Waiting Times

Chapter 7

These slides are based in part on slides that come with Cachon & Terwiesch book *Matching Supply with Demand* http://cachon-terwiesch.net/3e/. If you want to use these in your course, you may have to adopt the book as a textbook or obtain permission from the authors Cachon & Terwiesch.

Learning Objectives

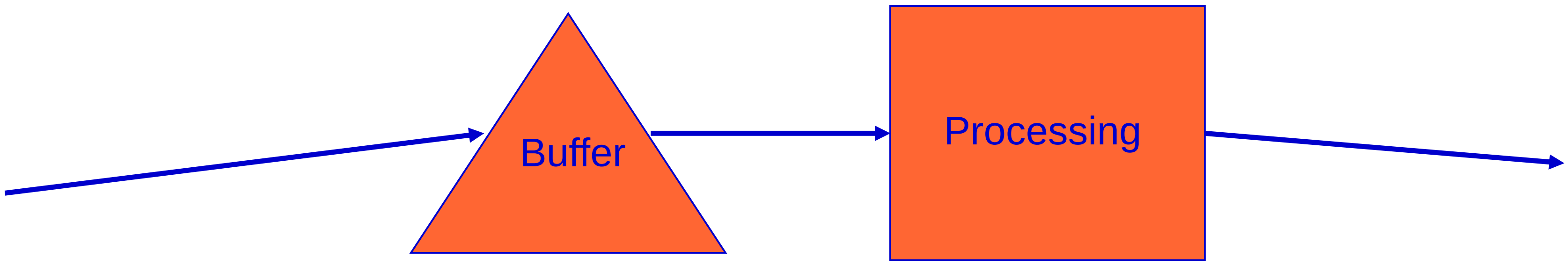
- Interarrival and Service Times and their variability
- Obtaining the average time spent in the queue
- Pooling of server capacities
- Priority rules

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



Queues

Input:
Passengers in an airport
Customers at a bank
Patients at a hospital
Callers at a call center

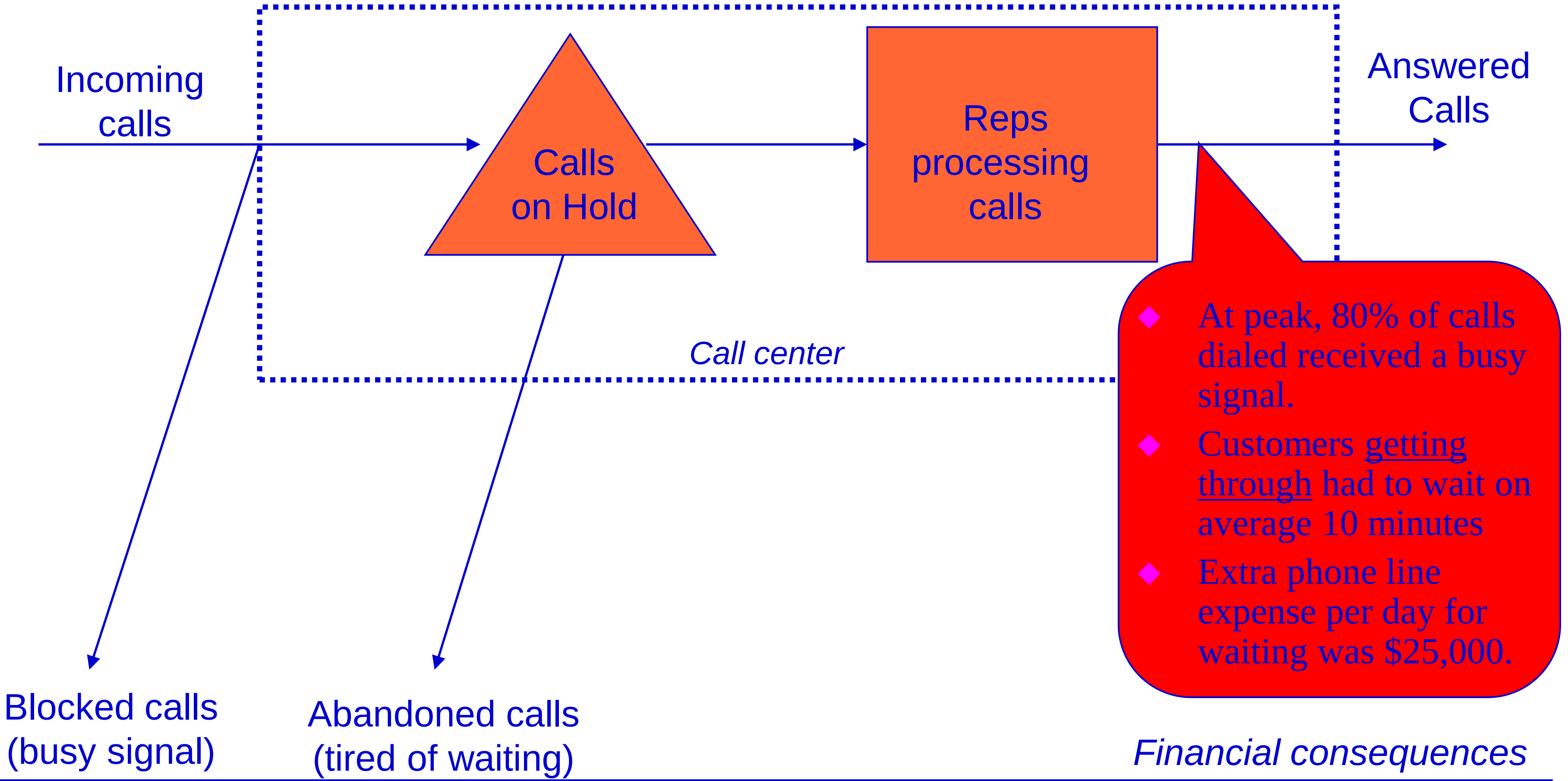
Arrival rate

Resources:
Check-in clerks at an airport
Tellers at a bank
Nurses at a hospital
Customer service representatives (CRS) at a call center

Capacity

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

An Example of a Simple Queuing System



Financial consequences

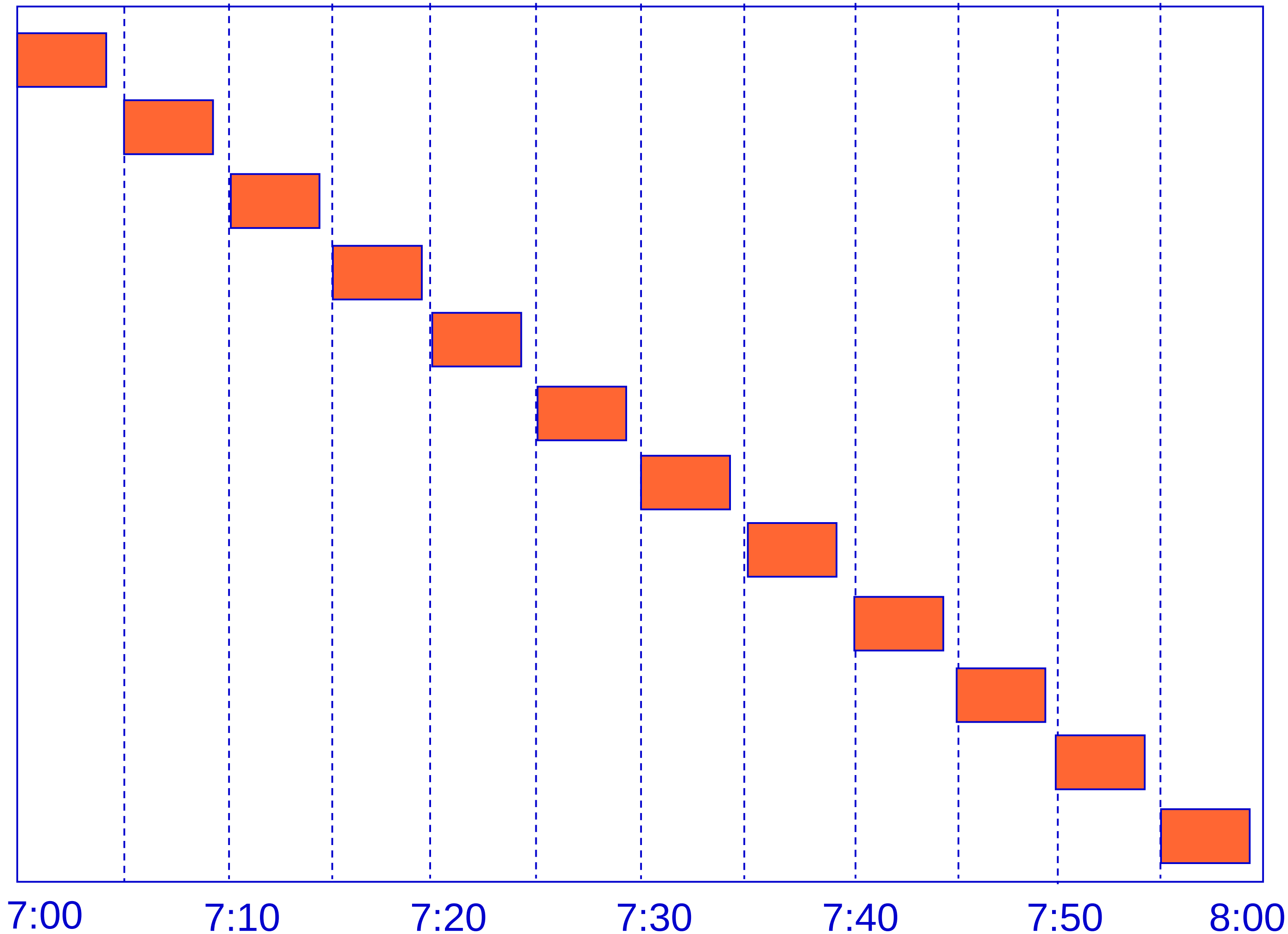
Lost throughput	Holding cost Lost goodwill Lost throughput (abandoned)	Cost of capacity Cost per customer	\$\$\$ Revenue \$\$\$

A Somewhat Odd Service Process

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

Arrival rate 0.2/min = 1/(4 mins) = 1 every five minutes, which implies interarrival time of 5 minutes.
Units of arrival rate 1/min whereas units of interarrival time is 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

