The "Fluid View", or Flow Models

- Introduction:
  - Legitimate models: Simple, General, Useful
  - Approximations (strong)
  - Tools
- Scenario analysis
  - vs. Simulation, Averaging, Steady-State
  - Typical scenario, or very atypical (eg. "catastrophy")
- Predictable Variability
  - Averaging scenarios, with small "CV"
  - A puzzle (the human factor ⇒ state dependent parameters)
  - Sample size needed increases with CV
  - Predictable variability could also turn unpredictable
- Hall: Chapter 2 (discrete events);
- 4 Pictures:
  - Cummulants
  - Rates (⇒ Peak Load)
  - Queues (⇒ Congestion)
  - Outflows (⇒ end of rush-hour)
- Scales (Transportation, Telephone (1976, 1993, 1999))
- Simple Important Models: EOQ, Aggregate Planning
- Skorohod's Deterministic Fluid Model (of a service station): teaching note
  - Phases of Congestion: under-, over- and critical-loading.
  - Rush Hour Analysis: onset, end
  - Mathematical Framework in approximations
- Queues with Abandonment and Retrials (=Call Centers; Time- and State-dependent Q's).
- Bottleneck analysis in a (feed-forward) Fluid Network, via National Cranberry
- Fluid Networks (Generalizing Skorohod): The Traffic Equations
- Addendum

The "Fluid View" or Flow Models of Service Networks

There is a rich body of literature on Fluid Models. It originates in many sources, it takes many forms, and it is very powerful when used properly. For example, the classical EOQ model takes a fluid view of an inventory system, and physicists have been analyzing *macroscopic models* for decades. Not surprisingly, however, the first explicit and influential advocate of the *Fluid View* to queueing systems is a Transportation Engineer (Gordon Newell, from Berkeley). To understand why this view was natural to Newell, just envision an airplane that is landing in an airport of a large city, at night – the view, in rush-hour, of the network of highways that surrounds the airport, as seen from the airplane, is precisely this fluid-view. (The influence of Newell is apparent in Hall's book, which again is not surprising: Hall graduated from Berkeley as Newell's PhD student, I believe.)

Illuminating quotes:

Oliver & Samuel, "Reducing letter delays in post-offices":
"Variation in mail flow are not so much due to random fluctuations about a known mean as they are time-variations in the mean itself.... Major contributor to letter delay within a post-office is the shape of the input flow rate: about 70% of all letter mail enters a post office within 4-hour period".
(Remark: In contrast, random fluctuations around / about a known mean are handled, for example, within the News-vendor paradigm; see also Yield/Revenue Management.)

Hall, page 187–8: "...a busy freeway toll plaza may have 8000 arrivals per hour, which would provide a coefficient of variation of just 0.011 for 1 hour. This means that a nonstationary Poisson arrivals pattern can be accurately approximated with a deterministic model". (Note: the calculations are based on a Poisson model, in which mean = variance.)

I shall now list the main (three) roles that fluid models play in the world of Service Engineering, as I perceive them: fluid models are interesting and useful in their own right, they provide simple approximations to complicated systems, and they constitute powerful technical tools in the analysis of stochastic systems.

1. Legitimate **models** for real systems, with prevalent predictable variability that dominates stochastic variability (verified, for example, by small CV, or by averaging).

   Examples (Newell, Hall, Harrison and Optional Readings):
   Industrial Eng.: Old EOQ-like models and the new BPR paradigm.

2.1

Inventory buildup diagrams (See the Trucks in National Cranberries).
Mean-value analysis (in Computer Science)
Transportation engineers often "think fluid" (see Newell's book).
Airport traffic (planes and people).
Vandergraft, Hall on staffing.
Service factories, for example mail-sorting.

Advantages of fluid-models:

- *Simple* to formulate (intuitive), fit (empirically) and analyze (elementary). (See the Homework on Empirical Models.)
- Cover a *broad spectrum* of features, relatively effortlessly.
- Often, they are *all that is needed* (for example, in capacity analysis, bottleneck identification, or utilization profiles, as in National Cranberries Cooperative and HW2.)

2. Useful **approximations**: first-order deterministic fluid approximations, via Functional (Strong) Laws of Large Numbers (FLLN), to support both performance analysis and control.
   - *Long-run*, detects trends. (See Chen and M.)
     - Identify bottlenecks (covered later, via National Cranberries.)
     - Traffic equations, for example in Jackson networks. (M.Sc. HW)
     - Stability and instability (currently very active).
   - *Short-run*, captures instantaneous (predictable) variability (Massey, Pats).
     - Identify phases in evolution (see Hall, pg. 189-191: stagnant = overloading with queues increasing then decreasing, back to stagnant.)
3. Technical **Tools** (articles by Jim Dai, and Sasha Stolyar - see our website).
   - Lyapounov functions: It is sometimes that case that sample paths of a stochastic system is attracted to *Fluid sample-paths*. This helps establish stability/instability, weak convergence or asymptotic-control optimality in a stochastic environment, but via a deterministic analysis.
   - Mathematical framework for analysis and approximations (reflection), which is amenable to the use of the continuous mapping theorem.

Further references on the Fluid View are provided in the reading packets within the syllabus.

Predictable Queues

Fluid Models

Service Engineering
Queueing Science

Eurandom

September 8, 2003

e.mail : avim@tx.technion.ac.il

Website: http://ie.technion.ac.il/serveng

3. Supporting Material (Downloadable)

Gans, Koole, and M.: “Telephone Call Centers: Tutorial, Review and Research Prospects.” *MSOM*.

Brown, Gans, M., Sakov, Shen, Zeltyn, Zhao: "Statistical Analysis of a Telephone Call Center: A Queueing-Science Perspective." Submitted.

Jennings, M., Massey, Whitt: "Server Staffing to Meet Time-Varying Demand." Management Science, 1996. - PRACTICAL

0. M., Massey, Reiman: "Strong Approximations for Markovian Service Networks." QUESTA, 1998.

1. M., Massey, Reiman, Rider: "Time Varying Multi-server Queues with Abandonment and Retrials", ITC-16, 1999.

2. M., Massey, Reiman, Rider and Stolyar: "Waiting Time Asymptotics for Time Varying Multiserver Queues with Abandonment and Retrials", Allerton Conference, 1999.

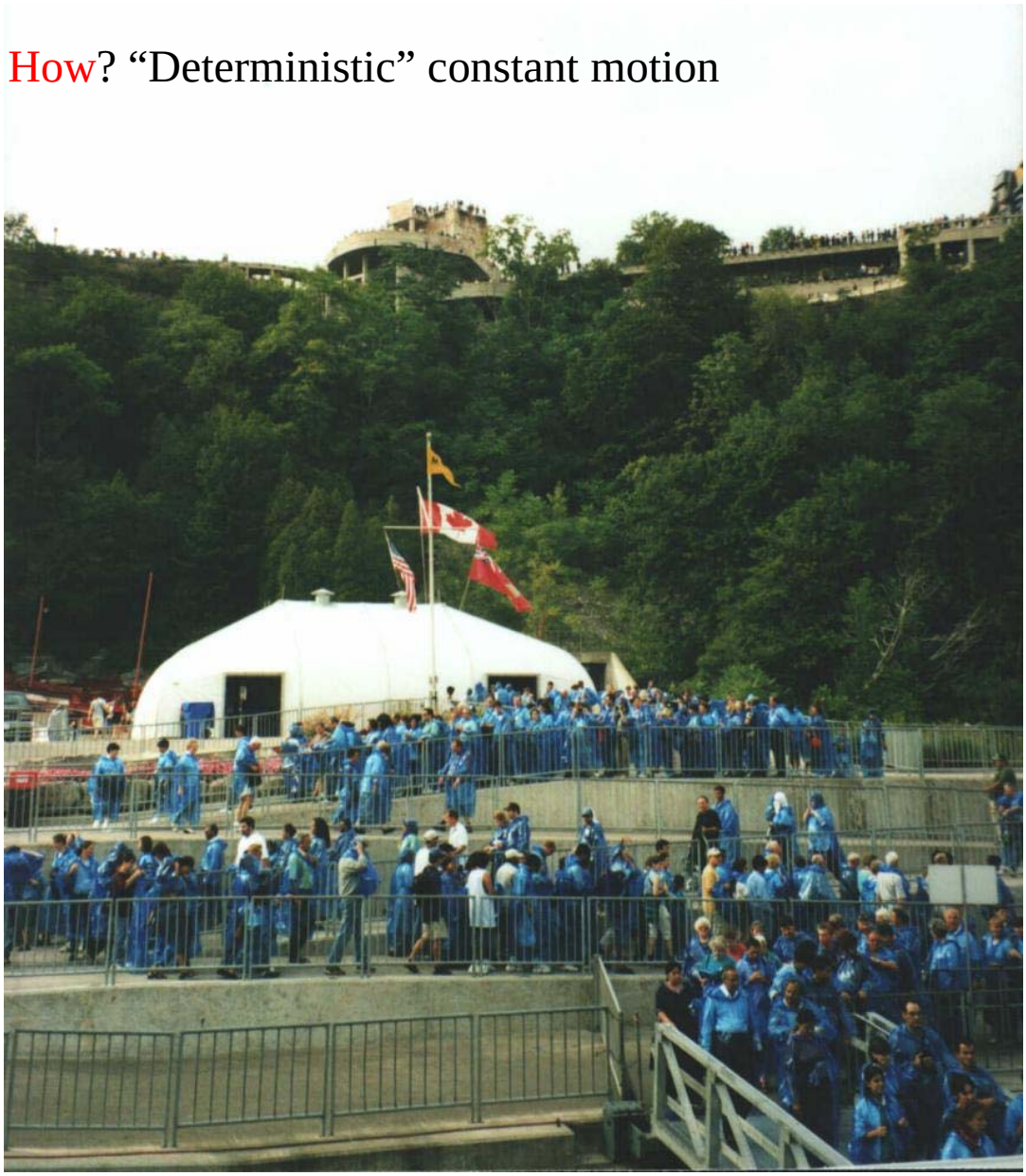
3. M., Massey, Reiman, Rider and Stolyar: "Queue Lengths and Waiting Times for Multiserver Queues with Abandonment and Retrials", Fifth INFORMS Telecommunications Conference, 2000

Labor-Day Queueing in Niagara Falls

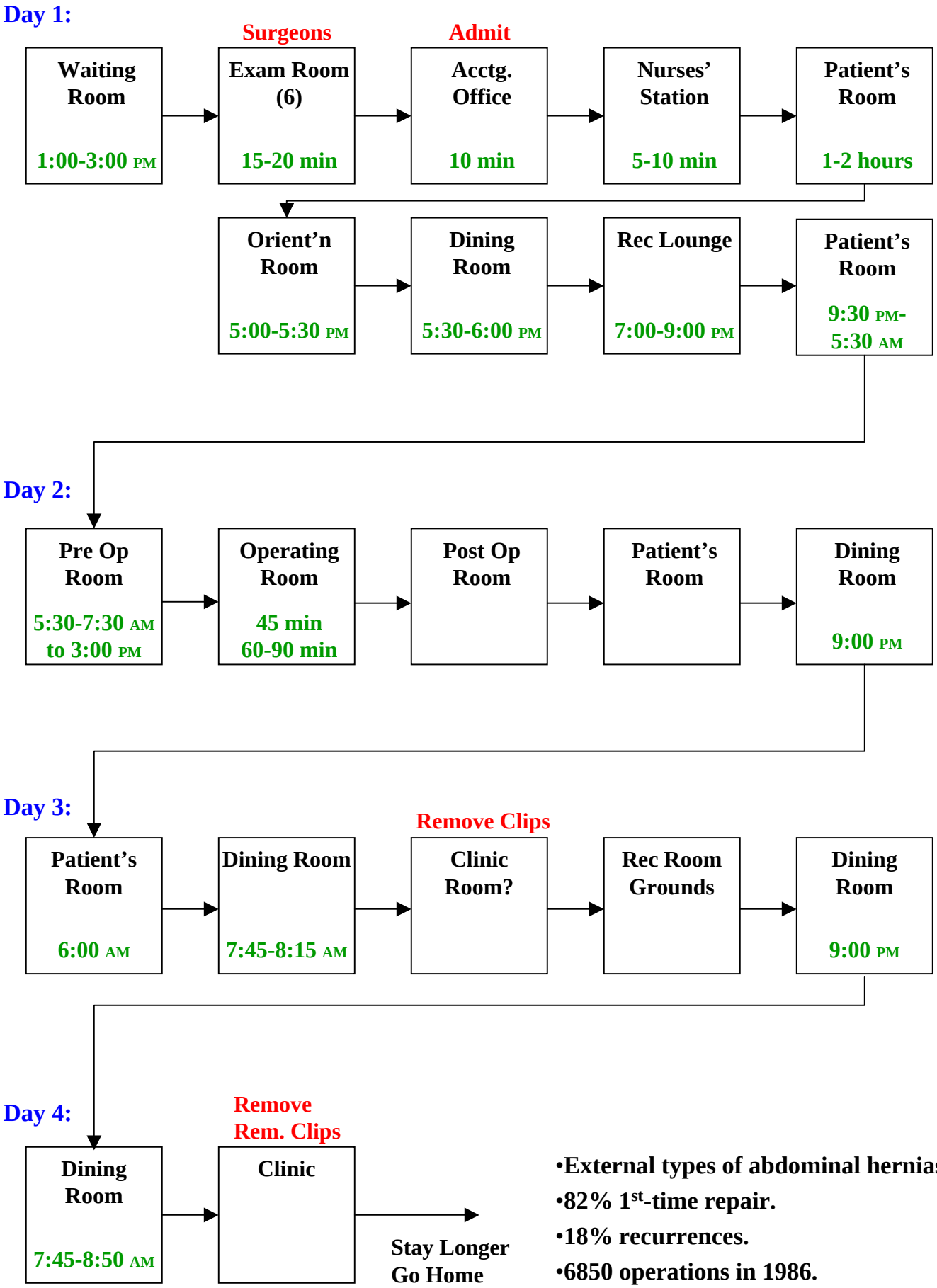
Three-station Tandem Network:
Elevators, Coats, Boats

