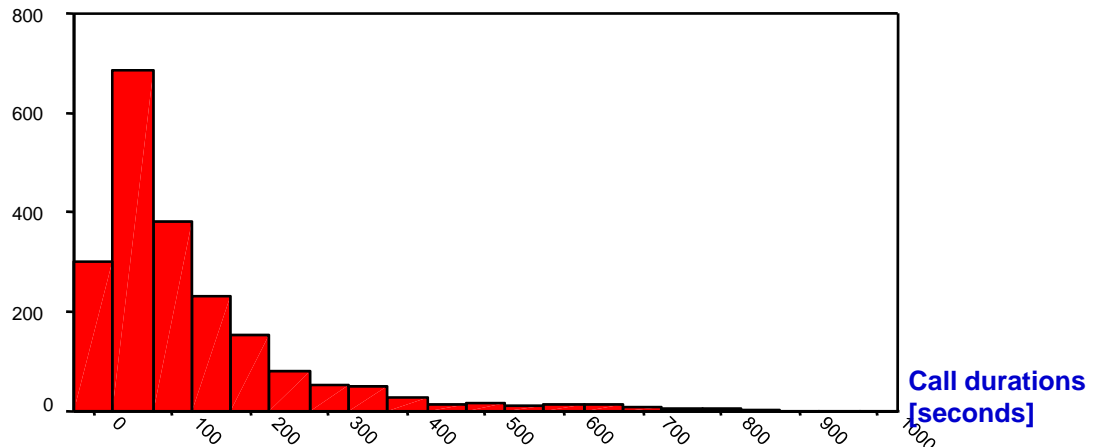
Seasonality and Variability

Call duration
[minutes]