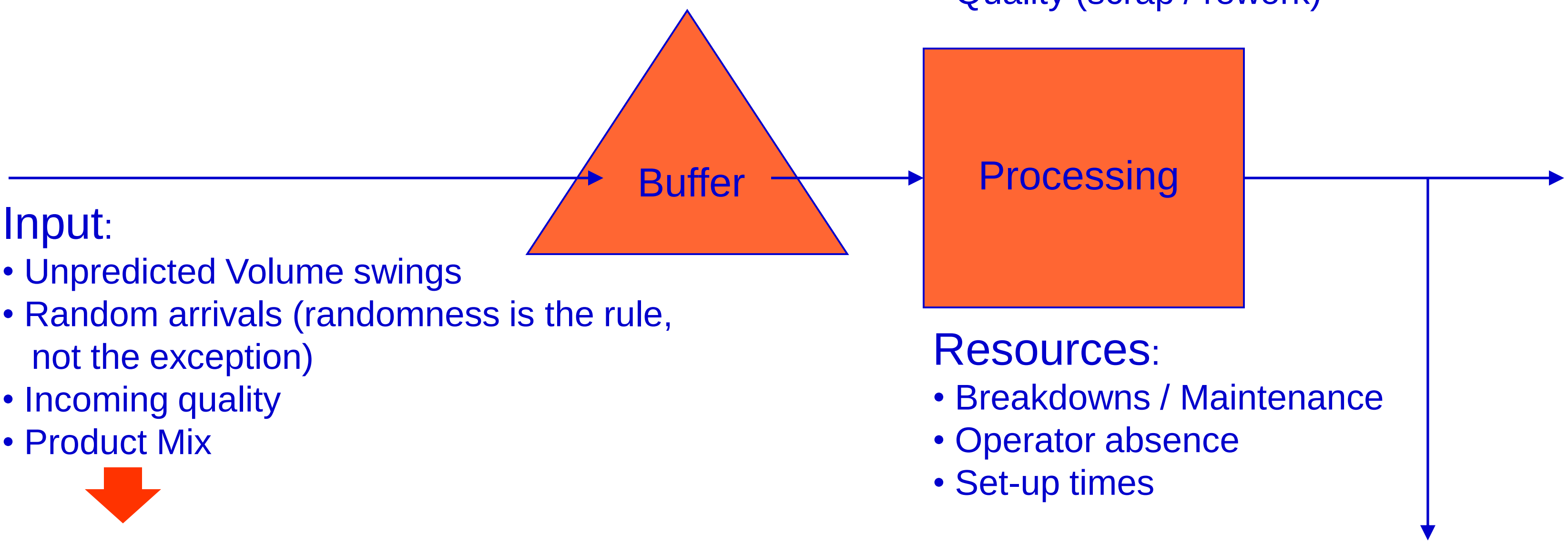


Where is Variability?

- There certainly is significant (actually infinite) amount of waiting when the arrival rate is greater than the service rate
  - Equivalently, the processing capacity is less than the arrival rate
- More interestingly, variability can cause long waiting times. Variability in
  - Arrival process
  - Processing times
  - Availability of resources; Absent, sick, broken or vacationing servers.
  - Types of customers; Priority versus regular customers.
  - Routing of flow units; Recall the Resume Validation Example.
  - Response of customers to waiting for a while; Wait more or abandon

Variability: Where does it come from? Examples

Tasks:
- Inherent variation
- Lack of Standard Operating Procedures
- Quality (scrap / rework)



Input:
- Unpredicted Volume swings
- Random arrivals (randomness is the rule, not the exception)
- Incoming quality
- Product Mix

Especially relevant in service operations (what is different in service industries?):
- *emergency room*
- *air-line check in*
- *call center*
- *check-outs at cashier*

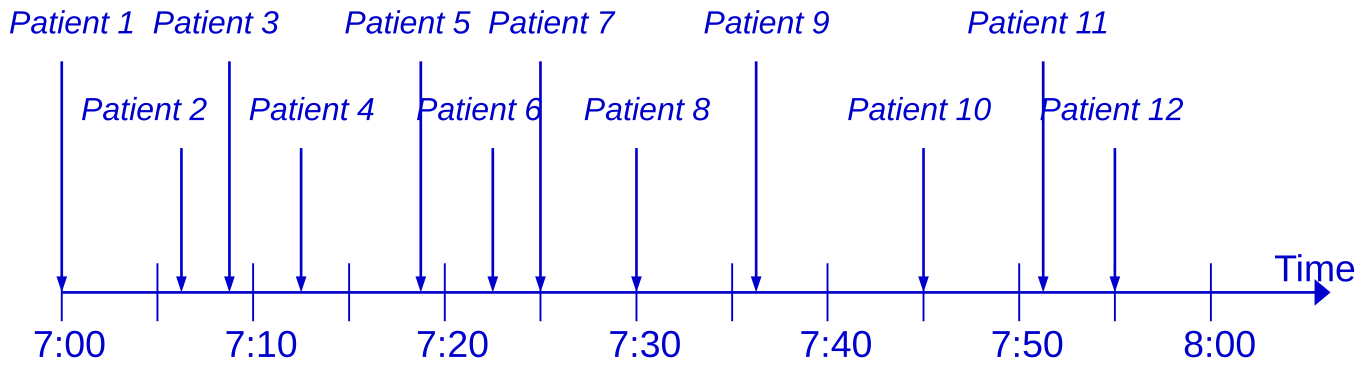
Resources:
- Breakdowns / Maintenance
- Operator absence
- Set-up times

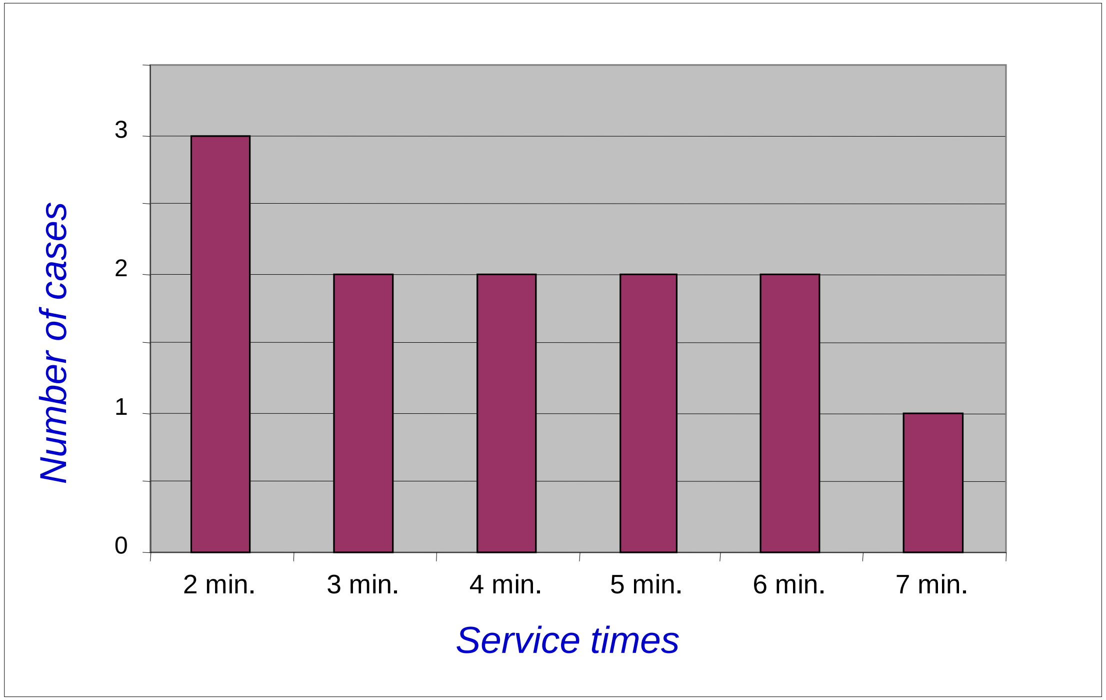
Routes:
- Variable routing
- Dedicated machines