Total wait of 15 minutes
from upper-right corner to boat

How? “Deterministic” constant motion

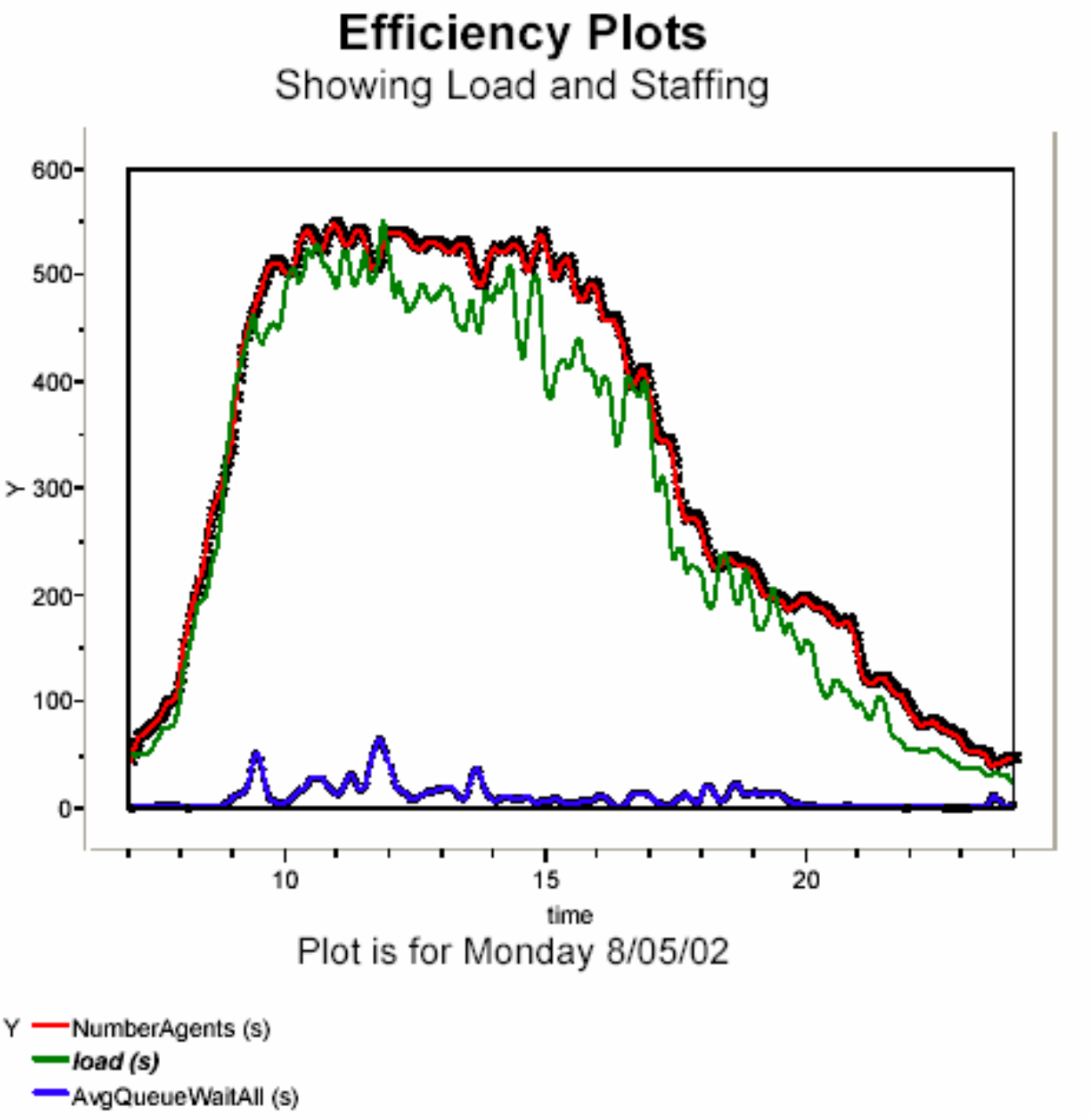


Shouldice Hospital: Flow Chart of Patients' Experience



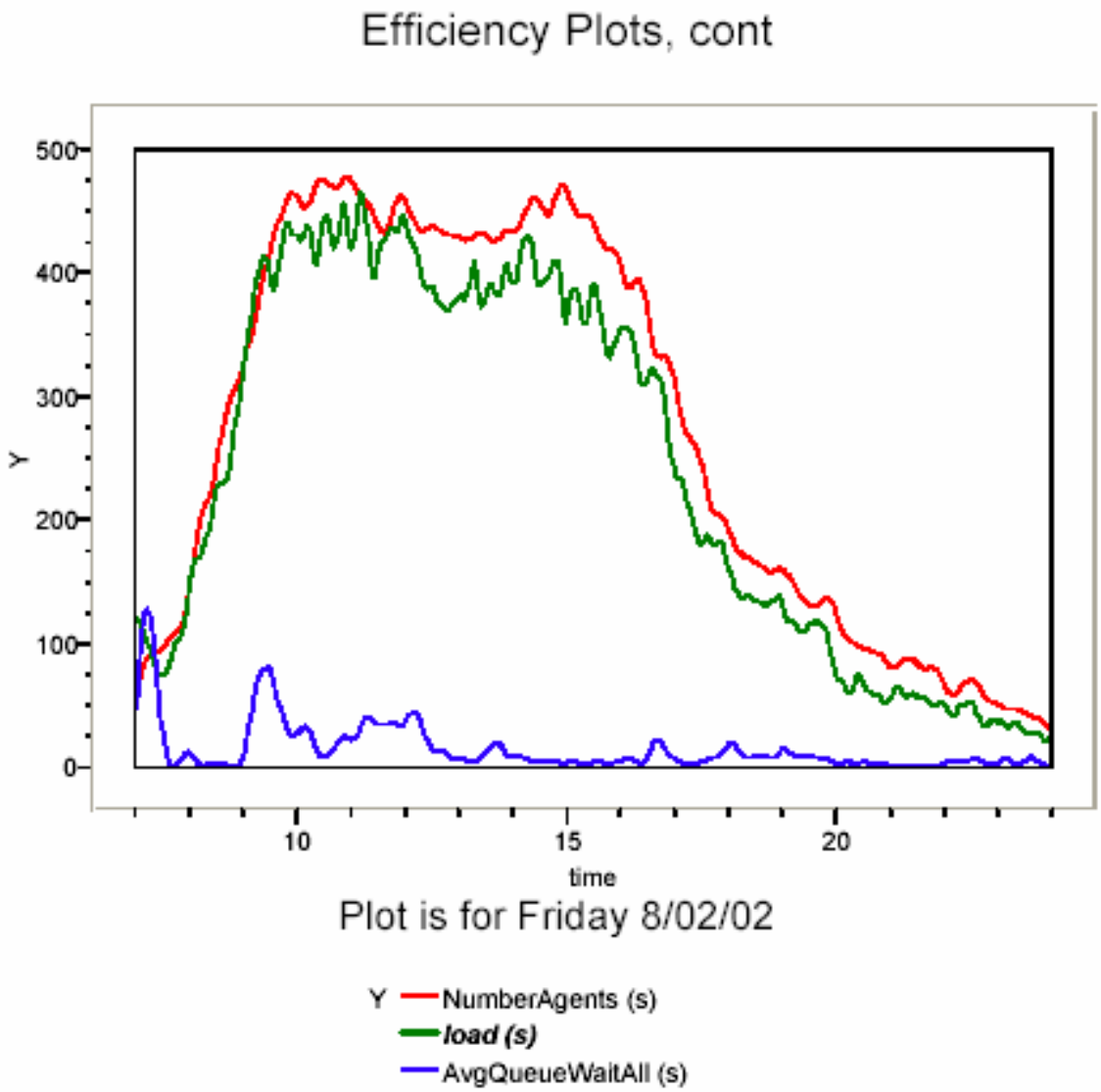
- External types of abdominal hernias.
- 82% 1st-time repair.
- 18% recurrences.
- 6850 operations in 1986.
- **Recurrence rate: 0.8% vs. 10% Industry Std.**

Matching Supply and Demand (Wharton)



"Agents" = *Estimate* of number of agents on-duty at that time.
[In each 150 second interval an agent is estimated to be on-active-duty for the entire interval if (s)he is on the phone sometime in that interval.]

Staffing Matters (on Fridays, 7:00 am)

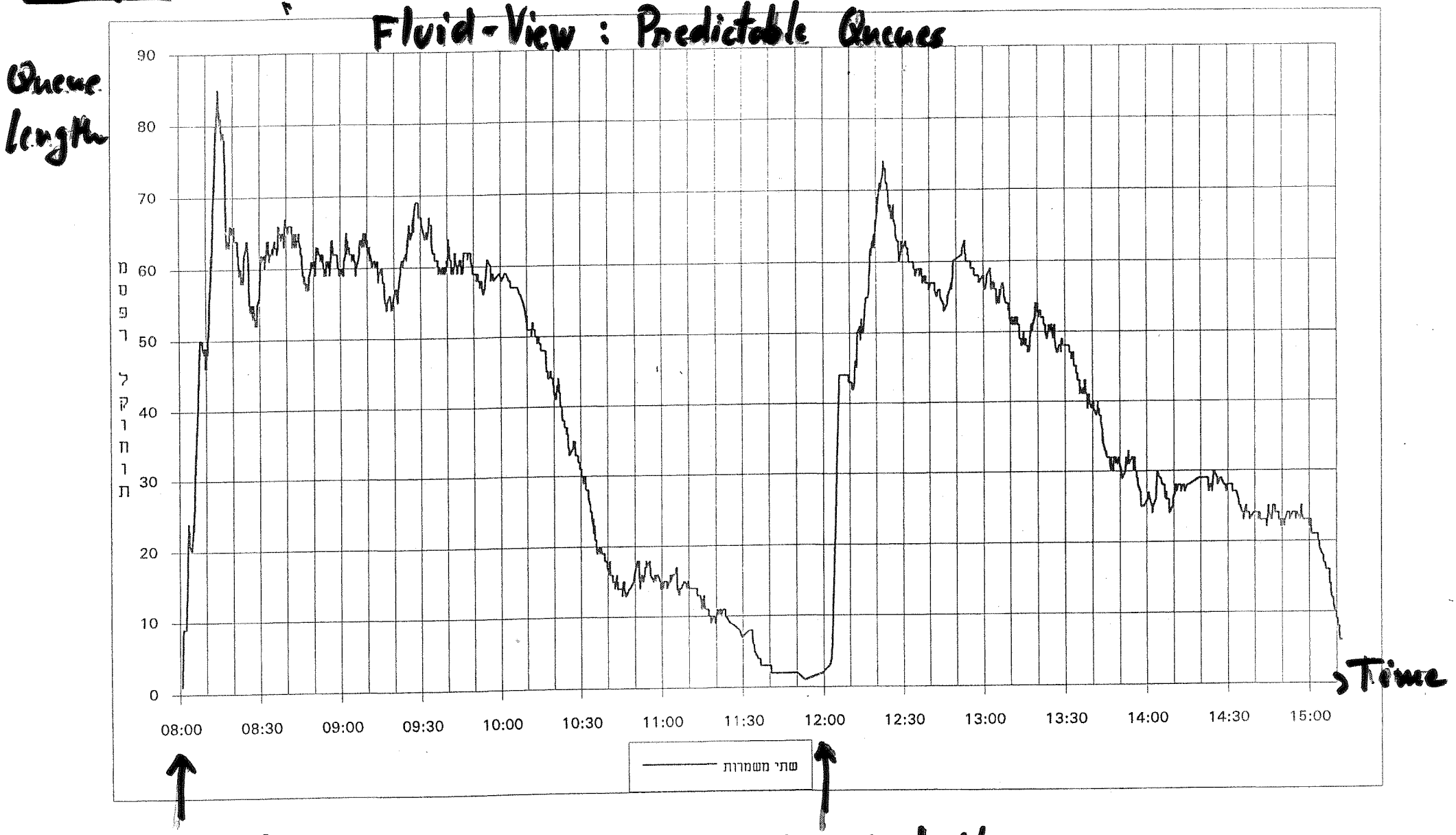


Note increased usage from 7-7:30 am (typical of Fridays).
Note increased average Queue-Wait during this time.
(Accompanied by a rise in abandonments to about 10%.)

Overall Utilization: 8/02/02 = 88%
8/05/02 = 89%

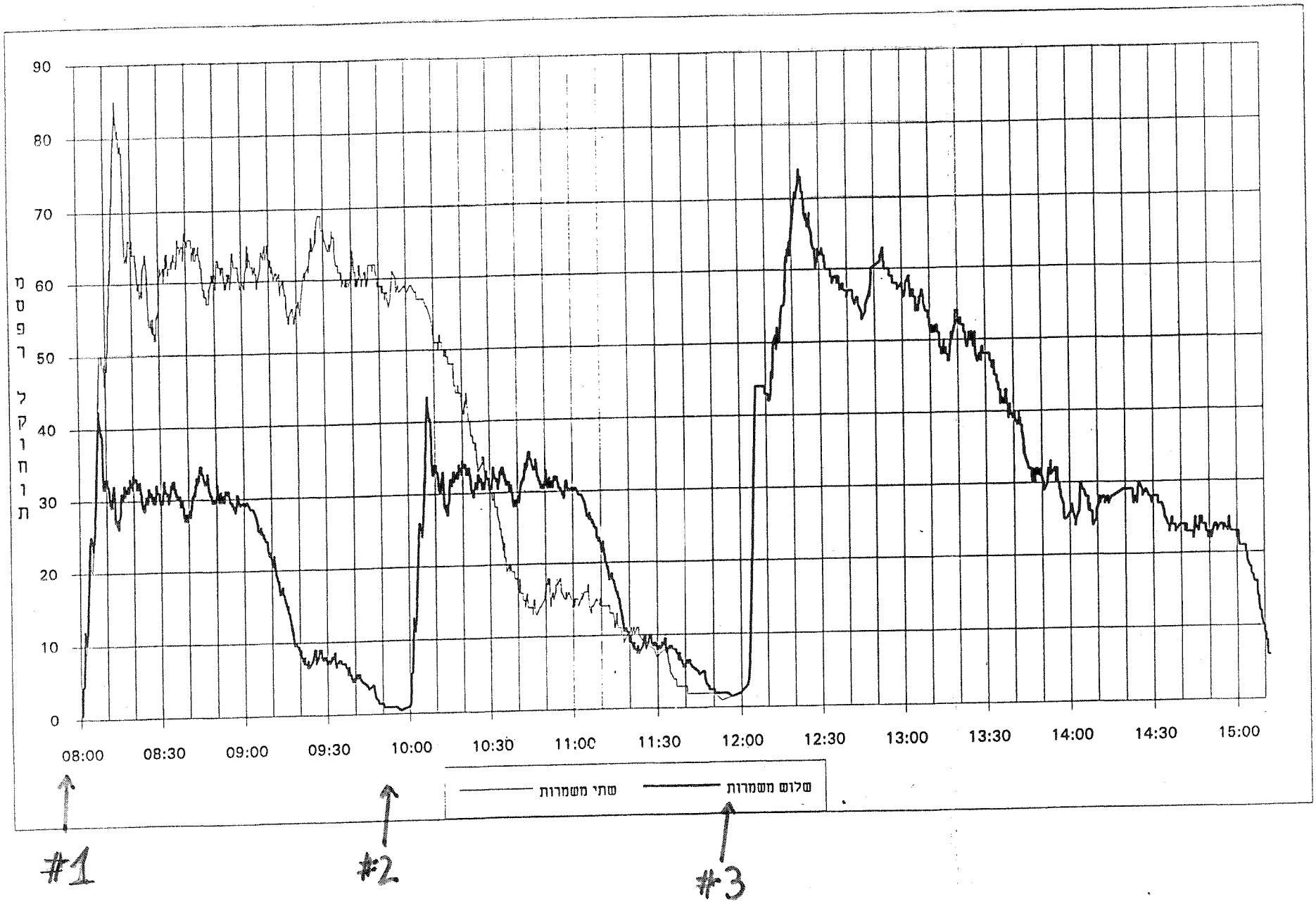
Scenario Analysis (ניתוח תרחישים) Government UN-Employment-office