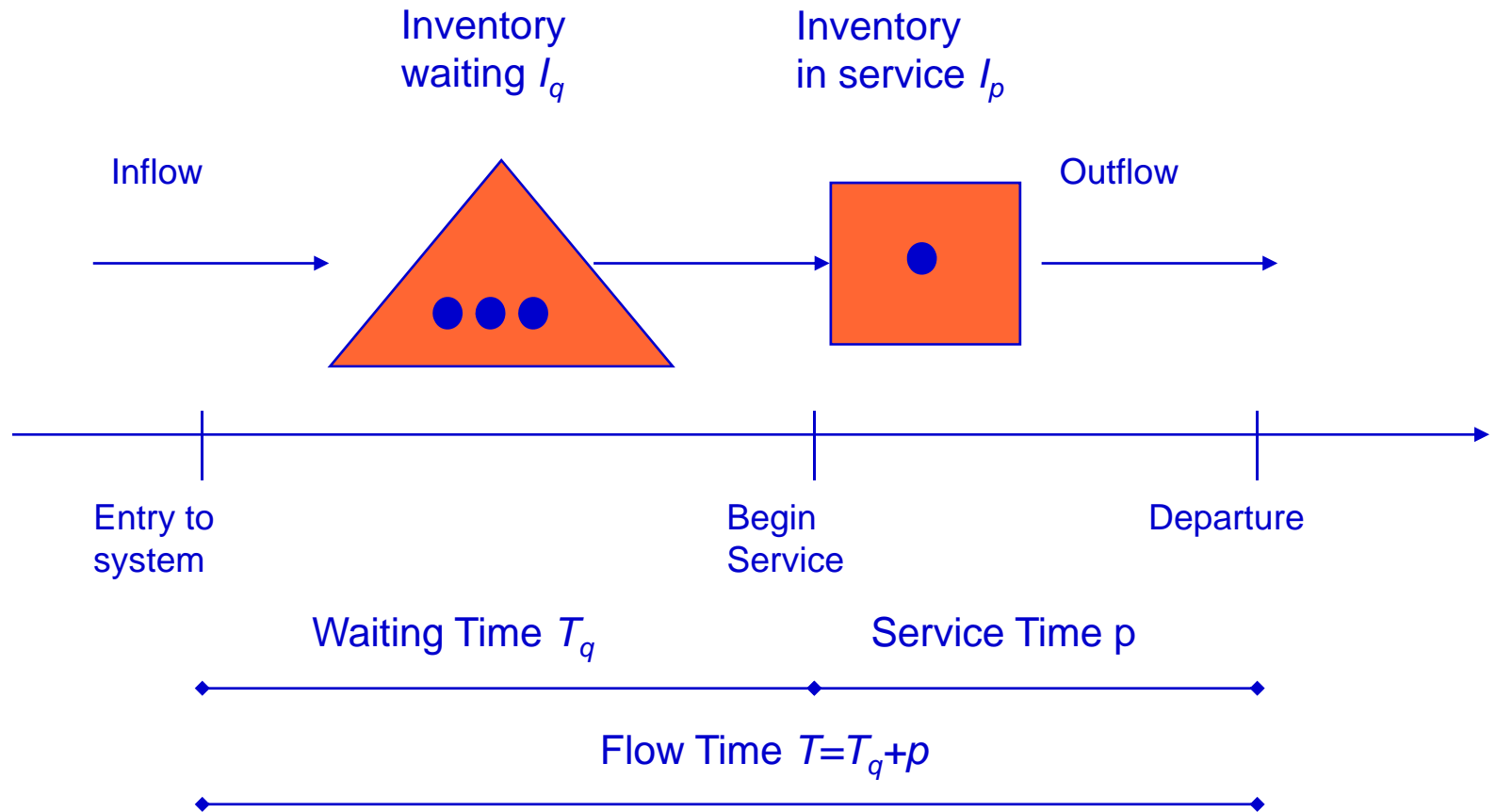


CV_p : Coefficient of variation
of service times

Frequency



Computing the expected waiting time T_q



Utilization

$$\text{Utilization } n = u = \frac{\text{FlowRate}}{\text{Capacity}} = \frac{1/a}{1/p} = \frac{p}{a}$$

Example: Average Activity time= $p=90$ seconds
Average Interarrival time= $a=300$ seconds

Utilization= $90/300=0.3=30\%$

The Waiting Time Formula

Waiting Time Formula for Exponential Arrivals

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

Diagram illustrating the components of the waiting time formula:

- Activity Time
- Utilization factor: $\left(\frac{\text{utilization}}{1 - \text{utilization}} \right)$
- Variability factor: $\left(\frac{CV_a^2 + CV_p^2}{2} \right)$
- Service time factor

Example: Average Activity time= $p=90$ seconds
Average Interarrival time= $a=300$ seconds
 $CV_a=1$ and $CV_p=1.333$

$$\text{Average time in queue} = 90 * \left(\frac{0.3}{1 - 0.3} \right) * \left(\frac{1^2 + 1.333^2}{2} \right) = 53.57 \text{ sec}$$

Bank Teller Example

- ◆ An average of 10 customers per hour come to a bank teller who serves each customer in 4 minutes on average. Assume exponentially distributed interarrival and service times.
 - (a) What is the teller's utilization?
 - (b) What is the average time spent in the queue waiting?
 - (c) How many customers would be waiting for this teller or would be serviced by this teller on average?
 - (d) On average, how many customers are served per hour?

Answer: $p=4$ mins; $a=6$ mins= $60/10$

$$\text{a) } u = \frac{p}{a} = \frac{4}{6} = 0.66$$

$$\text{b) } T_q = 4 * \left(\frac{0.66}{1-0.66} \right) * \left(\frac{1^2 + 1^2}{2} \right) = 8$$

$$\text{c) } I = \left(\frac{1}{a} \right) * (T_q + p) = \left(\frac{1}{6} \right) * (8 + 4) = 2$$

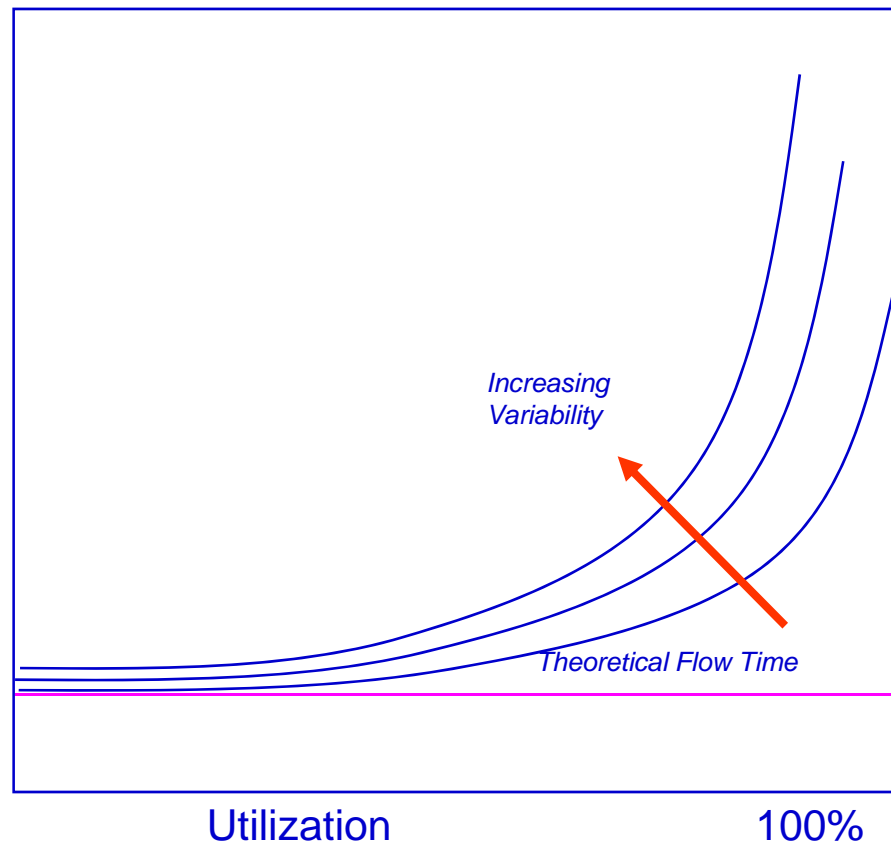
d) If teller is busy w.p. $2/3$, outputs 15 per hour.

