Queueing Theory

Dr. Antonio A. Trani

CEE 5614
Analysis of Air Transportation Systems

Fall 2012

Material Presented in this Section

Topics
Queueing Models + Background + Analytic solutions for various disciplines + Applications to infrastructure planning

The importance of queueing models in infrastructure planning and design cannot be overstated.

Queueing models offer a simplified way to analyze critical areas inside an airport terminal to evaluate levels of service and operational performance.

Principles of Queueing Theory

Historically starts with the work of A.K. Erlang while estimating queues for telephone systems

Applications are very numerous:

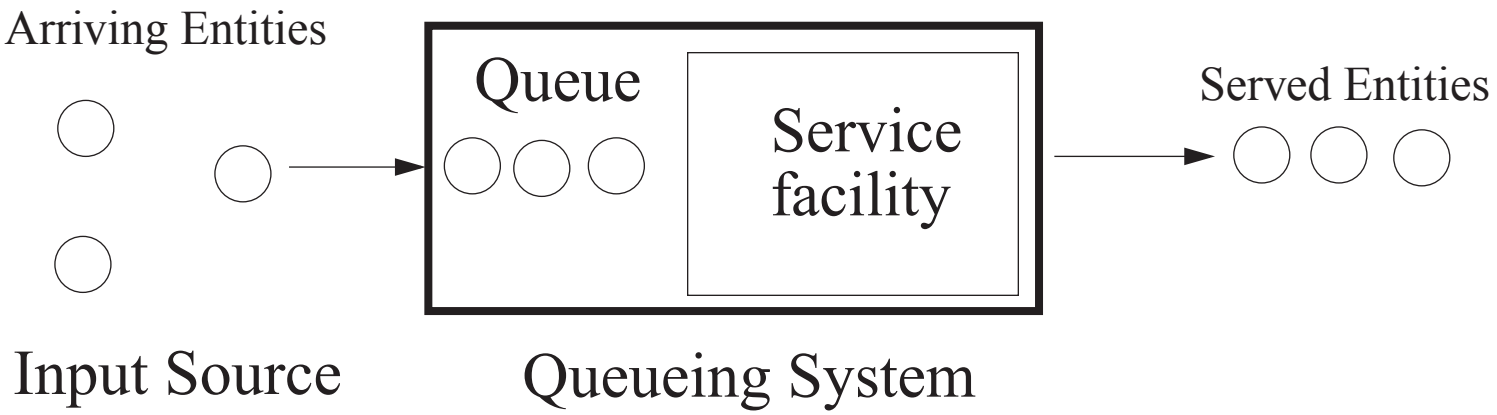
- Transportation planning (vehicle delays in networks)
- Public health facility design (emergency rooms)
- Commerce and industry (waiting line analysis)
- Communications infrastructure (switches and lines)

Elements of a Queue

a) Input Source

b) Queue

c) Service facility



Specification of a Queue

- Size of input source
- Input function
- Queue discipline
- Service discipline
- Service facility configuration
- Output function (distribution of service times)

Sample queue disciplines

- FIFO - first in, first out
- Time-based disciplines
- Priority discipline

What Does a Queue Represent?

Queues represent the state of a system such as the number of people inside an airport terminal, the number of trucks waiting to be loaded at a construction site, the number of ships waiting to be unloaded in a dock, the number of aircraft holding in an imaginary racetrack flight pattern near an airport facility, etc.

The important feature seems to be that the analysis is common to many realistic situations where a flows of traffic (including pedestrians movind inside airport terminals) can be described in terms of either continuous flows or discrete events.

Types of Queues

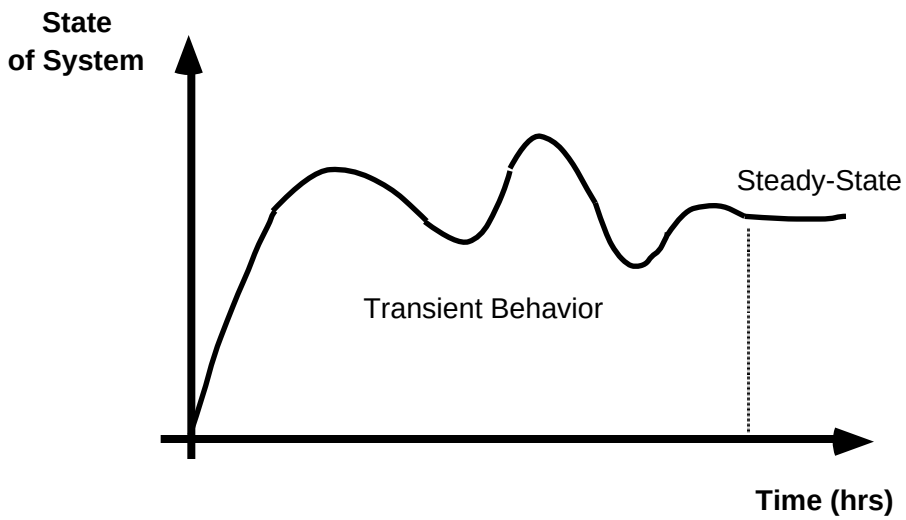
Deterministic queues - Values of arrival funtion are not random variables (continuous flow) but do vary over time.

- Example of this process is the hydrodynamic approximation of pedestrian flows inside airport terminals
- "Bottleneck" analysis in transportation processes employs this technique

Stochastic queues - deal with random variables for arrival and service time functions.

- Queues are defined by probabilistic metrics such as the expected number of entities in the system, probability of *n* entities in the system and so on

Generalized Time Behavior of a Queue



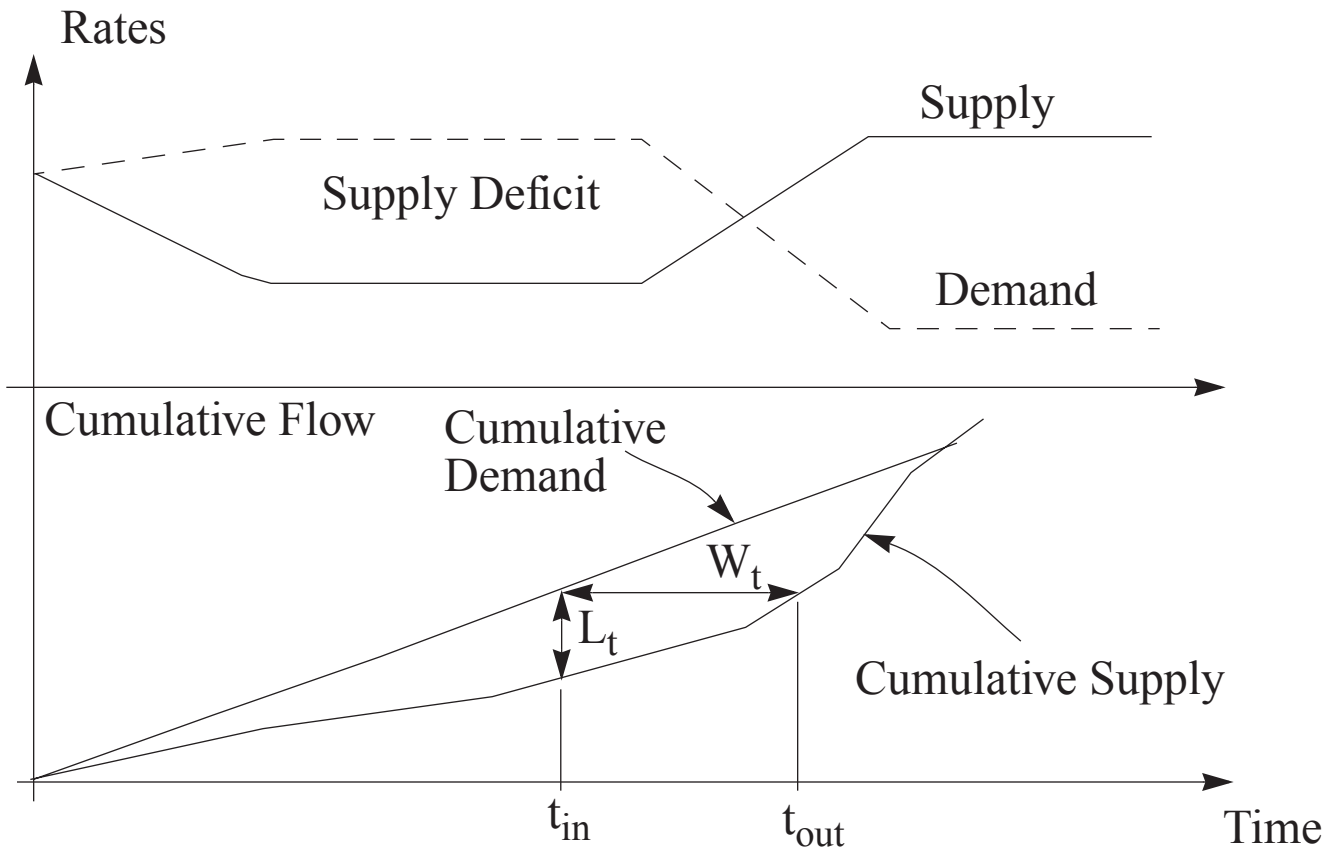
The state of the system goes through two well defined regions of behavior: a) transient and b) steady-state

Deterministic Queues

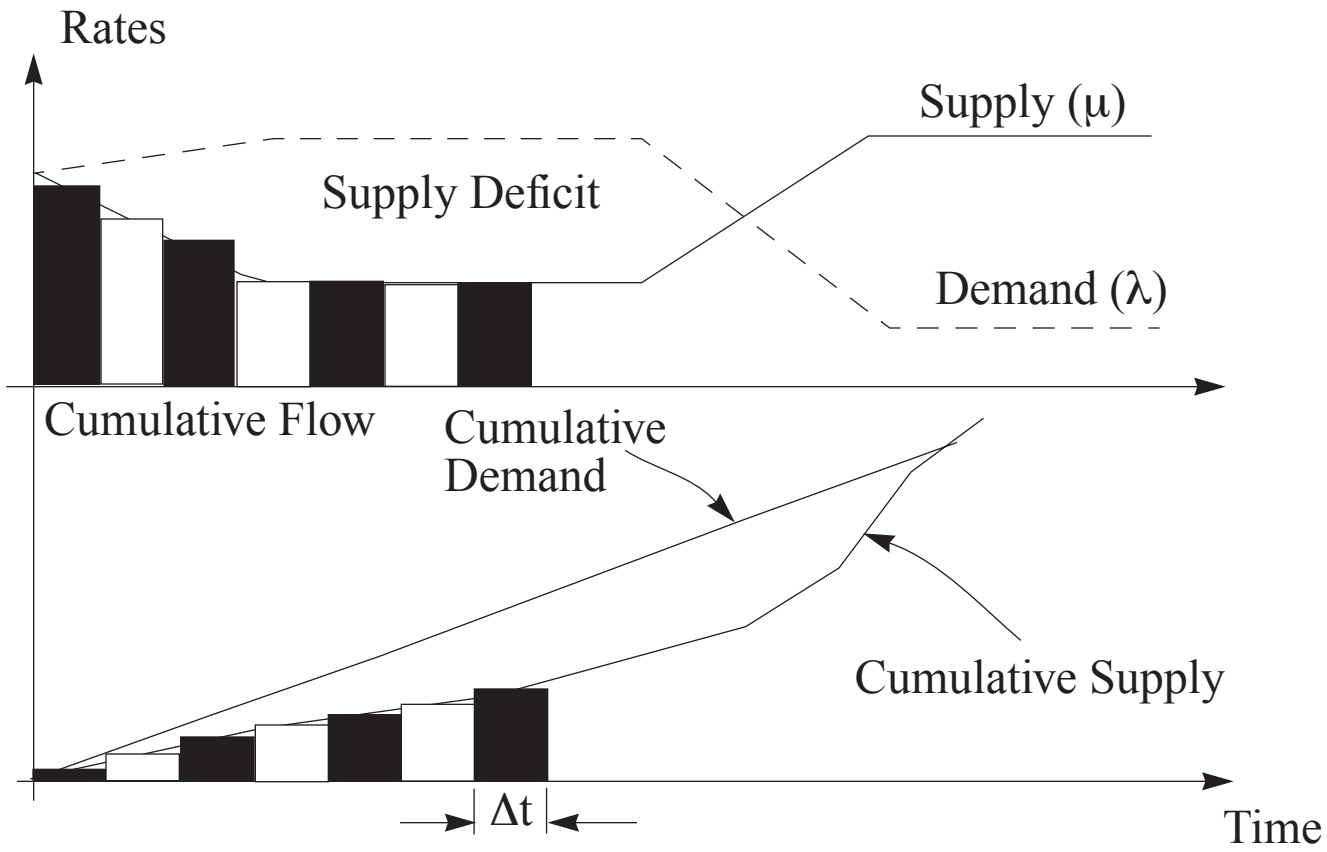
Deterministic Queues are analogous to a continuous flow of entities passing over a point over time. As Morlok [Morlok, 1976] points out this type of analysis is usually carried out when the number of entities to be simulated is large as this will ensure a better match between the resulting cumulative stepped line representing the state of the system and the continuous approximation line

The figure below depicts graphically a deterministic queue characterized by a region where demand exceeds supply for a period of time

Deterministic Queues (Continuous)



Deterministic Queues (Discrete Case)



Deterministic Queues (Parameters)

a) The queue length, L_t, (i.e., state of the system) corresponds to the maximum ordinate distance between the cumulative demand and supply curves

b) The waiting time, W_t, denoted by the horizontal distance between the two cumulative curves in the diagram is the individual waiting time of an entity arriving to the queue at time t_{in}

c) The total delay is the area under bounded by these two curves

d) The average delay time is the quotient of the total delay and the number of entities processed

Deterministic Queues

e) The average queue length is the quotient of the total delay and the time span of the delay (i.e., the time difference between the end and start of the delay)

Assumptions

Demand and supply curves are derived derived from known flow rate functions (λ and μ) which of course are functions of time.

The diagrams shown represent a simplified scenario arising in many practical situations such as those encountered in traffic engineering (i.e., bottleneck analysis).

Fundamental Equations Deterministic Queueing

$$\frac{dL}{dt} = \lambda(t) - \mu(t)$$

$\lambda(t) =$ arrival rate (entities/time unit)

$\mu(t) =$ service rate (entities/time unit)

$\lambda(t) =$ demand function

$\mu(t) =$ service function

In finite difference form we can solve the equation for L_t numerically

$$L_t = L_{t-\Delta t} + \left(\frac{dL}{dt}\right)\Delta t$$

$$L_t = L_{t-\Delta t} + \left(\lambda(t) - \mu(t)\right)\Delta t$$

Example : Freeway Bottleneck Analysis

A four lane freeway has a total directional demand of 4,000 veh/hr during the morning peak period. One day an accident occurs at the freeway that blocks the right-hand side lane for 30 minutes (at time t=1.0 hours). The capacity per lane is 2,200 veh/hr.

a) Find the maximum number of cars queued.

b) Find the average delay imposed to all cars during the queue.

Mathematical Equations to Work the Problem

$$\lambda(t) = 4000 \ \forall \ \text{t}$$

$$\mu(t) = \begin{cases} 4400 \text{ if t} < 1.0 \text{ hrs} \\ 2200 \text{ if } 1 <= \text{t} < 1.5 \text{ hrs} \\ 4400 \text{ if t} > 1.5 \text{ hrs} \end{cases}$$

$$\frac{dL}{dt} = \lambda(t) - \mu(t)$$

First-order differential equation to be solved

In finite difference form we can solve the equation for L_t numerically

$$L_t = L_{t-\Delta t} + \left(\frac{dL}{dt}\right)\Delta t$$

$$L_t = L_{t-\Delta t} + \left(\lambda(t) - \mu(t)\right)\Delta t$$

Hand Calculations and Solution

- Integrate the values of demand and capacity in a piecewise fashion (over time)
- For example for interval 0 < t <1.0 hrs

$$\int_0^{1.0} \lambda(t)\,dt = 4000t + C_1$$

$C_1 = 0$ from initial conditions (at t=0 queue is zero)

$$\int_0^{1.0} \mu(t)\,dt = 4400t + C_2$$

$C_2 = 0$ from initial conditions (at t=0 queue is zero)

Hand Calculations and Solution

- For interval 1.0 <= t <1.5 hrs

$$\int_{1.0}^{1.5} \lambda(t)dt = 4000t$$

$$\int_{1.0}^{1.5} \mu(t)dt = 2200t + C_3$$

$$\int_{1.0}^{1.5} \mu(t)dt = 2200t + 1800$$

$C_3 = 1800$ from initial conditions (at t=1 there are 4000 cars that arrived)

Hand Calculations and Solution

- For interval 1.5 <= t <3.75 hrs

$$\int_{1.5}^{3.75} \lambda(t)\,dt = 4000t$$

$$\int_{1.5}^{3.75} \mu(t)\,dt = 4400t + C_4$$

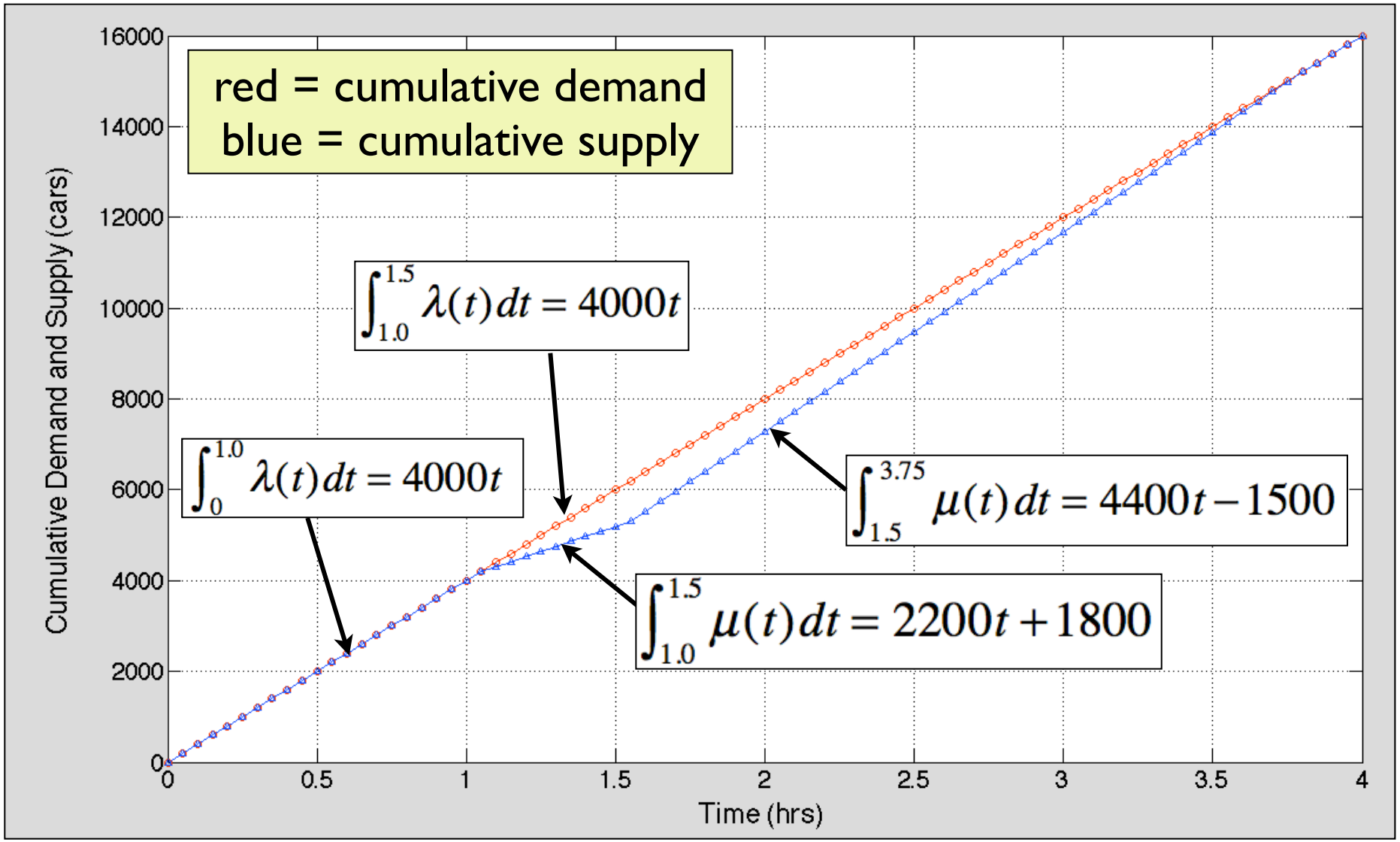
$$\int_{1.5}^{3.75} \mu(t)\,dt = 4400t - 1500$$

$C_4 = -1500$ from initial conditions (at t=1.5 there are 5100 cars that arrived)

$$4000 + 2200(0.5) = 4000t + C_4$$

$$C_4 = -1500$$

Cumulative Flow Diagram (Integral of demand and supply functions)



Excel Numerical Solution

Time (hours)	L (t) cars	lambda(t) cars/hr	mu(t) cars/hr	lambda(t) - mu(t) cars/hr	Δt Hours	[lambda(t) - mu(t)] Δt cars-hr
0.00	0	4000	4400	-400	0.05	-20
0.05	0	4000	4400	-400	0.05	-20
0.10	0	4000	4400	-400	0.05	-20
0.15	0	4000	4400	-400	0.05	-20
0.20	0	4000	4400	-400	0.05	-20
0.25	0	4000	4400	-400	0.05	-20
0.30	0	4000	4400	-400	0.05	-20
0.35	0	4000	4400	-400	0.05	-20
0.40	0	4000	4400	-400	0.05	-20
0.45	0	4000	4400	-400	0.05	-20
0.50	0	4000	4400	-400	0.05	-20
0.55	0	4000	4400	-400	0.05	-20
0.60	0	4000	4400	-400	0.05	-20
0.65	0	4000	4400	-400	0.05	-20
0.70	0	4000	4400	-400	0.05	-20
0.75	0	4000	4400	-400	0.05	-20
0.80	0	4000	4400	-400	0.05	-20
0.85	0	4000	4400	-400	0.05	-20
0.90	0	4000	4400	-400	0.05	-20
0.95	0	4000	4400	-400	0.05	-20
1.00	0	4000	4400	-400	0.05	-20
1.05	0	4000	2200	1800	0.05	90
1.10	90	4000	2200	1800	0.05	90
1.15	180	4000	2200	1800	0.05	90
1.20	270	4000	2200	1800	0.05	90
1.25	360	4000	2200	1800	0.05	90
1.30	450	4000	2200	1800	0.05	90
1.35	540	4000	2200	1800	0.05	90
1.40	630	4000	2200	1800	0.05	90
1.45	720	4000	2200	1800	0.05	90
1.50	810	4000	2200	1800	0.05	90
1.55	900	4000	4400	-400	0.05	-20
1.60	880	4000	4400	-400	0.05	-20

$\lambda(t)$

$\mu(t)$

$$\left(\lambda(t)-\mu(t)\right)\Delta t$$

$$\frac{dL}{dt}=\lambda(t)-\mu(t)$$

$$L_t = L_{t-\Delta t} + \left(\lambda(t)-\mu(t)\right)\Delta t$$

Observations

- We used the Euler numerical integration algorithm (fixed step size)
- When the value of:

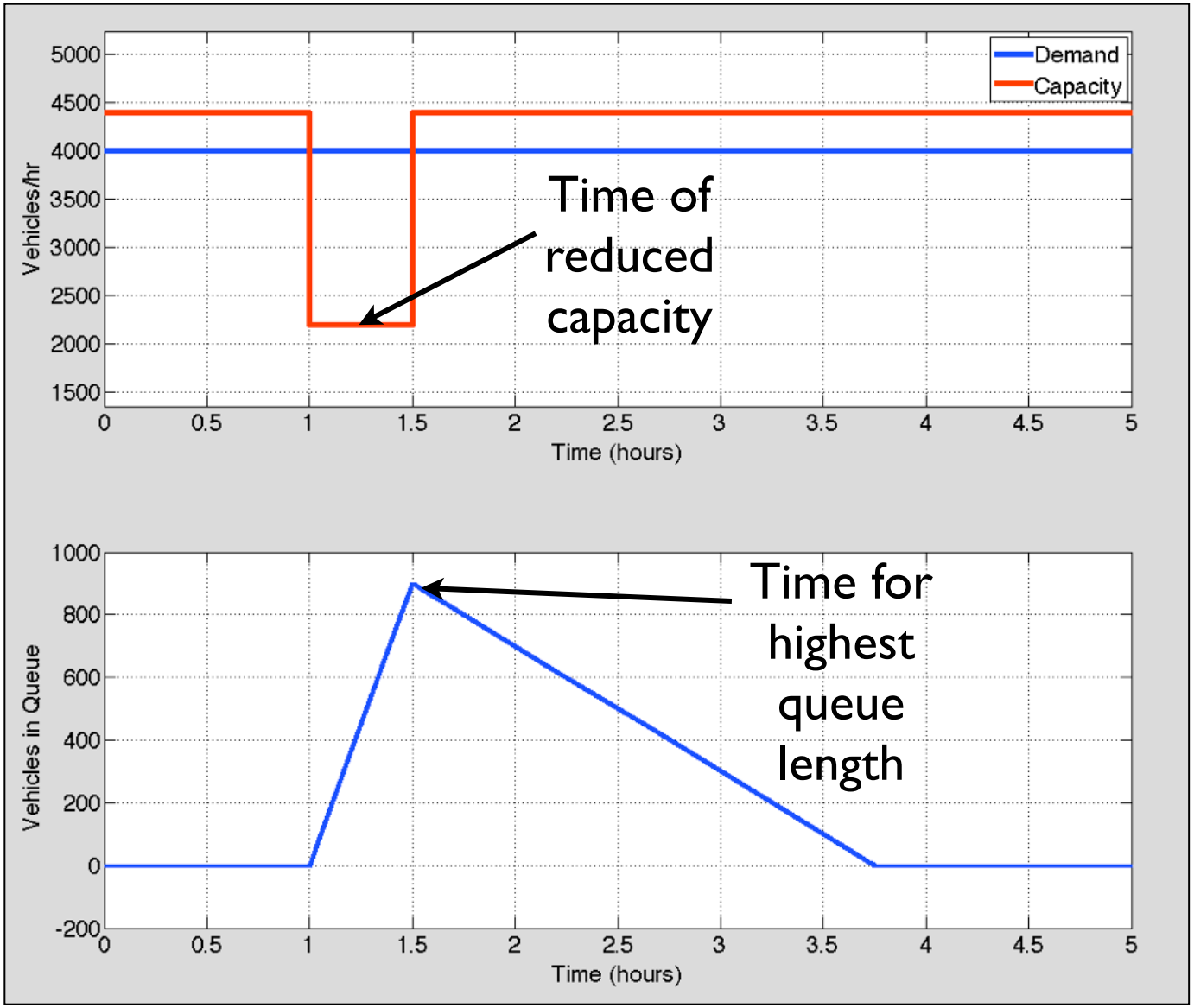
$$(\lambda(t) - \mu(t))\,\Delta t < 0$$

- is negative, we do not add the value to the queue length (queues cannot be negative)
- The queue starts at t = 1.0 hours
- The queue length peaks at t = 1.5 hours
- The queue ends at t = 3.75 hrs

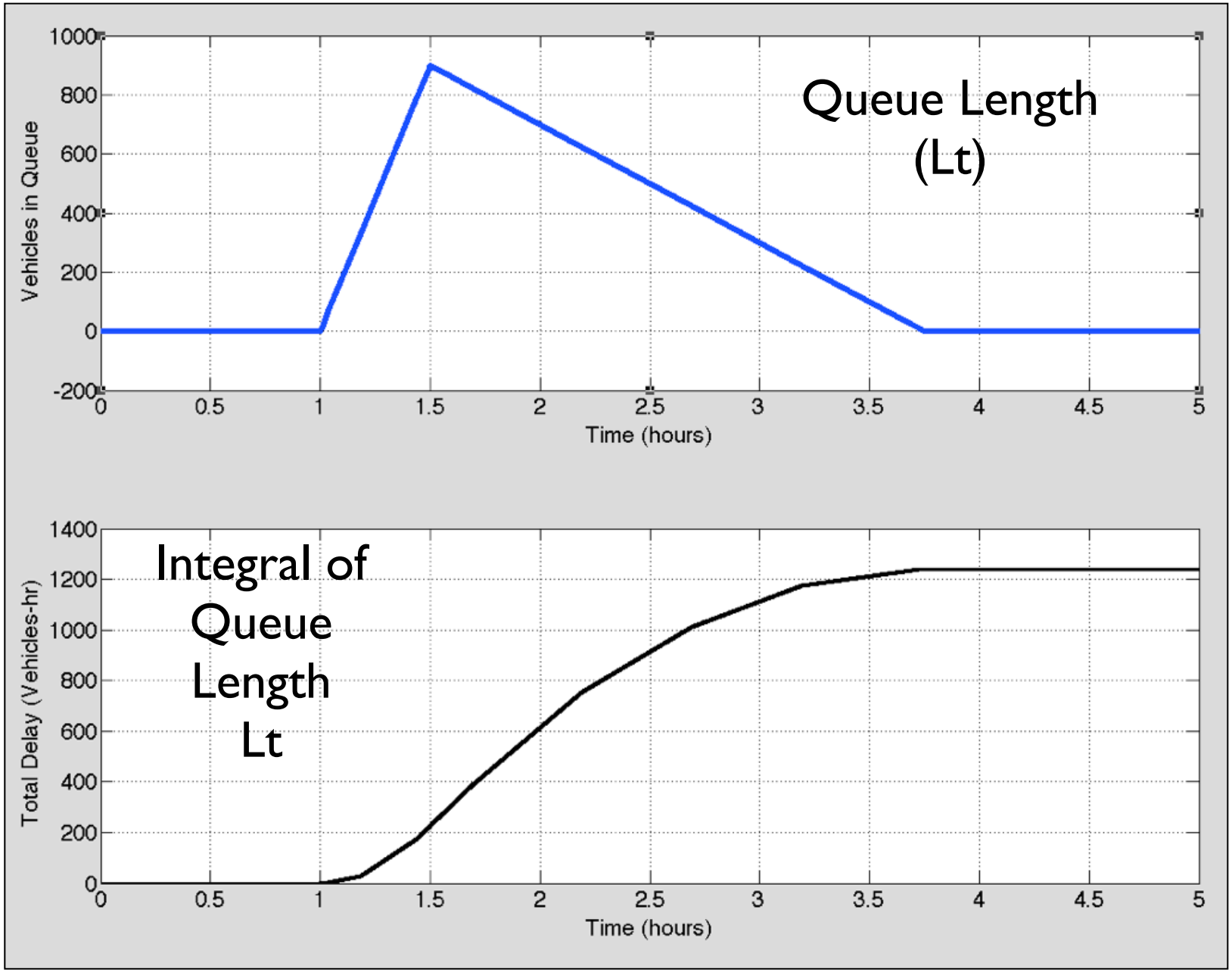
Graphical Solution to Highway "Bottleneck" Problem

Flow Rates (demand and supply)

Integral of (demand - supply) over time



Graphical Solution to Highway



Solutions

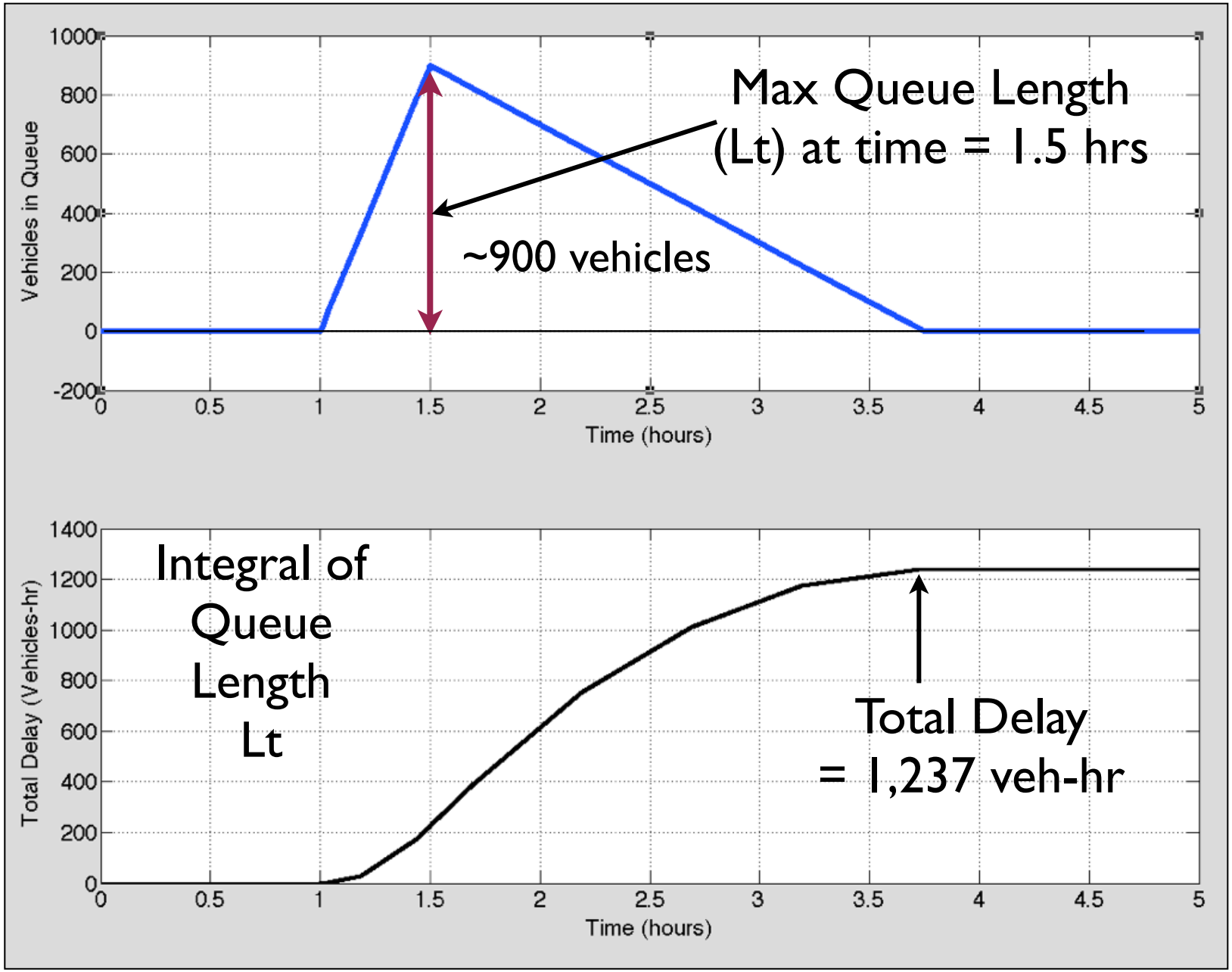
a) Find the maximum number of cars queued

By inspection the maximum number of cars queueing at the bottleneck are 900 passengers.

b) Find the average delay imposed to all cars during the queue.

Calculate the area under the second curve (in the previous slide) and then divide by the number of cars that were delayed

Graphical Solution to Highway



Some Statistics about the Problem

Average arrival rate (cars/hr) = 4000

Average capacity (cars/hr) = 3771

Simulation Period (hours) = 5 (hours)

Total delay (car-hr) = 1237

Max queue length (cars) = 900

Example : Lumped Service Model (Passengers at a Terminal Facility)

In the planning program for renovating an airport terminal facility it is important to estimate the requirements for the ground access area. It has been estimated that an hourly capacity of 1500 passengers can be adequately be served with the existing facilities at a medium size regional airport.

Due to future expansion plans for the terminal, one third of the ground service area will be closed for 2 hours in order to perform inspection checks to ensure expansion compatibility. A recent passenger count reveals an arrival function as shown below.

Example Problem (Airport Terminal)

$\lambda = 1500$ for $0 < t < 1$ t in hours

$\lambda = 500$ for $t > 1$

where, λ is the arrival function (demand function) and t is the time in hours. Estimate the following parameters:

- The maximum queue length, $L(t)_{max}$
- The total delay to passengers, T_d
- The average length of queue, L
- The average waiting time, W
- The delay to a passenger arriving 30 minutes hour after the terminal closure

Example Problem (Airport Terminal)

Solution:

The demand function has been given explicitly in the statement of the problem. The supply function as stated in the problem are deduced to be,

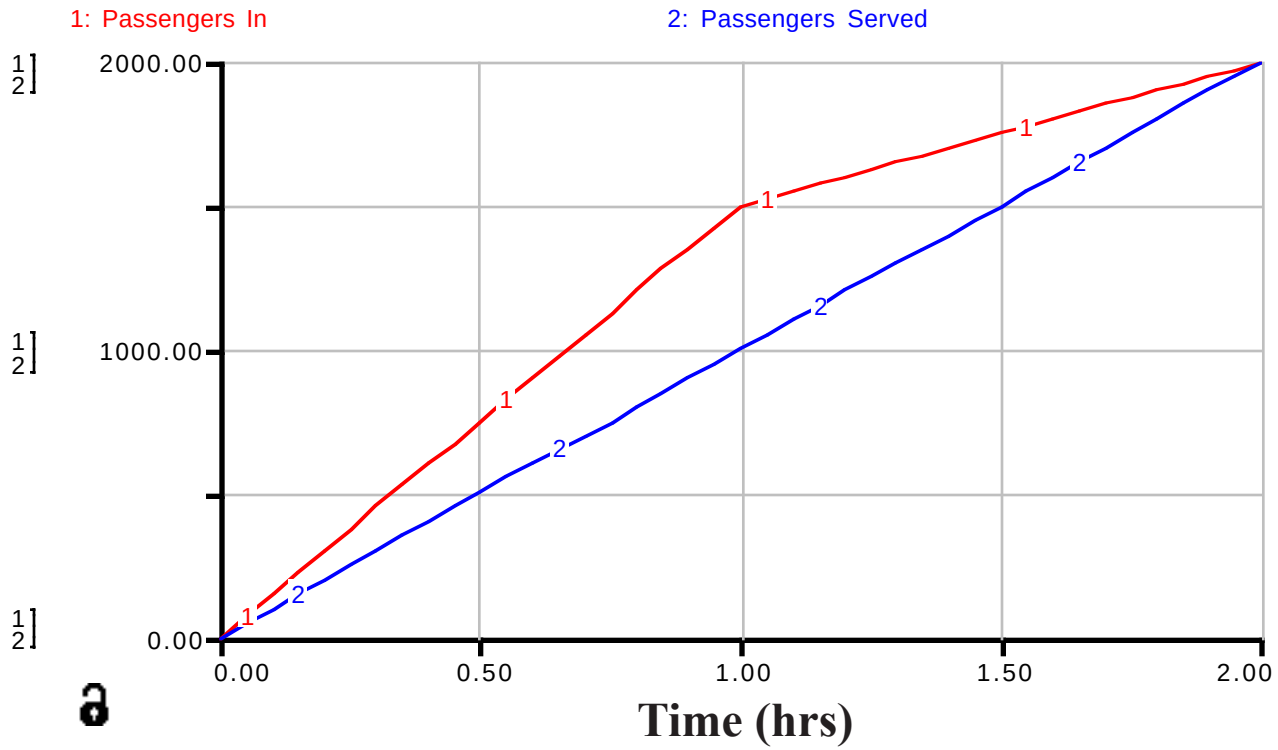
$\mu = 1000$ if $t < 2$

$\mu = 1500$ if $t > 2$

Plotting the demand and supply functions might help understanding the problem

Example Problem (Airport Terminal)

Demand and Supply Functions for Example Problem



Example Problem (Airport Terminal)

To find the average queue length (L) during the period of interest, we evaluate the total area under the cumulative curves (to find total delay)

$T_d = 2\ [(1/2)(1500\text{-}1000)] = 500$ passengers-hour

Find the maximum number of passengers in the queue, $L(t)_{max}$,

$L(t)_{max} = 1500 - 1000 = 500$ passengers at time t=1.0 hours

Find the average delay to a passenger (W)

Example Problem (Airport Terminal)

$$W = \frac{T_d}{N_d} = 15 \text{ minutes}$$

where, T_d is the total delay and N_d is the number of passengers that where delayed during the queueing incident.

$$L = \frac{T_d}{t_q} = 250 \text{ passengers}$$

where, T_d is the total delay and t_d is the time that the queue lasts.

Example Problem (Airport Terminal)

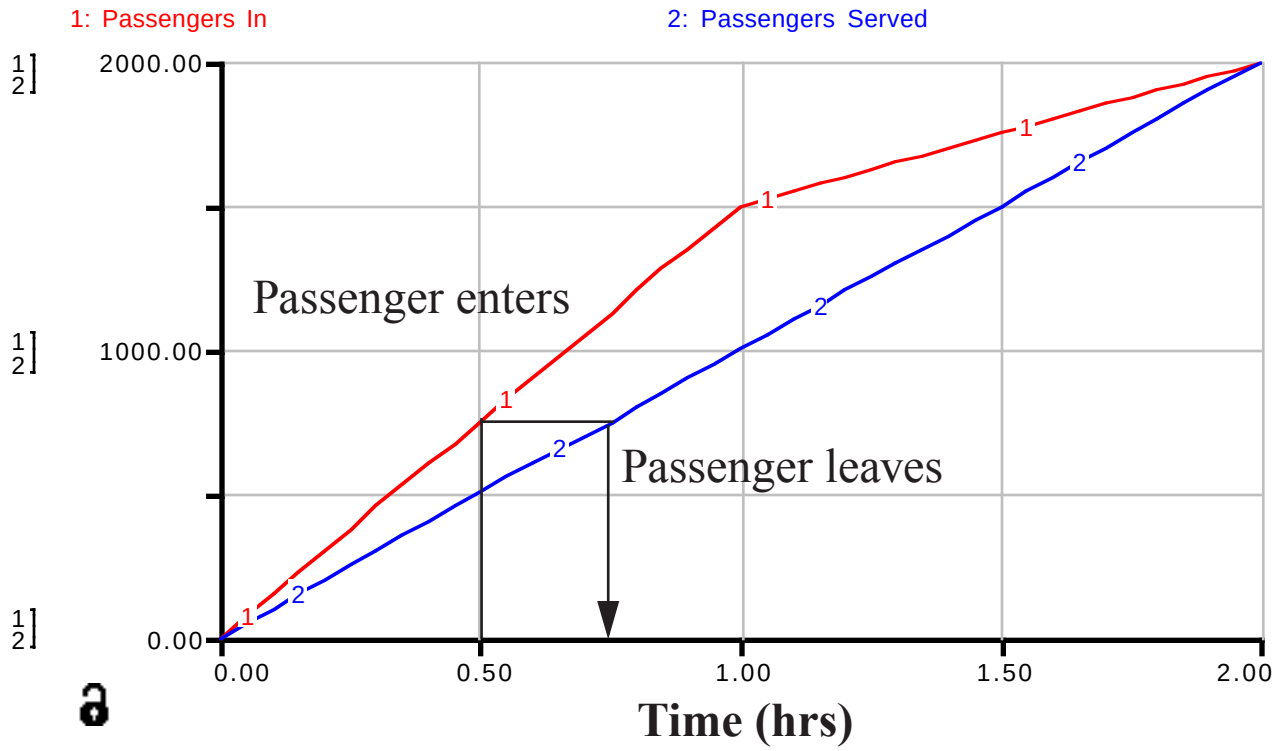
Now we can find the delay for a passenger entering the terminal 30 minutes after the partial terminal closure occurs. Note that at t = 0.5 hours 750 passengers have entered the terminal before the passenger in question. Thus we need to find the time when the supply function, μ(t), achieves a value of 750 so that the passenger "gets serviced". This occurs at,

$$\mu(t + \Delta t) = \lambda(t) = 750 \tag{2.1}$$

therefore Δt is just 15 minutes (the passenger actually leaves the terminal at a time t+Δt equal to 0.75 hours). This can be shown in the diagram on the next page.