Fluid-View : Predictable Queues

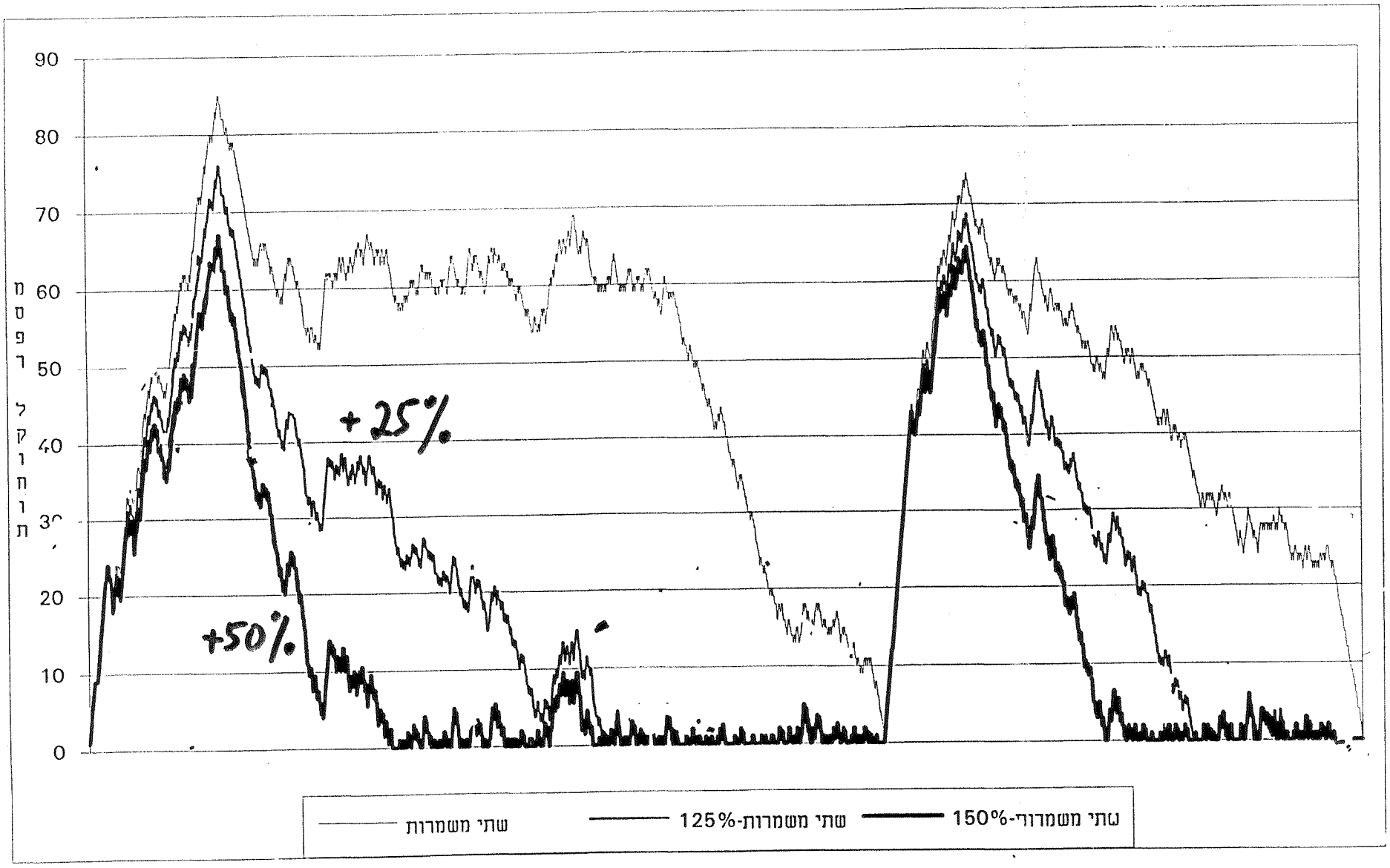


3 Shifts

3 משמרות



↑ workforce

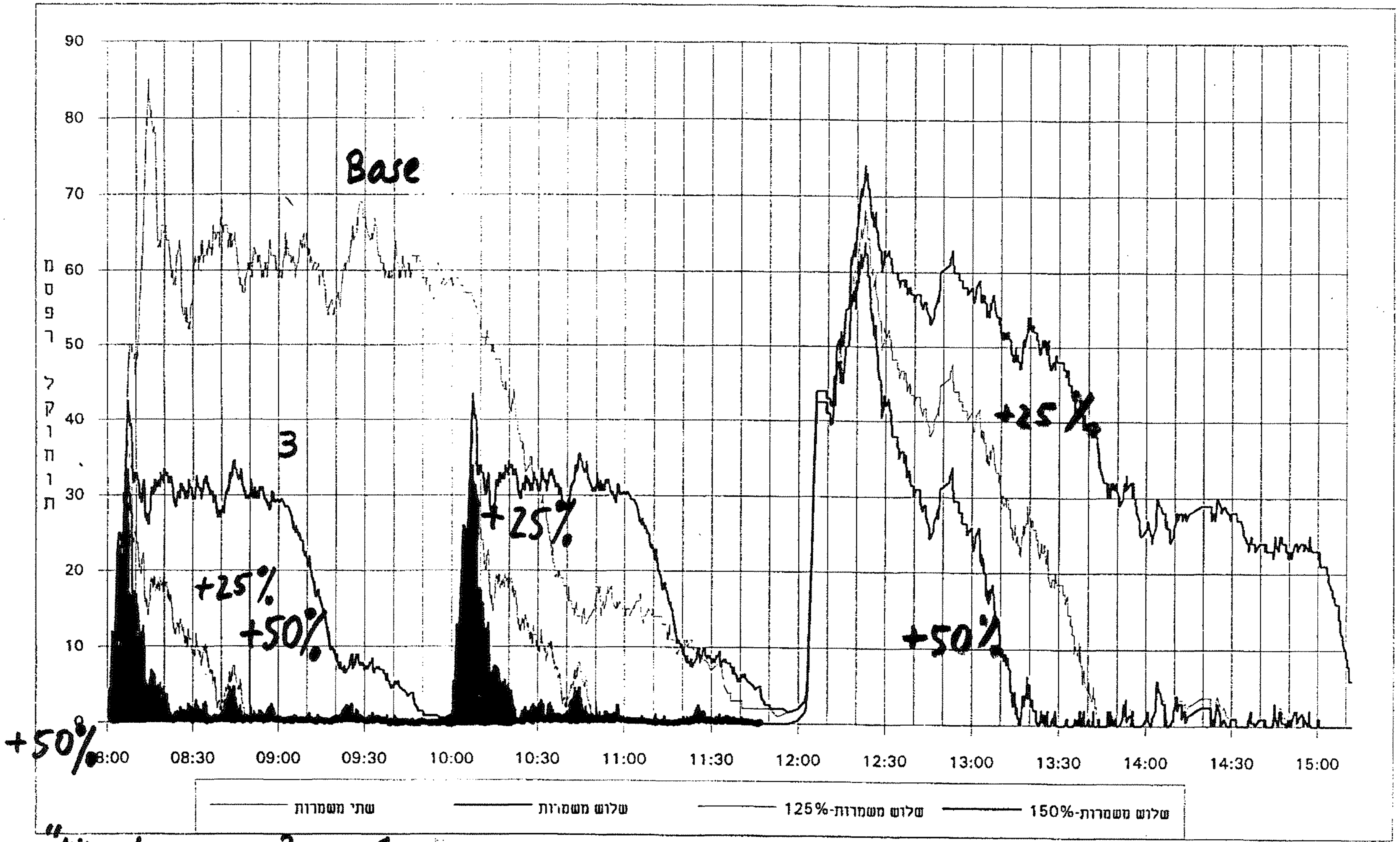


Reduces duration of peak, not size! (Explain? optional HW)

Scenario Analysis
(ניתוח תרחישים)

תוספת כ"א
+ 3 משמרות

How to model?

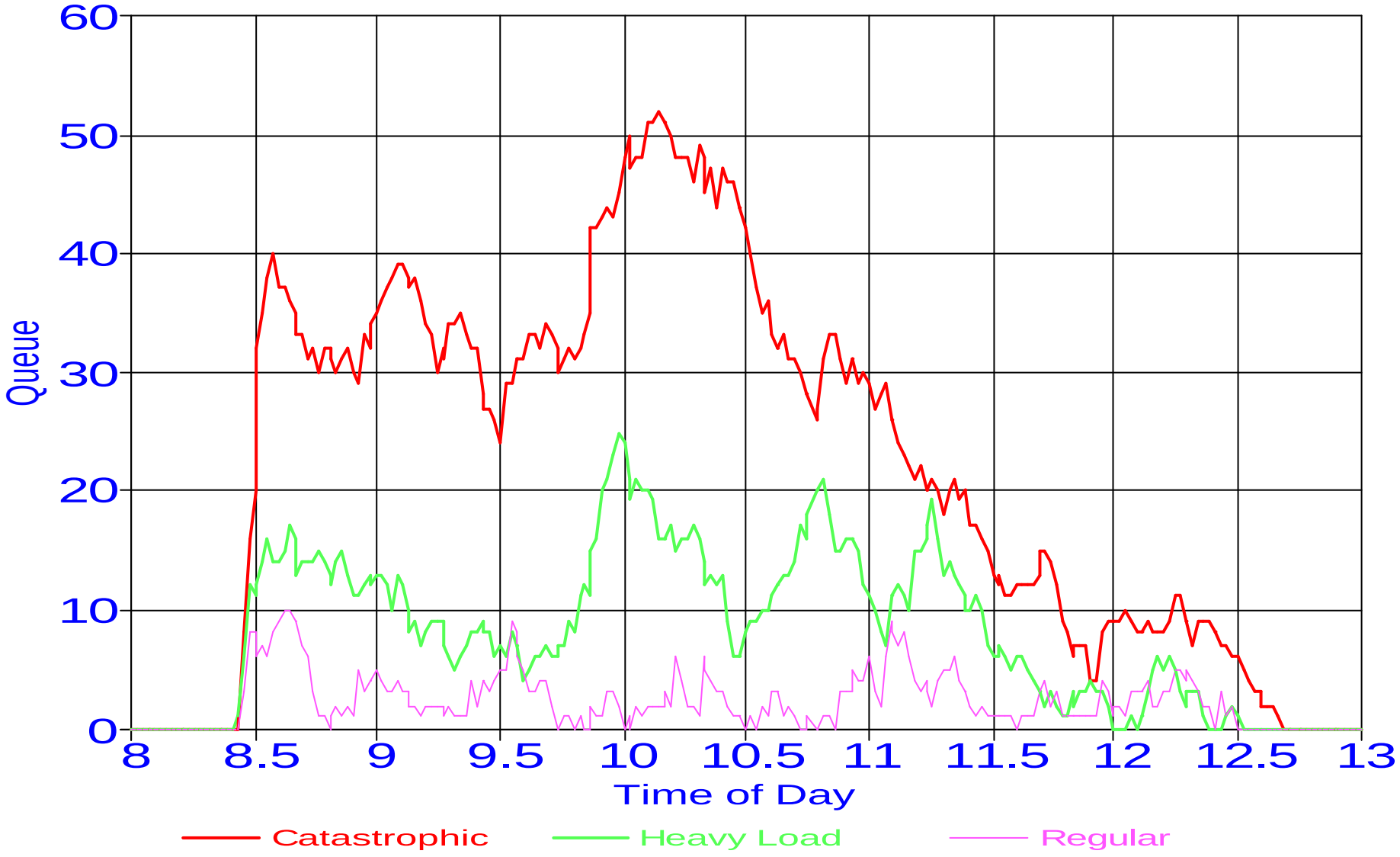


מקבר משק תור ש"א
שטח = שעות לקוח ממתין
(≈ מחיר המתנה)

Must have: Scenario ≈ Reality well represented!
Else?

I-Method

Bank Queue

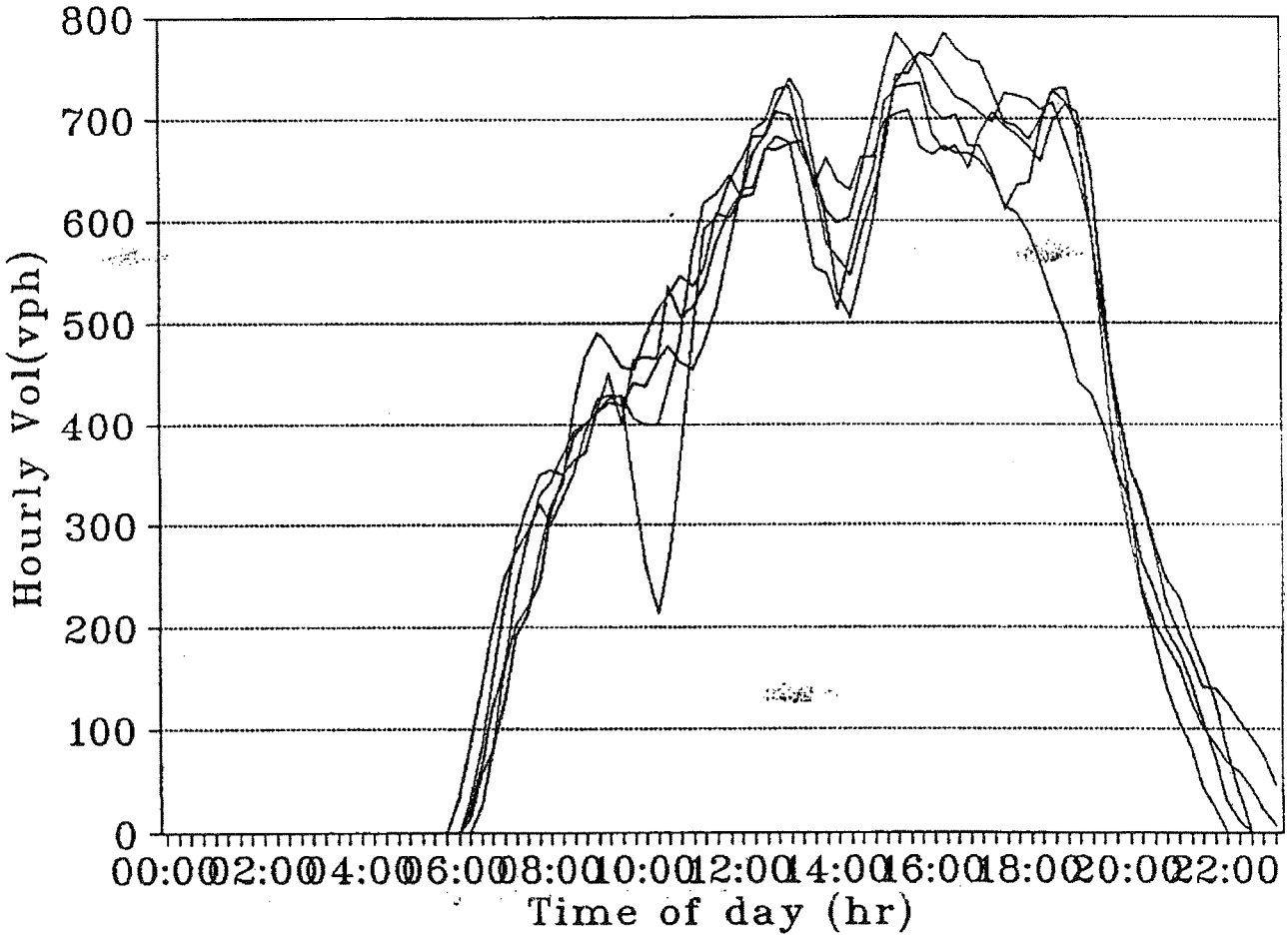


Discrete

Units ?