If teller is idle w.p. $1/3$, outputs 0 per hour.

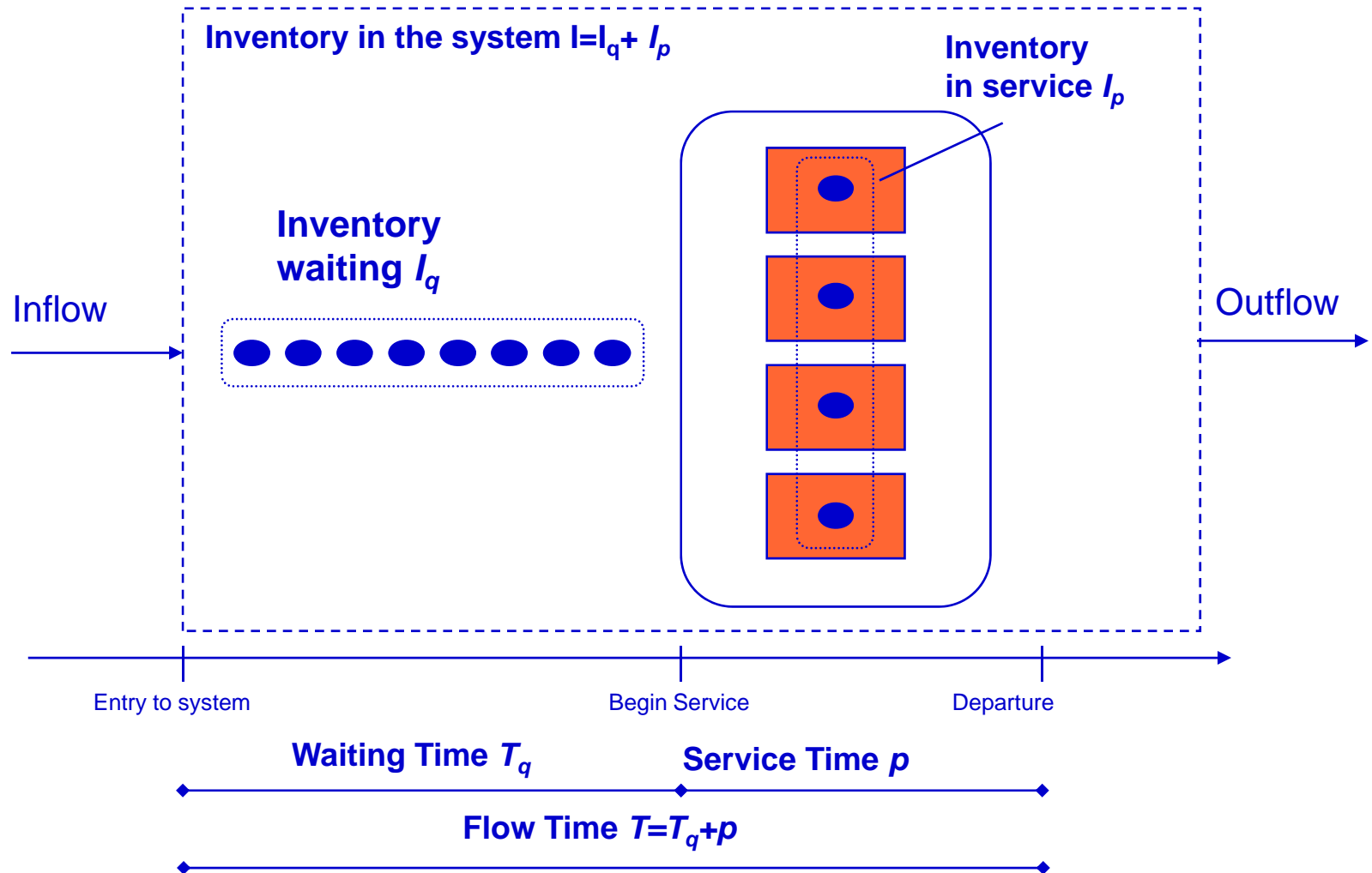
Average output = $(2/3) * 15 + (1/3) * 0 = 10$ per hour = Average input !!!