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	2
Averages	5	4

Interarrival time

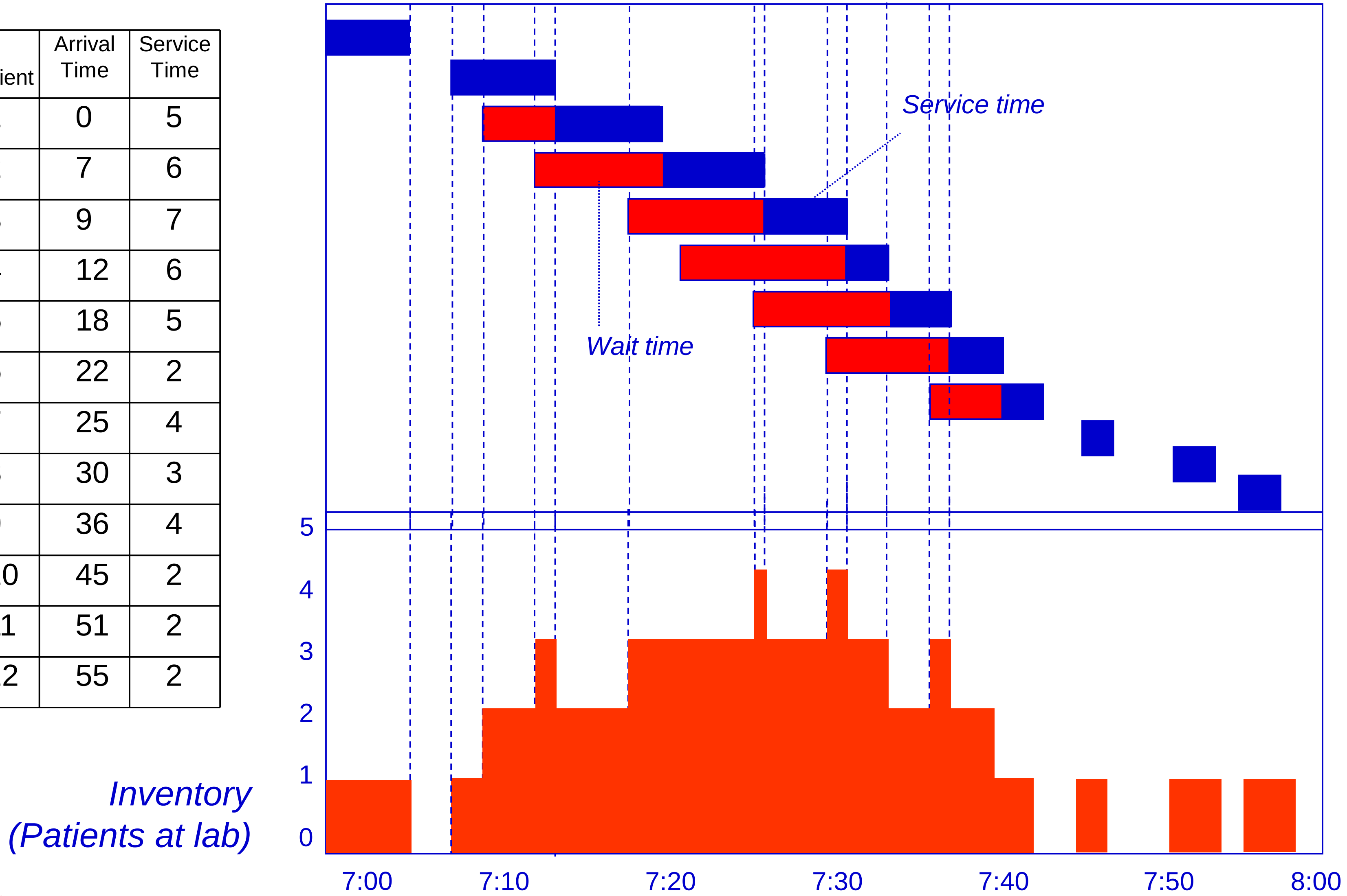




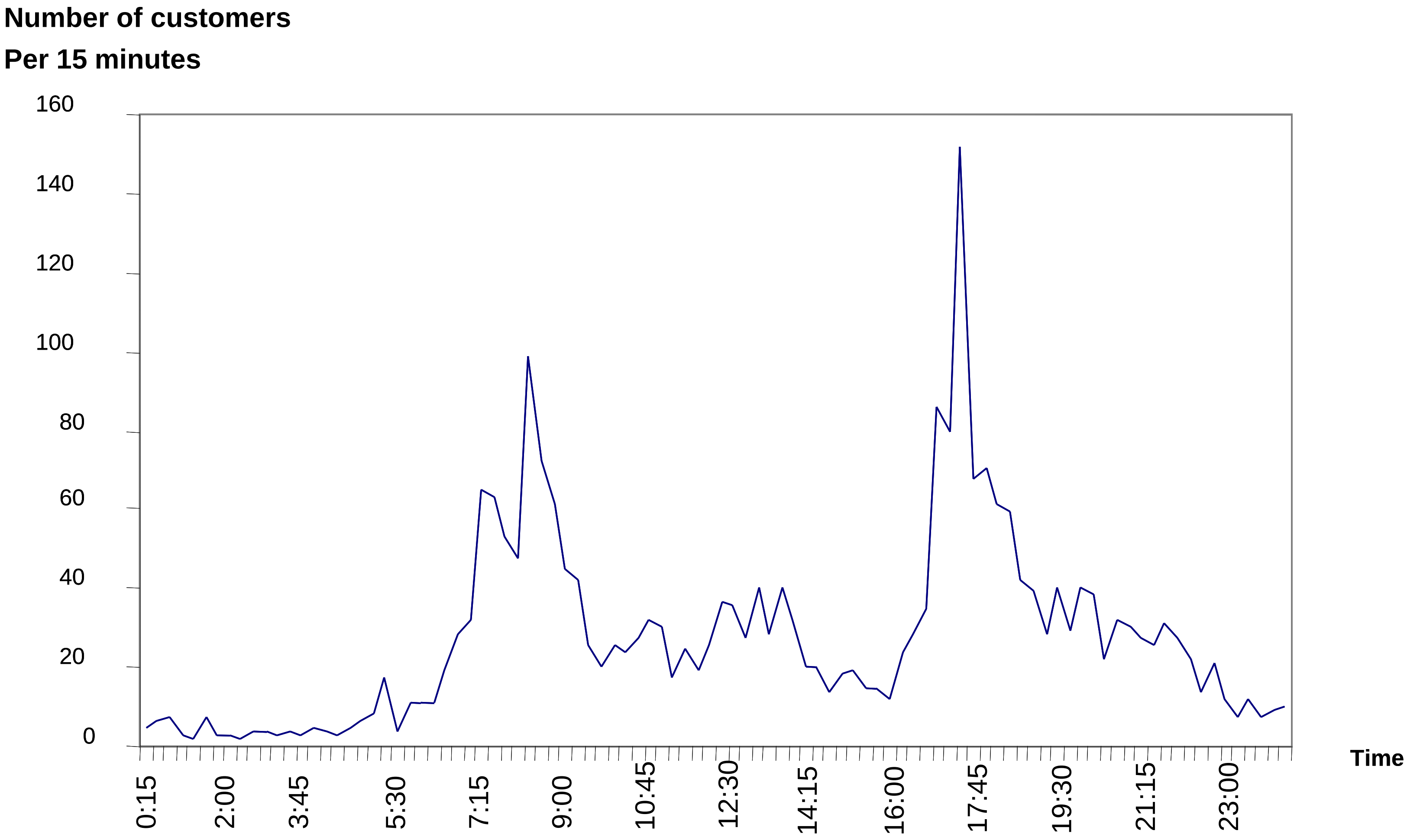
Variability Leads to Waiting Time

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

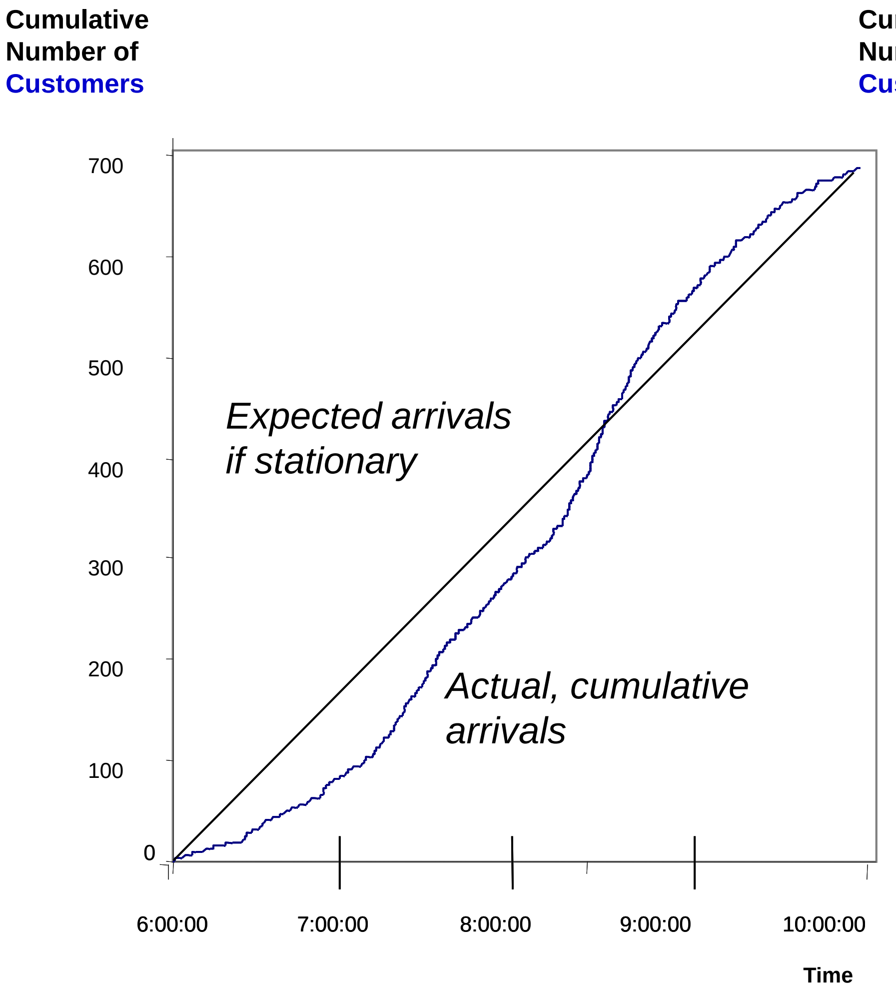
Patient	Arrival Time	Service Time
1	0	5
2	7	6
3	9	7
4	12	6
5	18	5
6	22	2
7	25	4
8	30	3
9	36	4
10	45	2
11	51	2
12	55	2

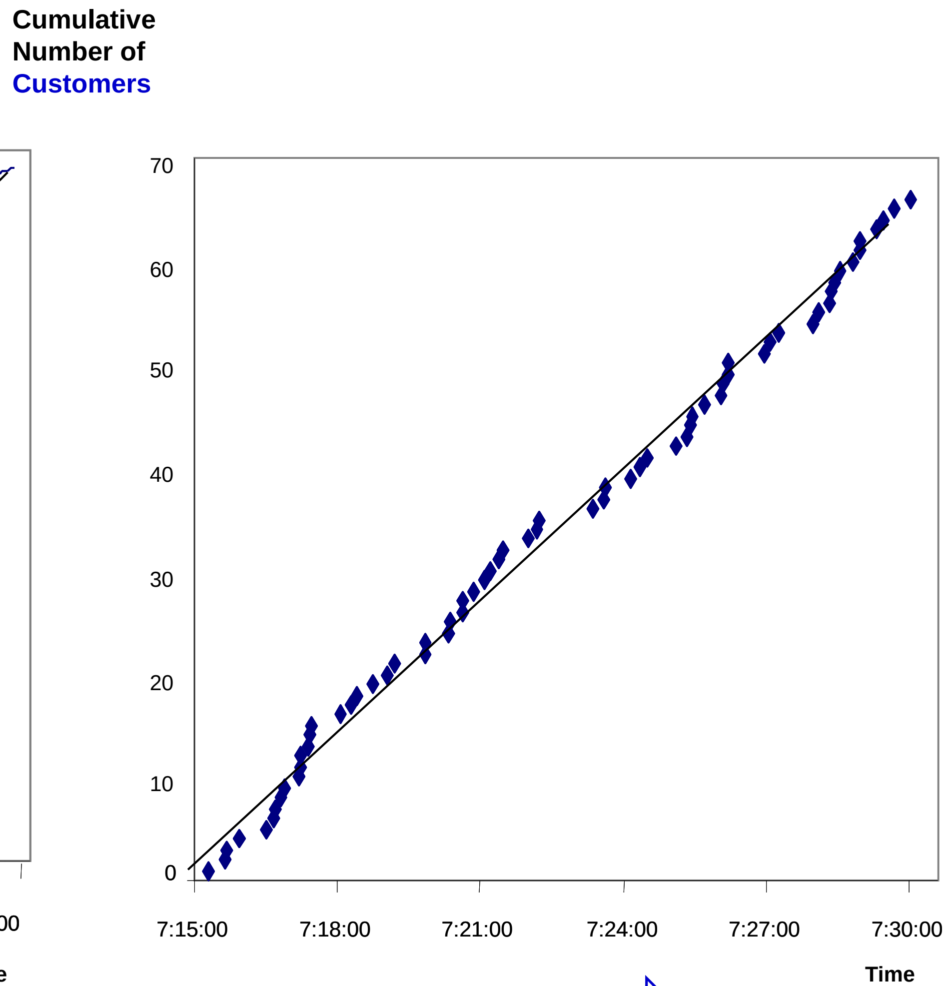


Is the incoming call rate stationary?