HERTZEL – BALFUR

KN010103-1019



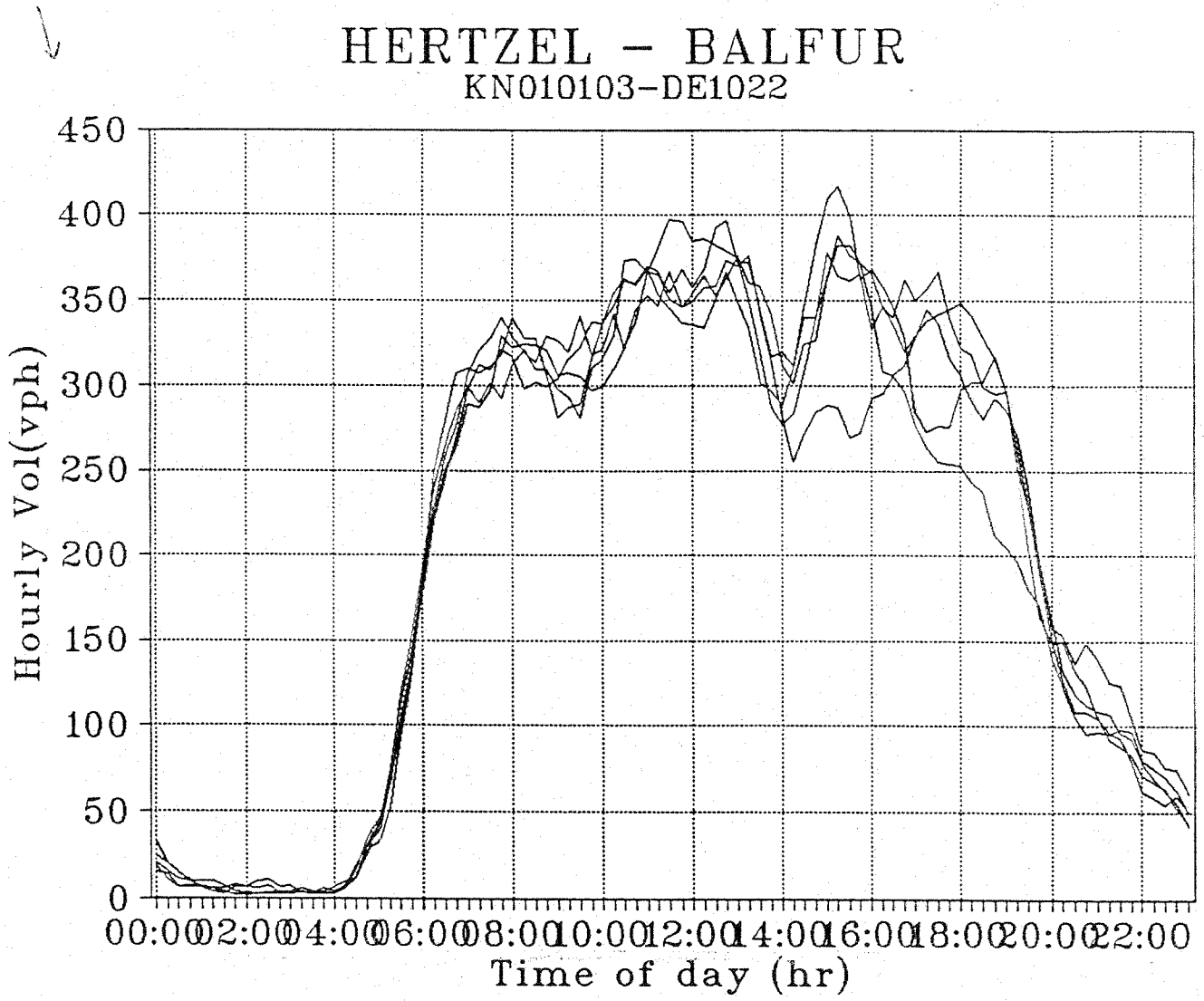
Data via one of six detectors

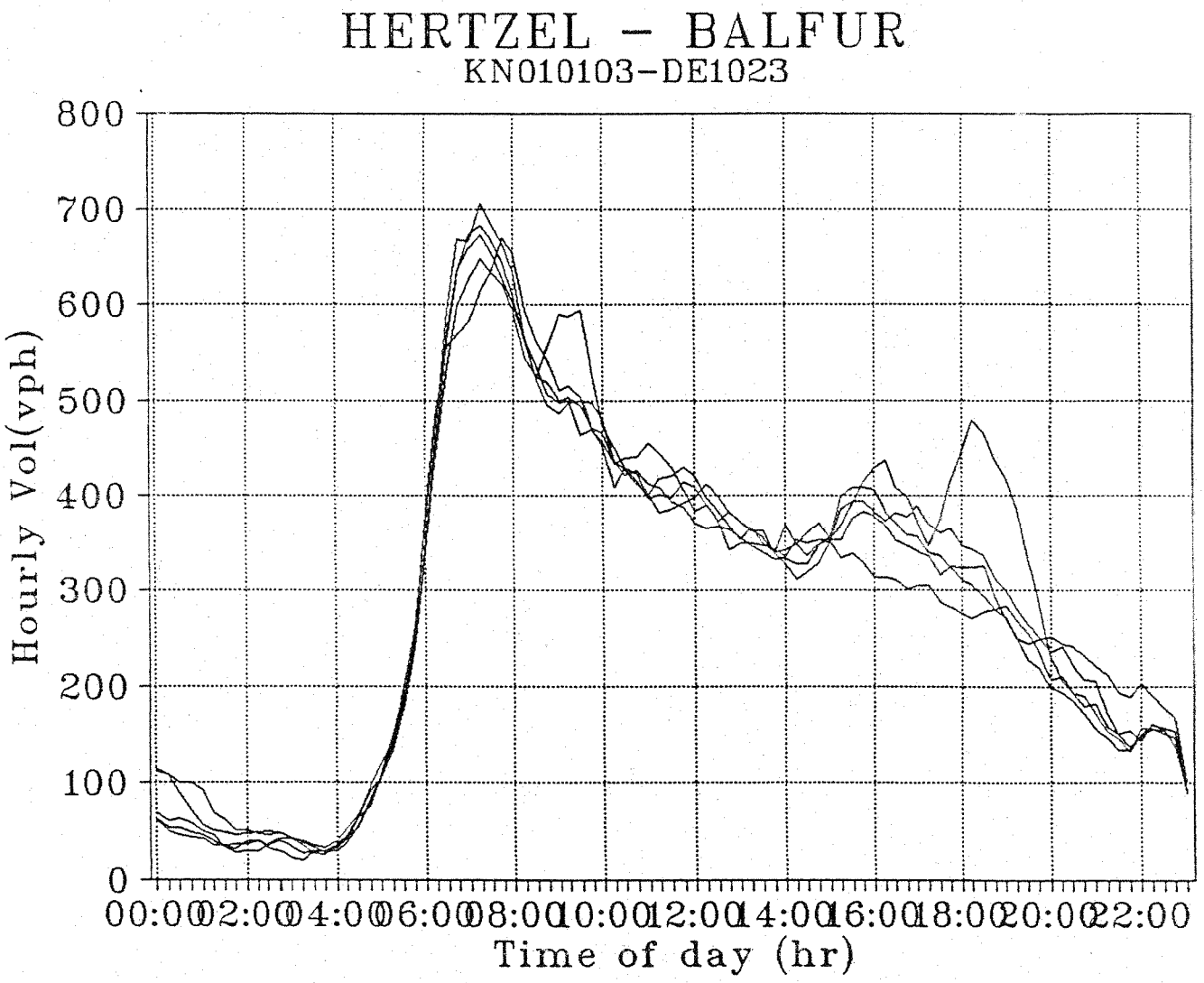
Each graph displays 1-day data (predictable variability)

⇒ Averaging days = smoothing

(or 6 detectors on a single day ?)

careful: scale





Predictably different!

Predictable Variability

השתנות בת-חזוי

איפה אתם בפקקים? עומסי התנועה הצפויים

Q-Science

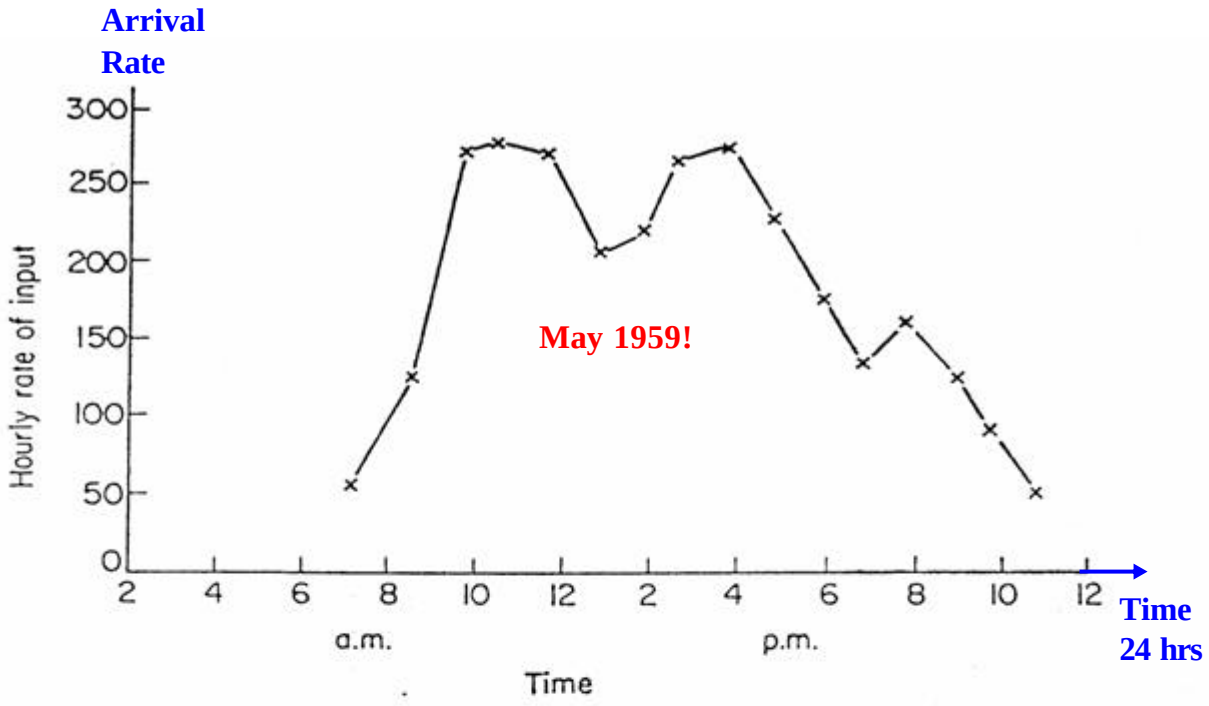
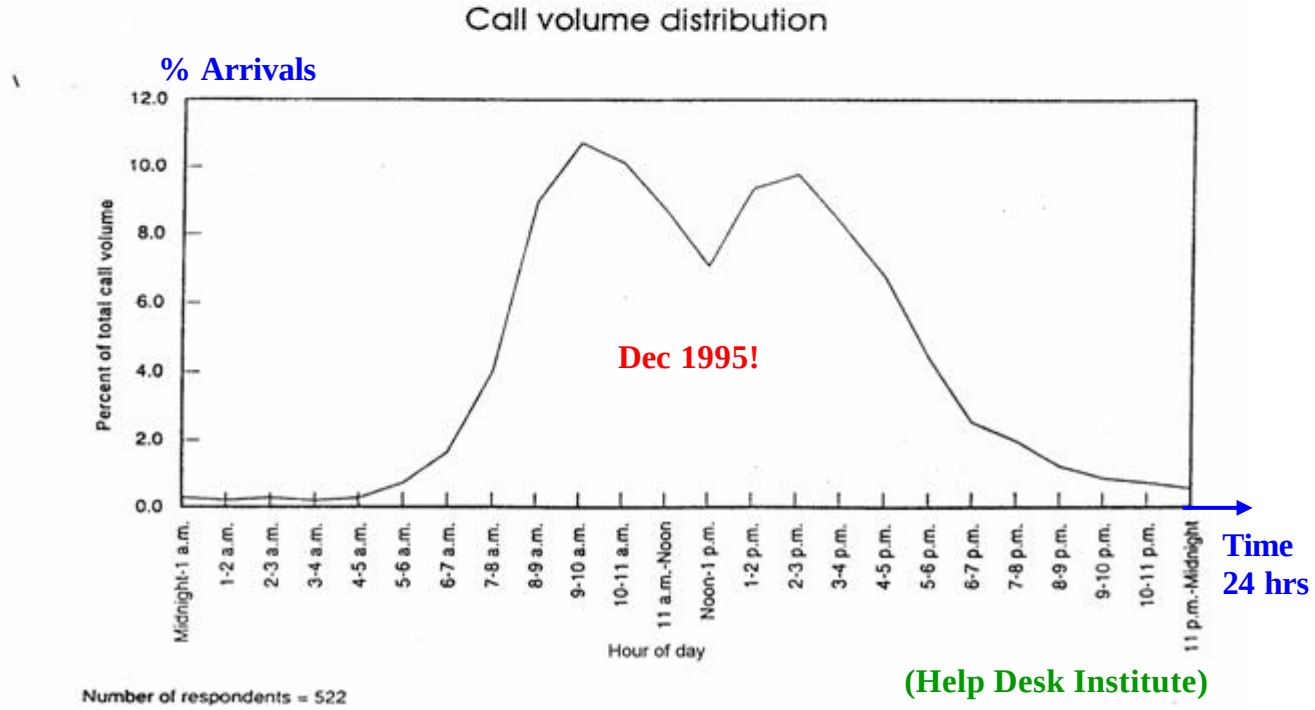


Fig. 15.1 The variation in the hourly input rates of reservations calls during a typical day (in May 1959)

(Lee A.M., Applied Q-Th)

1995 Help Desk and Customer Support Practices Report

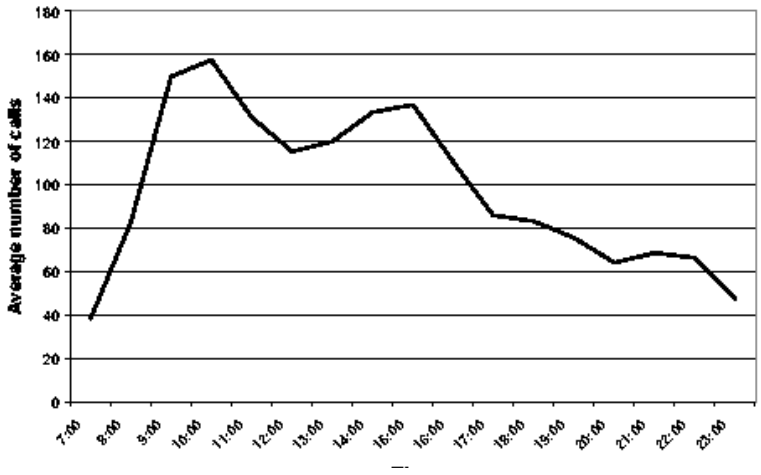


(Help Desk Institute)

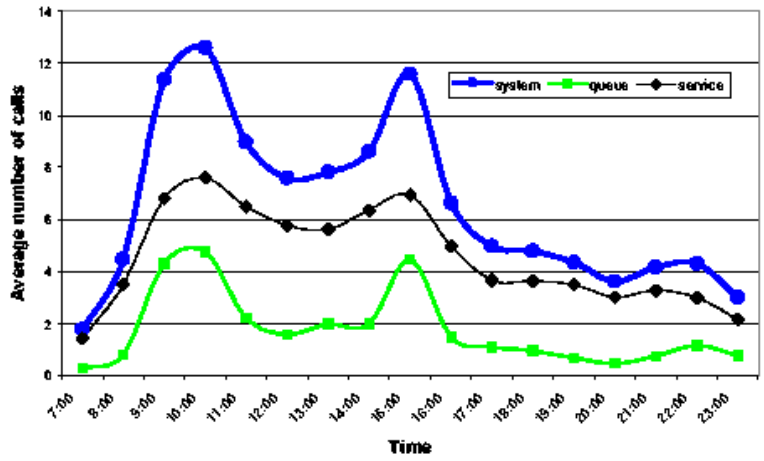
Time-Varying Queues: Predictable Variability