The Flow Time Increase Exponentially in Utilization

Average flow
time T



Computing T_q with m Parallel Servers



Utilization with m servers

$$\text{Utilization } n = u = \frac{\text{FlowRate}}{\text{Capacity}} = \frac{1/a}{m/p} = \frac{p}{a m}$$

Example: Average Activity time= $p=90$ seconds

Average Interarrival time= $a=11.39$ secs over 8-8:15

$m=10$ servers

Utilization= $90/(10 \times 11.39)=0.79=79\%$

Waiting Time Formula for Parallel Resources

Approximate Waiting Time Formula for Multiple (m) Servers

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

Example: Average Activity time= $p=90$ seconds
Average Interarrival time= $a=11.39$ seconds
 $m=10$ servers
 $CV_a=1$ and $CV_p=1.333$

$$\text{Time in queue} \cong \left(\frac{90}{10} \right) * \left(\frac{0.79^{\sqrt{2(10+1)-1}}}{1 - 0.79} \right) * \left(\frac{1^2 + 1.333^2}{2} \right) = 24.94 \text{ sec}$$

$$T = T_q + p = 24.94 + 90 = 114.94 \text{ sec} = 1.916 \text{ min}$$

Online Retailer Example

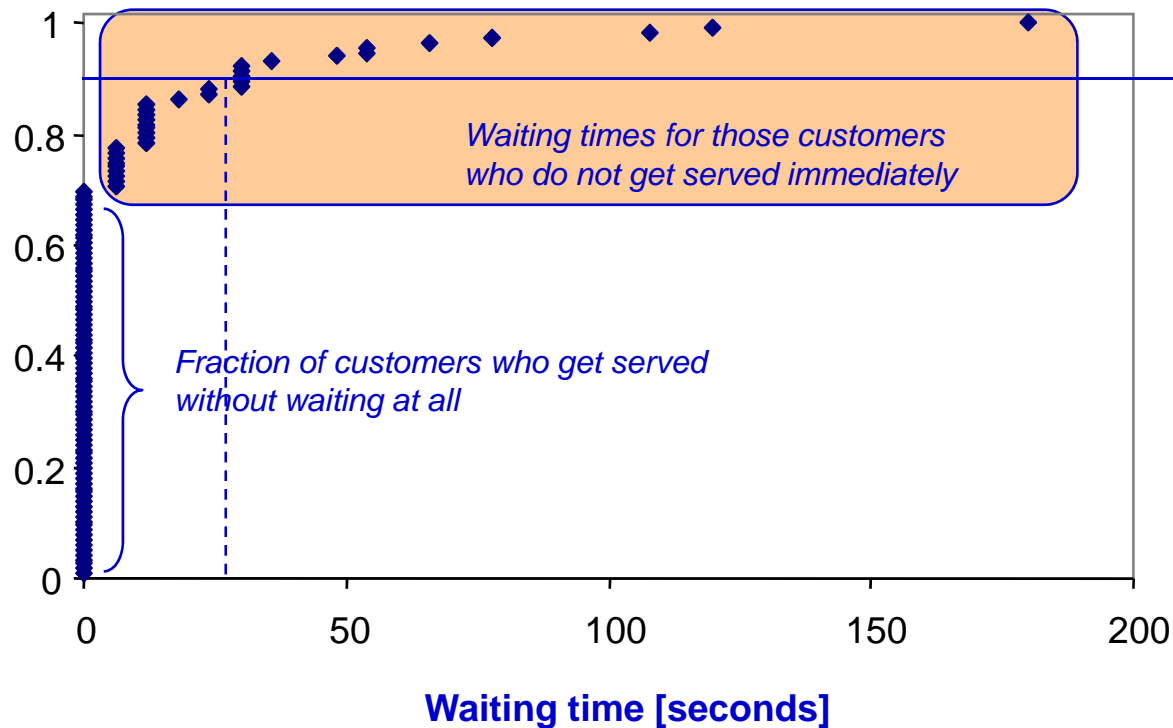
Customers send emails to a help desk of an online retailer every 2 minutes, on average, and the standard deviation of the inter-arrival time is also 2 minutes. The online retailer has three employees answering emails. It takes on average 4 minutes to write a response email. The standard deviation of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

- (a) Estimate the average customer wait before being served.
- (b) How many emails would there be -- on average -- that have been submitted to the online retailer, but not yet answered?

Answer: $a=2$ mins; $CV_a=1$; $m=3$; $p=4$ mins; $CV_p=0.5$
a) Find T_q . b) Find $I_q=(1/a)T_q$

Service Levels in Waiting Systems

Fraction of customers who have to wait x seconds or less



- Target Wait Time (TWT)
- Service Level = $\text{Probability}\{\text{Waiting Time} \leq \text{TWT}\}$; needs distribution of waiting time
- Example: Deutsche Bundesbahn Call Center
 - now (2003): 30% of calls answered within 20 seconds
 - target: 80% of calls answered within 20 seconds

Bank of America's Service Measures

Customer Service and Support

A Passion to Delight

Our Guiding Principles: *Commitment, Passion, Learning, Integrity, Respect, Balance, Family, Fun and Service Excellence About Us*

Customer Service and Support is an integral part of Bank of America, employing more than 9,500 highly skilled associates in contact centers located in twenty cities across the United States. These associates provide service and financial solutions to more than 130 million phone customers and 1.74 million e-mail customers each year, making our contact centers among the busiest in the country.

Customer Service and Support is working to build a world-class customer service organization. The nine guiding principles listed above and the Bank of America Spirit provide the foundation for our daily work routine. Our associates are brand ambassadors whose hard work and determination will be the driving force behind our goal to make Bank of America the most admired company in the world.