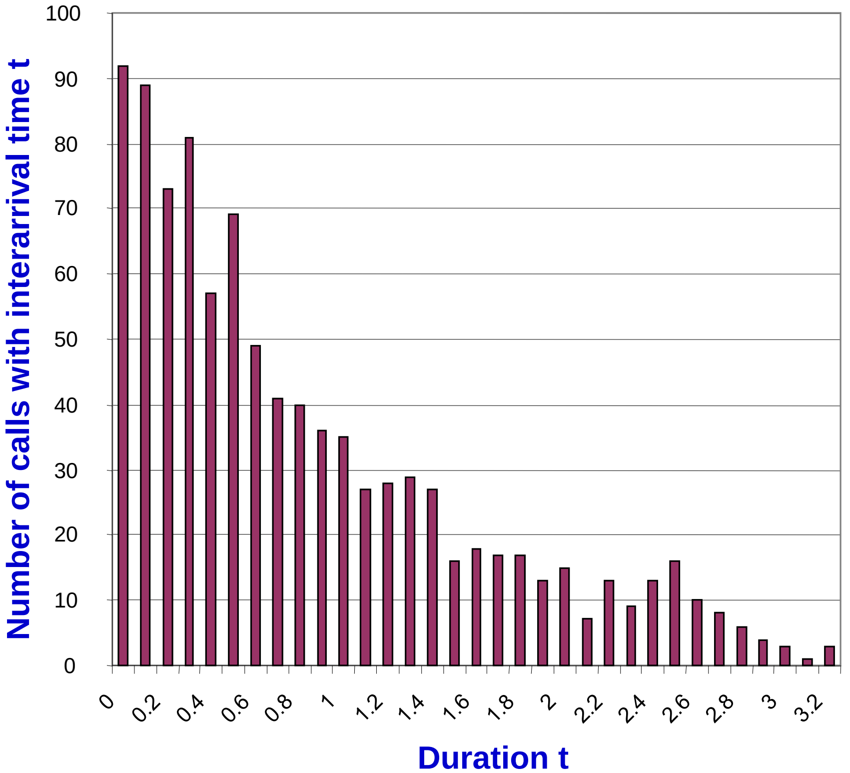
How to test for stationary?

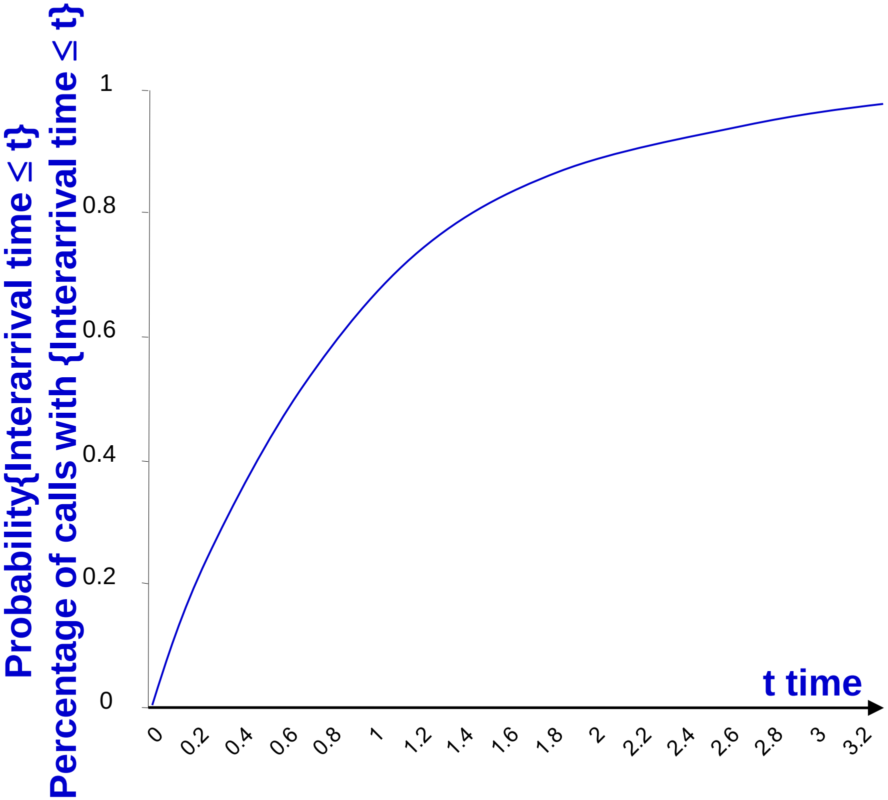




Not stationary over a day, try over an hour or over 30 minutes.

Exponential distribution for Interarrival times



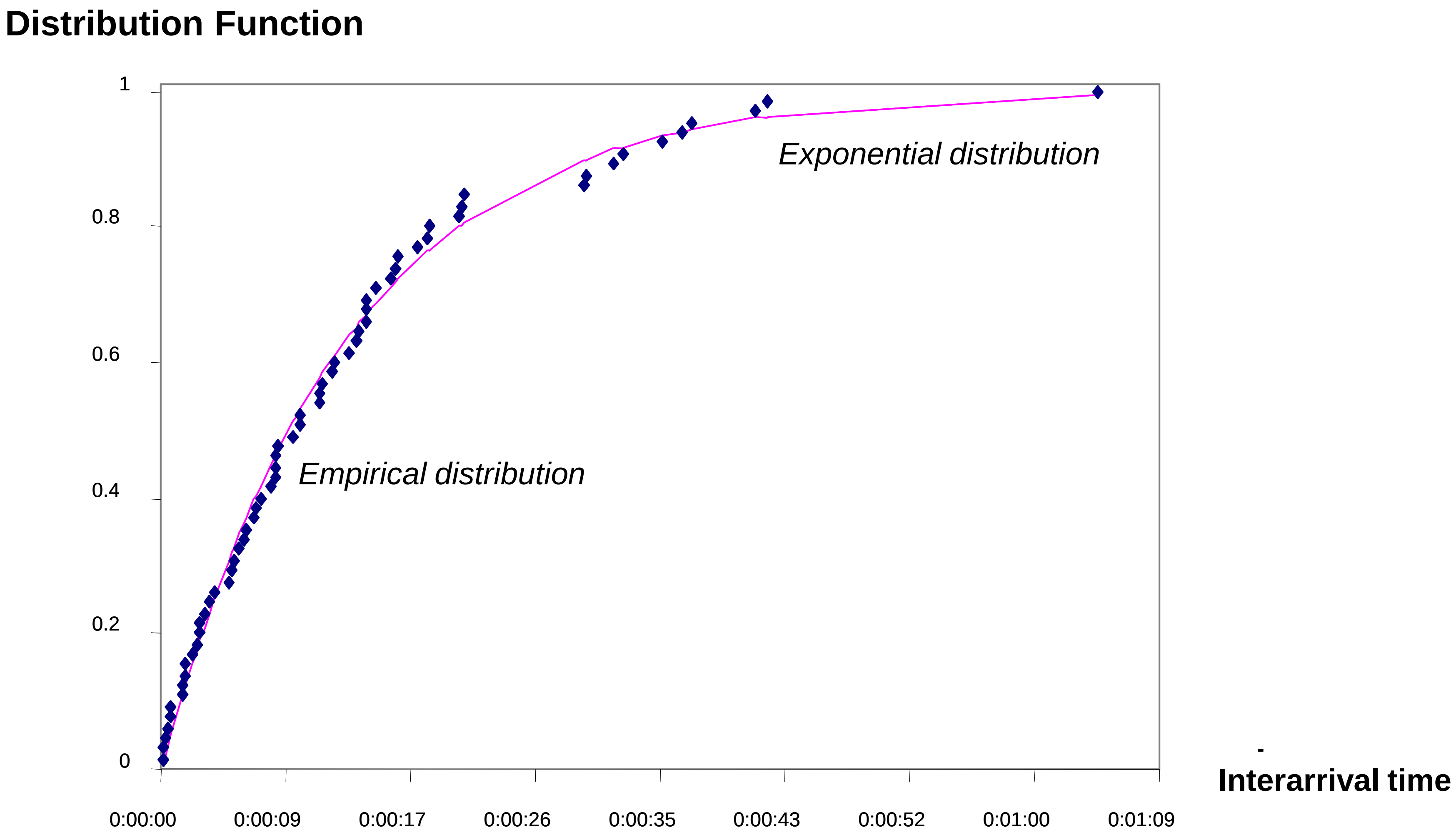


$\mathrm{Prob}(IA \leq t) = 1 - \exp\left(-\frac{t}{a}\right)$, IA interarrival time