Example Problem (Airport Terminal)

Demand and Supply Functions for Example Problem



Remarks About Deterministic Queues

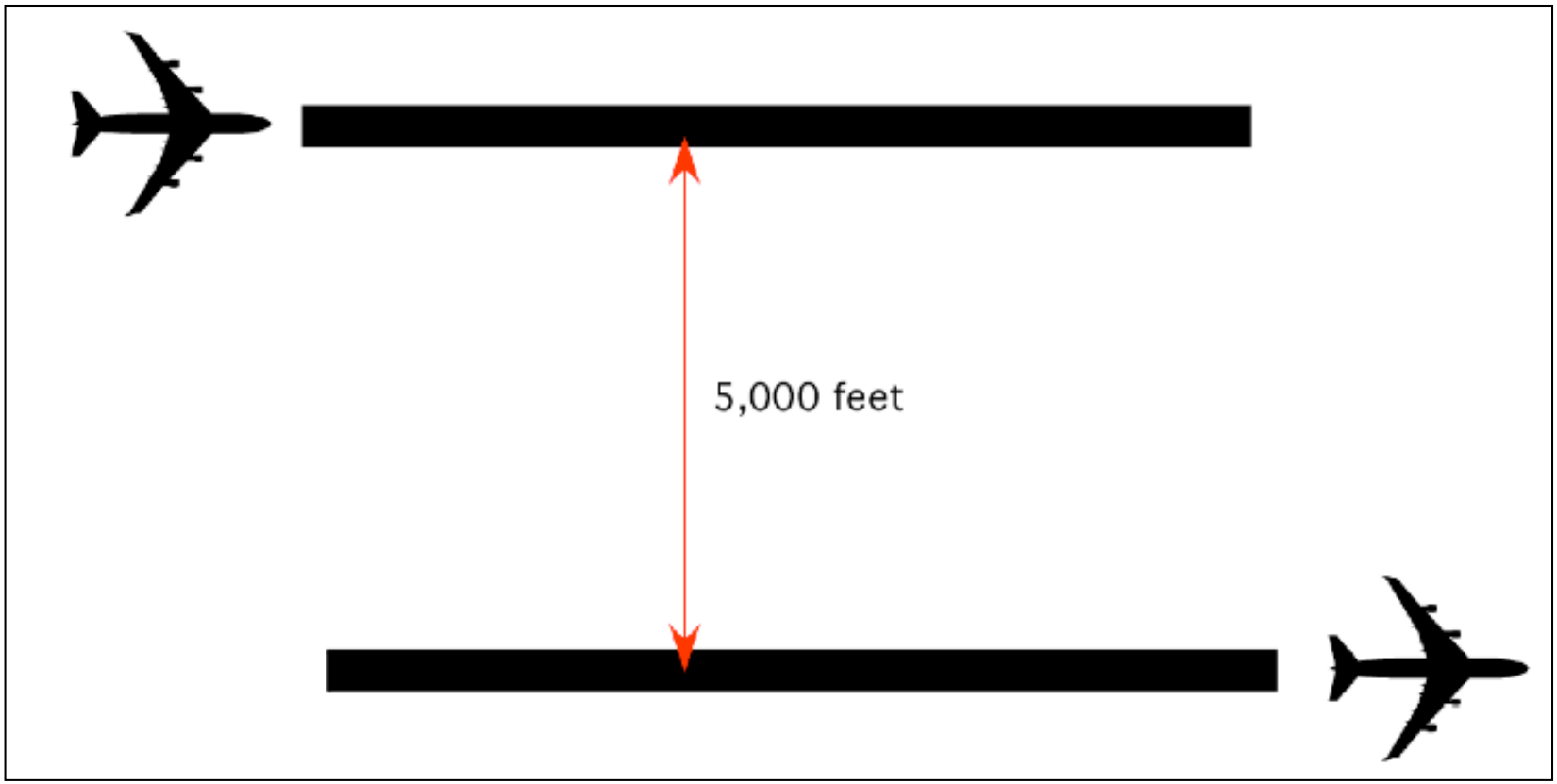
- Introducing some time variations in the system we can easily grasp the benefit of the simulation
- Most of the queueing processes at airport terminals are non-steady thus analytic models seldom apply
- Data typically exist on passenger behaviors over time that can be used to feed these deterministic, non-steady models
- The capacity function is perhaps the most difficult to wuantify because human performance is affected by the state of the system (i.e., queue length among others)

Example: Aircraft Delays Using Deterministic Queueing Model

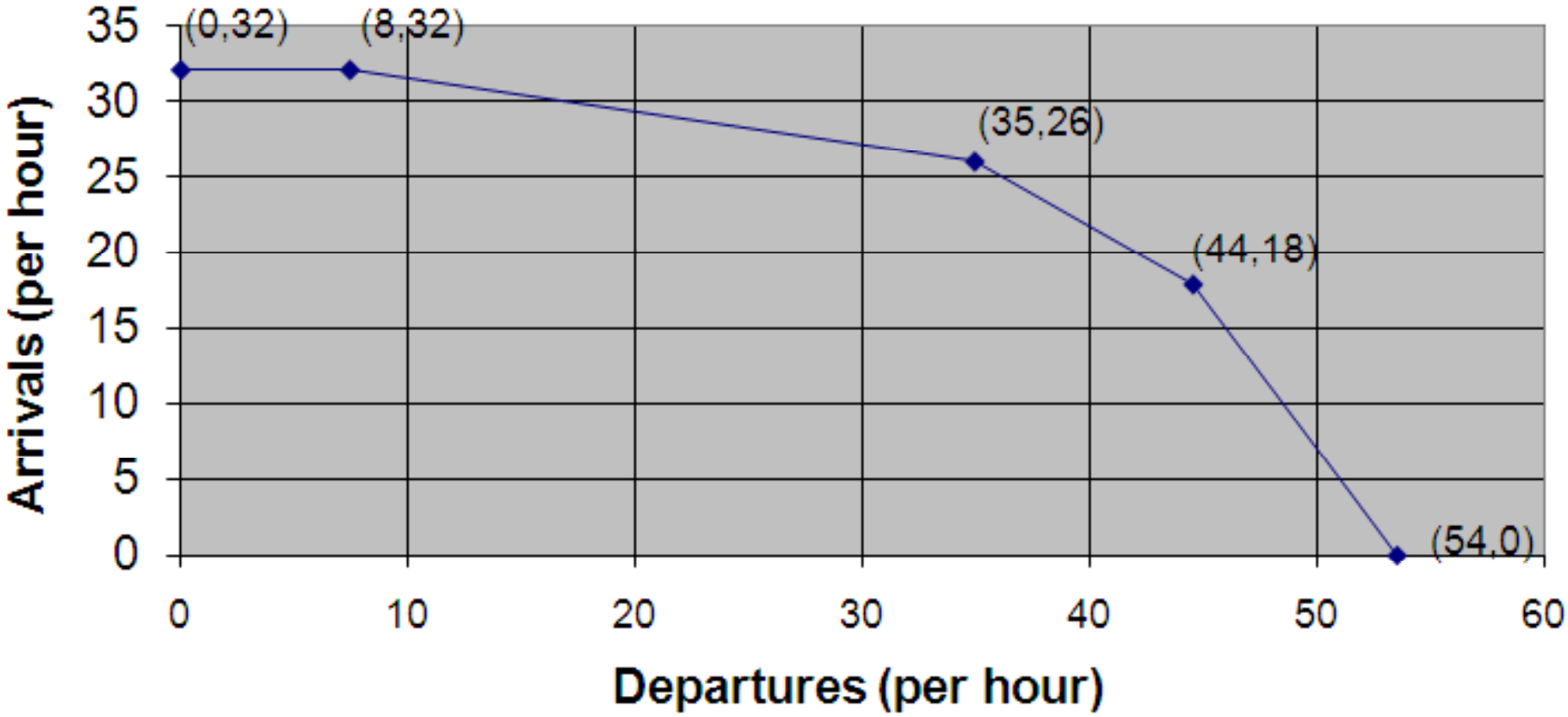
- An airport has two parallel runways separated 5,000 feet away and oriented East-West
- The saturation capacity analysis for one of the runways yields the Pareto diagram shown in the following figure
- Assume that the fleet mix operating at both runways is the same
- The diagram assumes that the runway is operated in mixed mode
- The analysis was done for IFR conditions

Airport Diagram

- Note: runways are used in segregated mode



Pareto Diagram for a **Single Runway** at the Airport (Mixed Mode)



One-runway Pareto Diagram. Mixed Runway Use. IFR Conditions. Numbers in the Plot Represent (Departure, Arrival) Pairs.

Airport Demand Function (Daily Demand)

Time (hrs) (Center of hourly interval)	Arrival demand (aircraft/hr)	Departure demand (aircraft/hr)
0.5	7	6
1.5	10	6
2.5	4	6
3.5	3	4
4.5	10	4
5.5	21	17
6.5	22	41
7.5	33	51
8.5	40	73
9.5	38	63
10.5	32	41
11.5	20	43

Airport Demand Function (Daily Demand) - Part 2

Time (hrs) (Center of hourly interval)	Arrival demand (aircraft/hr)	Departure demand (aircraft/hr)
12.5	32	34
13.5	23	23
14.5	37	26
15.5	40	29
16.5	25	38
17.5	23	71
18.5	20	62
19.5	37	62
20.5	36	43
21.5	29	36
22.5	20	36
23.5	13	11

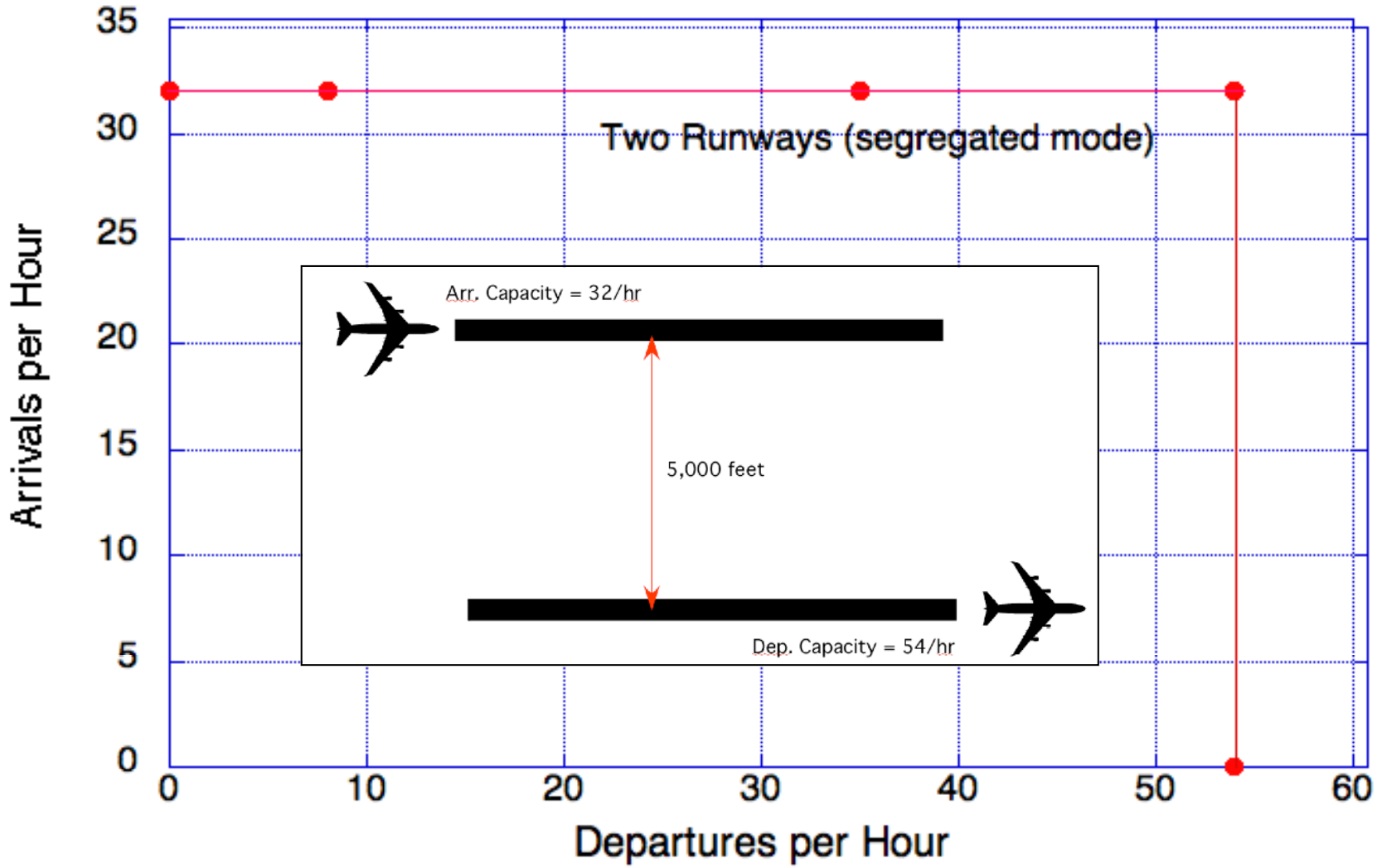
Relevant Questions

a) Draw the Pareto capacity diagram for the complete airport runway system (i.e., both runways) if the runways are used in **segregated** mode in IFR conditions.

b) If the airport is operated in segregated mode, determine the **average delay to arriving aircraft** if the arrival demand function proposed by the airlines is shown in Table 1. Assume IFR conditions prevail in the design day.

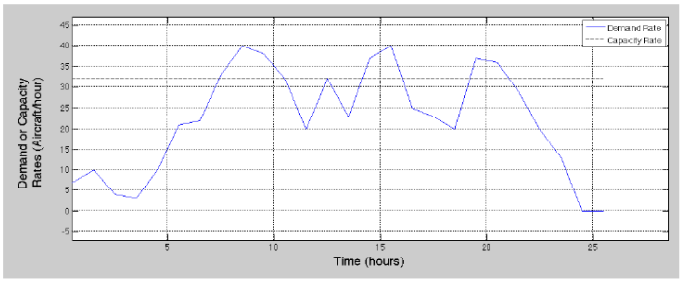
c) If the airport is operated in segregated mode, determine the **average delay to departing aircraft** if the departure demand function proposed by the airlines is shown in Table 1. Assume IFR conditions prevail in the design day.

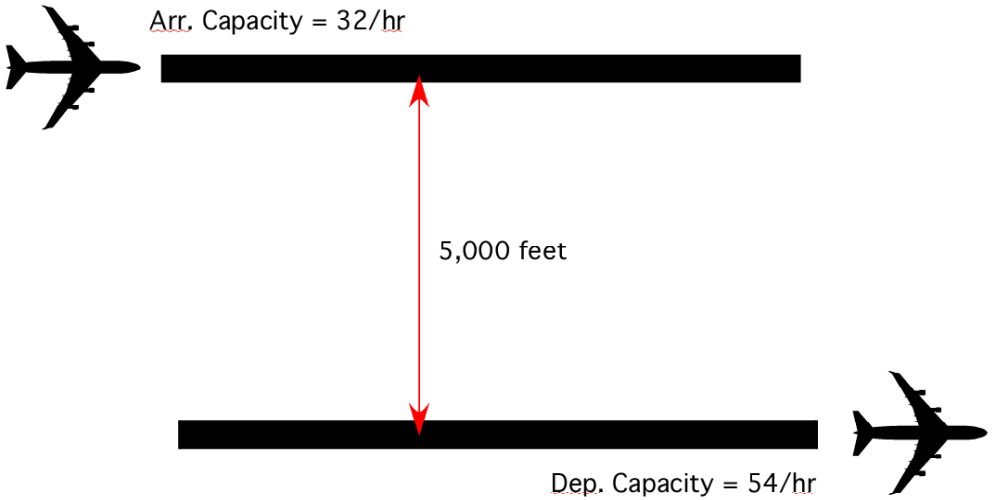
Part (a) Pareto Diagram for Complete Airport (Segregated Mode)

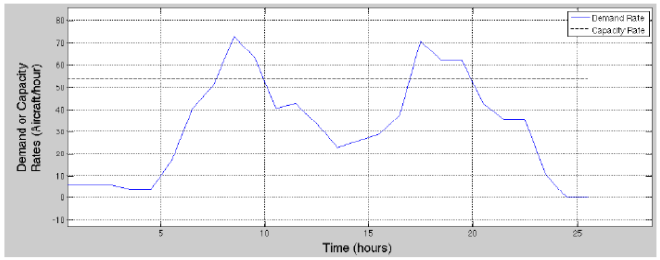


Part (b) Delays for Arriving Aircraft

- The problem is decomposed into two separate analyses:
  - delay calculation for arrivals and,
  - delay calculation for departures

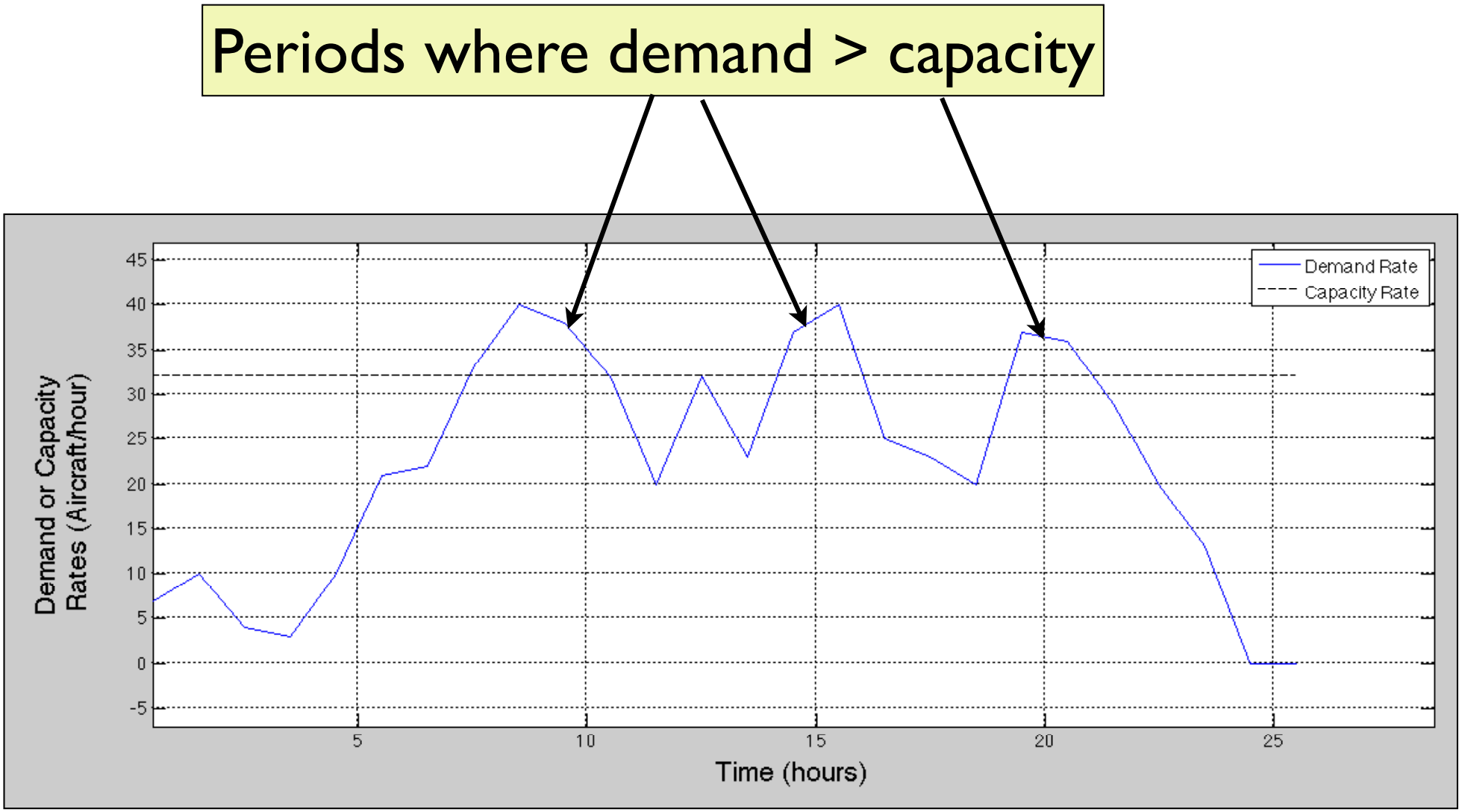






Part (b) Delays for Arriving Aircraft

- Arrival demand and arrival runway capacity



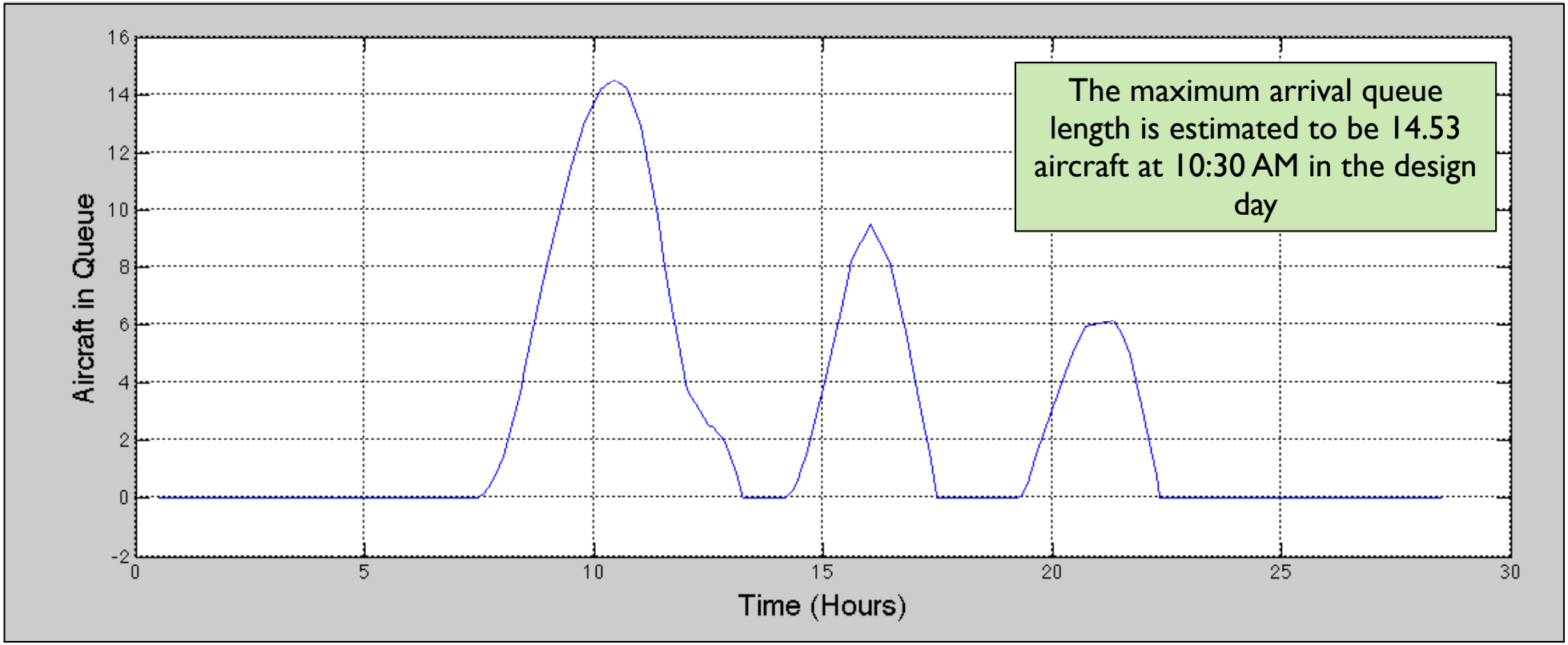
Part (b) Queue Length Function

- Numerical integration solution for queue length function (L_t)

$$L_t = L_{t-\Delta t} + \left(\frac{dL}{dt}\right)\Delta t$$

$$L_t = L_{t-\Delta t} + \left(\lambda(t) - \mu(t)\right)\Delta t$$

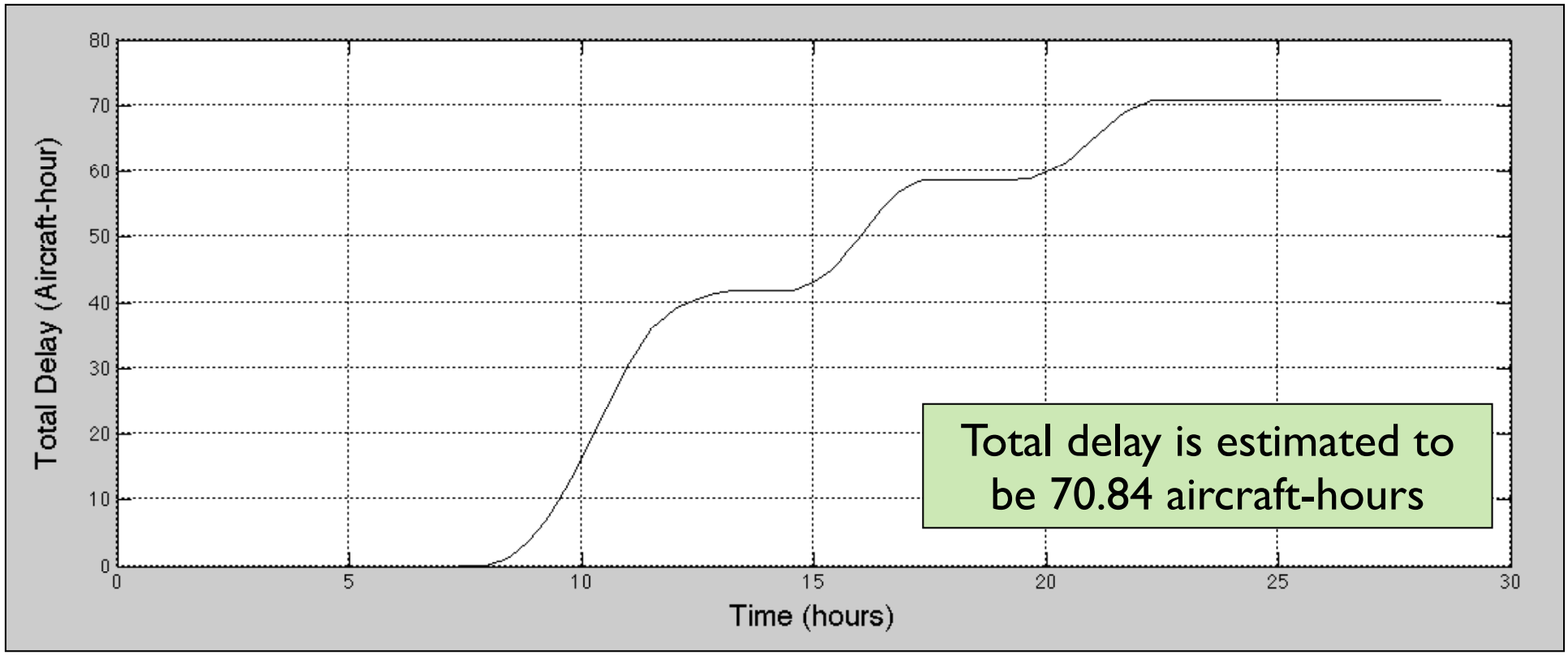
Use deterministic queue Matlab code:
det_queue_enhanced.m
(see Matlab files for CEE 5614)