Customer Service and Support is focused on building better, stronger and deeper relationships with our customers. Our associates have a passion for reaching a Higher Standard, achieving results and winning for our customers. It is important to all of us that we strive to provide the highest level of service to ensure that all of our customers are "delighted" with their Bank of America experience.

Functional Scope Areas

- Customer Service and Support
- National Consumer Service Centers
 - Consumer and Consumer Card
 - Dealer Financial Services
 - IBCC
 - NDS
 - Plus
 - Prime
- Associate Experience and Communications
- Client Service and Support
 - Associate Banking
 - Commercial
 - Merchant and Commercial Card Services
 - Premier
 - Small Business
- Multicultural Services
- Customer Service Process and Operations
 - Resolution Services and Support
- Risk Management
- Customer Delight
- Strategy and Marketing
- Customer Contact Management

Factoids:

Annualized

Customer Calls Received by VRU in 2002	508,500,000
Customer Calls Handled by VRU in 2002	503,500,000
Customer Calls Offered to Associates in 2002	147,000,000
Customer Calls Handled by Associates in 2002	130,000,000
Avg. Speed to Answer	96.54 secs
E-mails Received in 2002	1,750,000
E-mails Processed in 2002	1,740,000
2002 Customer Delight	54.3%
Certified Green Belts through 3/03	203
Certified Black Belts through 3/03	2
Associate Satisfaction in 2002	72%
Associate Retention in 2002	78%

2003 Performance Plan

Bank of America Vision:

Be recognized as the world's most admired company

Customer Service and Support Vision:

A Passion to Delight

To reach our goal of being the world's most admired company, we must do the following:

- Execute on our Hoshin Plan
- Live the Bank of America Spirit
- Communicate accurately and consistently
- Execute reliable, repeatable, consistent processes
- Focus on delivering world-class service for our customers

The focus for 2003 is: **65 / 75 / 64**

- **65%** Customer Delight
- **75%** Associate Delight
- **\$64** million in productivity benefits (Shareholder Delight)

Waiting Lines: Points to Remember

- Variability is the norm, not the exception
 - understand where it comes from and eliminate what you can
 - accommodate the rest
- 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
- Adding capacity is expensive, although some safety capacity is necessary
- Next case:
 - application to call center
 - use $CV=1$
 - careful in interpreting March / April call volume

Summary of the formulas

1. Collect the following data:

- number of servers, m
- activity time, p
- interarrival time, a
- coefficient of variation for interarrival (CV_a) and processing time (CV_p)

2. Compute utilization: $u = \frac{p}{a \times m}$

3. Compute expected waiting time

$$T_q = \left(\frac{\text{Activity time}}{m} \right) \times \left(\frac{\text{utilization} \cdot n^{\sqrt{2(m+1)}-1}}{1 - \text{utilization} \cdot n} \right) \times \left(\frac{CV_a^2 + CV_p^2}{2} \right)$$

4. Based on T_q , we can compute the remaining performance measures as

Flow time $T = T_q + p$

Inventory in service $I_p = m \cdot u$

Inventory in the queue $I_q = T_q / a$

Inventory in the system $I = I_p + I_q$

Staffing levels

Cost of direct labor per serviced unit

$$\text{Cost of Direct Labor} = \frac{\text{Total wages per time}}{\text{Flow rate} = \text{Arrival rate}} = \frac{m \times (\text{wages per time})}{1/a} = \frac{p \times (\text{wages per time})}{u}$$

Ex: \$10/hour wage for each CSR; $m=10$

Activity time= $p=90$ secs; Interarrival time= 11.39 secs

1-800 number line charge \$0.05 per minute

$$\text{Utilization} = u = \frac{p}{m \times a} = \frac{90}{10 \times 11.39} = 0.79$$

$$\text{Cost of Direct Labor} = \frac{1.5 \text{ min/call} \times (16.66 \text{ cents/min})}{0.79} = 31.64 \text{ cents/call}$$