(with Jennings, Massey, Whitt)

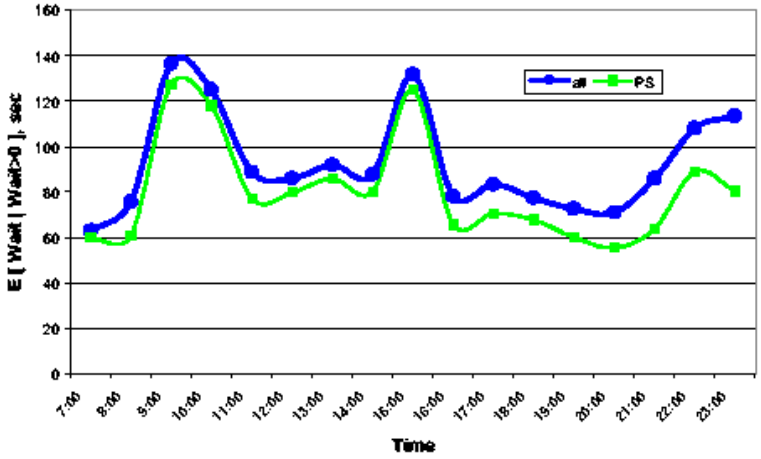
Arrivals



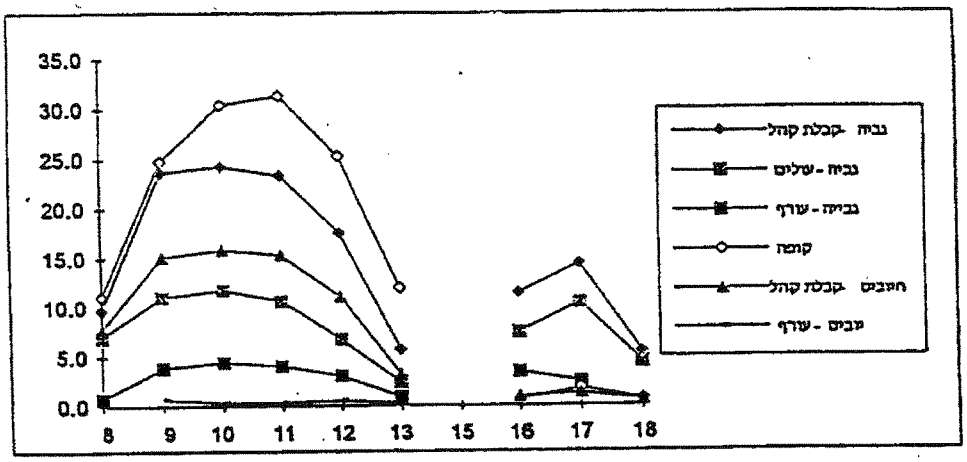
Queues



Waiting

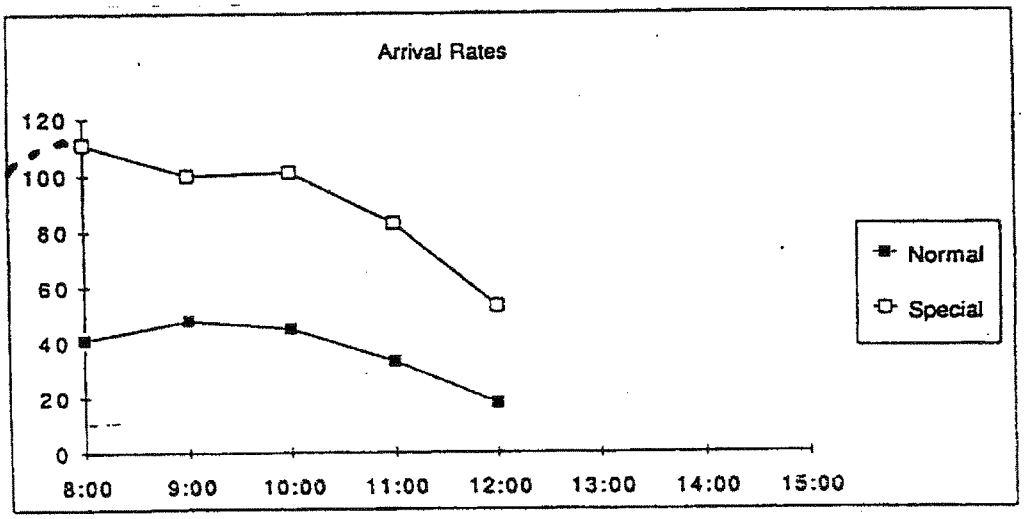


$\lambda(t)$



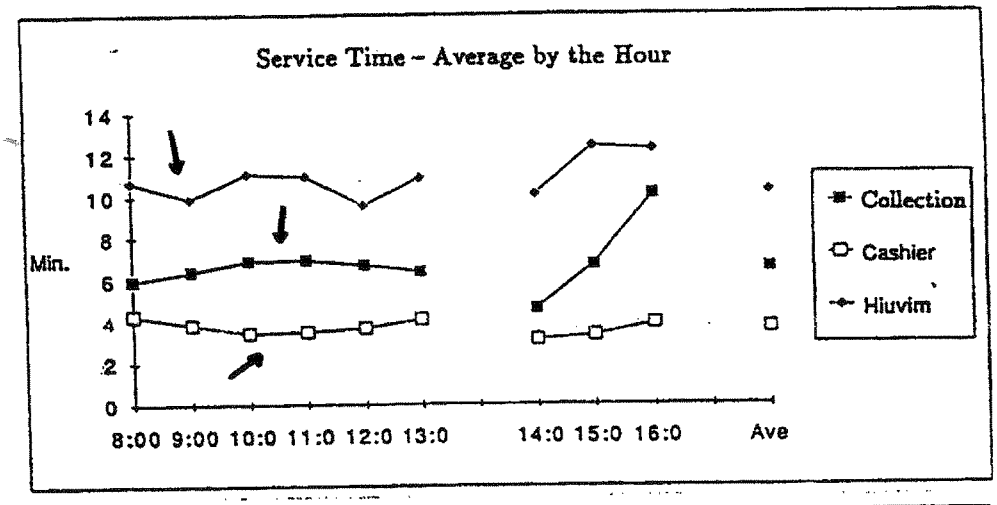
Arrival Rates

Routine



Special

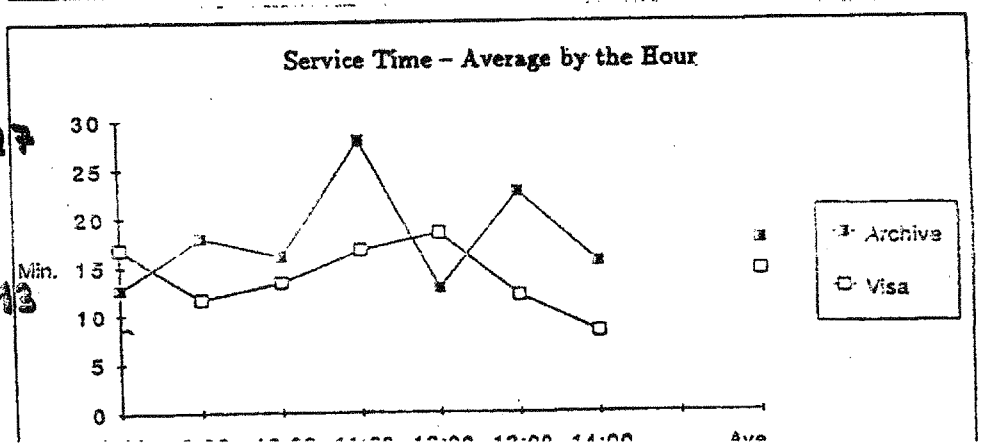
$\mu(t)^{-1}$



Service Times =

1 / service rate

3 patterns

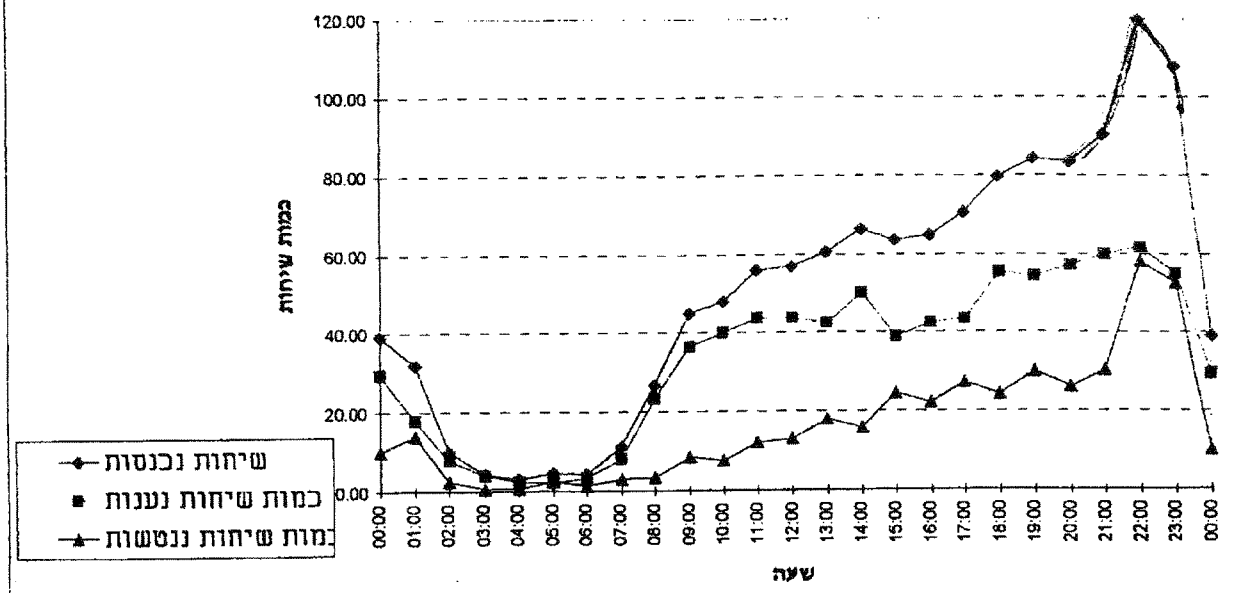


?

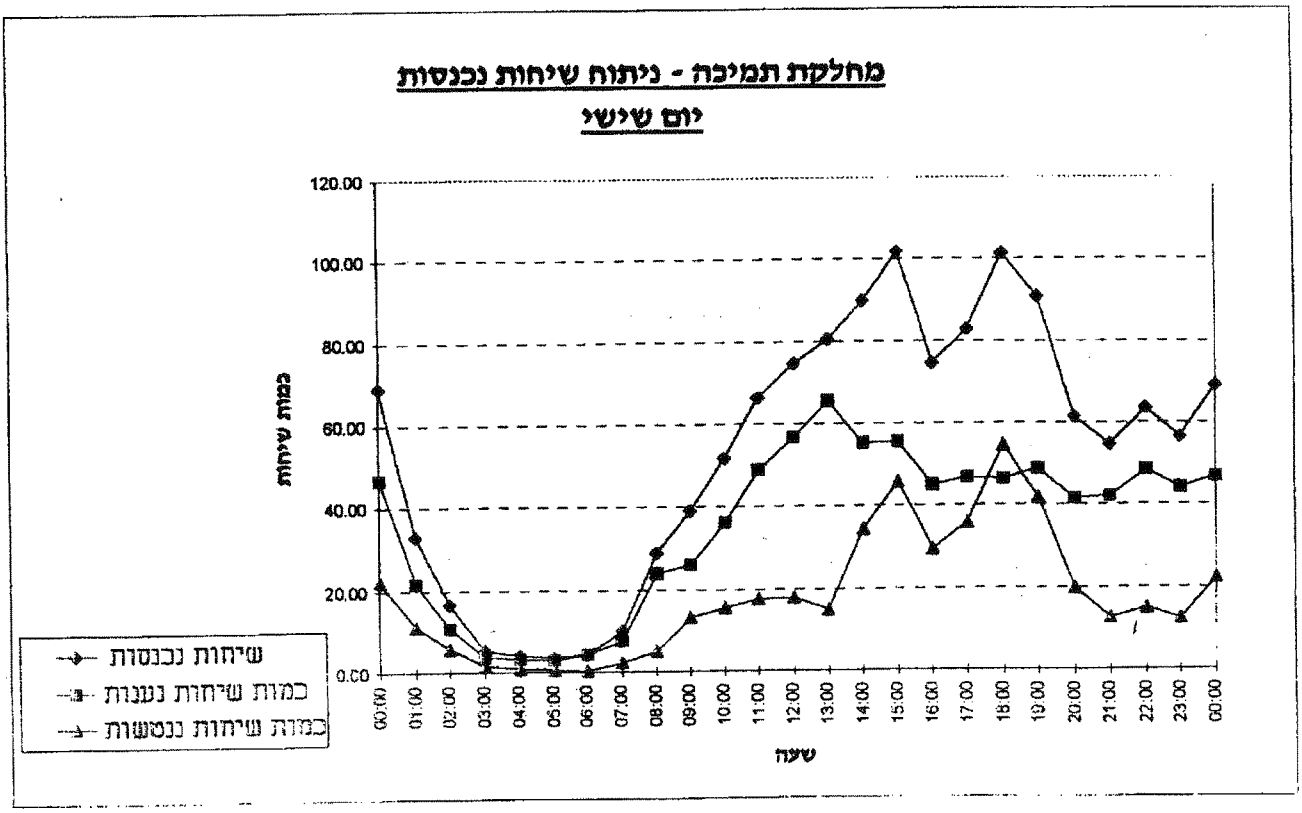
מחלקת תמיכה

מחלקת תמיכה - ניתוח שיחות נכנסות
ימי חול

Peak at 22:00



מחלקת תמיכה - ניתוח שיחות נכנסות
יום שישי

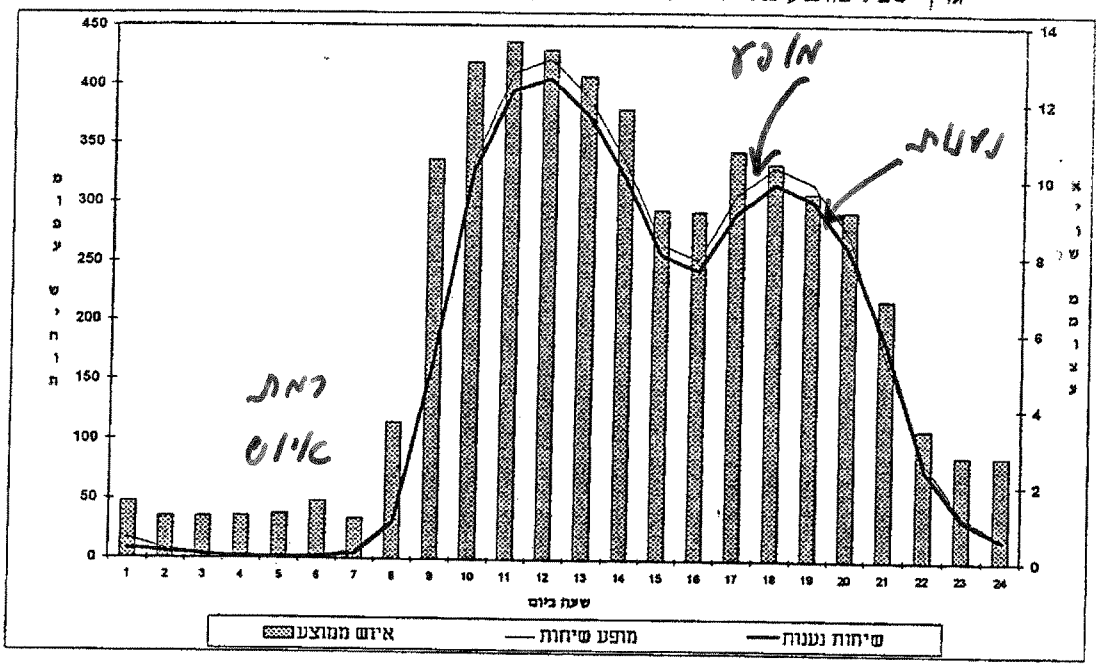


⇒ do not mix "apples & oranges" : clusters analysis (in Data Mining)

ה. דמה יומית

התמונה המוצגת בתת סעיף זה הינה בחתך שעתי במהלך יום עבודה.