$\mathrm{E}(IA) = a$, expected time between two arrivals in a row

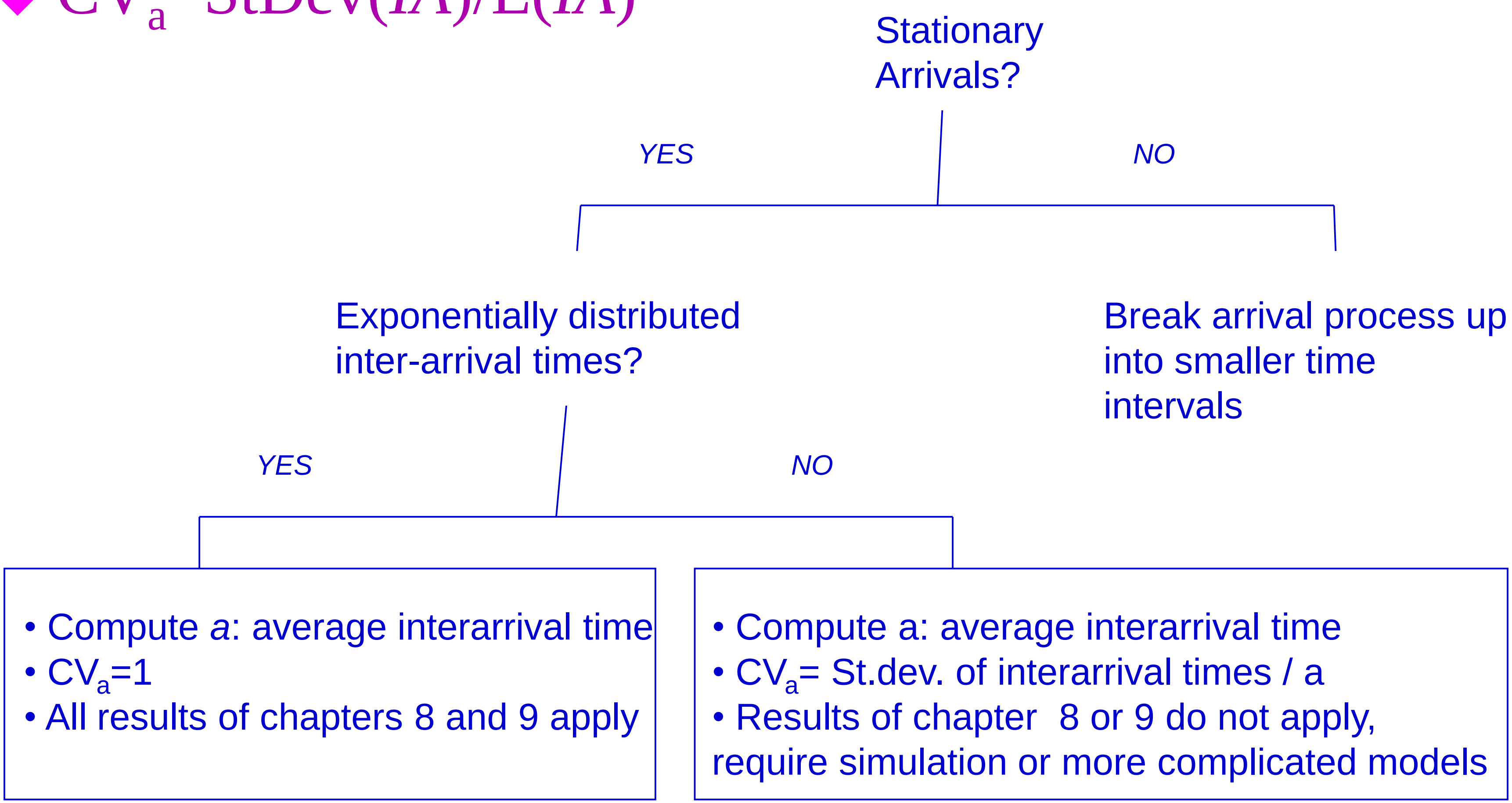
$\mathrm{StDev}(IA) = a$, expected and StDev are the same for exponential distribution

Comparing empirical and theoretical distributions

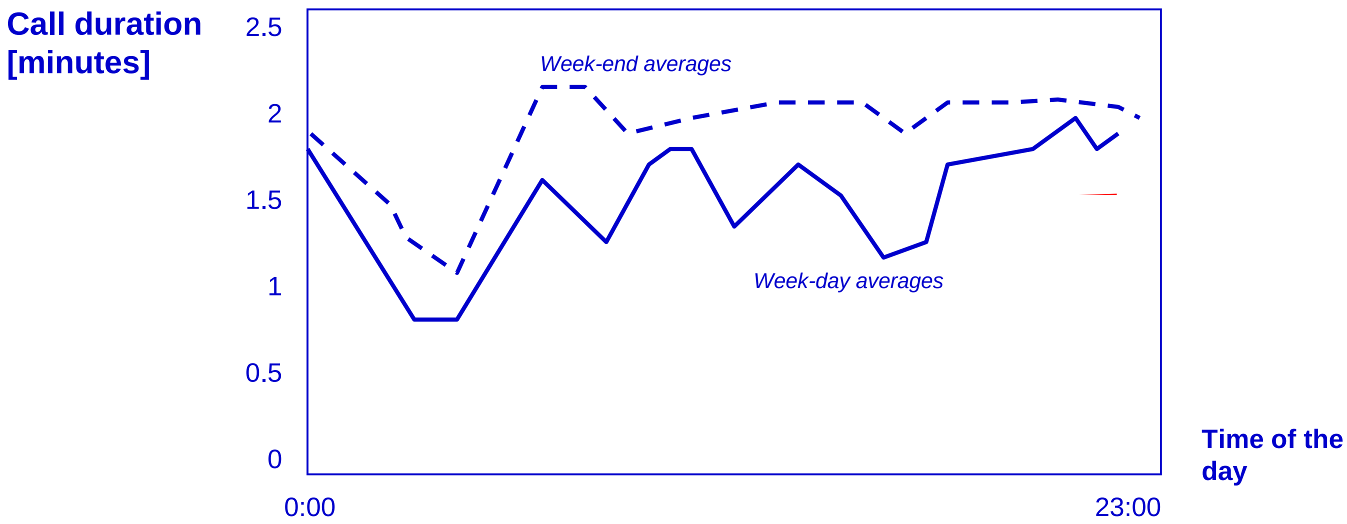


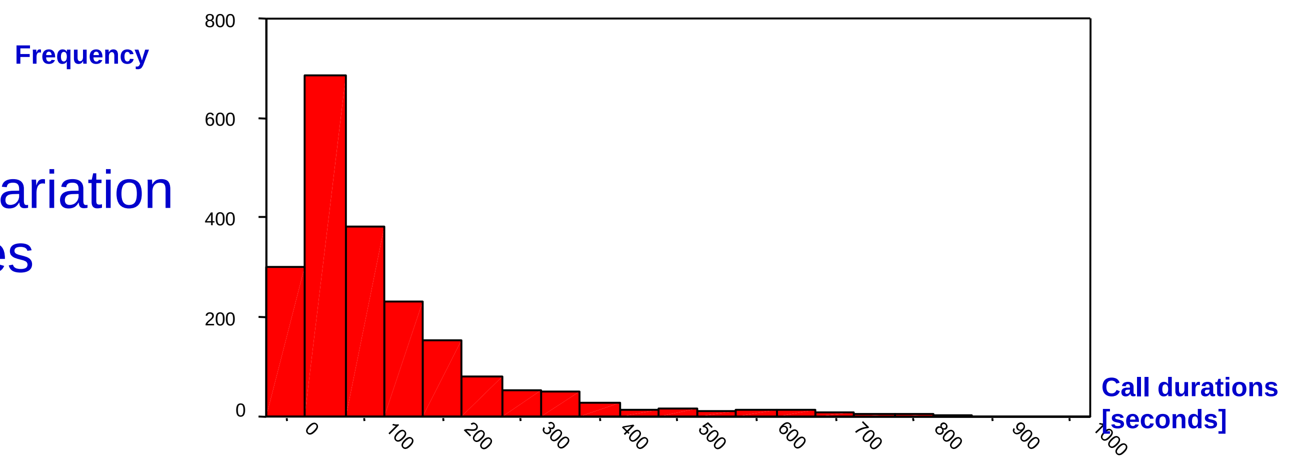
Analyzing the Arrival Process

- $CV_a = StDev(IA)/E(IA)$



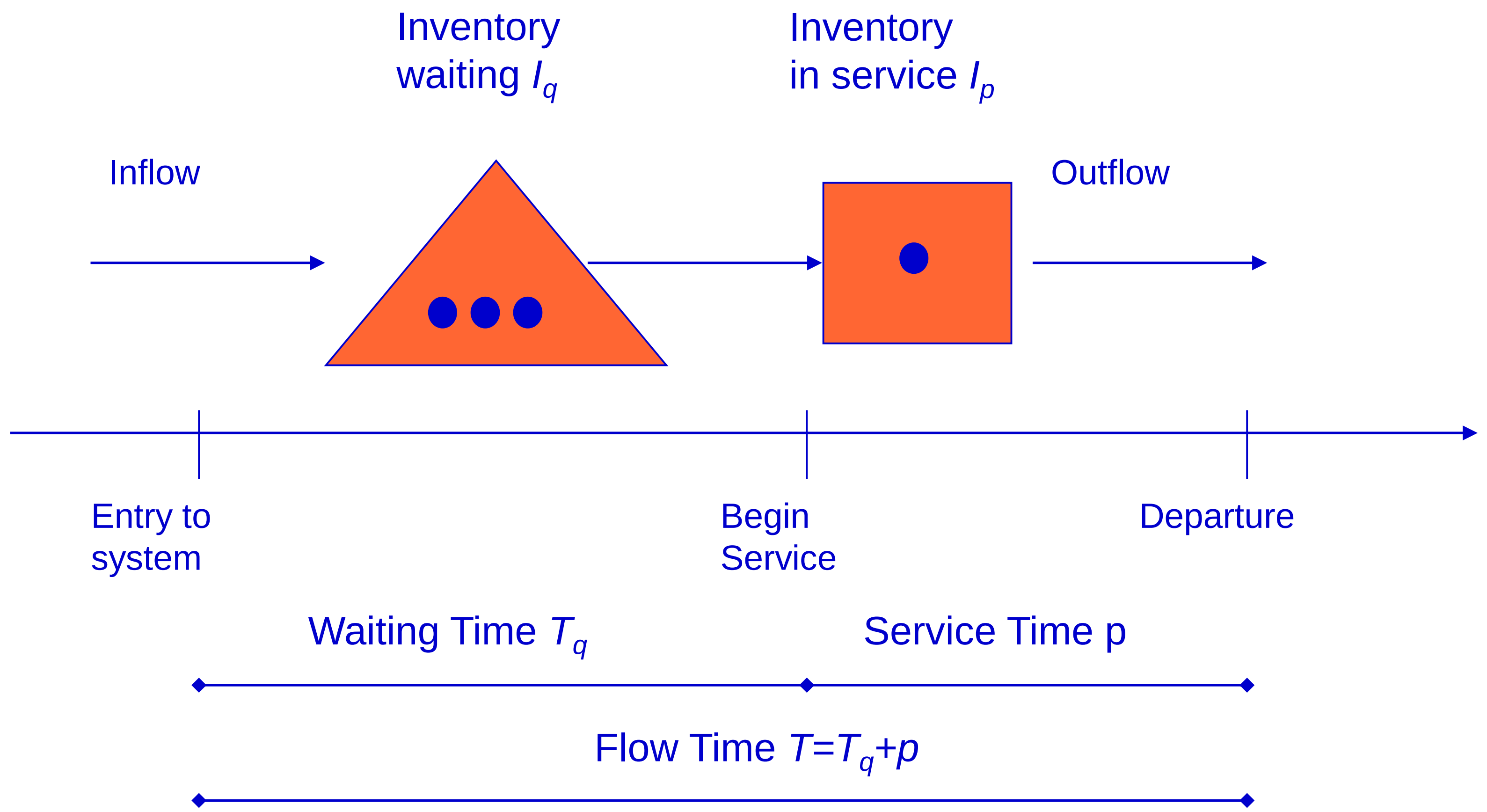
Analyzing the Service Times
Seasonality and Variability