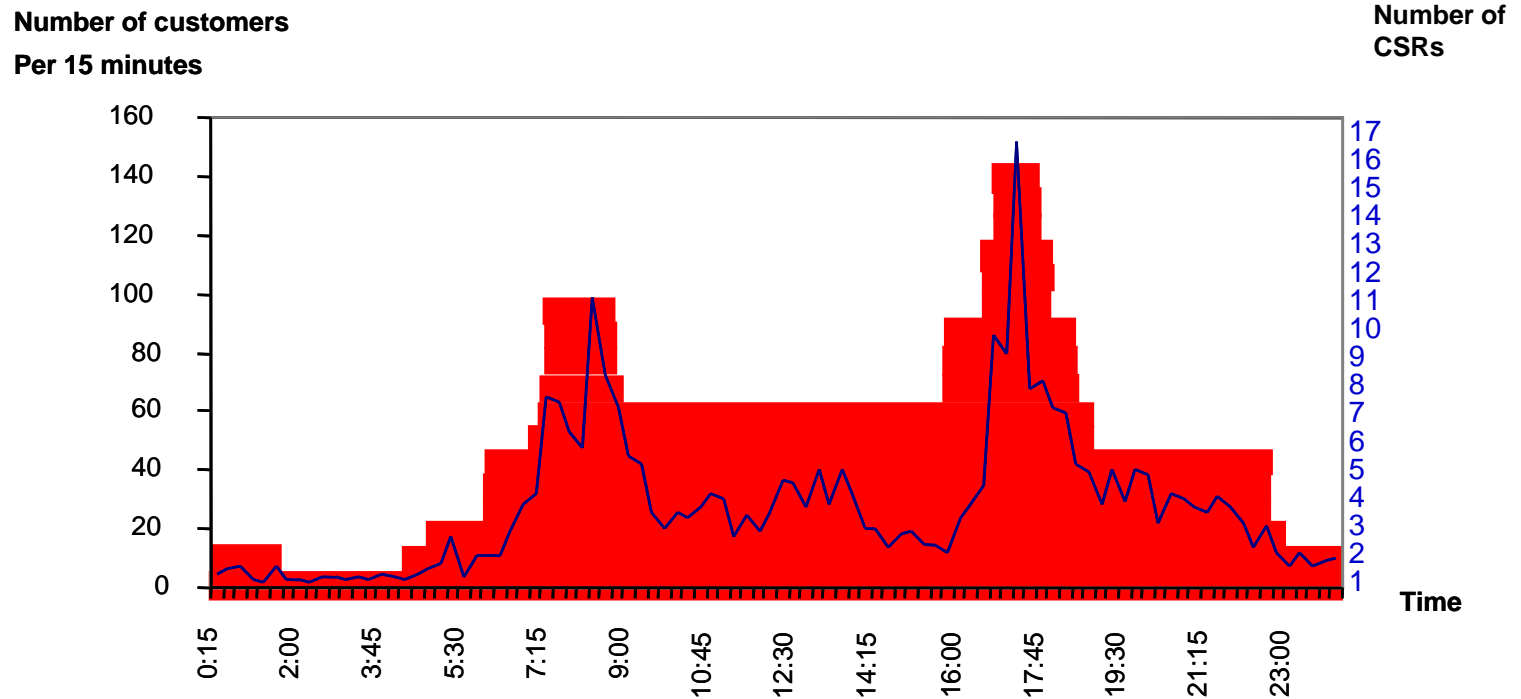
$$\text{Recall } T = 1.916; \text{ Cost of line charge per call} = 1.916 \times 0.05 = \$0.0958 / \text{call}$$

Cost of Staffing Levels

m	u	Labor cost per call	Line cost charge per call	Total cost per call
8				1.3458
9				0.4201
10	0.790	0.3164	0.0958	0.4122
11				0.4323
12				0.4593
13				0.4887
14				0.5193
15				0.5503