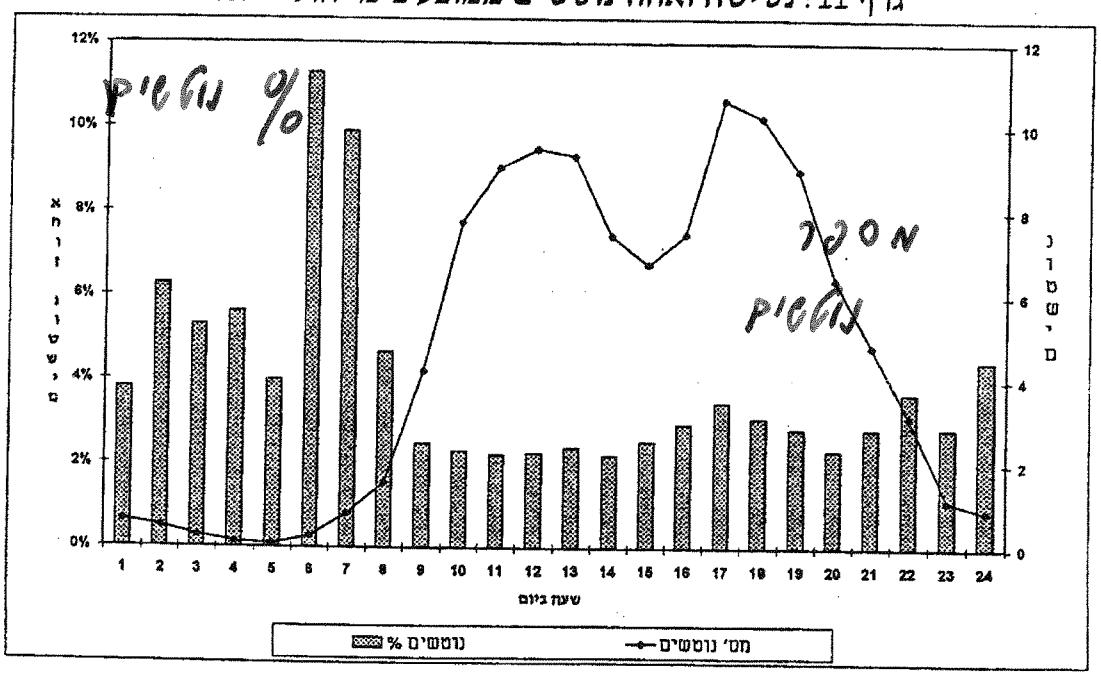
גרף 10 : **מופע פניות מול מספר עמדות פעילות בו זמנית - יומי**



הסבר: מספר השיחות הנענות ומופע הפניות השעתי הממוצע מוצגים בקו ומתייחסים לסקלה השמאלית. מספר העמדות הפעילות בו זמנית מיוצג על ידי העמודות המתייחסות לסקלה הימנית.

Systematic!

גרף 11 : **נטישה ואחוז נוטשים ממוצע בימי חול - יומי**



?

הסבר: העמודה בגרף מתייחסת לסקאלה השמאלית בגרף ומתארת את אחוז הנוטשים לפי שעה ביום. הקו מתייחס לסקאלה הימנית ומתאר את מספר הנוטשים.

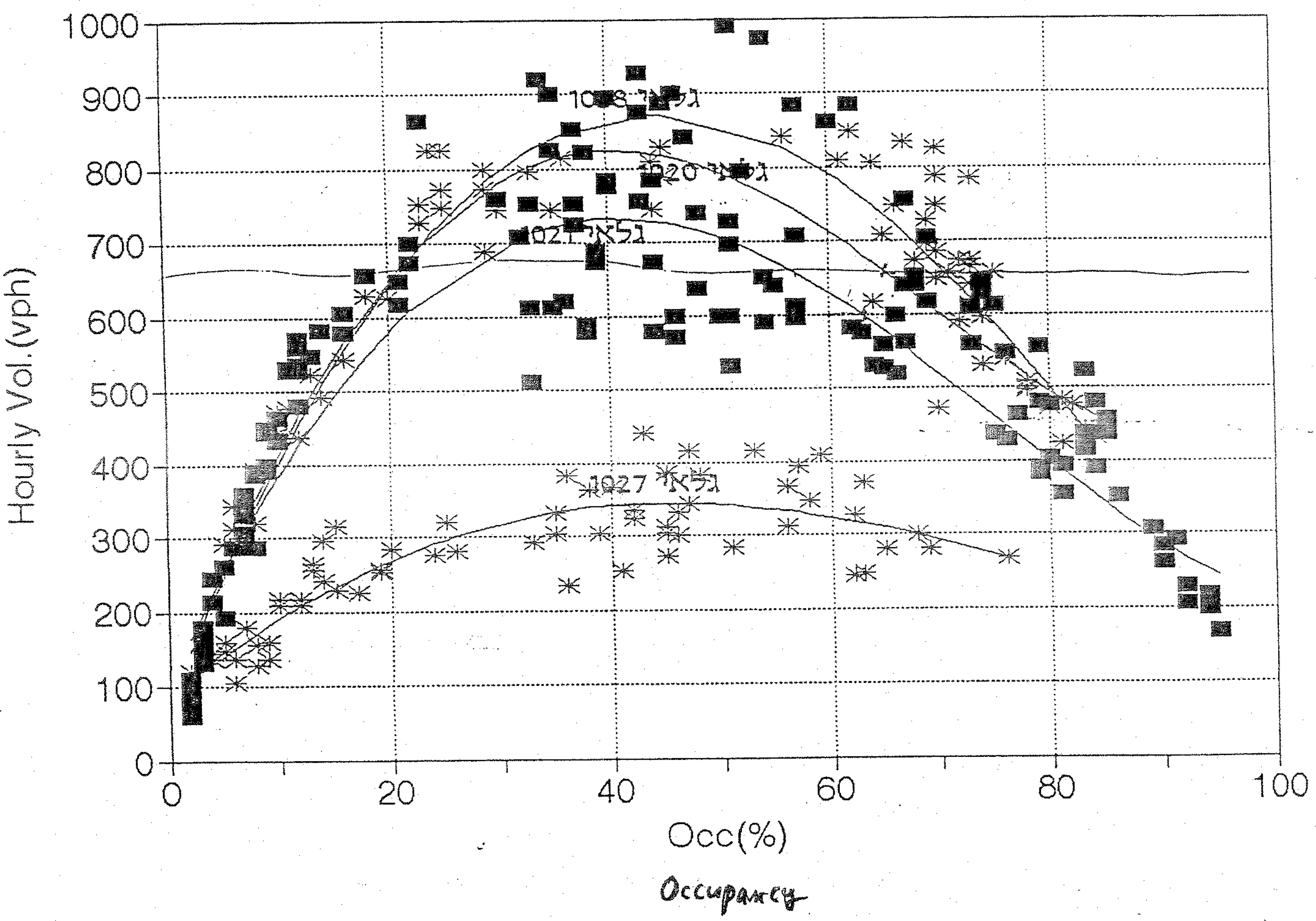
10

Puzzle: arises in transportation

(and communications, in a more subtle way)

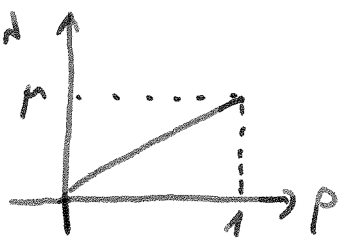
HERZEL - BALFUR

KN010103-4 1020-1-7-8 27-28/9/93



תרגיל: הסבר לתופעה הנחשפת ע"י הגרף „הרצל-בלפור"?

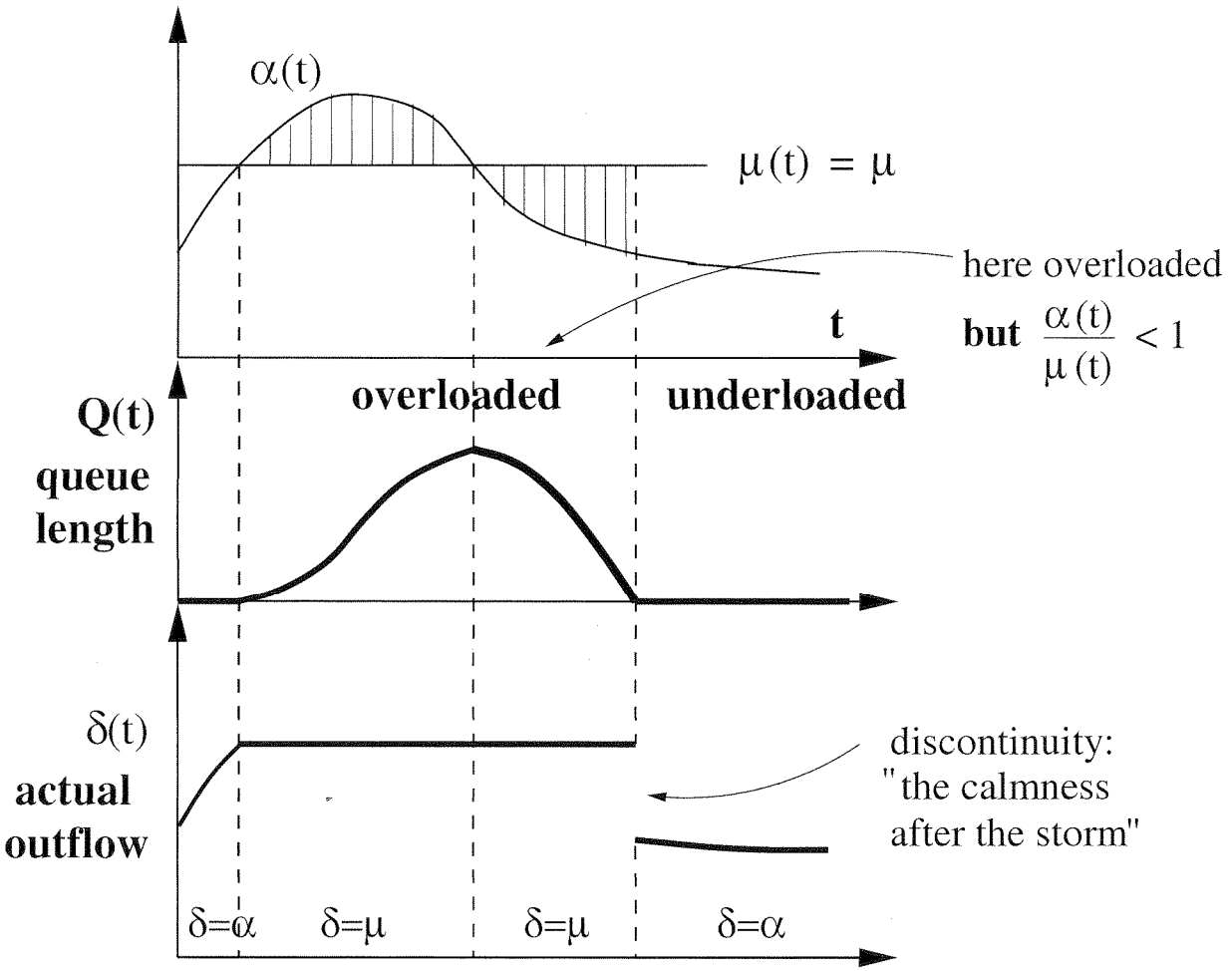
השוואה רגע: $\lambda = \mu \cdot \rho$



Phases of Congestion (Rush Hour Analysis)

Peak Load

Peak Congestion



Onset

End of Rush Hour

Face-to-Face

Services

Peak load

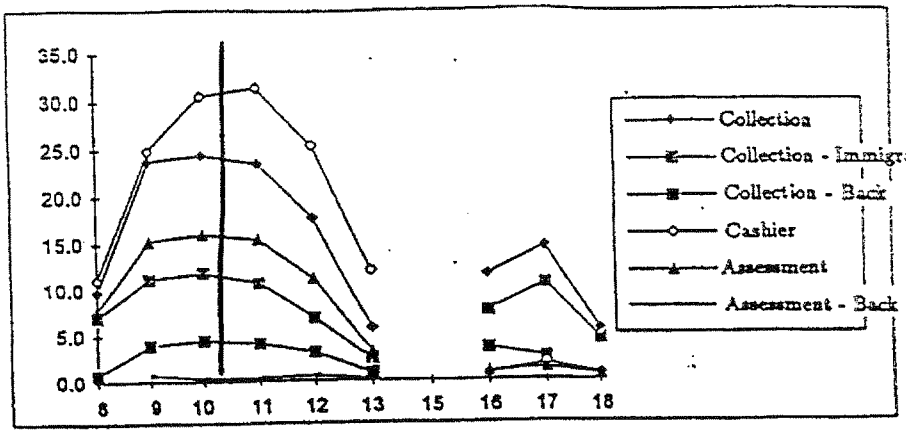
Phenomenon:

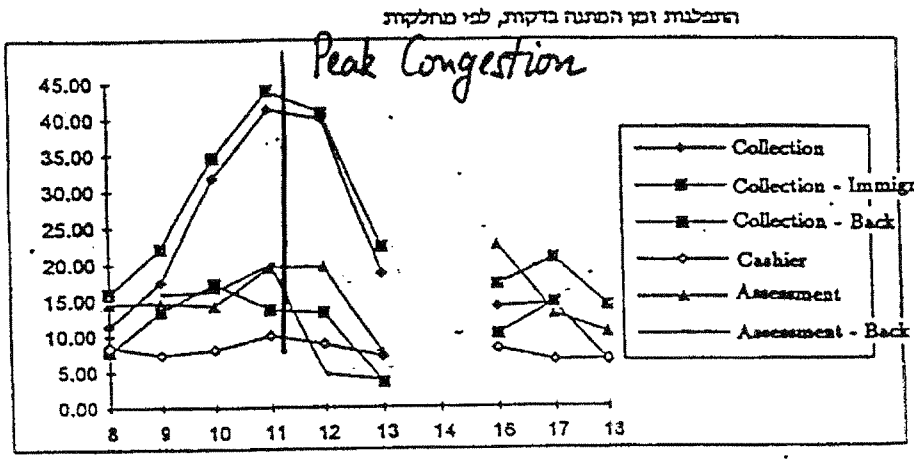
Peak congestion lags behind peak load

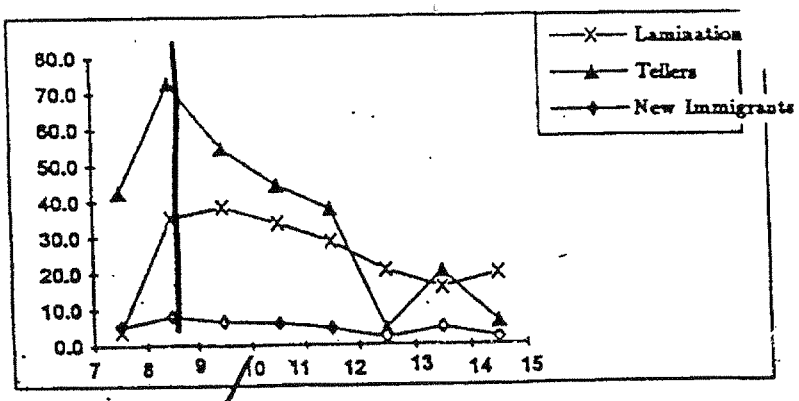
How to "explain"?