Part (b) Integral of Queue Length Function

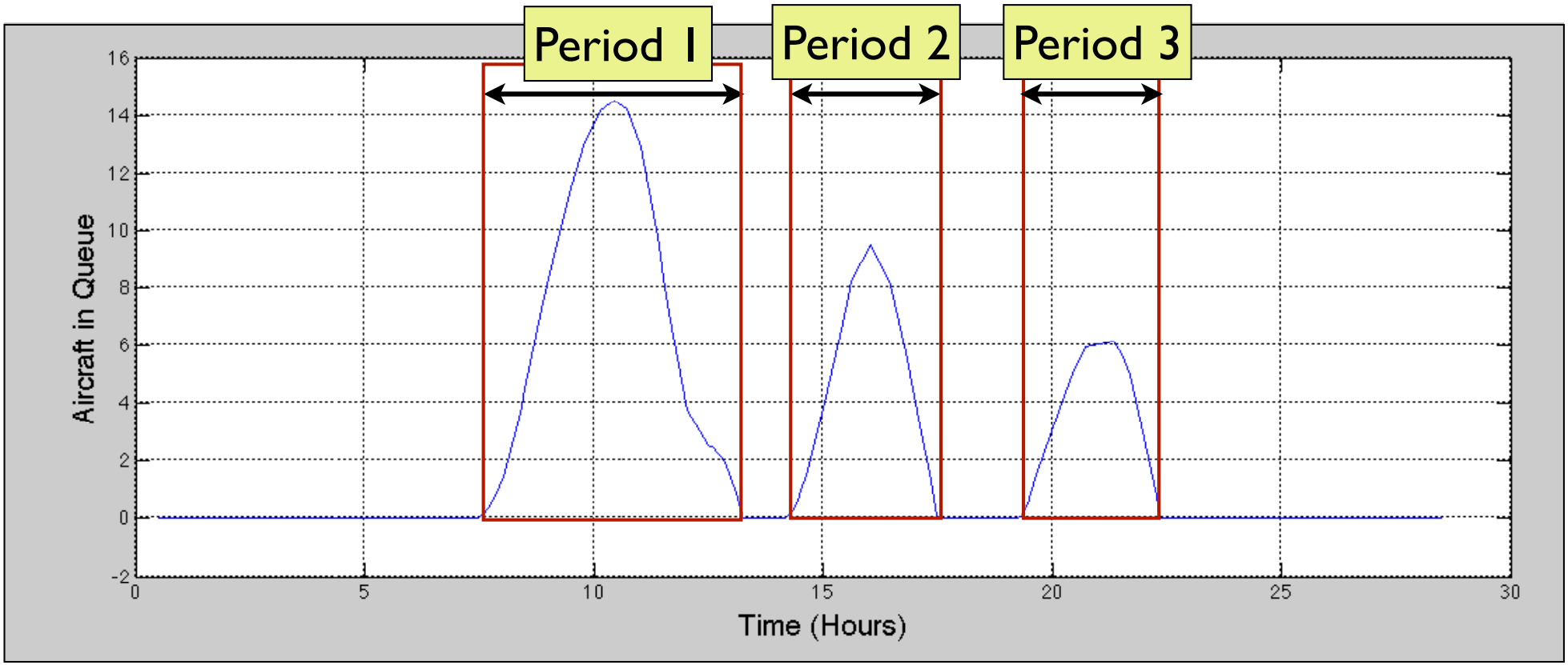
- Numerical integration solution for the area under the curve of the queue length function (L_t)

Use deterministic queue Matlab code:
det_queue_enhanced.m
(see Matlab files for CEE 5614)

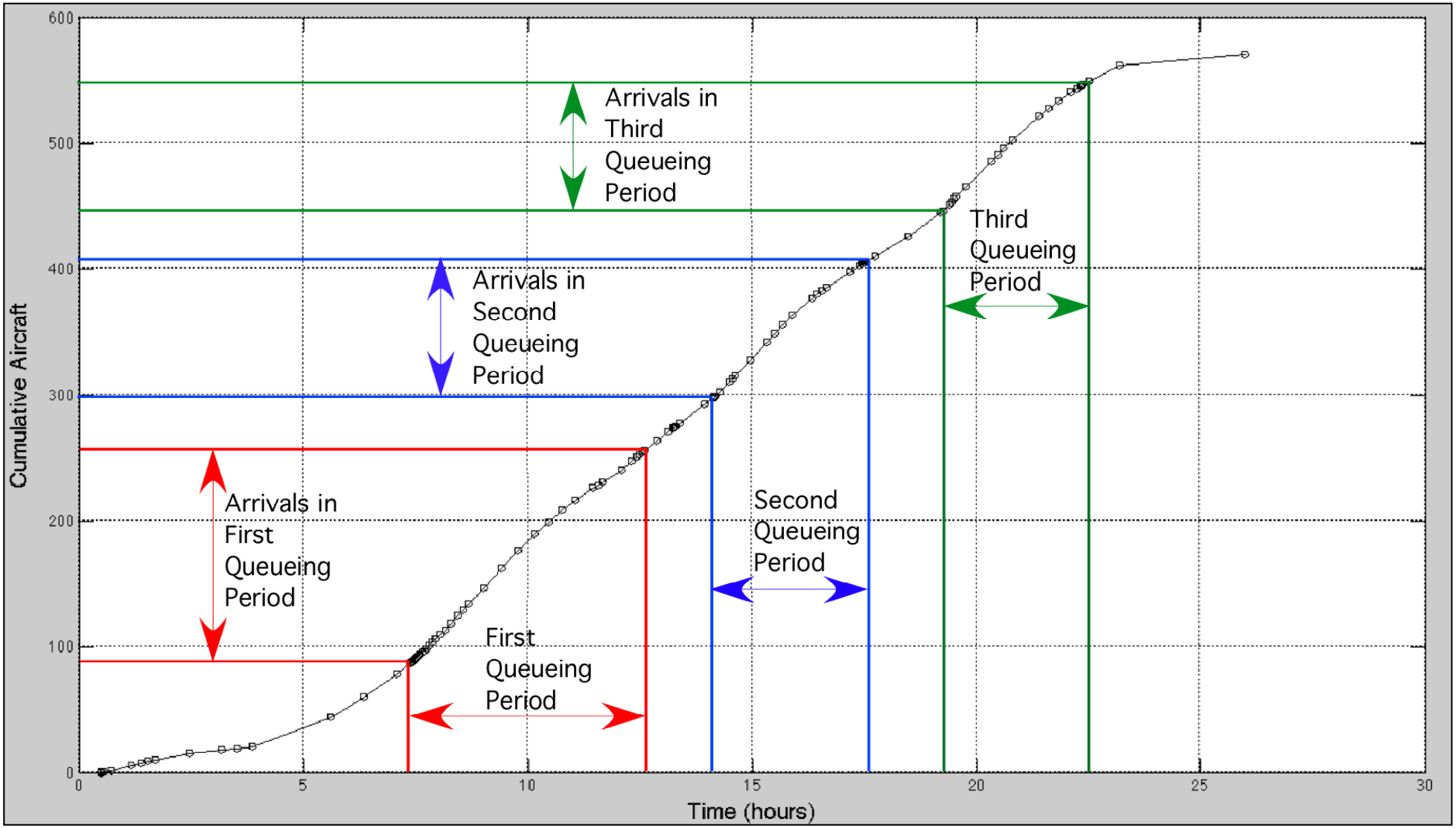


Part (b) Find the Average Arrival Delay

- Three queues were observed in the period of analysis (one day)
- Total number of aircraft delayed are calculated for the periods when the queues persist



Part (b) Cumulative Arrivals Function



Part (b) Cumulative Arrivals

Time (hrs)	Cumulative Arrivals	Arrivals in Queueing Period
7.4	87	274-87 = 187
13.25	274	
14.15	298	405-298=107
17.5	405	
19.2	445	546-445=101
22.34	546	
	Total Arrivals Queued	**395**

Part (b) Find the Average Arrival Delay

- Total delay = 70.84 aircraft-hours
- Total aircraft delayed = 395 aircraft
- Average delay per aircraft = 0.179 hours (or 10.7 minutes)

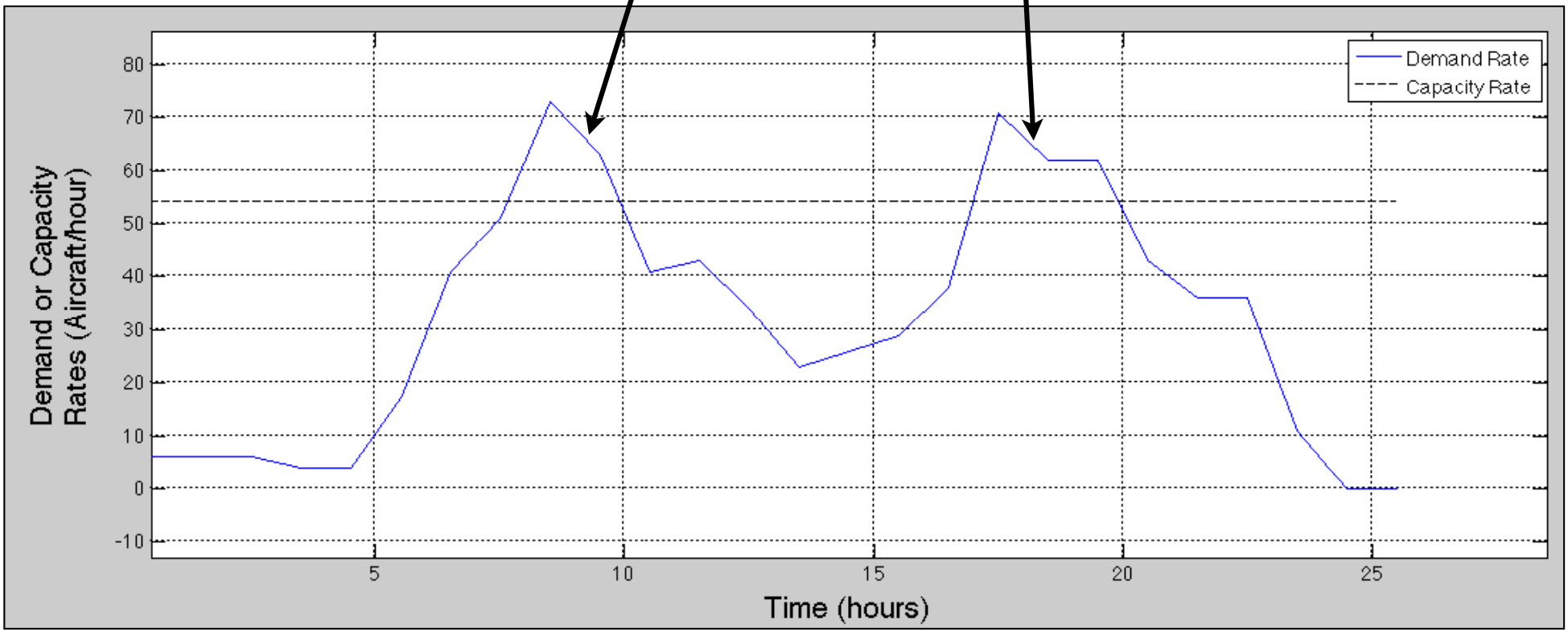
Part (c) Departure Delays

- Repeat the same process for departures
- Use departure saturation capacity found in the initial analysis (54 operations per hour)

Part (c) Departure Demand Function

- Departure demand and departure runway capacity

Periods where departure demand > dep. capacity



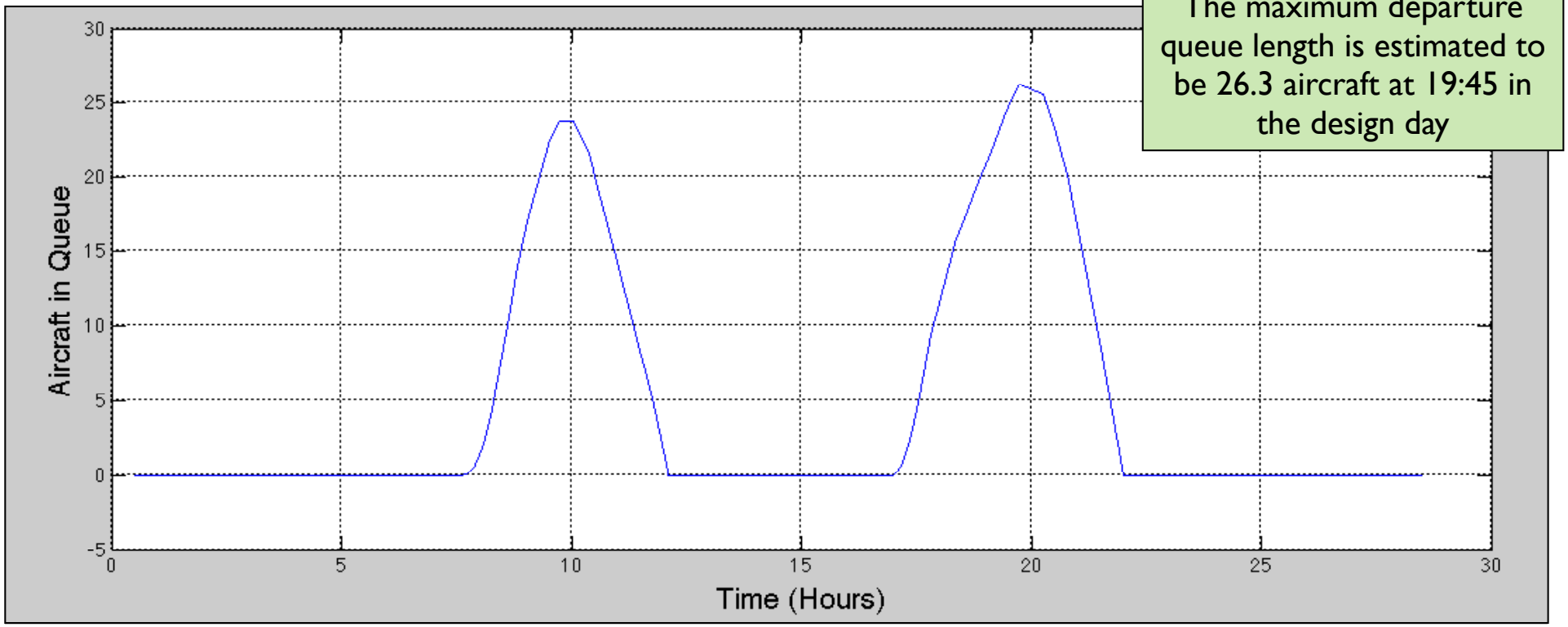
Part (c) Departure Queue Length Function

- Numerical integration solution for queue length function (L_t)

$$L_t = L_{t-\Delta t} + \left(\frac{dL}{dt}\right)\Delta t$$

$$L_t = L_{t-\Delta t} + \left(\lambda(t) - \mu(t)\right)\Delta t$$

Use deterministic queue Matlab code:
det_queue_enhanced.m
(see Matlab files for CEE 5614)

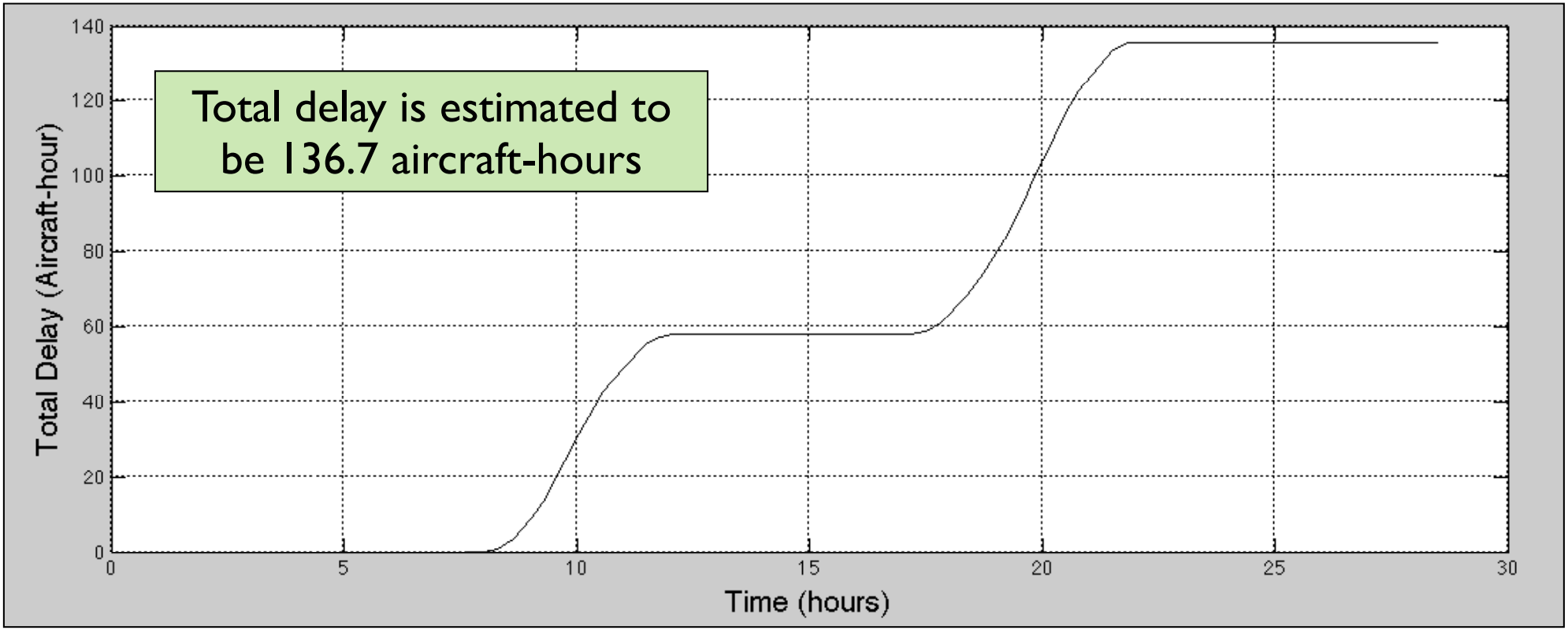


The maximum departure queue length is estimated to be 26.3 aircraft at 19:45 in the design day

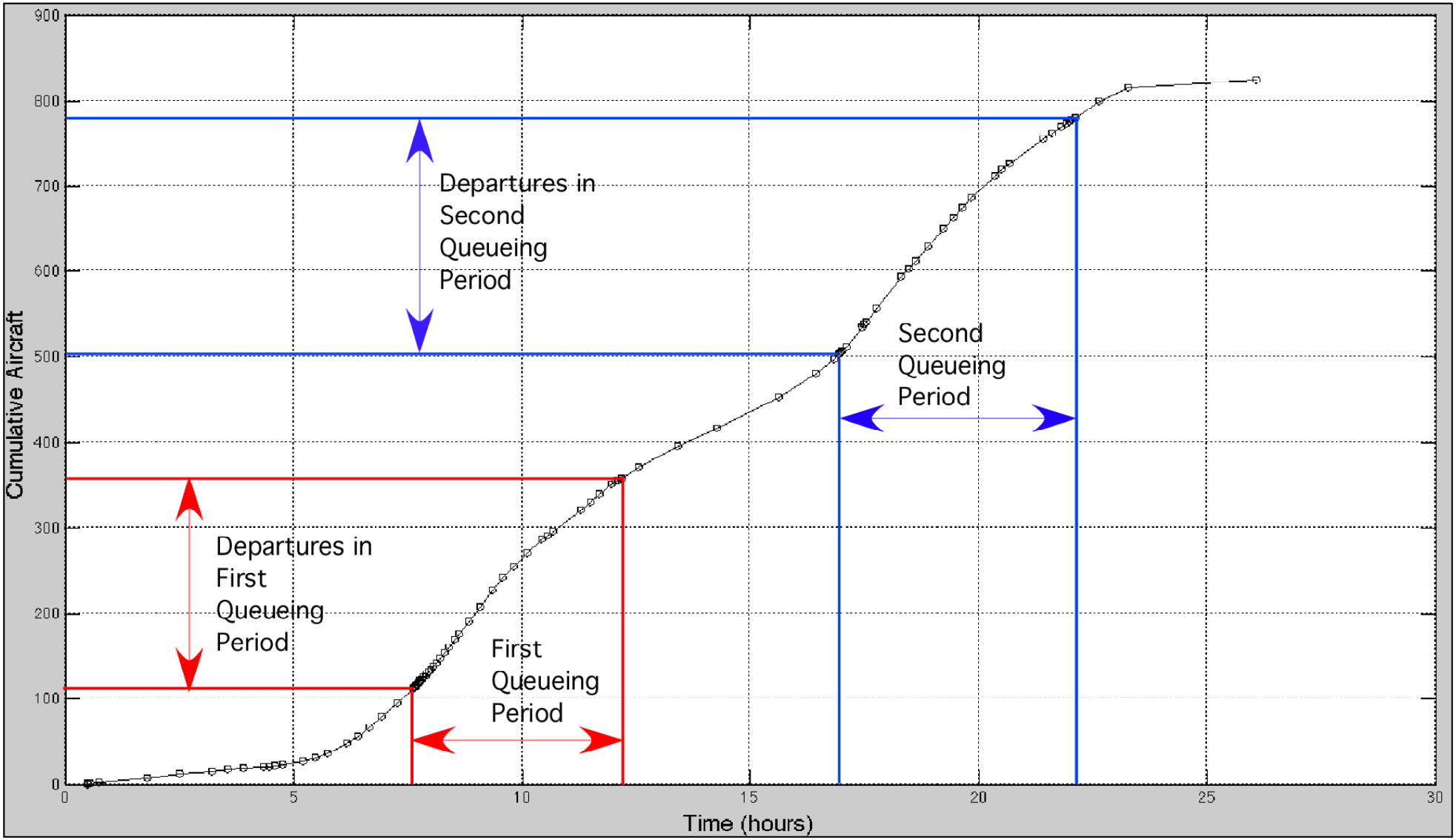
Part (c) Integral of Departure Queue Length Function

- Numerical integration solution for the area under the curve of the queue length function (L_t)

Use deterministic queue Matlab code:
det_queue_enhanced.m
(see Matlab files for CEE 5614)