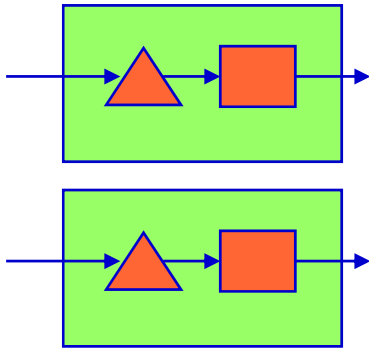
The optimal staffing level $m=10$

Staffing Levels under various Interarrival Times

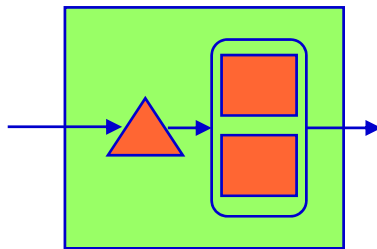


The Power of Pooling

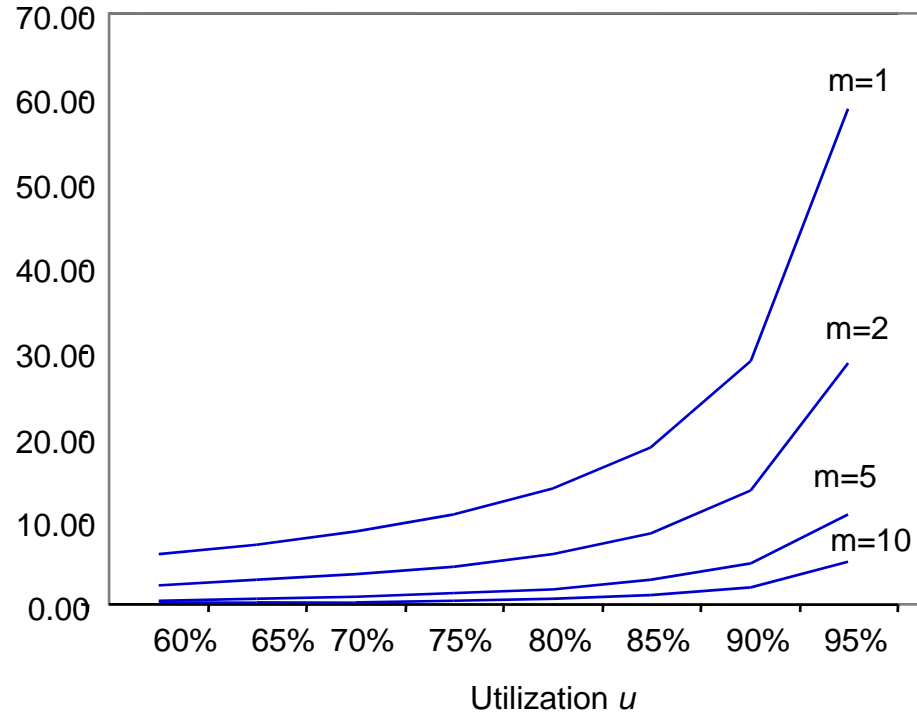
Independent Resources
 $2 \times (m=1)$



Pooled Resources
($m=2$)



Waiting
Time T_q



Implications:

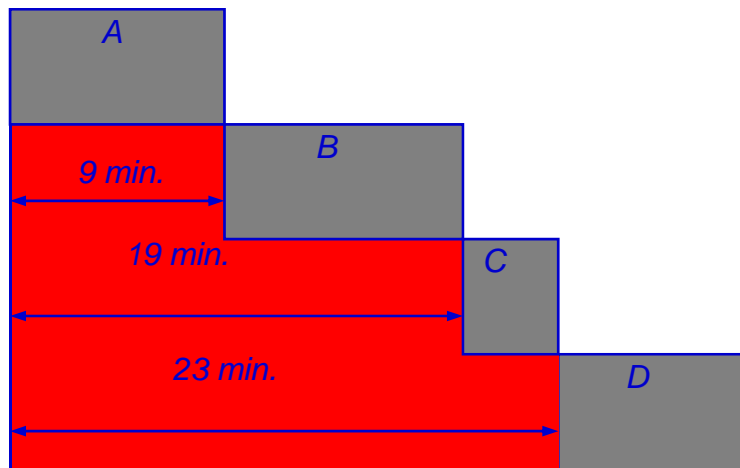
+ balanced utilization

+ Shorter waiting time (pooled safety capacity)

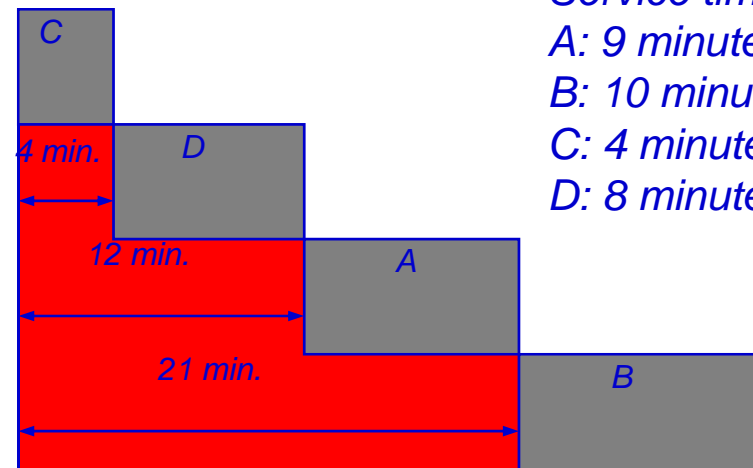
- Change-overs / set-ups

Service-Time-Dependent Priority Rules

- Flow units are sequenced in the waiting area (triage step)
- Shortest Processing Time (SPT) Rule
 - Minimizes average waiting time
 - Do you want to wait for a short process or a long one?