CV_p: Coefficient of variation of service times

Computing the expected waiting time T_q



Utilization

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

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

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

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

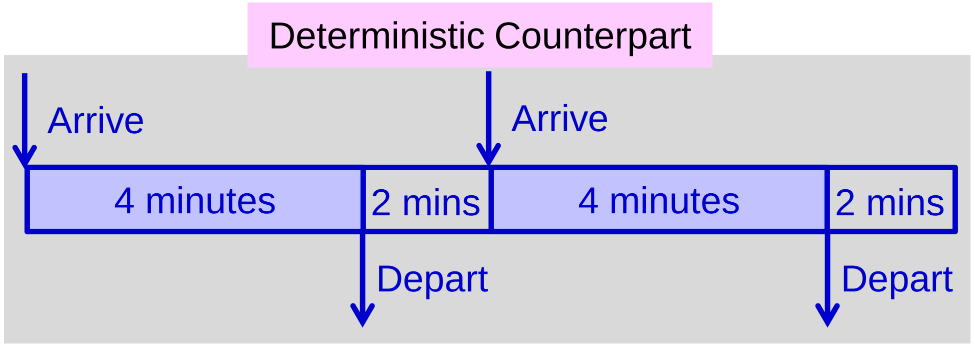
$$\text{Average time in queue} = 90 * \left(\frac{0.3}{1-0.3}\right) * \left(\frac{1^2 + 1.333^2}{2}\right) = 53.57 \sec$$

Waiting Time Formula above is a restatement of Pollaczek-Khinchin (PK) Formula:

$$\text{PK Formula: } T_q = \frac{1}{a}\frac{1}{1-p/a}\frac{\textit{Second moment of activity time}}{2}$$
$$= p\,\frac{p/a}{1-p/a}\,\frac{1}{p^2}\,\frac{p^2+\text{Variance of activity time}}{2}$$
$$= p\,\frac{p/a}{1-p/a}\,\frac{1+CV_p^2}{2} = p\,\frac{u}{1-u}\,\frac{1+CV_p^2}{2}$$

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

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

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

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

d) If teller is busy wp 2/3, outputs 15 per hour.

If teller is idle wp 1/3, outputs 0 per hour.

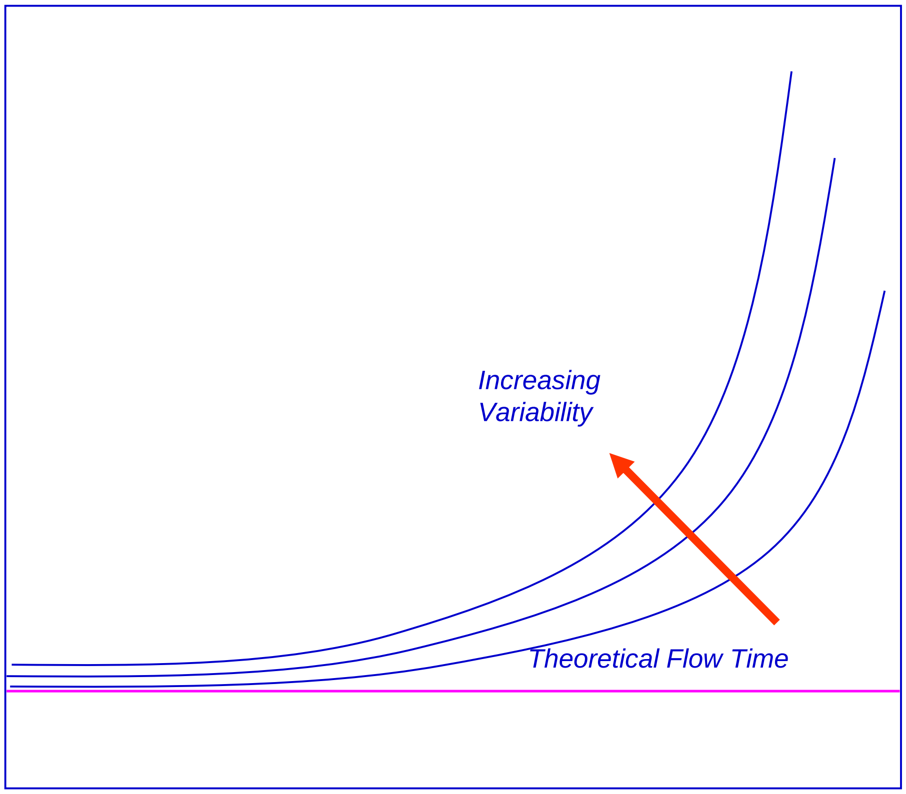
Average output $= (2/3)*15 + (1/3)*0 = 10$ per hour = Average input!!!

The Flow Time Increase Exponentially in Utilization

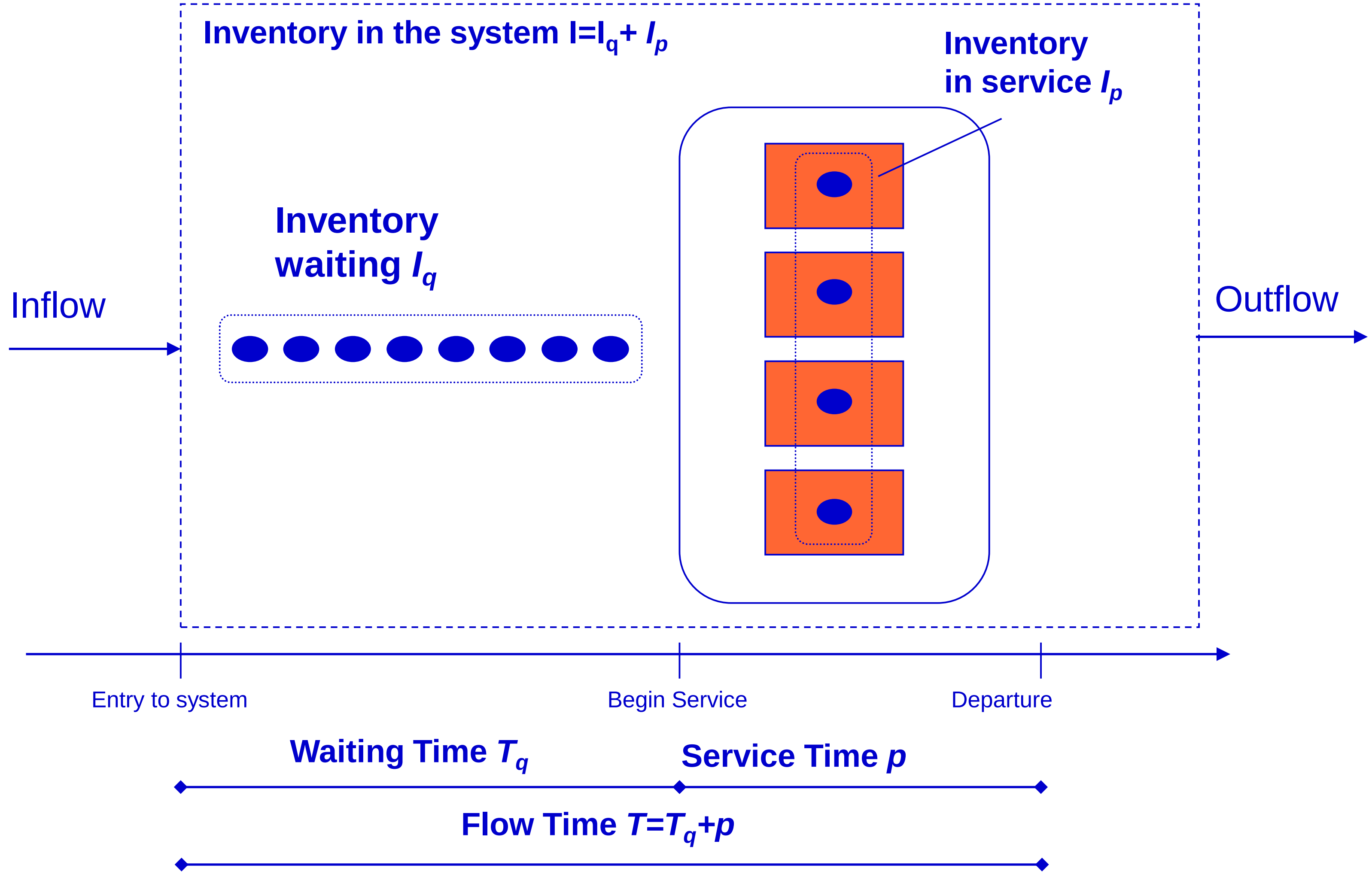
Average flow time T

$$= p + Time \text{ in queue}$$

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



Computing T_q with m Parallel Servers



Utilization with m servers

$$Utilization = u = \frac{FlowRate}{Capacity} = \frac{1/a}{m*(1/p)} = \frac{p}{a\,m}$$

Example: Average Activity time=p=90 seconds
Average Interarrival time=a=11.39 secs over 8-8:15
m=10 servers

Utilization=90/(10x11.39)=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

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

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

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.