Part (c) Cumulative Departures Function



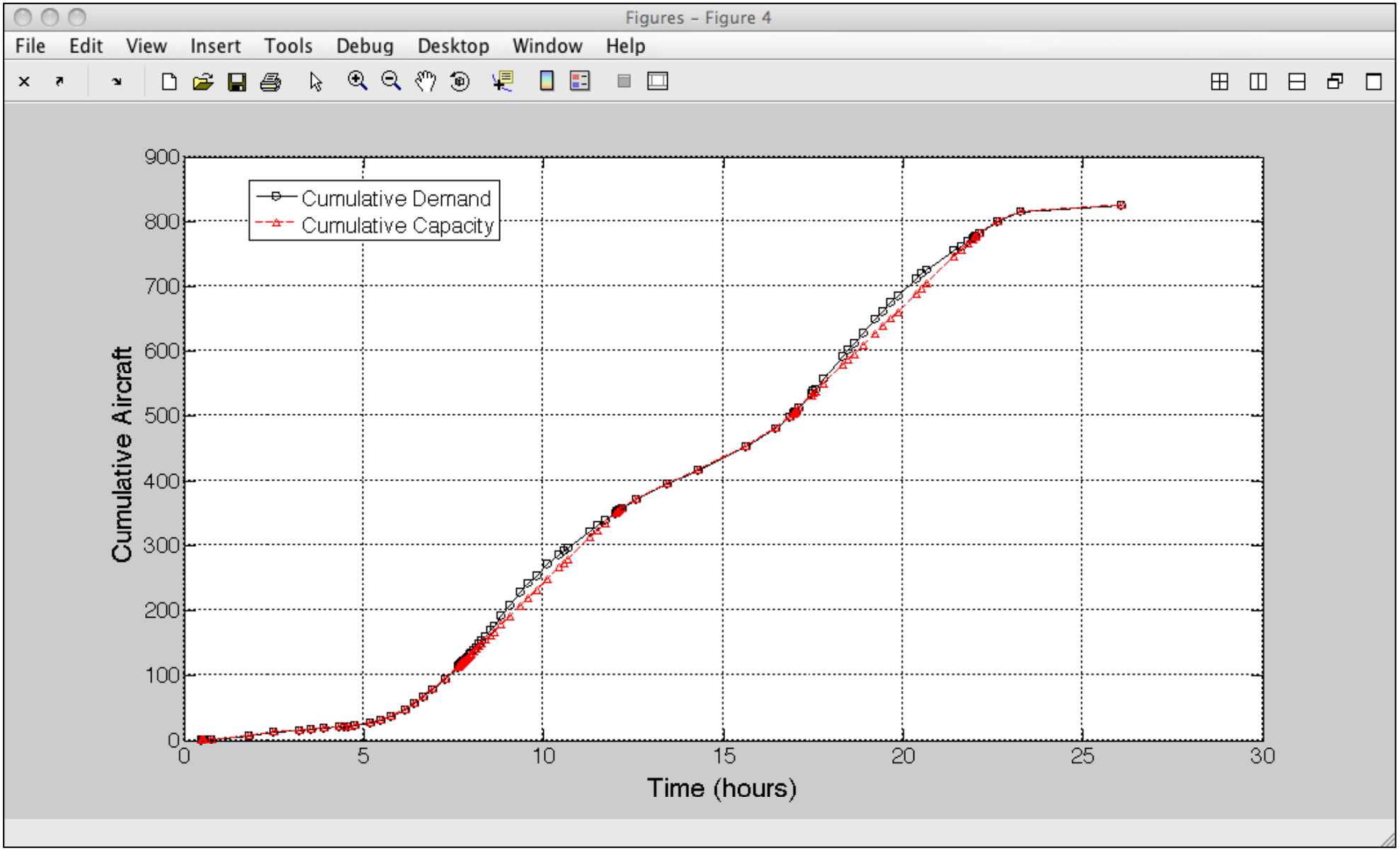
Part (c) Cumulative Departures

Time (hrs)	Cumulative Departures	Departures in Queueing Period
7.7	113	354-113 = 241
12.1	354	
17	506	776-506=270
22	776	
	Total Departures Queued	**511**

Part (b) Find the Average **Departure** Delay

- Total delay = 135.7 aircraft-hours
- Total aircraft delayed = 511 aircraft
- Average delay per aircraft = 0.27 hours (or 16 minutes)

Cumulative Demand and Supply Diagrams (for Departure Operations)



Matlab Source Code for Deterministic Queueing Model (main file)

```
% Deterministic queueing simulation
% T. Trani (Rev. Mar 99)
global demand capacity time

% Enter demand function as an array of values over time

% general demand - capacity relationships
%
% demand = [70 40 50 60 20 10];
% capacity = [50 50 30 50 40 50];
% time = [0 10 20 30 40 50];

demand = [1500 1000 1200 500 500 500];
capacity = [1200 1200 1000 1000 1200 1200];
time = [0.00  1.00  1.500  1.75  2.00  3.00];
```

```
% Compute min and maximum values for proper scaling in plots
mintime                                  = min(time);
maxtime                                  = max(time);
maxd                                      = max(demand);
maxc                                     = max(capacity);
mind                                      = min(demand);
minc                                     = min(capacity);
scale                                    = round(.2 *(maxc+maxd)/
2)
minplot = min(minc,mind) - scale;
maxplot                                  = max(maxc,maxd) +
scale;

po = [0 0];                              % intial number of
passengers
to = mintime;
tf = maxtime;
tspan = [to tf];

% where:
```

```
% to is the initial time to solve this equation
% tf is the final time
% tspan is the time span to solve the simulation

[t,p] = ode23('fqueue_2',tspan,po);

% Compute statistics

Ltmax                                   = max(p(:,1));
tdelay                                  = max(p(:,2));
a_demand                                = mean(demand);
a_capacity                              = mean(capacity);

  clc
  disp([blanks(5),'Deterministic Queueing Model '])
  disp(' ')
  disp(' ')
  disp([blanks(5),'  Average arrival rate (entities/time) = ',
num2str(a_demand)])
```

```
  disp([blanks(5),'  Average capacity (entities/time) = ',
num2str(a_capacity)])
  disp([blanks(5),'  Simulation Period (time units) = ', num2str(maxtime)])

  disp(' ')

  disp(' ')
  disp([blanks(5),'  Total delay (entities-time) = ', num2str(tdelay)])
  disp([blanks(5),'  Max queue length (entities) = ', num2str(Ltmax)])
  disp(' ')

pause

% Plot the demand and supply functions

plot(time,demand,'b',time,capacity,'k')
xlabel('Time (minutes)')
ylabel('Demand or Cpacity (Entities/time)')
axis([mintime maxtime minplot maxplot])
```

```
grid

pause
% Plot the results of the numerical integration procedure
subplot(2,1,1)
plot(t,p(:,1),'b')
xlabel('Time')
ylabel('Entities in Queue')
grid

subplot(2,1,2)
plot(t,p(:,2),'k')
xlabel('Time')
ylabel('Total Delay (Entities-time)')
grid
```

Matlab Source Code for Deterministic Queueing Model (function file)

```
% Function file to integrate numerically a differential equation
% describing a deterministic queueing system

function pprime = fqueue_2(t,p)
global demand capacity time

% Define the rate equations
demand_table = interp1(time,demand,t);
capacity_table = interp1(time,capacity,t);

if (demand_table < capacity_table) & (p > 0)
    pprime(1) = demand_table - capacity_table;    % rate of change in state
variable
elseif demand_table > capacity_table
    pprime(1) = demand_table - capacity_table;    % rate of change in state
variable
```

```
else
        pprime(1) = 0.0; % avoids accumulation of entities
end

pprime(2) = p(1);                                   % integrates the delay
curve over time
pprime = pprime';
```

Output of Deterministic Queueing Model

Deterministic Queueing Model

Average arrival rate (entities/time) = 866.6667
Average capacity (entities/time) = 1133.3333
Simulation Period (time units) = 3

Total delay (entities-time) = 94.8925
Max queue length (entities) = 89.6247

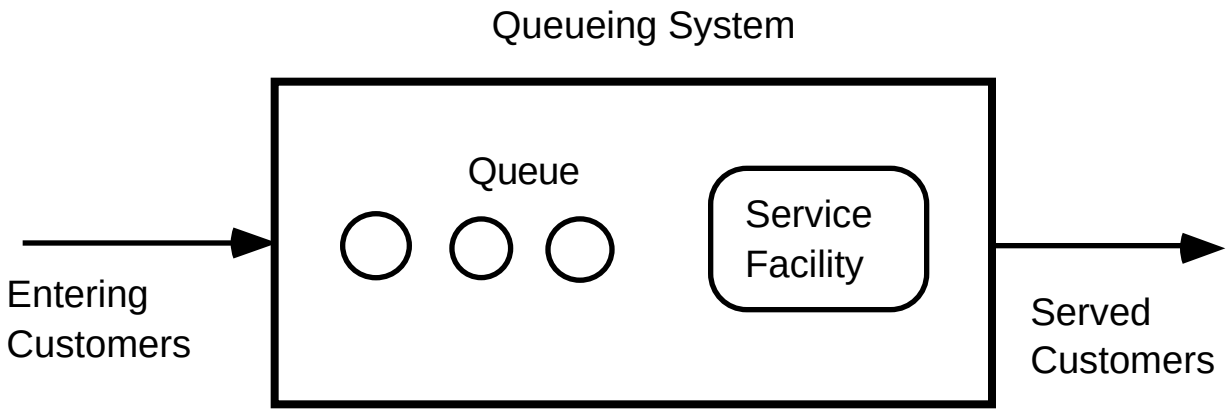
Stochastic Queueing Theory (Nomenclature per Hillier and Lieberman)

Virginia Tech

These models can only be generalized for simple arrival and departure functions since the involvement of complex functions make their analytic solution almost impossible to achieve. The process to be described first is the so-called **birth and death process** that is completely analogous to the arrival and departure of an entity from the queueing system in hand.

Before we try to describe the mathematical equations it is necessary to understand the basic principles of the stochastic queue and its nomenclature.

Fundamental Elements of a Queueing System



Nomenclature

Queue length = No. of customers waiting for service

L(t) = State of the system - customers in queue at time t

N(t) = Number of customers in queueing system at time t

P(t) = Prob. of exactly n customers are in queueing system at time t

s = No. of servers (parallel service)

λ_n = Mean arrival rate

μ_n = Mean service rate for overall system

Other Definitions in Queueing Systems

Virginia Tech

If λ_n is constant for all n then $(1/\lambda)$ it represents the interarrival time. Also, is μ_n is constant for all $n > 1$ (constant for each busy server) then $\mu n = m$ service rate and $(1/\mu)$ is the service time (mean).

Finally, for a multiserver system $s\mu$ is the total service rate and also $\rho = l/s\mu$ is the utilization factor. This is the amount of time that the service facility is being used.

Stochastic Queueing Systems

The idea behind the queueing process is to analyze steady-state conditions. Lets define some notation applicable for steady-state conditions,

N = No. of customers in queueing system

P_n = Prob. of exactly n customers are in queueing system

L = Expected no. of customers in queueing system

L_q = Queue length (expected)

W = Waiting time in system (includes service time)

W_q = Waiting time in queue

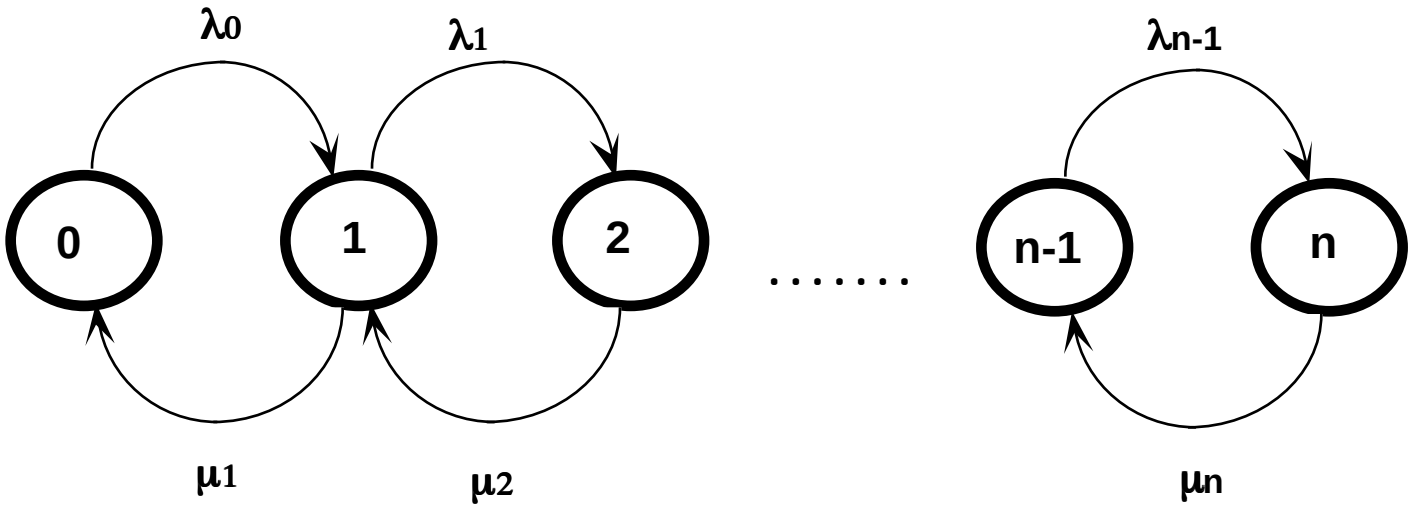
There are some basic relationships that have beed derived in standard textbooks in operations research [Hillier and Lieberman, 1991]. Some of these more basic relationships are:

$L = \lambda W$

$Lq = \lambda Wq$

The analysis of stochastic queueing systems can be easily understood with the use of "Birth-Death" rate diagrams as illustrated in the next figure. Here the transitions of a system are illustrated by the state conditions 0, 1, 2, 3,.. etc. Each state corresponds to a situation where there are n customers in the system. This implies that state 0 means that the system is idle (i,e., no customers), system at state 1 means there is one customer and so forth.

Rate Diagram for Birth-and-Rate Process



Note: Only possible transitions in the state of the system are shown.

Stochastic Queueing Systems

Virginia Tech

For a queue to achieve steady-state we require that all rates in equal the rates out or in other words that all transitions out are equal to all the transitions in. This implies that there has to be a balance between entering and leaving entities.

Consider state 0. This state can only be reached from state 1 if one departure occurs. The steady state probability of being in state 1 (P_1) represents the portion of the time that it would be possible to enter state 0. The mean rate at which this happens is $\mu_1 P_1$. Using the same argument the mean occurrence rate of the leaving incidents must be $\lambda_0 P_0$ to the balance equation,

Stochastic Queueing Systems

$\mu_1 P_1 = \lambda_0 P_0$

For every other state there are two possible transitions. Both into and out of the state.

$\lambda_0 P_0 = P_1 \mu_1$

$\lambda_0 P_0 + \mu_2 P_2 = \lambda_1 P_1 + \mu_1 P_1$

$\lambda_1 P_1 + \mu_3 P_3 = \lambda_2 P_2 + \mu_2 P_2$

$\lambda_2 P_2 + \mu_4 P_4 = \lambda_3 P_3 + \mu_3 P_3$

until, $\lambda_{n-1} P_{n-1} + \mu_{n+1} P_{n+1} = \lambda_n \mu_n + \mu_n P_n$

Stochastic Queueing Systems

Since we are interested in the probabilities of the system in every state n want to know the P_n's in the process. The idea is to solve these equations in terms of one variable (say P_0) as there is one more variable than equations.

For every state we have,

$$P_1 = \lambda_0 / \mu_1 \, P_0$$

$$P_2 = \lambda_1 \lambda_0 / \mu_1 \, \mu_2 \; P_0$$

$$P_3 = \lambda_2 \, \lambda_1 \, \lambda_0 / \mu_1 \, \mu_2 \, \mu_3 \, P_0$$

$$P_{n+1} = \lambda_n \; ..... \; \lambda_1 \, \lambda_0 / \mu_1 \, \mu_2 \; ..... \; \mu_{n+1} \; P_0$$

Stochastic Queueing Systems

Let C_n be defined as,

$$C_n = \lambda_{n-1} \ldots\ldots \lambda_1\lambda_0 / \mu_1 \mu_2 \ldots\ldots \mu_n$$

Once this is accomplished we can determine the values of all probabilities since the sum of all have to equate to unity.

$$\sum_{i=0}^{n} P_n = 1$$

$$P_0 + \sum_{i=1}^{n} P_n = 1$$

Stochastic Queueing Systems

$$P_0 + \sum_{i=1}^{n} C_n P_0 = 1$$

Solving for P_0 we have,

$$P_0 = \left[\frac{1}{1 + \sum_{i=1}^{n} C_n} \right]$$

Now we are in the position to solve for the remaining queue parameters, L the average no. of entities in the system, Lq, the average number of customers in the

queue, W, the average waiting time in the system and Wq the average waiting time in the queue.

$$P_n = C_n P_0$$

$$L = \sum_{n=1}^{\infty} n P_n$$

$$L_q = \sum_{n=s}^{\infty} (n-s) P_n$$

$$W = \frac{L}{\lambda}$$

$$W_q = \frac{L_q}{\lambda}$$

This process can then be repeated for specific queueing scenarios where the number of customers is finite, infinite, etc. and for one or multiple servers. All systems can be derived using “birth-death” rate diagrams.

Stochastic Queueing Systems

Depending on the simplifying assumptions made, queueing systems can be solved analytically.

The following section presents equations for the following queueing systems when poisson arrivals and negative exponential service times apply:

a) Single server - infinite source (constant λ and μ)

b) Multiple server - infinite source (constant λ and μ)

c) Single server - finite source (constant λ and μ)

d) Multiple server - finite source (constant λ and μ)

Stochastic Queueing Systems Nomenclature

The idea behind the queueing process is to analyze steady-state conditions. Lets define some notation applicable for steady-state conditions,

N = No. of customers in queueing system

P_n = Prob. of exactly n customers are in queueing system

L = Expected no. of customers in queueing system

L_q = Queue length (expected)

W = Waiting time in system (includes service time)

W_q = Waiting time in queue

Stochastic Queueing Systems

There are some basic relationships that have beed derived in standard textbooks in operations research [Hillier and Lieberman, 1991]. Some of these more basic relationships are:

$$L = \lambda W$$

$$Lq = \lambda Wq$$

Stochastic Queueing Systems

Single server - infinite source (constant λ and μ)

Assumptions:

a) Probability between arrivals is negative exponential with parameter λ_n

b) Probability between service completions is negative exponential with parameter μ_n

c) Only one arrival or service occurs at a given time

Single server - Infinite Source (Constant λ and μ)

$\rho = \lambda/\mu$ Utilization factor

$$P_0 = \frac{1}{1 + \sum_{n=1}^{\infty} \rho^n} = \left(\sum_{n=0}^{\infty} \rho^n \right)^{-1} = \left(\frac{1}{1-\rho} \right)^{-1} = 1 - \rho$$

$$P_n = \rho^n P_0 = (1-\rho)\rho^n \quad \text{for n} = 0,1,2,3,.....$$

$L = \frac{\lambda}{\mu - \lambda}$ expected number of entities in the system

$L_q = \frac{\lambda^2}{(\mu - \lambda)\mu}$ expected no. of entities in the queue

$W = \frac{1}{\mu - \lambda}$ average waiting time in the queueing system

$W_q = \frac{\lambda}{(\mu - \lambda)\mu}$ average waiting time in the queue

$P(W > t) = e^{-\mu(1-\rho)t}$ probability distribution of waiting times

(including the service potion in the SF)

Multiple Server

Infinite source (constant λ and μ)

Assumptions:

a) Probability between arrivals is negative exponential with parameter λ_n

b) Probability between service completions is negative exponential with parameter μ_n

c) Only one arrival or service occurs at a given time

Multiple Server - Infinite Source (constant λ , μ)

$\rho = \lambda / s\mu$ utilization factor of the facility

$$P_0 = 1 \Big/ \left(\sum_{n=0}^{s-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^s}{s!} \left(\frac{1}{1-(\lambda/s\mu)} \right) \right)$$

idle probability

$$P_n = \begin{cases} \dfrac{(\lambda/\mu)^n}{n!} P_0 & 0 \le n \le s \\ \dfrac{(\lambda/\mu)^n}{s!\, s^{n-s}} P_0 & n \ge s \end{cases}$$

probability of n entities in the system

$$L = \frac{\rho P_0 \left(\frac{\lambda}{\mu}\right)^s}{s!(1-\rho)^2} + \frac{\lambda}{\mu}$$

expected number of entities in system

$$L_q = \frac{\rho P_0 \left(\frac{\lambda}{\mu}\right)^s}{s!(1-\rho)^2}$$ expected number of entities in queue

$$W_q = \frac{L_q}{\lambda}$$ average waiting time in queue

$$W = \frac{L}{\lambda} = W_q + \frac{1}{\lambda}$$ average waiting time in system

Finally the probability distribution of waiting times is,

$$P(W > t) = e^{-\mu t}\left[1 + \frac{P_0\left(\frac{\lambda}{\mu}\right)^s}{s!(1-\rho)}\left(\frac{1 - e^{-\mu t(s-1-\lambda/\mu)}}{s-1-\lambda/\mu}\right)\right]$$

if $s - 1 - \lambda/\mu = 0$ then use

$$\frac{1 - e^{-\mu t(s-1-\lambda/\mu)}}{s-1-\lambda/\mu} = \mu t$$

Single Server - Finite Source (constant λ and μ)

Virginia Tech

Assumptions:

a) Interarrival times have a negative exponential PDF with parameter λ_n

b) Probability between service completions is negative exponential with parameter μ_n

c) Only one arrival or service occurs at a given time

d) M is the total number of entities to be served (calling population)

Single Server - Finite Source (constant λ and μ)

$\rho = \lambda/\mu$ utilization factor of the facility

$$P_0 = 1/\sum_{n=0}^{M}\left(\frac{M!}{(M-n)!}(\lambda/\mu)^n\right)$$ idle probability

$$P_n = \frac{M!}{(M-n)!}(\lambda/\mu)^n P_0$$ for $n = 1, 2, 3, \ldots M$ probability of n entities in the system

$$L_q = M - \frac{\mu+\lambda}{\lambda}(1-P_0)$$ expected number of entities in queue

$L = M - \frac{\mu}{\lambda}(1 - P_0)$ expected number of entities in system

$W_q = \frac{L_q}{\bar{\lambda}}$ average waiting time in queue

$W = \frac{L}{\bar{\lambda}}$ average waiting time in system

where:

$\bar{\lambda} = \lambda(M - L)$ average arrival rate

Multiple Server Cases

Finite source (constant λ and μ)

Assumptions:

a) Interarrival times have a negative exponential PDF with parameter λ_n

b) Probability between service completions is negative exponential with parameter μ_n

c) Only one arrival or service occurs at a given time

d) M is the total number of entities to be served (calling population)

Multiple Server - Finite Source (constant λ and μ)

$\rho = \lambda/\mu s$ utilization factor of the facility

$$P_0 = 1/\left[\sum_{n=0}^{s-1}\left(\frac{M!}{(M-n)!n!}(\lambda/\mu)^n\right) + \sum_{n=s}^{M}\left(\frac{M!}{(M-n)!s!s^{n-s}}(\lambda/\mu)^n\right)\right]$$

idle probability

$$P_n = \begin{cases} \dfrac{M!}{(M-n)!n!}(\lambda/\mu)^n P_0 & 0 \le n \le s \\ \dfrac{M!}{(M-n)!s!s^{n-s}}(\lambda/\mu)^n P_0 & \text{if } s \le n \le M \\ 0 & n \ge M \end{cases}$$

$$L_q = \sum_{n=s}^{M} (n-s)P_n$$ expected number of entities in queue

$$L = \sum_{n=0}^{M} nP_n = \sum_{n=0}^{s-1} nP_n + L_q + s\left(1 - \sum_{n=0}^{s-1} P_n\right)$$

expected number of entities in system

$W_q = \frac{L_q}{\bar{\lambda}}$ average waiting time in queue

$W = \frac{L}{\bar{\lambda}}$ average waiting time in system

where:

$\bar{\lambda} = \lambda(M - L)$ average arrival rate

Example (3): Level of Service at Security Checkpoints

The airport shown in the next figures has two security checkpoints for all passengers boarding aircraft. Each security check point has two x-ray machines. A survey reveals that on the average a passenger takes 45 seconds to go through the system (negative exponential distribution service time).