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



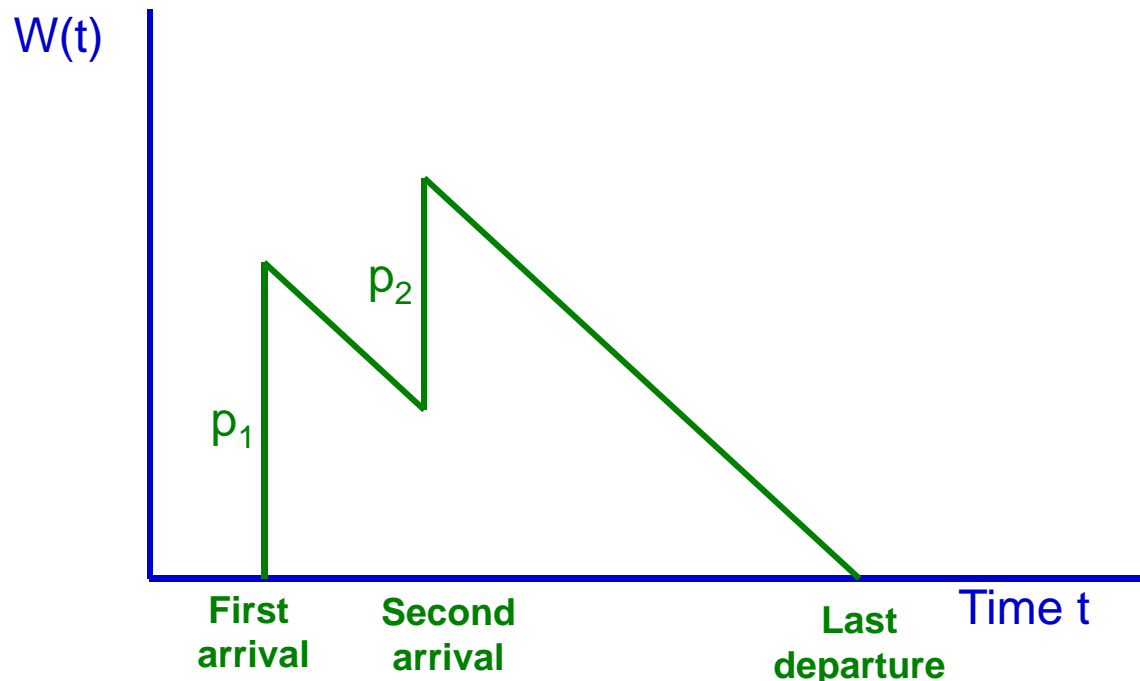
Total wait time: $4+12+21=37$ min

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

- Problem of having “true” processing times

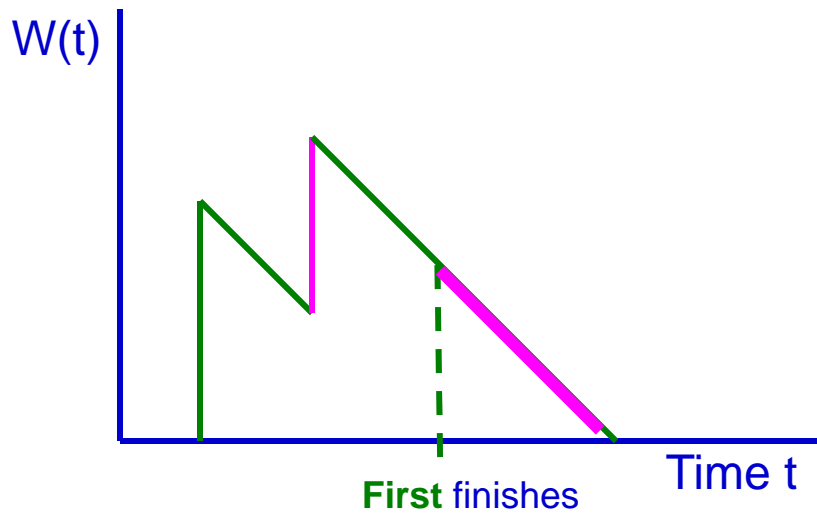
Service-Time-Independent Priority Rules

- Sequence based on importance
 - emergency cases; identifying profitable flow units
- First-Come-First-Serve
 - easy to implement; perceived fairness
- The order in which customers are served does **Not** affect the average waiting time.
 - $W(t)$: Work in the system
 - An arrival at t waits until the work $W(t)$ is completed

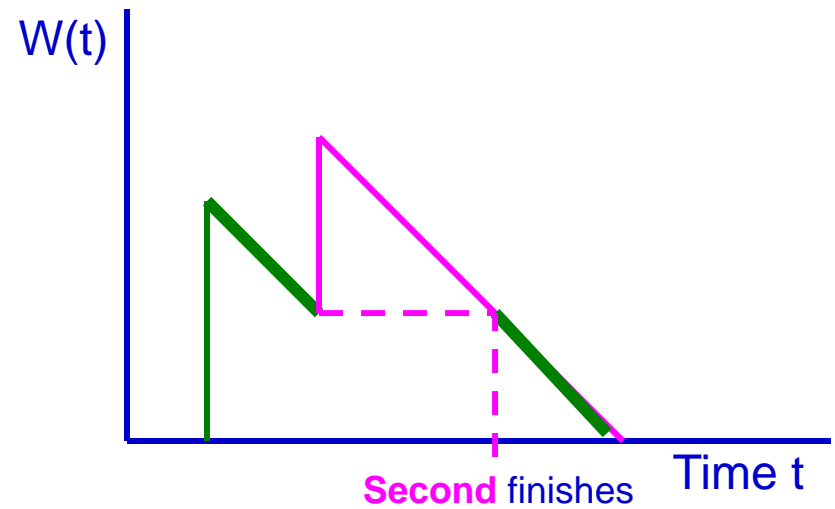


Service-Time-Independent Order does not affect the waiting time

Order: 1-1-2



Order: 1-2-1



Work is conserved even when the processing order changes.

No matter what the order is, the third arrival finds the same amount of work $W(t)$.

Summary

- ◆ Interarrival and Service Times and their variability
- ◆ Obtaining the average time spent in the queue
- ◆ Pooling of server capacities
- ◆ Priority rules