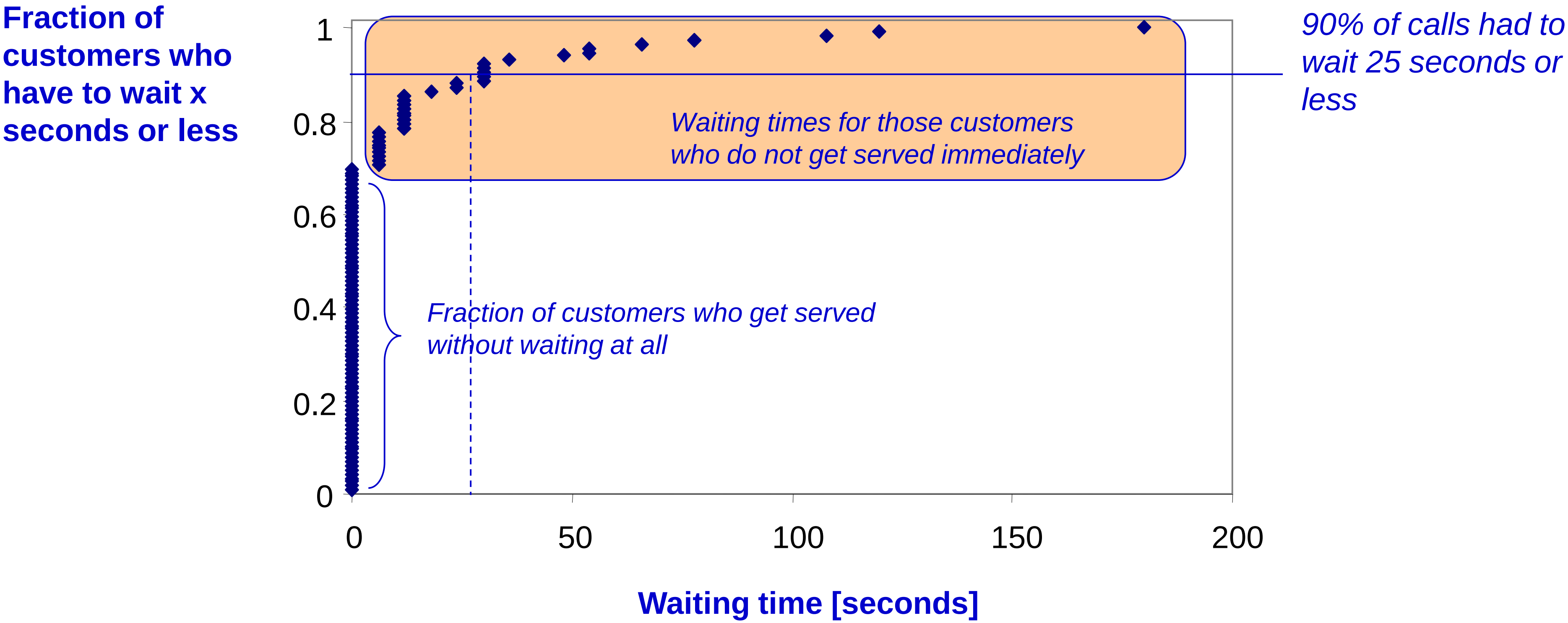
(a) Estimate the average customer wait before being served.

(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



- Target Wait Time (TWT)
- Service Level = Probability{Waiting Time≤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)

7

customerservice.bankofamerica.com

Bank of America **Higher Standards**

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 quantitative feeling of the system
  - analyze specific recommendations / scenarios

- Adding capacity is expensive, although some safety capacity is necessary

- Next case:
  - application to call center
  - 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: $u = \dfrac{p}{a \times m}$

3. Compute expected waiting time

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

4. Based on T_q , we can compute the remaining performance measures as
   Flow time $T=T_q+p$
   Inventory in service $I_p=m*u$
   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

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

Because *ma=p/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

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

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

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

Cost of direct labor per serviced unit
Another example with 9 servers

Ex: m=9. $10/hour wage for each CSR; Activity time=p=90 secs;
Interarrival time=11.39 secs; 1-800 number line charge $0.05 per minute

$$\text{Utilization} = u = \frac{p}{m \times a} = \frac{90}{9 \times 11.39} = 0.878$$

$$\text{Cost of Direct Labor} = \frac{1.5 \text{ min/call } \times (16.66 \text{ cents/min})}{0.878} = 28.474 \text{ cents/call}$$

$$\text{Time in queue} \cong \left(\frac{90}{9}\right) * \left(\frac{0.878^{\sqrt{2(9+1)}-1}}{1-0.878}\right) * \left(\frac{1^2 + 1.333^2}{2}\right) = 10 * 5.218 * 1.39 = 72.54 \sec$$

$$T = T_q + p = 72.54 + 90 = 162.54 \sec = 2.709 \min$$

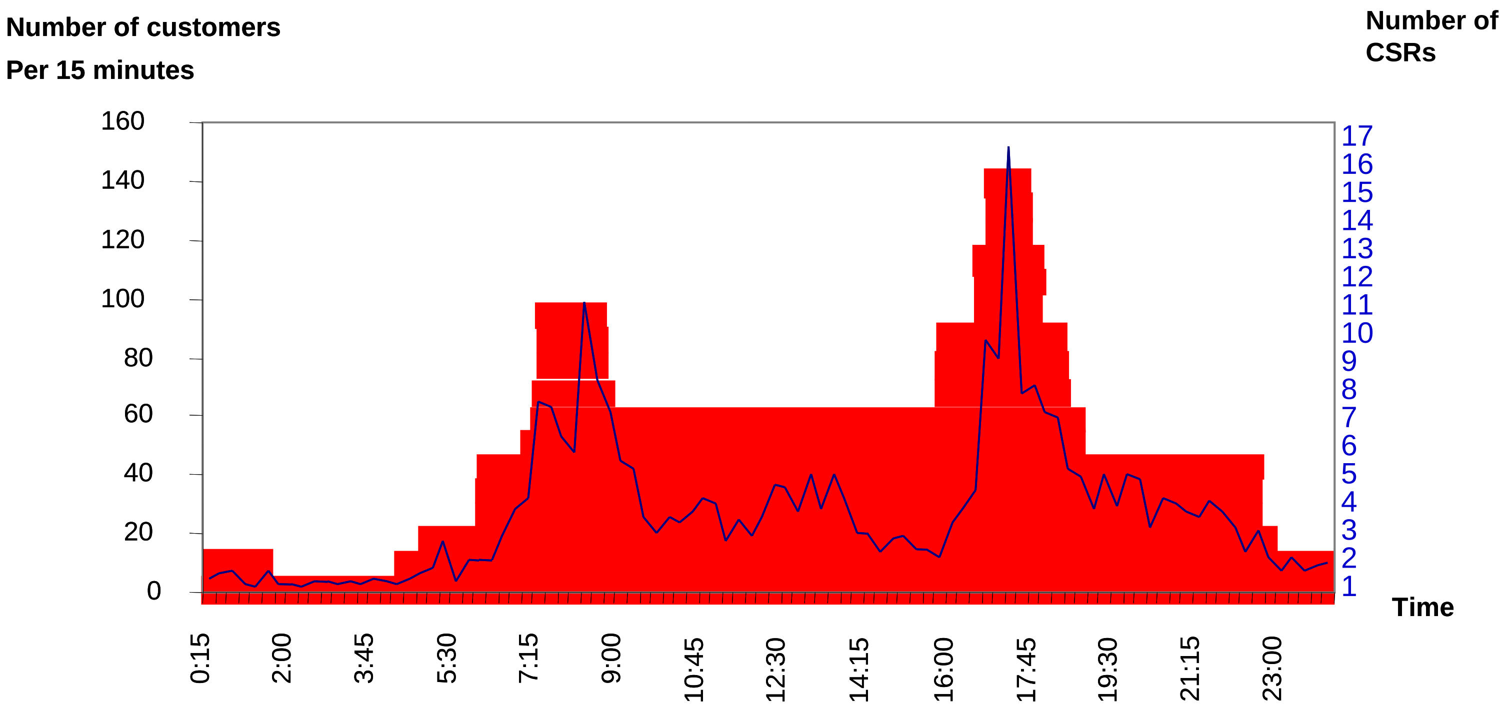
$$\text{With } T = 2.709; \text{ Cost of line charge per call} = 2.709 * 0.05 = \$0.1354 / \text{call}$$

Cost of Staffing Levels

m	u	Labor cost per call	Line cost charge per call	Total cost per call
8				1.3458
9	0.878	0.2847	0.1354	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