The **arrival rate** is known to be random (this equates to a Poisson distribution) with a mean arrival rate of one passenger every 25 seconds.

In the design year (2010) the demand for services is expected to grow by 60% compared to that today.

Relevent Operational Questions

a) What is the current utilization of the queueing system (i.e., two x-ray machines)?

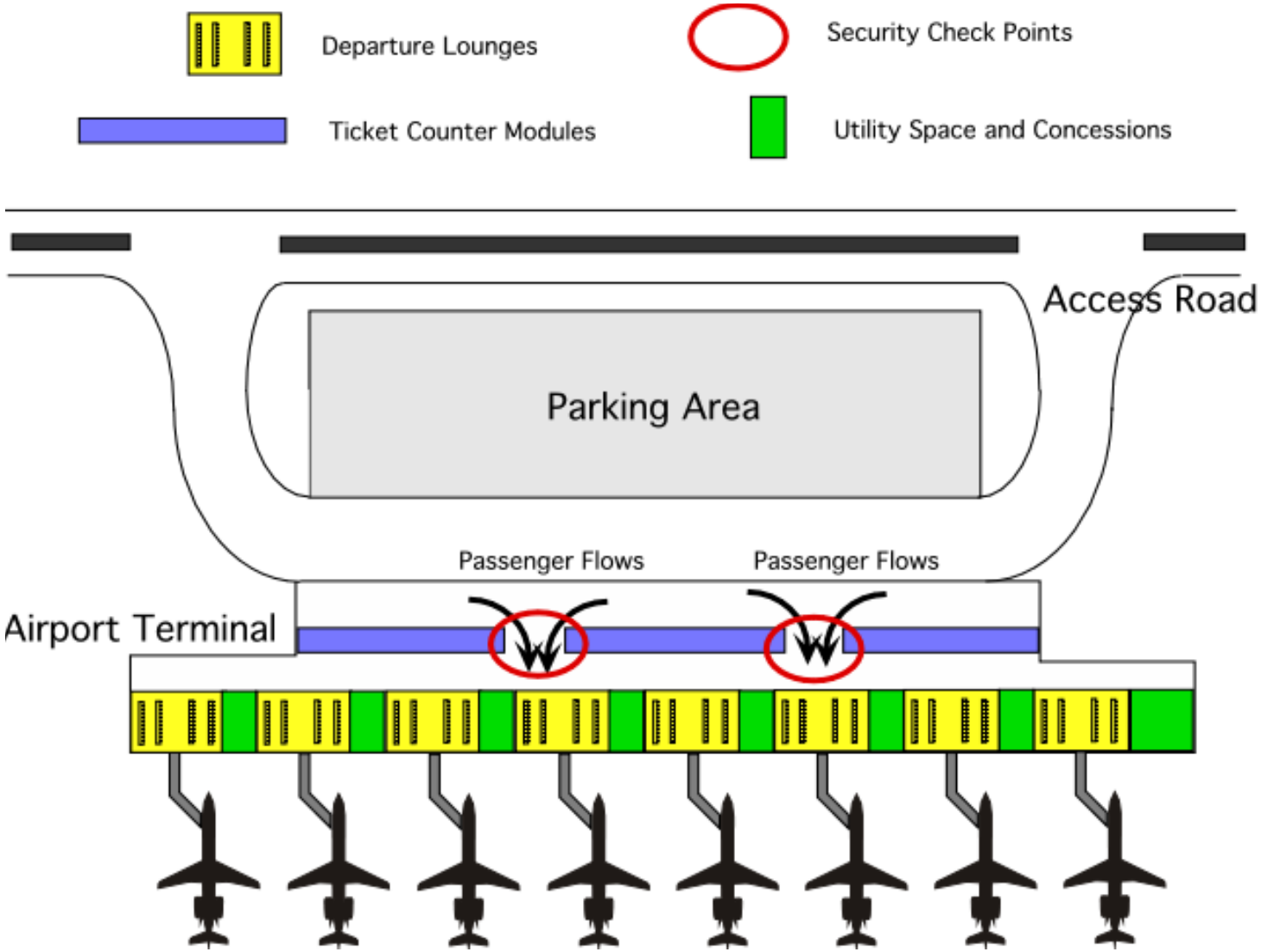
b) What should be the number of x-ray machines for the design year of this terminal (year 2020) if the maximum tolerable **waiting time in the queue** is 2 minutes?

c) What is the expected number of passengers at the checkpoint area on a typical day in the design year (year 2020)? Assume a 60% growth in demand.

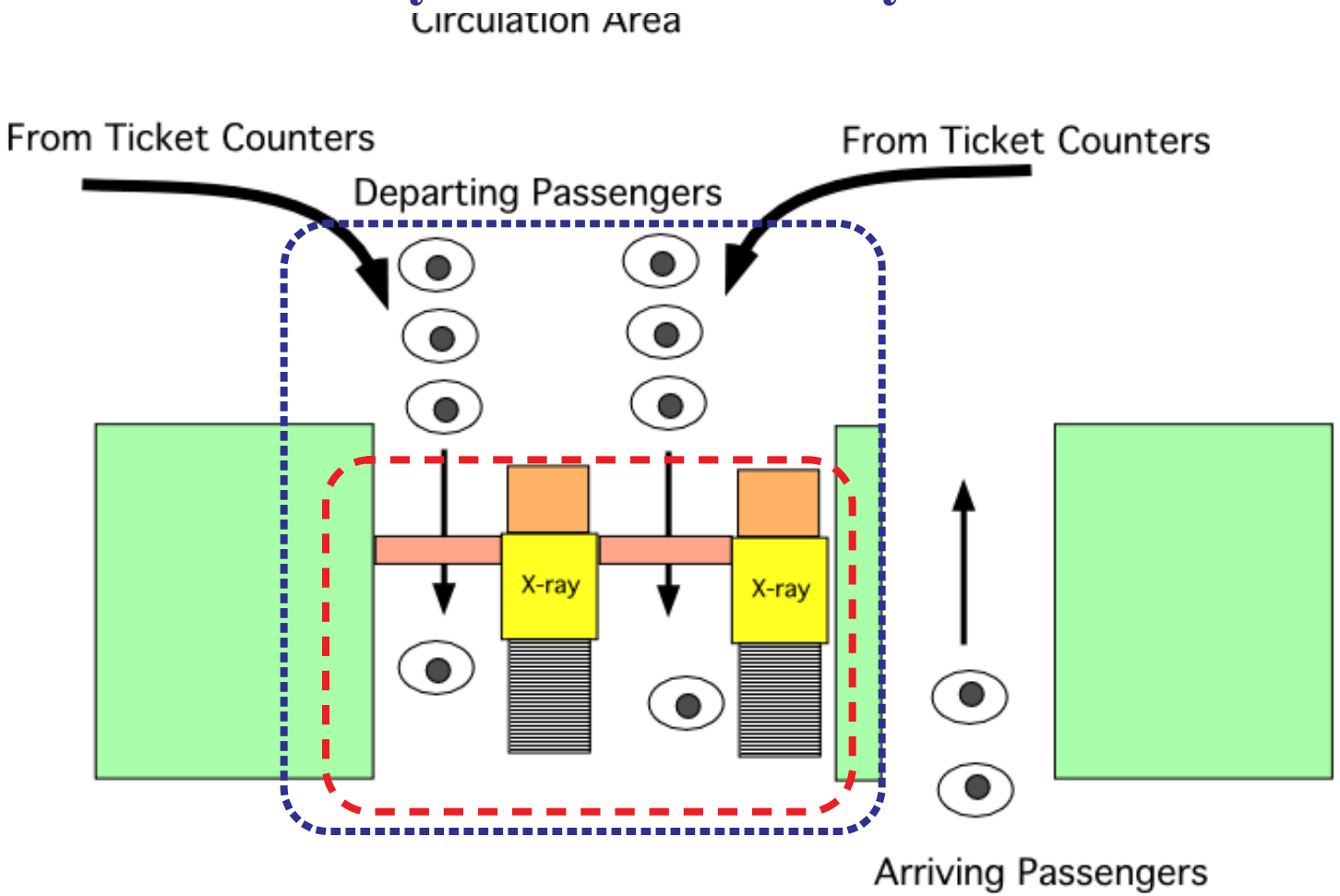
d) What is the new utilization of the future facility?

e) What is the probability that more than 4 passengers wait for service in the design year?

Airport Terminal Layout



Security Check Point Layout



Security Check Point Solutions

a) Utilization of the facility, ρ. Note that this is a multiple server case with infinite source.

$\rho = \lambda / (s\mu) = 144/(2*80) = 0.90$

Other queueing parameters are found using the steady-state equations for a multi-server queueing system with infinite population are:

Idle probability = 0.052632
Expected No. of customers in queue (Lq) = 7.6737
Expected No. of customers in system (L) = 9.4737
Average Waiting Time in Queue = 192 s
Average Waiting Time in System = 237 s

b) The solution to this part is done by trail and error (unless you have access to design charts used in queueing models. As a first trial lets assume that the number of x-ray machines is 3 (s=3).

Finding Po, $$P_0 = \sum_{n=0}^{s-1} \frac{(\lambda/\mu)^2}{n!} + \frac{(\lambda/\mu)^s}{s!}\left(\frac{1}{1-(\lambda/s\mu)}\right)$$

Po = .0097 or less than 1% of the time the facility is idle

Find the waiting time in the queue,

***Wq* = 332** s

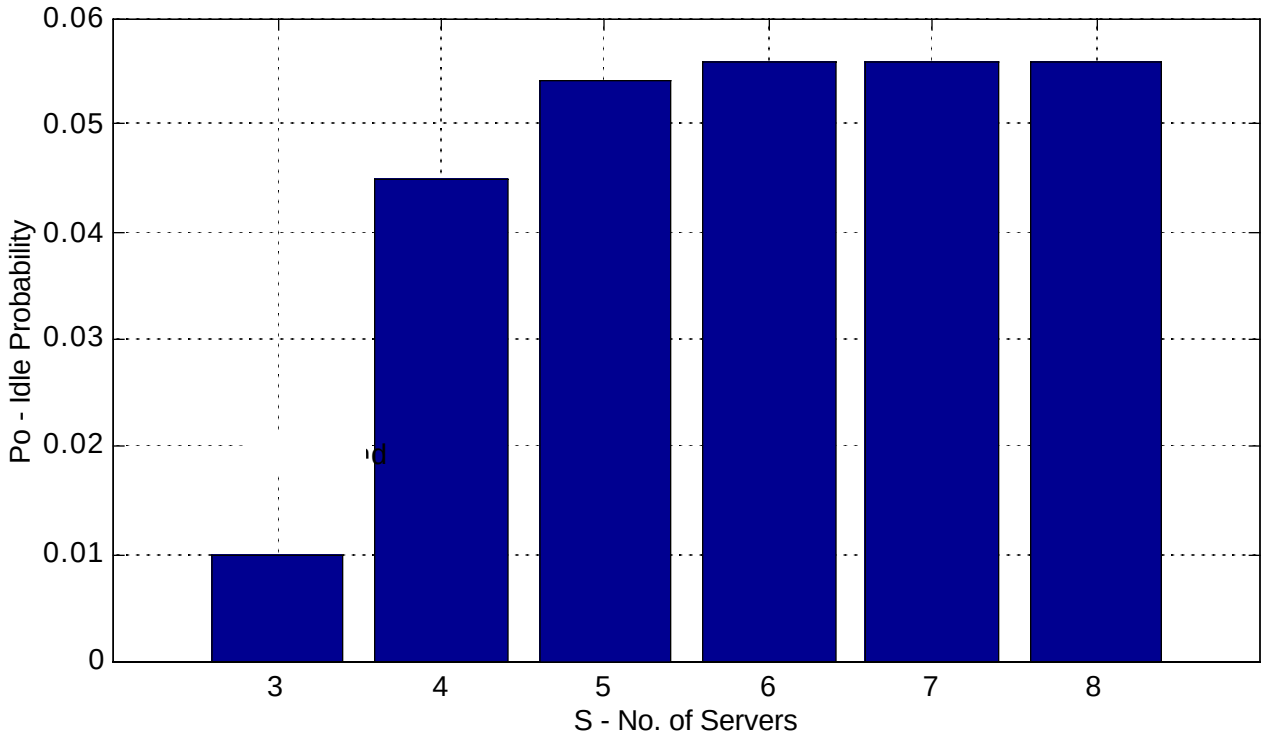
Since this waiting time violates the desired two minute maximum it is suggested that we try a higher number of x-ray machines to expedite service (at the expense of

cost). The following figure illustrates the sensitivity of P_o and L_q as the number of servers is increased.

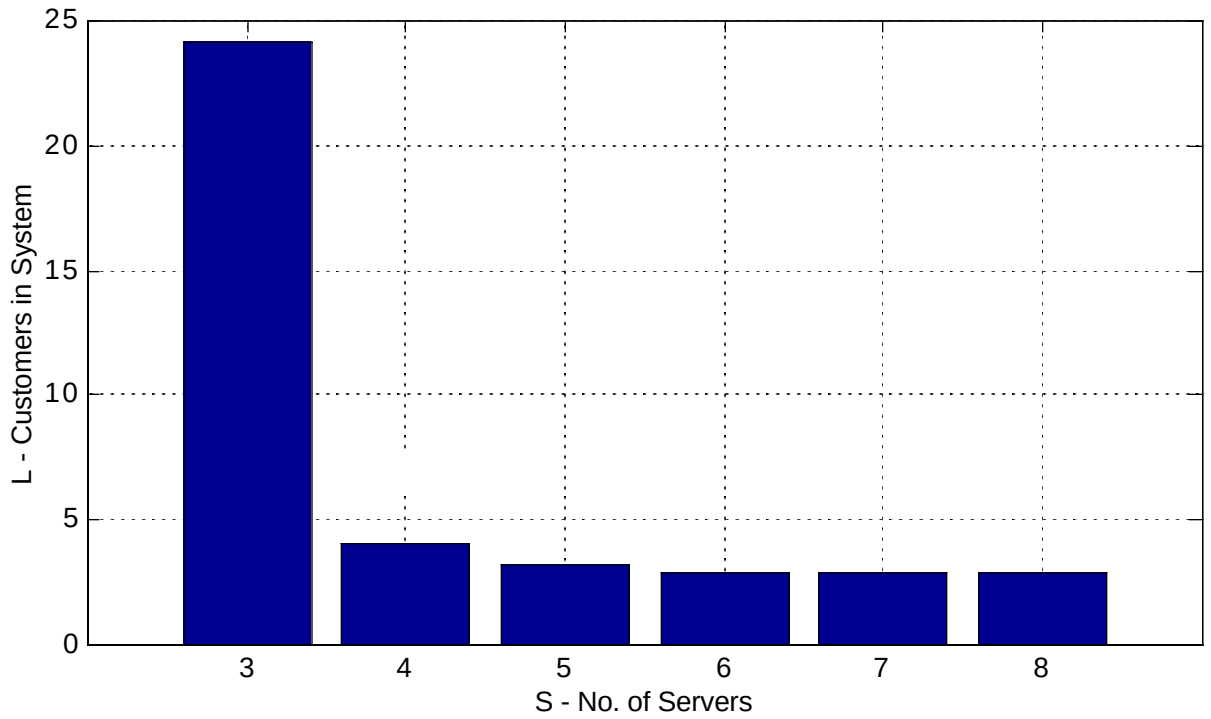
Note that four x-ray machines are needed to provide the desired average waiting time, *Wq*.

Sensitivity of *Po* with *S*

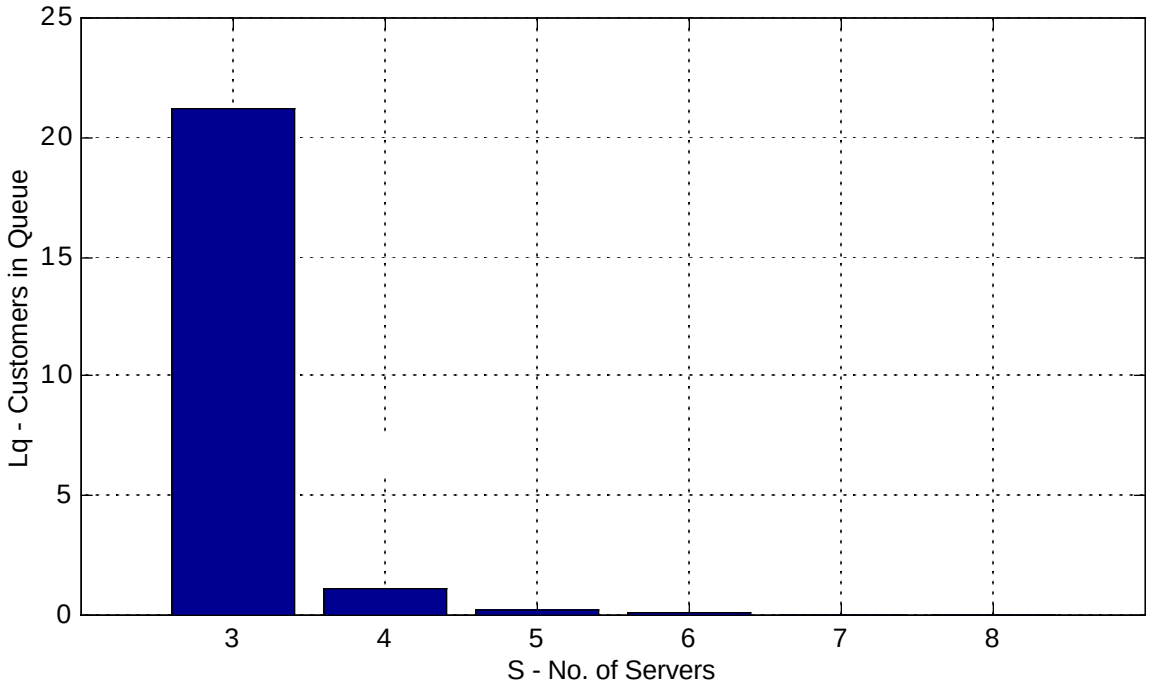
Note the quick variations in Po as S increases.



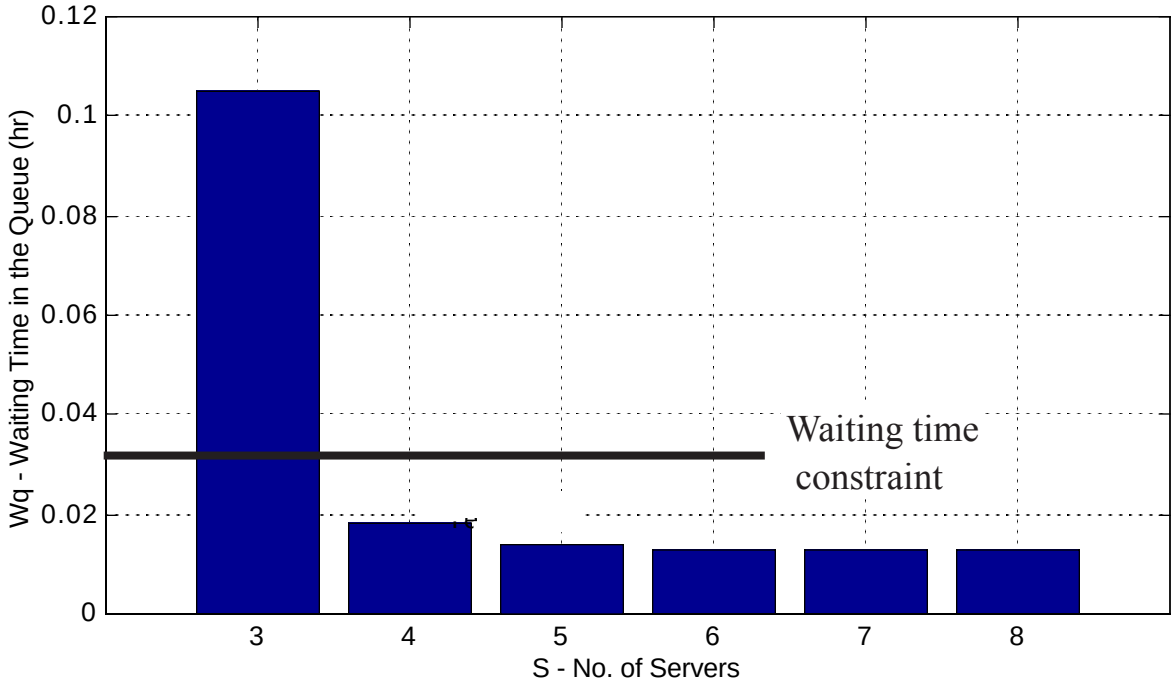
Sensitivity of *L* with *S*



Sensitivity of *Lq* with *S*



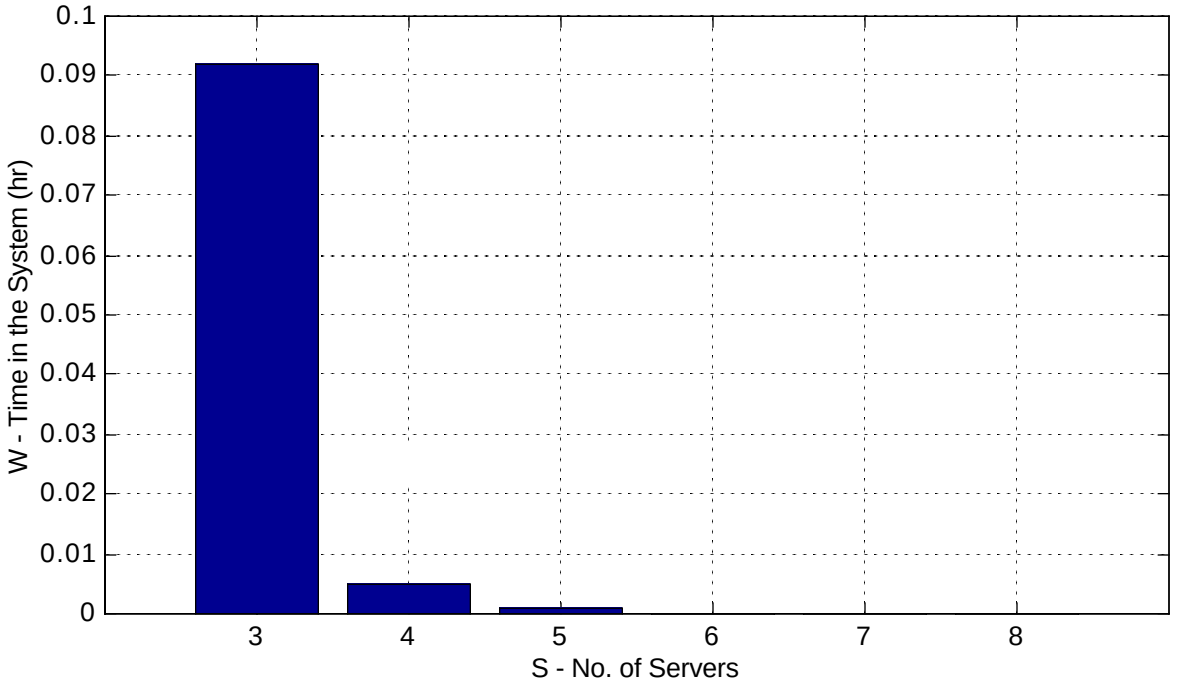
Sensitivity of *Wq* with *S*



This analysis demonstrates that 4 x-ray machines are needed to satisfy the 2-minute operational design constraint.