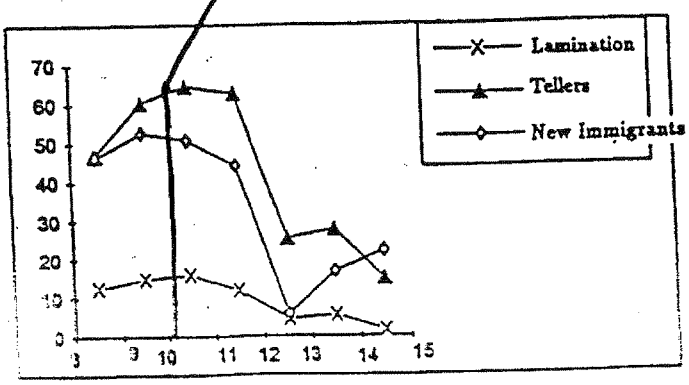
Fluid-view suffices

Peak Congestion Lags Behind Peak Load





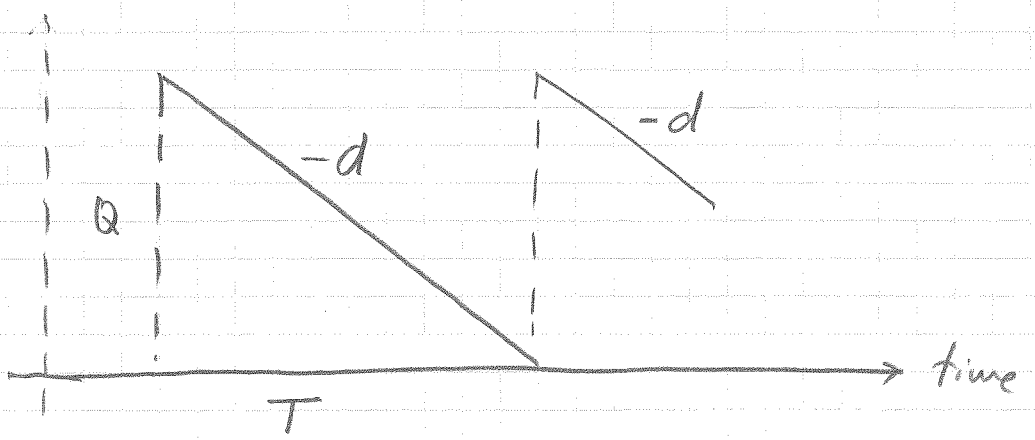




Simple (yet important, and classical) Application of

(Rate) Fluid Models: the EOQ Formula

- Tradeoff between inventory holding costs and ordering costs.



eg: $Q = 100$ units, $d = 25$ units per week

$\Rightarrow T = 100/25 = 4$ weeks : $T = Q/d$

Data: demand rate d (eg. stamps)

Dec. Var: order quantity Q (eg. go to post office)

Parameters: h = unit holding costs (h large $\Rightarrow Q \downarrow$)

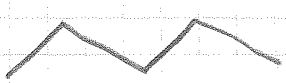
c = ordering costs (c large $\Rightarrow Q \uparrow$)

$$\text{average cost (over cycle)} = \underbrace{\frac{1}{2} Q \cdot h}_{L} + \frac{c}{T} = \frac{1}{2} Q h + \frac{cd}{Q}$$

Optimal Q^* where derivative $= 0$: $\frac{1}{2} h = \frac{cd}{Q^2}$ $\left(\Rightarrow \frac{1}{2} Q h = \frac{cd}{Q}\right)$

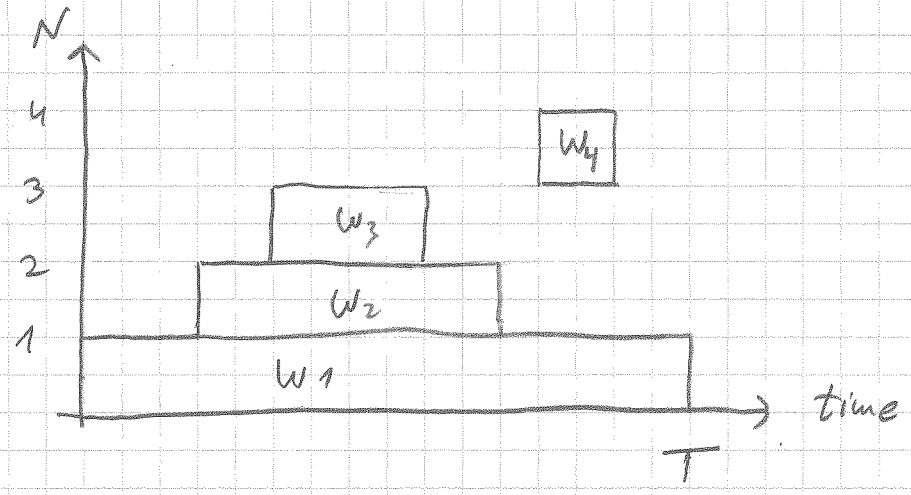
$$\boxed{Q^* = \sqrt{\frac{2cd}{h}}}$$ classical EOQ formula

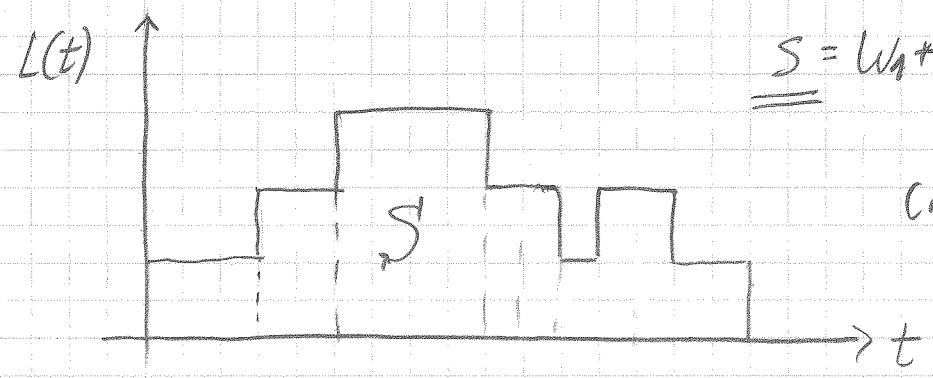
(d large $\Rightarrow$, c large $\Rightarrow$, h large $\Rightarrow$?)

Extension: finite production rate  , Q = batch size

Little: Review (Transition to Cumulatives)

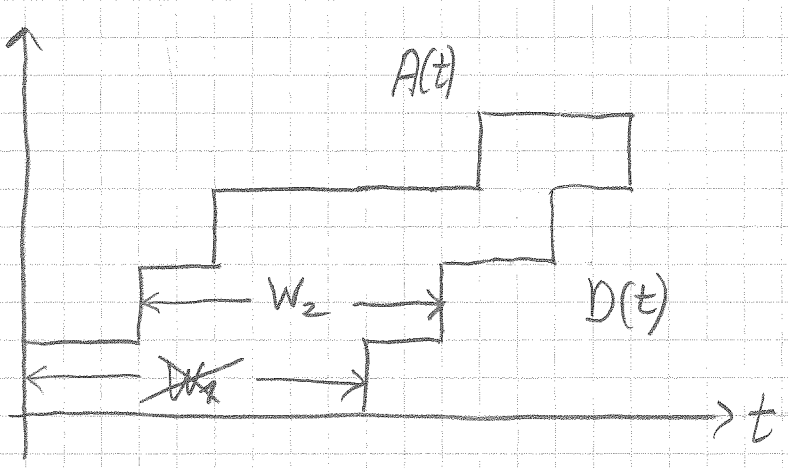
Gantt





$S = W_1 + \ldots + W_N$ total waiting (in hours × customers)

"Cumulative" picture



- 1st to arrive is last to leave
- 2nd to arrive is 2nd to leave

Divide waiting over time: $S/T = L$ (manager)

" " among customers: $S/N = W$ (customer)

Server: $N/T = \lambda$ (server)

$\Rightarrow$ $\boxed{L = \lambda W}$ Little's Law

Hall, Chapter 2 : Measurments, Empirical Models in Discrete Units

Definitions 2.2

$A(t)$ = *cumulative arrivals* from time 0 to time t

$D_s(t)$ = *cumulative departures from the system* from time 0 to time t

$D_q(t)$ = *cumulative departures from the queue* from time 0 to time t

• Arrivals

• Departures

The starting time, time 0, can be set at any time that is convenient to the analysis. For example, if a store opens at 9:30 A.M., time 0 would be 9:30 and $A(t)$ would be the number of customers who arrived between 9:30 and time t.

Consider the following data:

Customer	Arrival time	Departure from queue	Departure from system
1	9:36	9:36	9:40
2	9:37	9:40	9:44
3	9:38	9:44	9:48
4	9:40	9:48	9:52
5	9:45	9:52	9:56

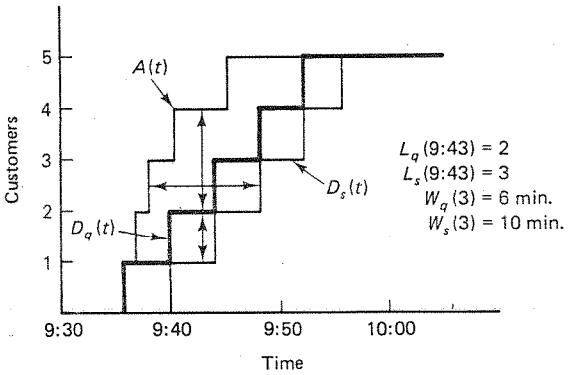


Figure 2.2 Cumulative arrival and departure diagram. Queue lengths are determined by vertical separation between curves. Waiting times are determined by horizontal separation.

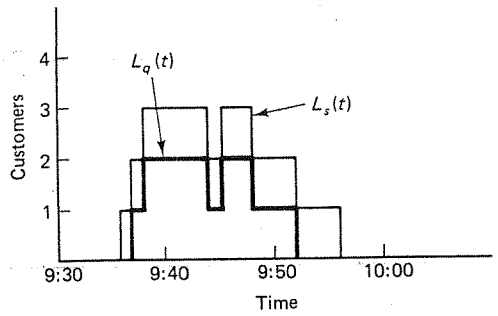


Figure 2.3 Customers in the system and in the queue versus time.

Definitions 2.3

$L_q(t)$ = number of customers in the queue at time t

$= A(t) - D_q(t)$

$L_s(t)$ = number of customers in the system at time t

$= A(t) - D_s(t)$

• Queues

Definition 2.8

L_q = average queue length (customers)

$$= \frac{\int_a^b L_q(t)dt}{b - a} \qquad (2.7)$$

Similarly L_s, σ_{L_q}, σ_{L_s}

2.2.1 Waiting Times

When Fig. 2.2 is read vertically, the queue size and number of customers in the system are identified. Reading Fig. 2.2 horizontally reveals the time in queue and the time in system.

Definitions 2.4

$$A^{-1}(n) = \text{time of the } n\text{th arrival}$$
$$D_q^{-1}(n) = \text{time of the } n\text{th departure from queue}$$
$$D_s^{-1}(n) = \text{time of the } n\text{th departure from system}$$

Whereas $A(t)$, $D_q(t)$, and $D_s(t)$ convert a time into a customer number, $A^{-1}(n)$, $D_q^{-1}(n)$ and $D_s^{-1}(n)$ take a customer number and convert it to a time. They correspond exactly to the data provided before:

n	$A^{-1}(n)$	$D_q^{-1}(n)$	$D_s^{-1}(n)$
1	9:36	9:36	9:40
2	9:37	9:40	9:44
3	9:38	9:44	9:48
4	9:40	9:48	9:52
5	9:45	9:52	9:56

Definitions 2.5

$$W_q(n) = \text{time in queue, for } n\text{th customer to arrive}$$
$$W_s(n) = \text{time in system, for } n\text{th customer to arrive}$$

When the discipline is FCFS, the waiting times, $W_q(n)$ and $W_s(n)$, are found by computing the horizontal distance between the steps in Fig. 2.2:

FCFS Waiting Time

$$W_q(n) = D_q^{-1}(n) - A^{-1}(n) \tag{2.1}$$
$$W_s(n) = D_s^{-1}(n) - A^{-1}(n) \tag{2.2}$$

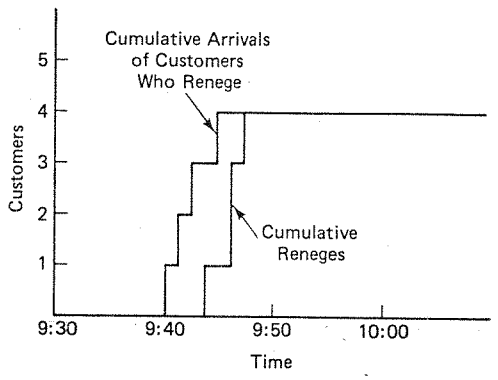


Figure 2.4 Cumulative diagram of reneges.

Definitions 2.6

D

$$W_q = \text{average waiting time in queue} = \frac{\sum_{n=1}^{N} W_q(n)}{N} \tag{2.3}$$

$$W_s = \text{average waiting time in system} = \frac{\sum_{n=1}^{N} W_s(n)}{N} \tag{2.4}$$

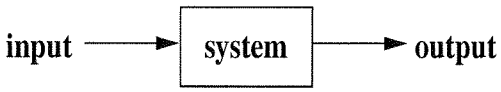
In Class # 2

Service Engineering

Created: April 1997
Last revision: October 2003

LITTLE'S LAW

A conservation law that applies to the following general setting:



Input: Continuous flow or discrete units (examples: granules of powder measured in tons, tons of paper, number of customers, \$1000's).

System: Boundary is all that is required (very general, abstract).

Output: Same as input, call it *throughput*.

Two possible scenarios:

- System during a "cycle" (empty → empty, finite horizon);
- System in steady state/in the long run (for example, over many cycles).