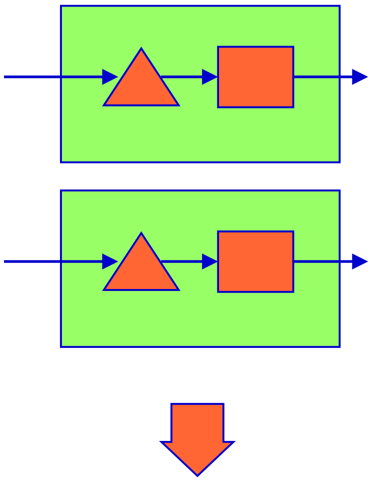
The optimal staffing level m=10

Staffing Levels under various Interarrival Times

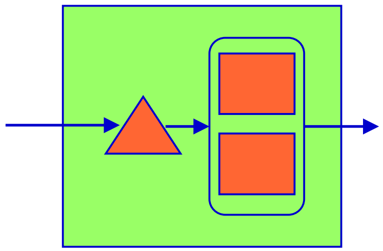


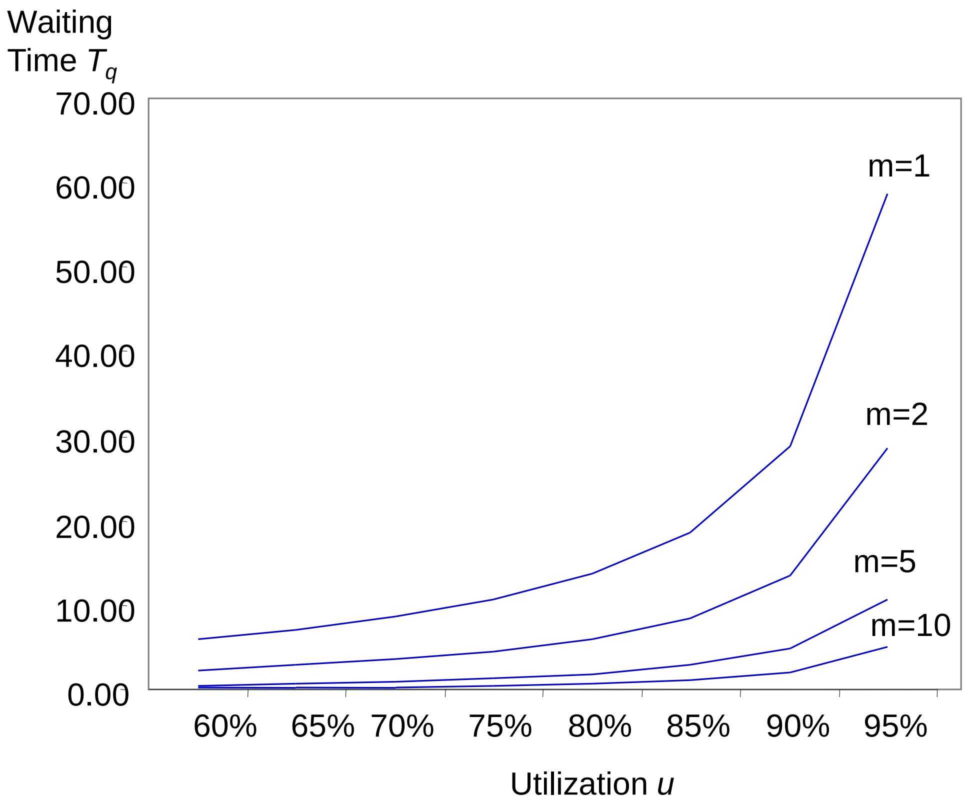
The Power of Pooling

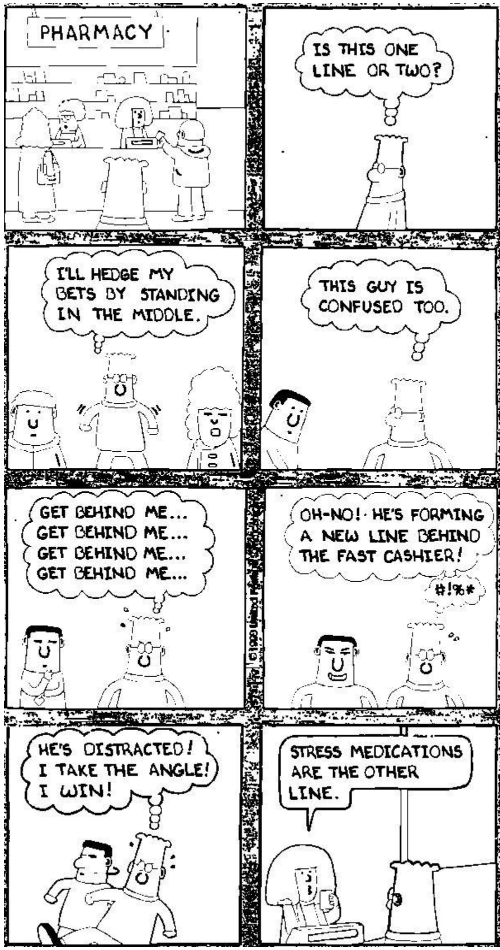
Independent Resources
2x(m=1)



Pooled Resources
(m=2)







Implications:

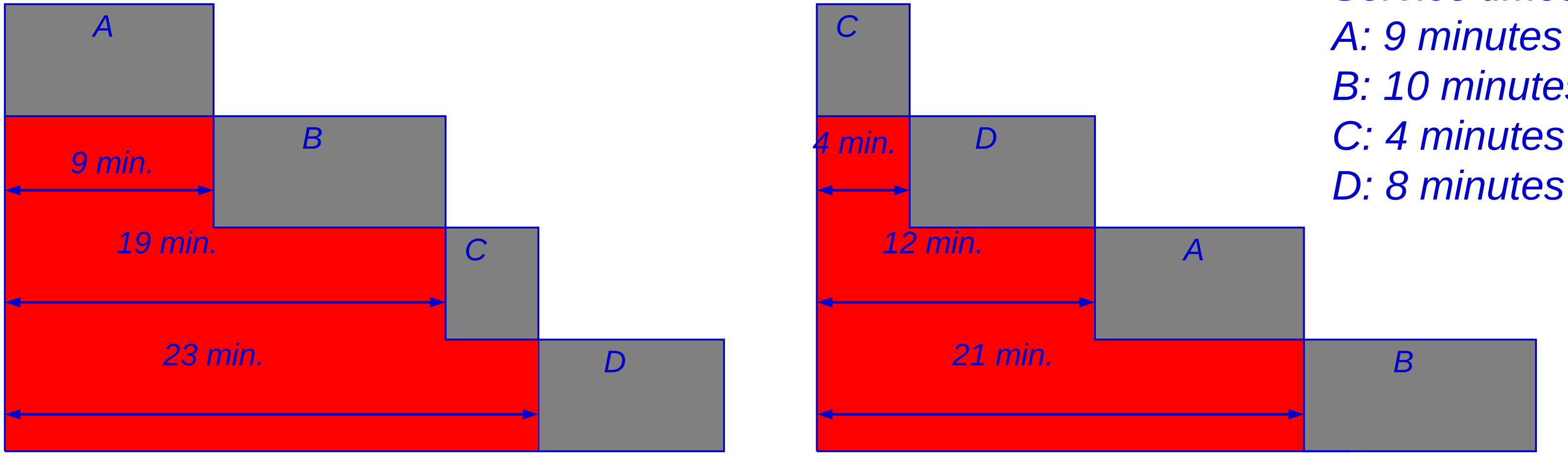
+ balanced utilization

+ 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?



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

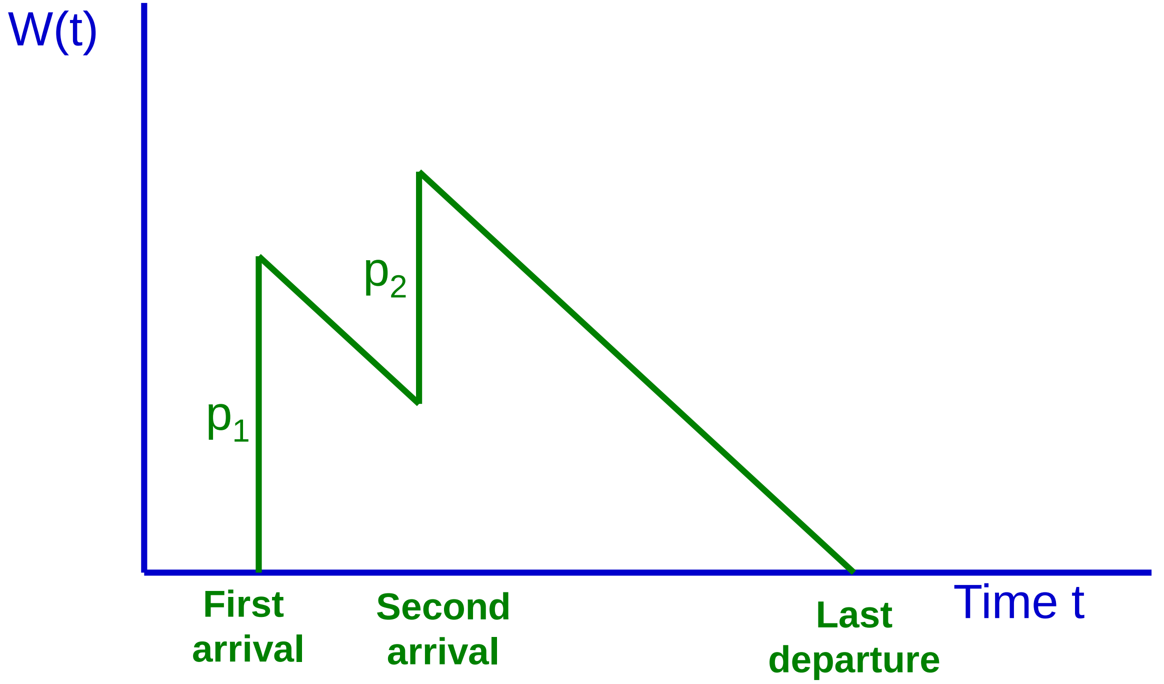
Total wait time: 9+19+23=51min

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

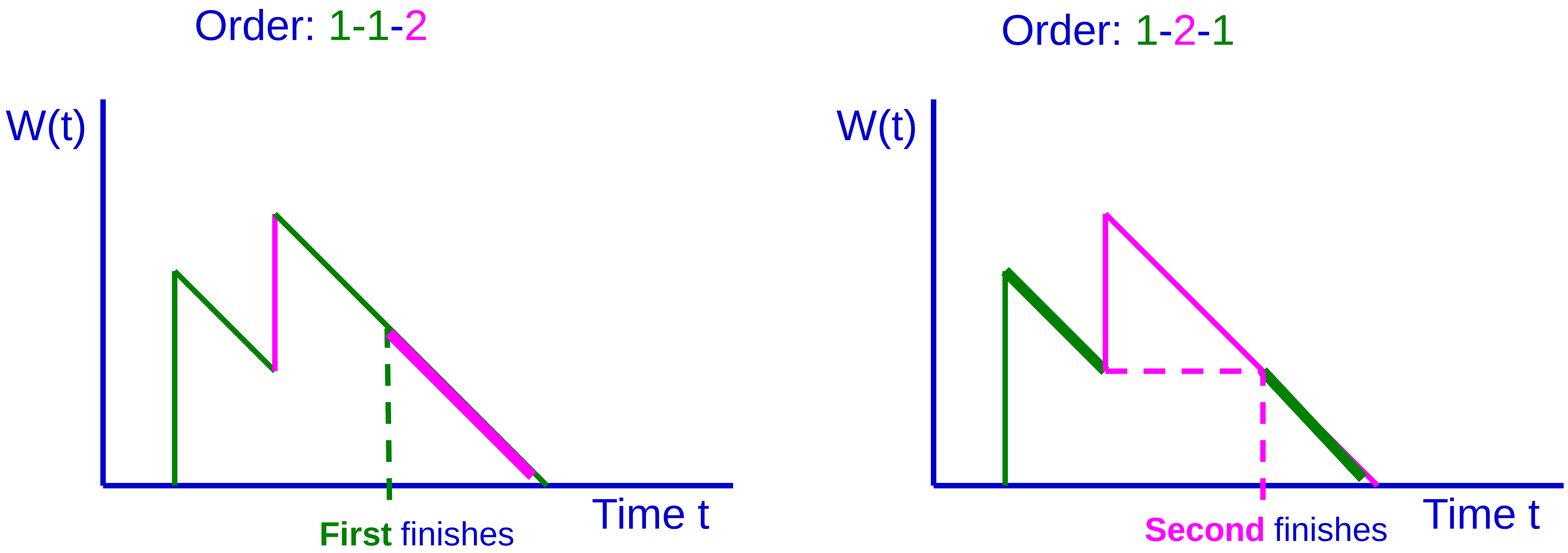
- Problem of having “true” processing times

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



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