Sensitivity of *W* with *S*

Note how fast the waiting time function decreases with *S*



Security Check Point Results

c) The expected number of passengers in the system is (with $S = 4$),

$$L = \frac{\rho P_0 \left(\frac{\lambda}{\mu}\right)^s}{s!(1-\rho)^2} + \frac{\lambda}{\mu}$$

***L* =** 4.04 passengers in the system on the average design year day.

d) The utilization of the improved facility (i.e., four x-ray machines) is

$\rho = \lambda / (s\mu) = 230/ (4*80) =$ **0.72**

e) The probability that more than four passengers wait for service is just the probability that more than eight passengers are in the queueing system, since four are being served and more than four wait.

$$P(n > 8) = 1 - \sum_{n=0}^{8} P_n$$

where,

$$P_n = \frac{(\lambda/\mu)^n}{n!} P_0 \quad \text{if } n \le s$$

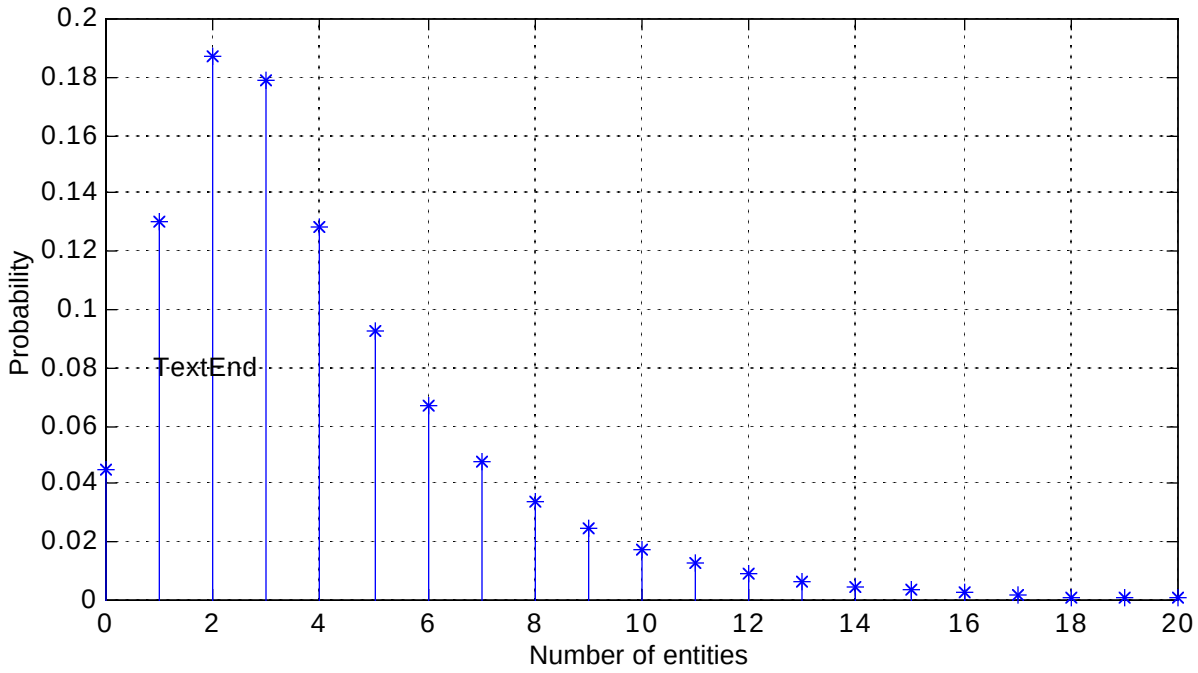
$$P_n = \frac{(\lambda/\mu)^n}{s! s^{n-s}} P_0 \quad \text{if } n > s$$

from where, $P_n > 8$ is 0.0879.

Note that this probability is low and therefore the facility seems properly designed to handle the majority of the expected traffic within the two-minute waiting time constraint.

PDF of Customers in System (L)

The PDF below illustrates the stochastic process resulting from poisson arrivals and neg. exponential service times



Matlab Computer Code

```
% Multi-server queue equations with infinite population

% Sc = Number of servers
% Lambda = arrival rate
% Mu = Service rate per server
% Rho = utilization factor
% Po = Idle probability
% L = Expected no of entities in the system
% Lq = Expected no of entities in the queue
% nlast - last probability to be computed

% Initial conditions

 S=5;
 Lambda=3;
 Mu = 4/3;
```

```
  nlast = 10;                                          % last probability
value computed

  Rho=Lambda/(S*Mu);

% Find Po
    Po_inverse=0;
    sum_den=0;

  for i=0:S-1 %                                        for the first term in the
denominator (den_1)
    den_1=(Lambda/Mu)^i/fct(i);
    sum_den=sum_den+den_1;
  end

    den_2=(Lambda/Mu)^S/(fct(S)*(1-Rho)); % for the second part of den
(den_2)
    Po_inverse=sum_den+den_2;
    Po=1/Po_inverse
```

```
% Find probabilities (Pn)

Pn(1) = Po;     % Initializes the first element of Pn column vector to be Po
n(1) = 0;       % Vector to keep track of number of entities in system

% loop to compute probabilities of n entities in the system

for j=1:1:nlast
      n(j+1) = j;
      if (j) <= S
           Pn(j+1) = (Lambda/Mu)^j/fct(j) * Po;
      else
            Pn(j+1) = (Lambda/Mu)^j/(fct(S) * Sc^(j-S)) * Po;
      end
end

% Queue measures of effectiveness

 Lq=(Lambda/Mu)^S*Rho*Po/(fct(S)*(1-Rho)^2)
```

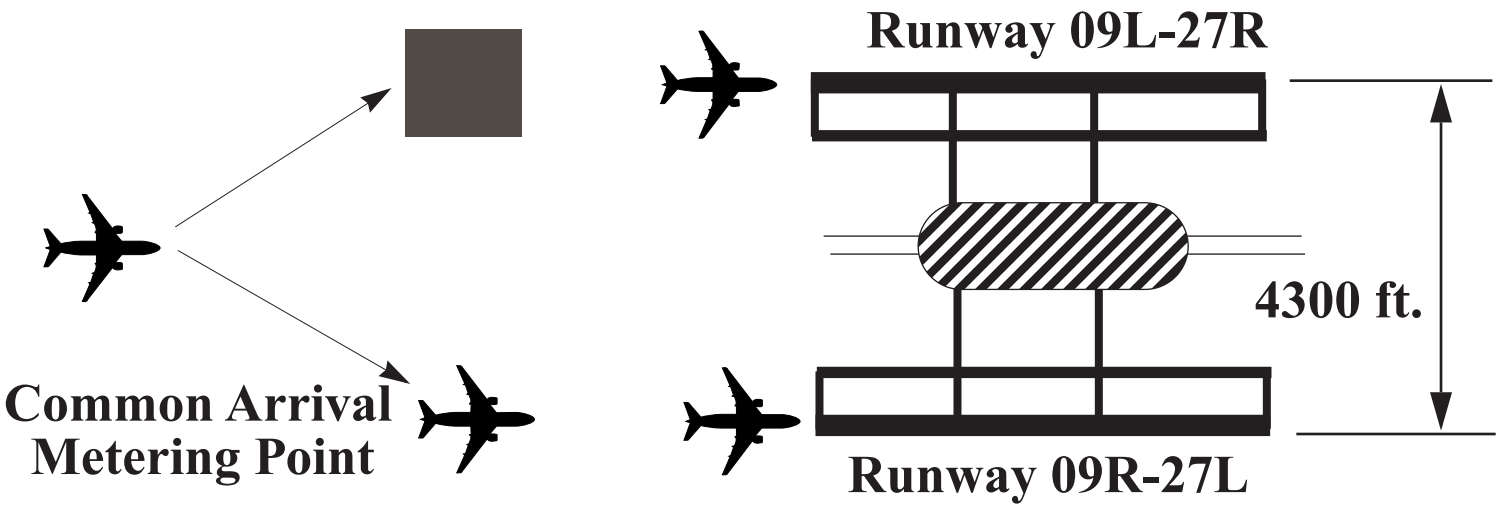
L=Lq+Lambda/Mu
Wq=Lq/Lambda
W = L/Lambda

plot(n,Pn)
xlabel('Number of entities')
ylabel('probability')

Example 4 - Airport Operations

Assume IFR conditions to a large hub airport with

- Arrival rates to metering point are 45 aircraft/hr
- Service times dictated by in-trail separations (120 s headways)

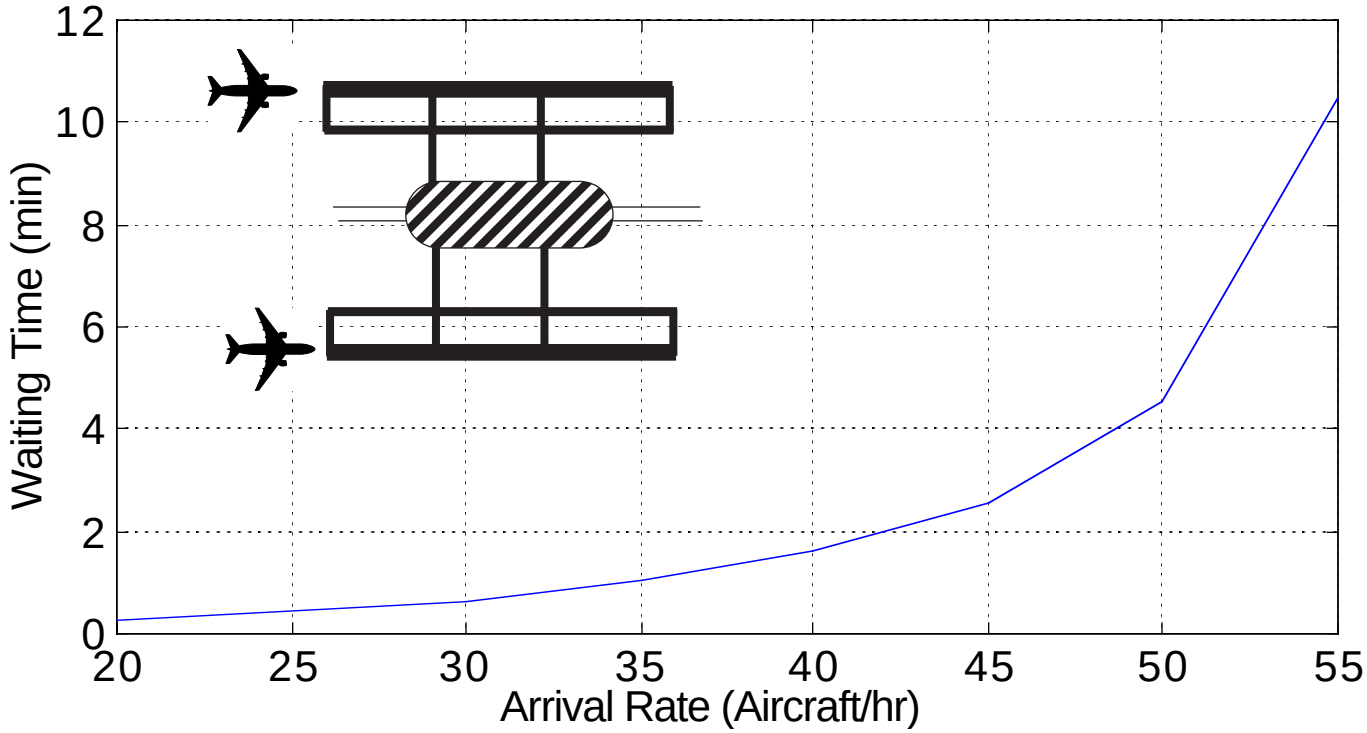


Some Results of this Simple Model

Parameter	Numerical Values
λ	45 aircraft/hr to arrival metering point
μ	30 aircraft per runway per hour
P_o	0.143
ρ	0.750
L	3.42 aircraft (includes those in service)
W_q	2.57 minutes per aircraft
W	4.57 minutes per aircraft

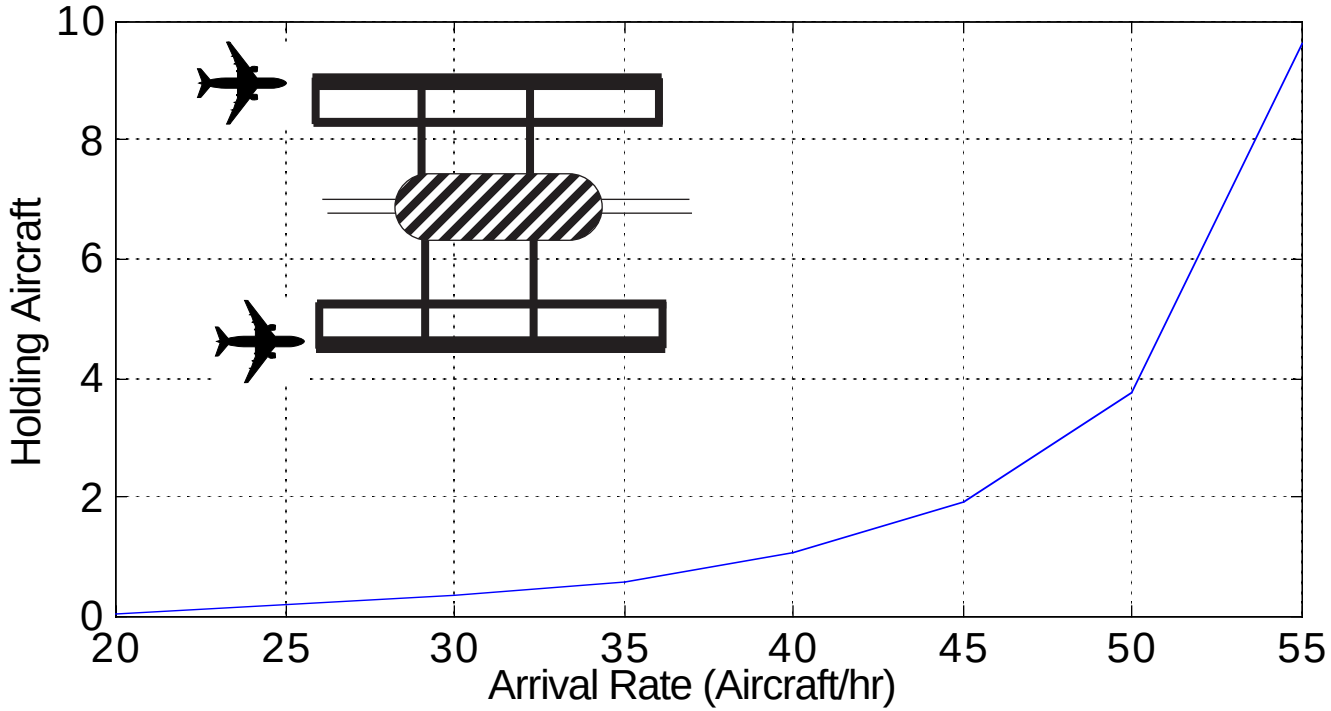
Sensitivity Analysis

Lets vary the arrival rate (λ) from 20 to 55 per hour and see the effect on the aircraft delay function.



Sensitivity of L_q with Demand

The following diagram plots the sensitivity of the expected number of aircraft holding vs. the demand function



Example # 5 Seaport Operations

- Seaport facility with 4 berths (a beth is an area where ships dock for loading/unloading)
- Arrivals are random with a mean of 2.5 arrivals per day
- Average service time for a ship is 0.9 days (assume a negative exponential distribution)

- **Find:**
- Expected waiting time and total cost of delays per year if the average delay cost is $12,000 per day per ship

Solution (use Stochastic Queueing Model - Infinite Population)

For the port example with parameters, $\lambda = 2.5$ ships per day and $\mu = 0.9$ ships per day (service rate) per berth.

Use stochastic queueing model equations (infinite population) to estimate queueing parameters,

- System utilization (%) = 69.4
- Idle probability (dim) = 0.05
- Expected No. of ships in queue (Lq) = 0.95
- Expected No. of ships in system (L) = 3.7

Average Waiting Time in Queue (days) = 0.38

Solution (Seaport Example)

- Average Waiting Time in System (includes service time) = 1.50 days
- The annual waiting cost (W_{cost}) is calculated using the following simple relationship,

$$W_{cost} = W_q \lambda N(C_{hour})$$

where;

W_q is the waiting cost per ship (days/ship), λ is the ship arrival rate to port (ships/day), N is the number of days in a year the port is open (days), and C_{hour} is the delay cost per unit of time per ship ($/day).

Solution (Seaport Example)

In the first year of operations (t=0), the port has an estimated delay cost,

$$W_{cost} = (0.382)\left[\frac{days}{ship}\right](2.5)\left[\frac{ships}{day}\right](365)[days]$$

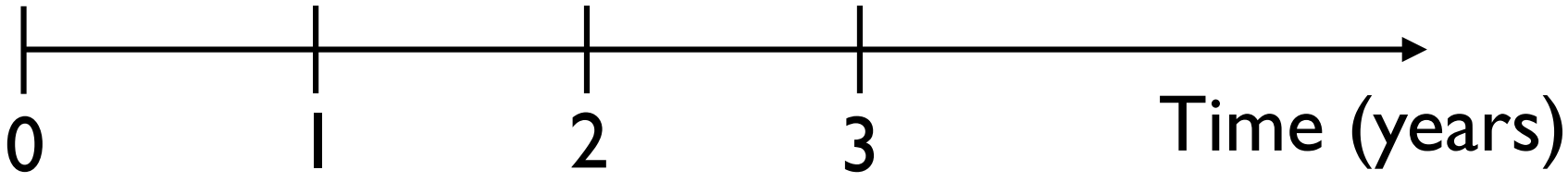
$$(12000)\left[\frac{dollars}{day}\right]$$

$$W_{cost} = 4{,}182{,}000 \text{ dollars}$$

Others Uses of Queueing Models (Facilities Planning)

- Queueing models can be used to estimate the life cycle cost of a facility
- Using the expected delays we can estimate times when a facility needs to be upgraded
- **For example,**
  - Suppose the demand function (i.e., number of ships arriving to port) for ships arriving to port increases 10% per year
  - Determine the year when new berths will be required if the port authority wants to maintain waiting times below 0.5 days.

Calculations for Seaport Example



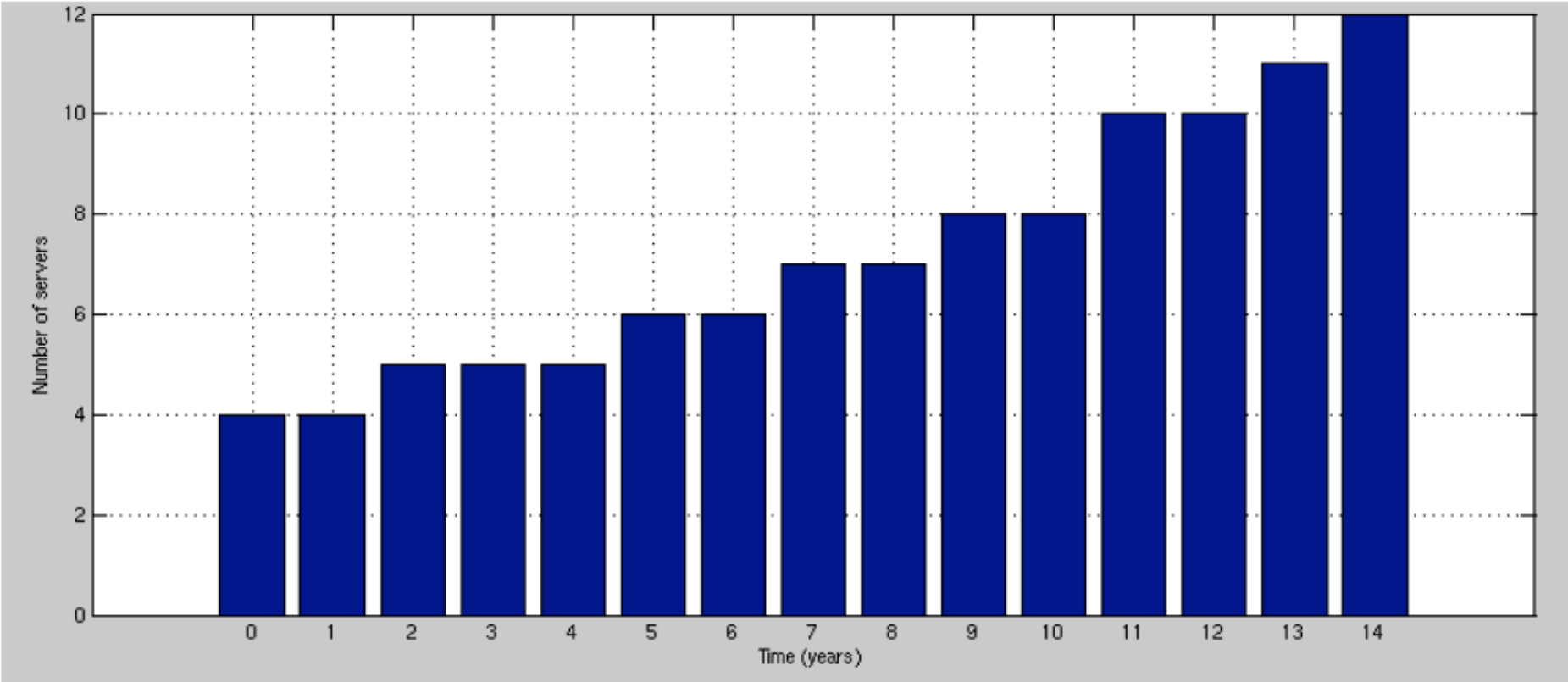
Repeat the process for future years. For example, in year t=1 the demand function is now 2.75 ships/day. The new waiting time in the queue is then 0.63 days.

Note that since the value of waiting time in port exceeds 0.5 days, we would have to construct a new berth facility either in year t=1 (**lag solution**) or in year t=0 (**lead solution**) to anticipate growth in the delay function.

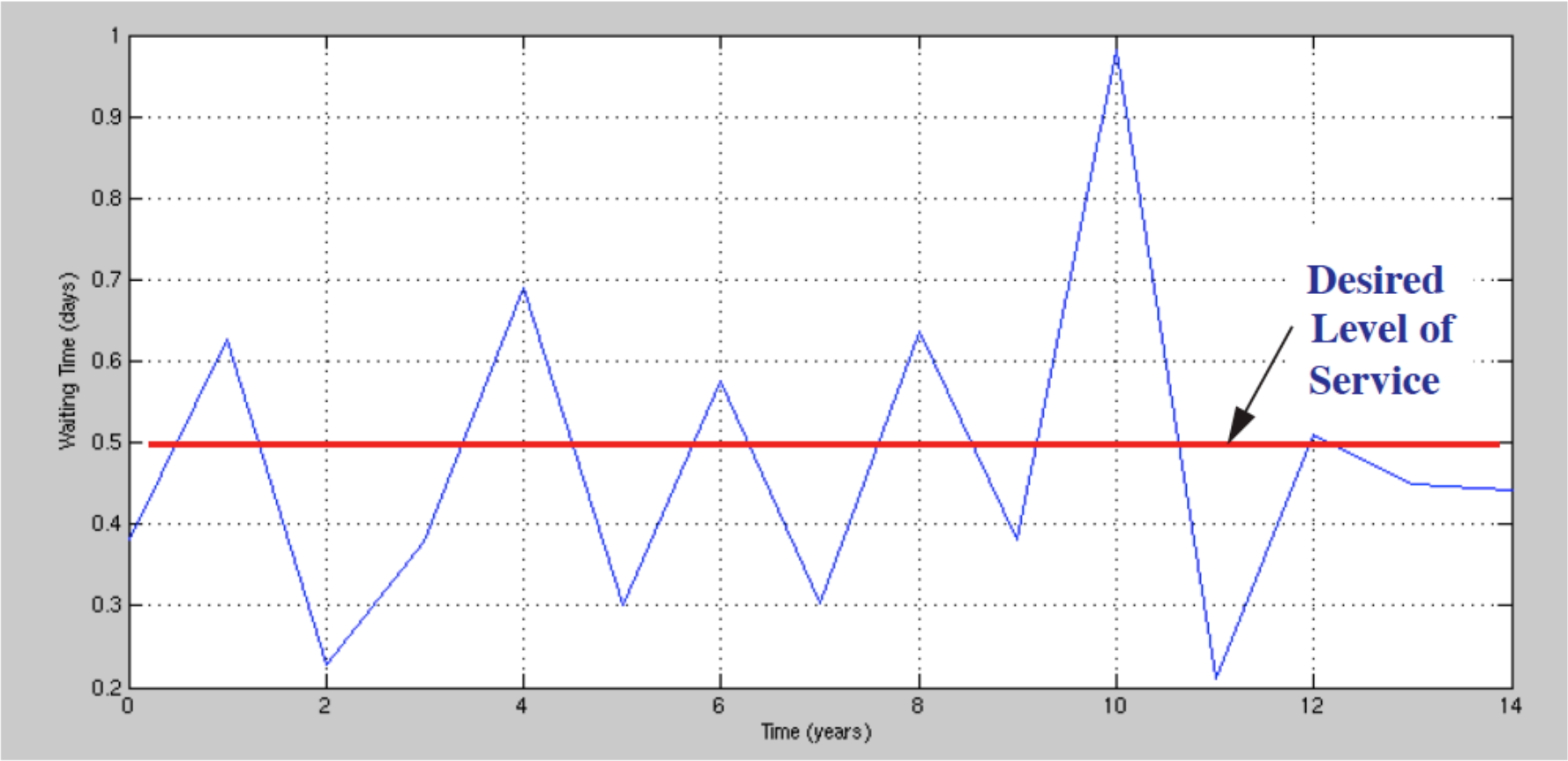
Note: For every iteration check the value of *Wq* to study when an update to the infrastructure is needed. The solution can be done iteratively or using trial and error calculations for every time period.

Calculations for Seaport Example

A solution to the problem is illustrated in the figure below. The solution presented here assumes that construction of berths occurs in the following year when the expected waiting time in port exceeds the desired Wq threshold.

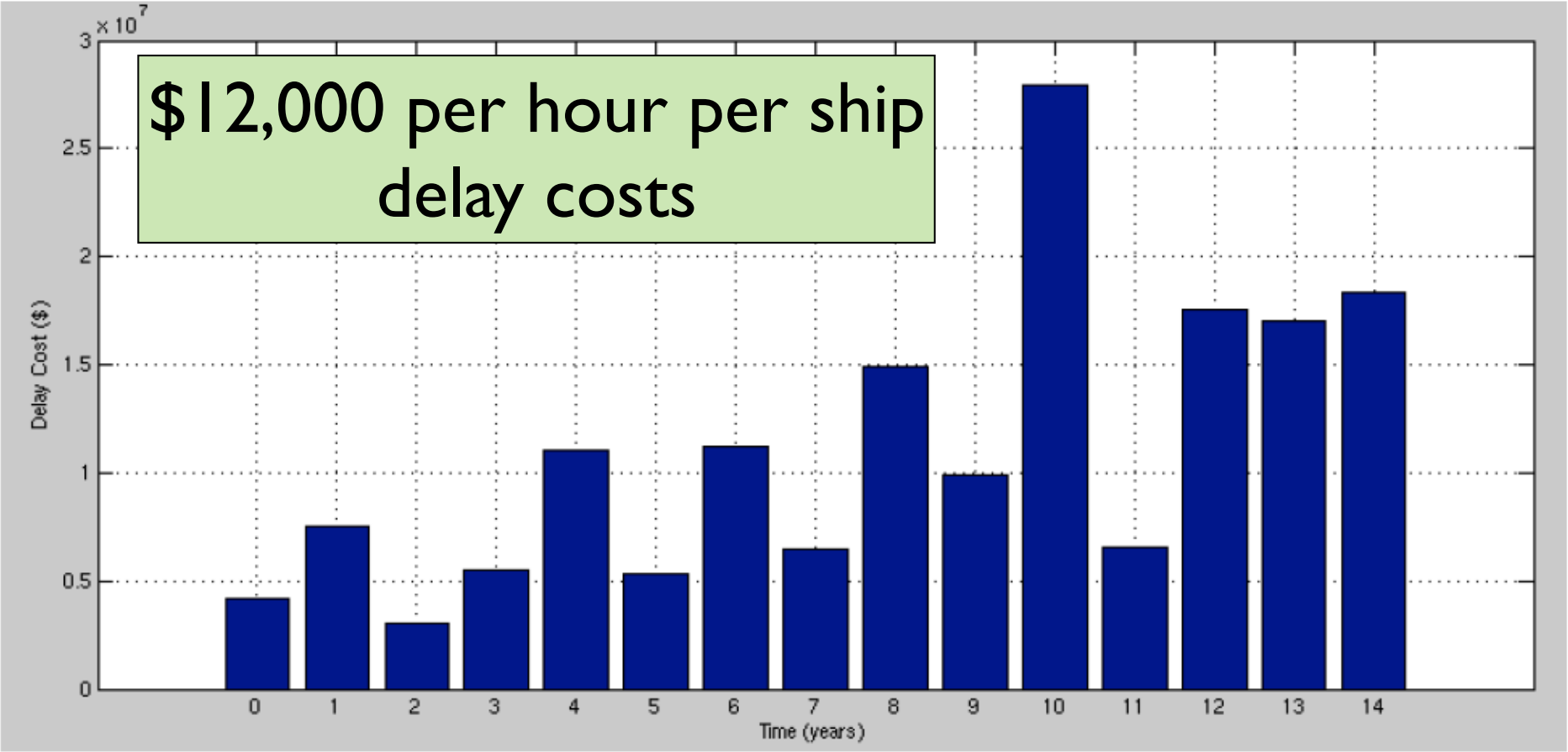


Calculations for Seaport Example



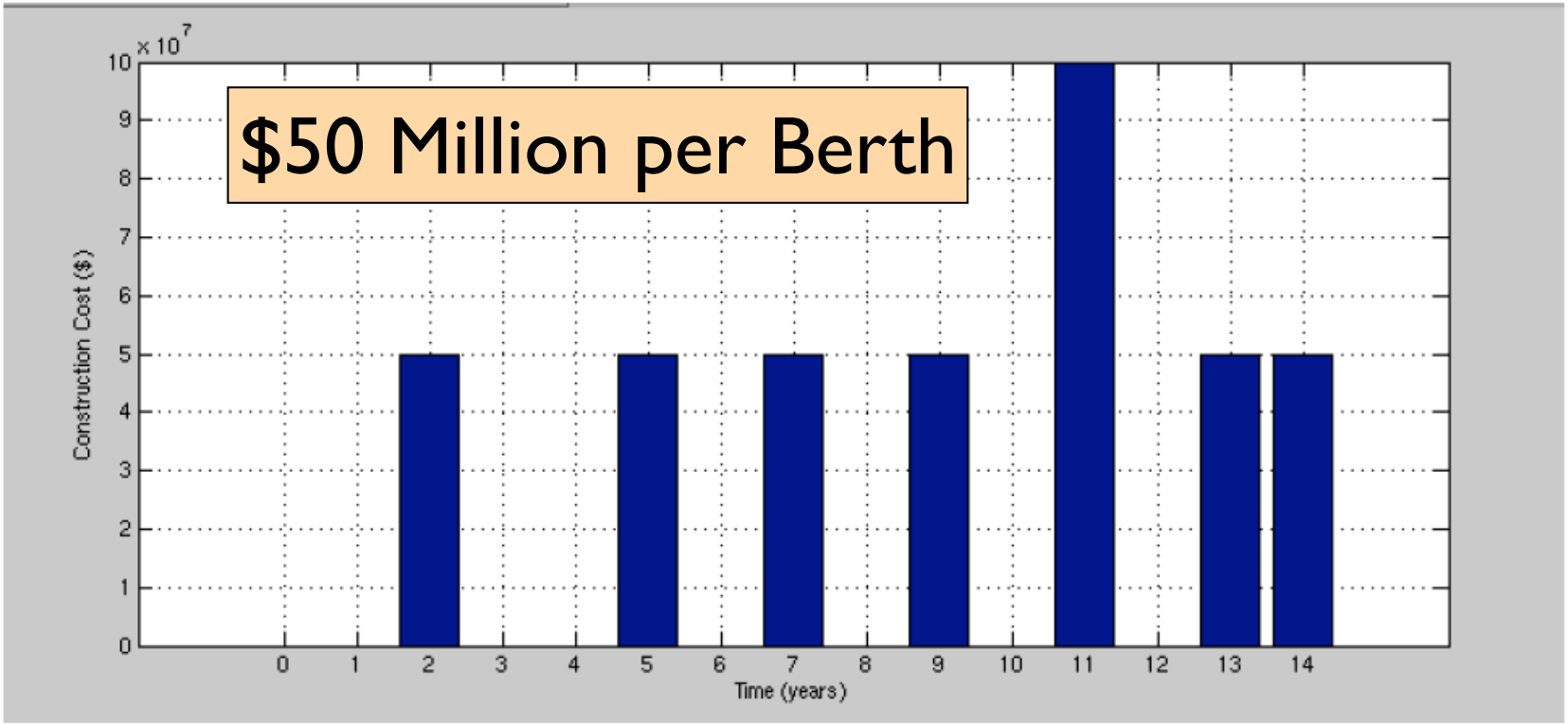
The figure above tracks the behavior ship delays over time. Note that the solution presented "lags" behind the solution that keeps Wq below 0.5 days per ship.

Undiscounted Annual Delay Costs (Lag Solution)



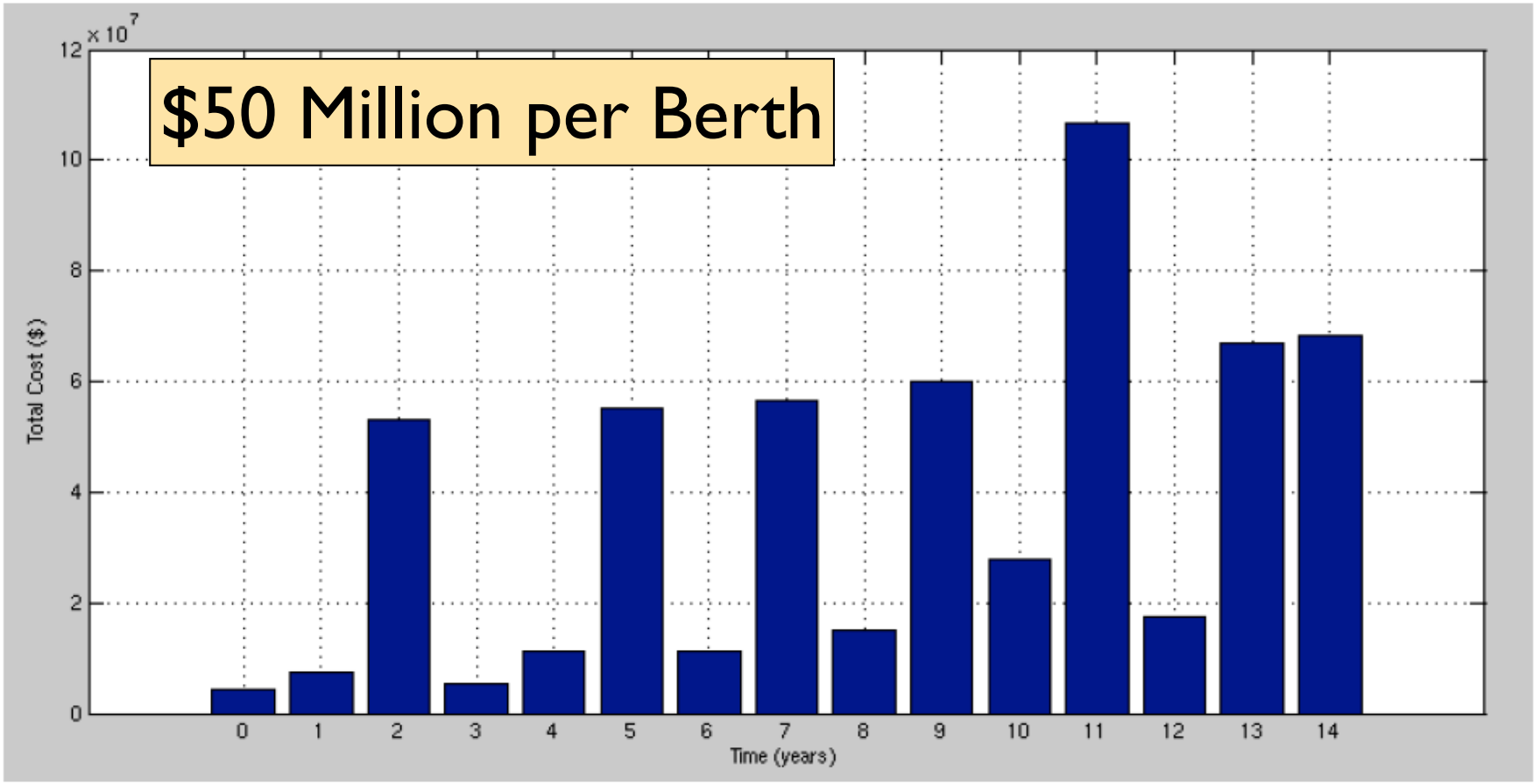
The bar plot shows the annual delay cost (undiscounted) for the lag solution.

Construction Cost Profile (Seaport) (Lag Solution)



The figure above illustrates the construction cost as a function of time for the "lag" solution. Note that in years 2,5,7,9, 13 and 14 one more berth is built. In year 11 two berths are needed. A total of at least 8 berths are needed.

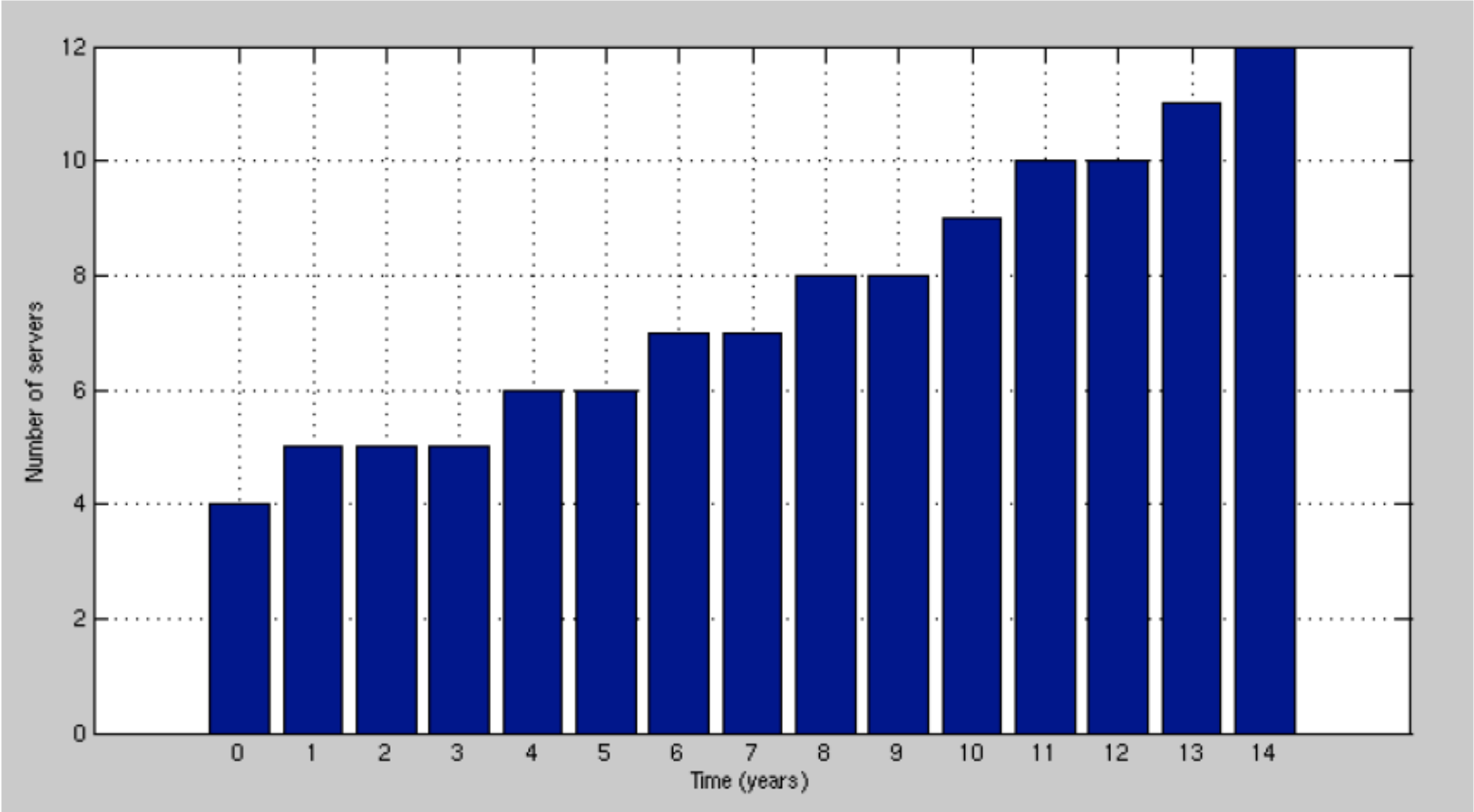
Total Annual Cost (Seaport - Lag Solution)



The total cost (construction plus delay costs) time behavior is illustrated above.

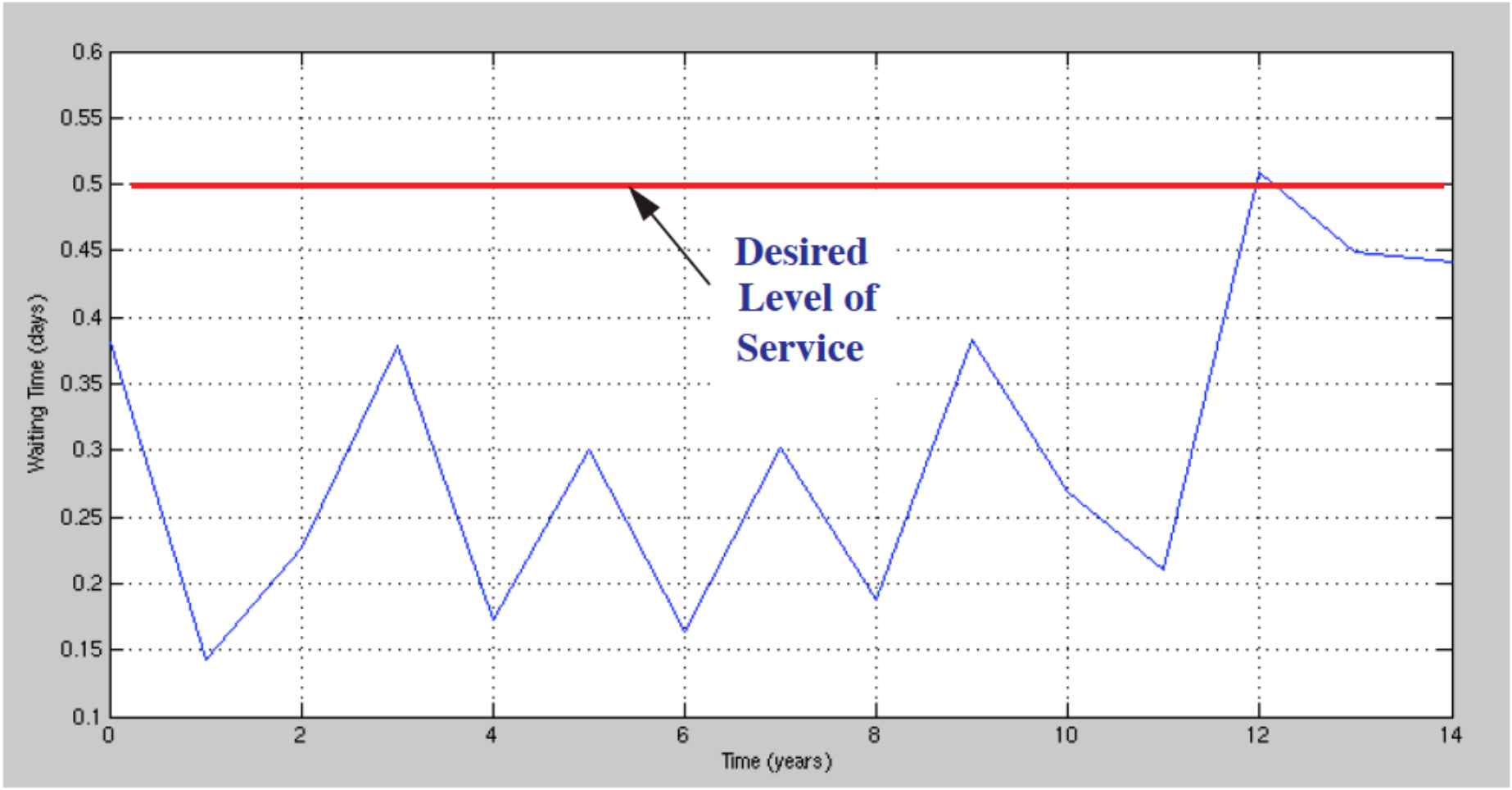
Lead Solution for Berth Construction

Another solution anticipates the violation of the level of service parameter (Wq). This "lead" solution also requires 8 berths throughout the life cycle of the system as illustrated below.



Lead Solution for Berth Construction

The lead solution yields the following level of service function (Wq as a function of time).



Comparing Both Solutions
Life Cycle Cost Analysis

Total life cycle cost (undiscounted) for lead solution is 5.134×10^8 dollars.

Total life cycle cost (undiscounted) for lag solution is 5.66×10^8 dollars.

The obvious conclusion is that building berths earlier in the life cycle saves money in the long run. Anticipating the violation of the level of service parameter (Wq) reduces the life cycle cost due to the reductions in delay cost.

Conclusions About Analytic Queueing Models

Virginia Tech

Advantages:

- Good traceability of causality between variables
- Good only for first order approximations
- Easy to implement

Disadvantages:

- Too simple to analyze small changes in a complex system
- Cannot model transient behaviors very well
- Large errors are possible because secondary effects are neglected
- Limited to cases where PDF has a close form solution