Quantities that are related via Little's law:

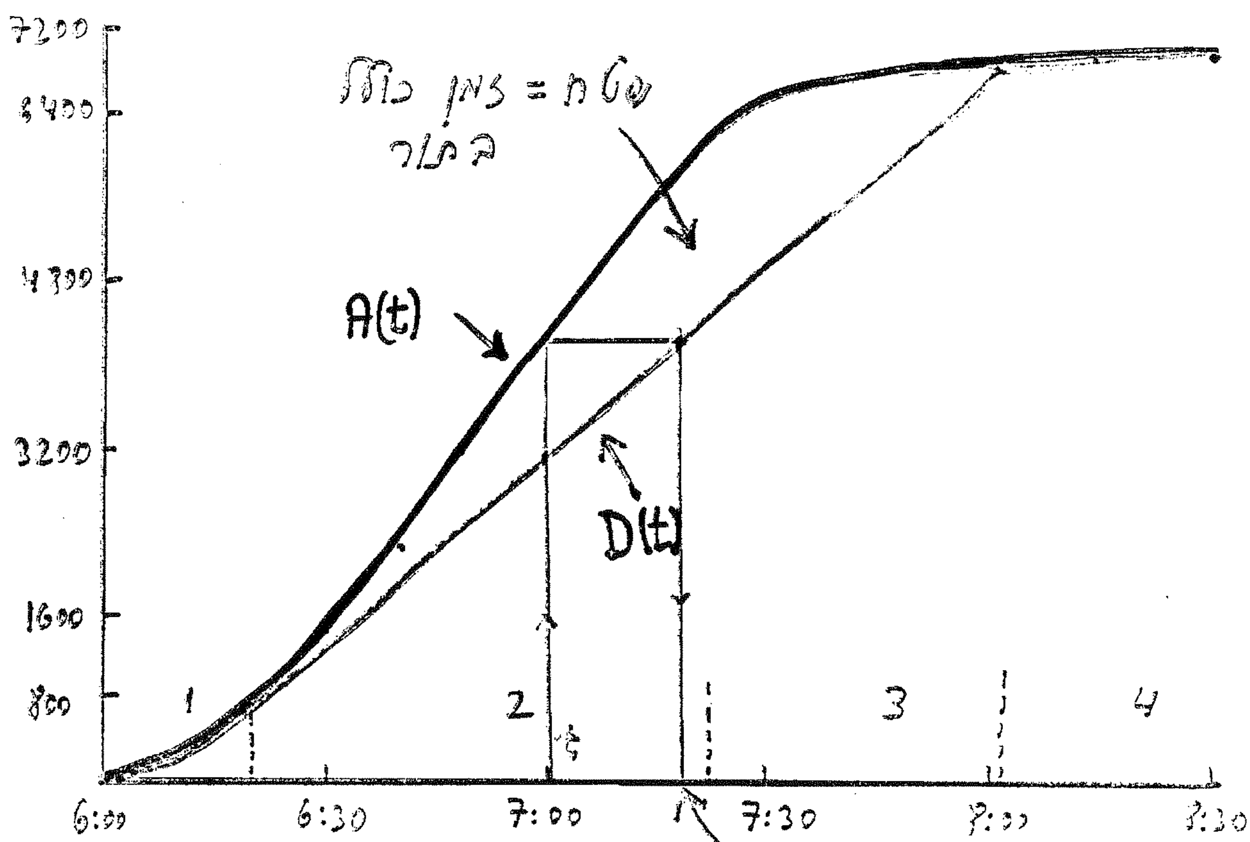
- λ = long-run average rate at which units *arrive*
  (= long-run average rate at which units *depart*) = *throughput-rate*, whose units are quantity/time-unit or #/time-unit;
- L = long-run average *inventory*/quantity/number in the system
  (eg. WIP: Work-In-Process, customers);
- W = long-run average time a unit spends in the system = *throughput time*
  (eg. hours) = sojourn time.

Little's Law $\boxed{L = \lambda W}$

Motivation 1: λ customers/hour, each charged \$1/hour while remaining in the system. Then $\lambda \times W$ is the rate at which the system generates cash which, in turn, "clearly" equals L.

18

עקומה מצטברת



הזמן בו עזבו בדיוק
כל המופיעים עד 7:00

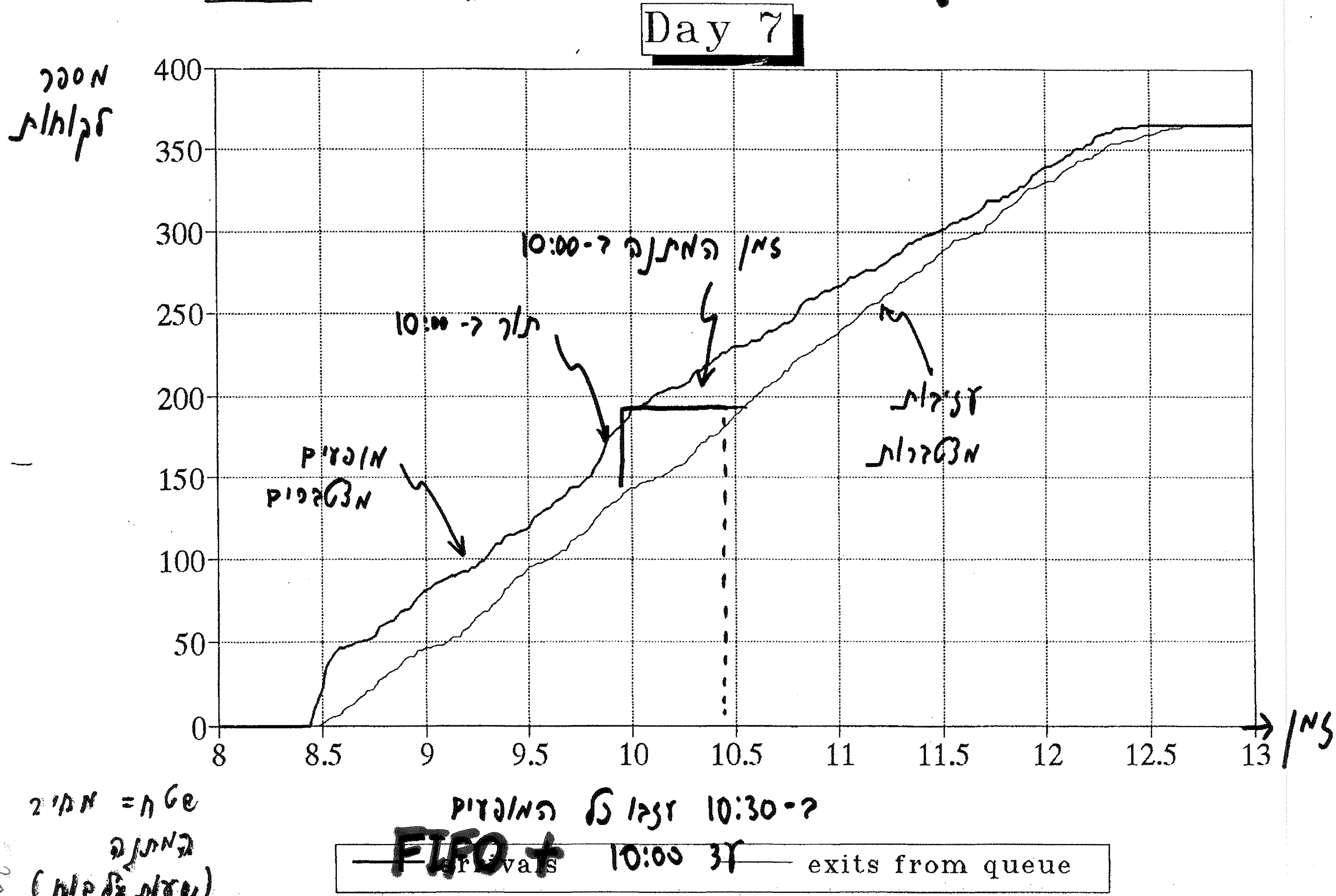
אורך תור = מרחק אנכי בין A(t) ל- D(t) : A(t)-D(t)

זמן המתנה (תחת FCFS) = מרחק אופקי

עקב FCFS, בדיוק זמן ההמתנה של
מופע ב- 7:00

גרף „הצטברות" אמפירי

? Discrete Units



Hall

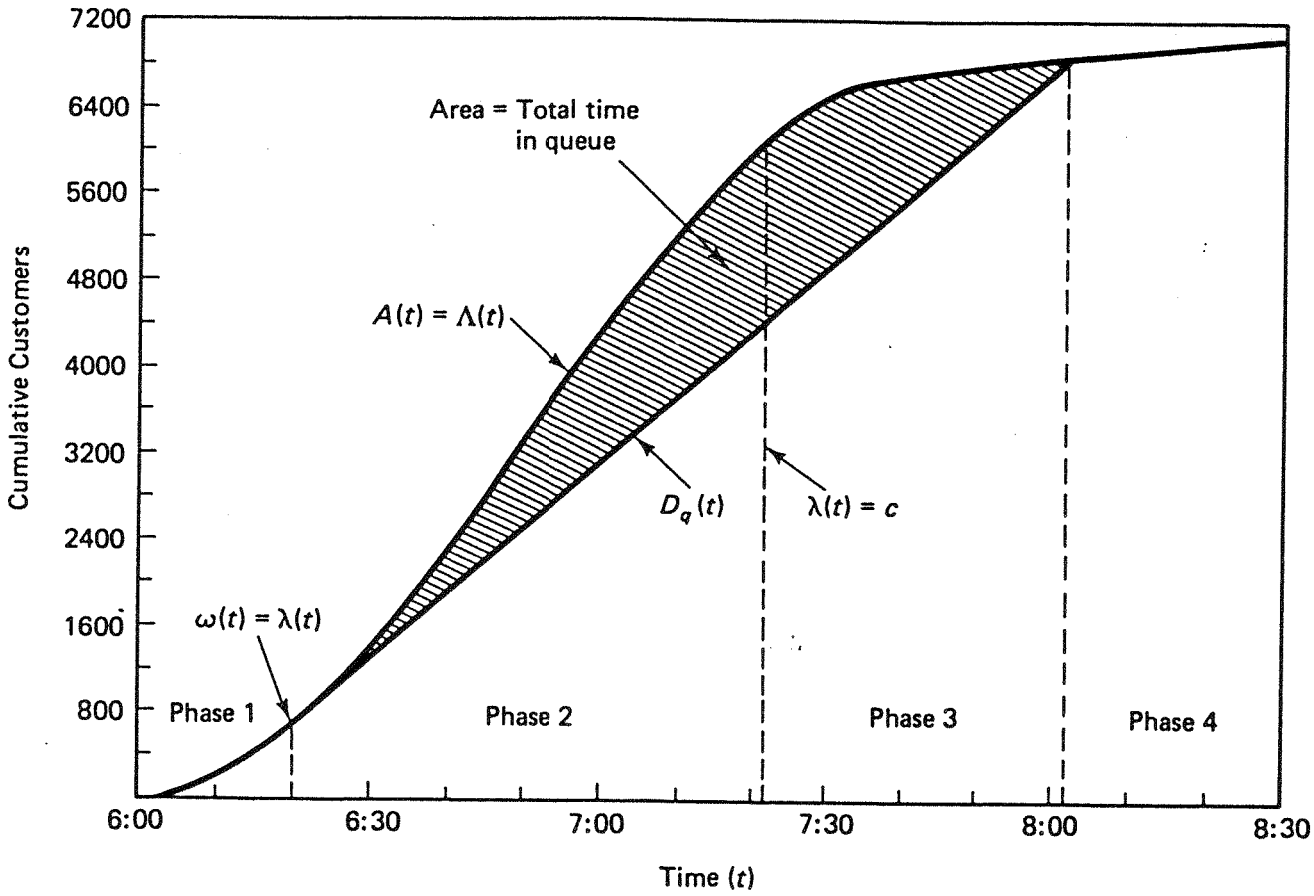


Figure 6.6 Cumulative diagram illustrating deterministic fluid model. When a queue exists, customers depart at a constant rate. Queues increase when the arrival rate exceeds the service capacity and decrease when the service capacity exceeds the arrival rate.

Hall, pg. 189-90:
1. stagnant
2. growth
3. decline
4. stagnant, etc.

Phases of Congestion: via Cumulants

21.

In Recitation

Service Engineering 2003

Example: Empirical Models
Analysis of a Face-to-Face Service Operation

Data from 12 days of work (two weeks) in a Face-to-Face service of a bank was collected. Several servers work simultaneously at a single station, the data for which is described below. The maximum number of servers is five.

The data from day 6 and day 12 is not considered here. (These days are Fridays and are different from the others.)

Figure 1 (see page 3) presents average waiting times on the considered days.

Using this figure the working days were divided into three categories.

- Catastrophic day: day 7.
- Heavily loaded days: days 8, 9 and 10.
- Regular days: days 1, 2, 3, 4, 5 and 11.

We now analyze the data according to the following categories: queues, arrivals, waiting times and staffing levels.

Queues: we see the *average* queue length for every category in Figure 2. Below we describe the queue pattern for every category.

Catastrophic day. The queue increases sharply when the working day starts (40 customers in the queue shortly after 8:30). At 9:30 the queue goes down to 25 customers and then grows rapidly again. Approximately at 10:10 we get the record queue for all days: more then 50 customers. Then the queue gradually decreases to zero at the end of the working day.

Heavily loaded days. The average queue sharply grows to 10 customers at the beginning of the day and then oscillates between 5 and 10 customers until 9:45. Then a growth to the level of 13-18 customers happens. After 11:00 the queue slowly decreases to zero.

Figure 3 shows sample path of queues on heavily loaded days.

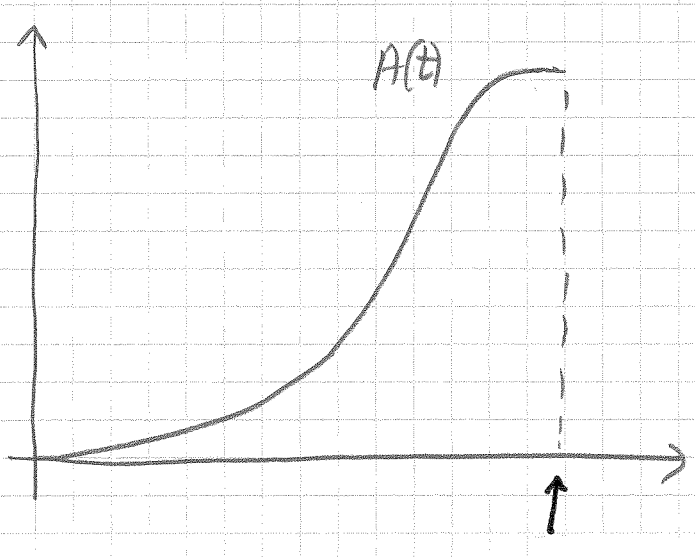
Regular days. The average queue jumps almost to 10 at the beginning, decreases close to zero before 9:30 and then over almost the whole working day, it oscillates in "steady state" near 5.

See Figure 4 for examples of sample paths on regular days.

We observe the following common features for the different categories.

- Sharp growth of the queues at the start of a working day.
- Queue decrease before 9:30.
- Queue growth before 10:00.
- Gradual decrease to zero at the end of a working day.

Aggregate Planning : via "Cumulative Pictures"



A(·) : seasonal

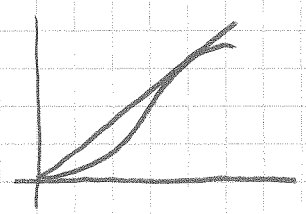
eg. airconditioners, fashion, arrivals to airport

Q : service rate ? (i.e. capacity)

- strategy: chase demand D = A costly variable workforce

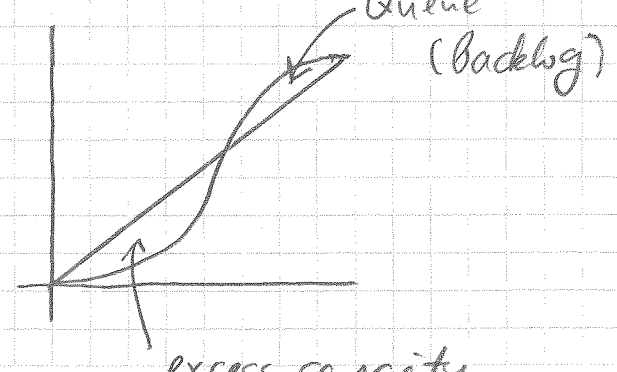
Suppose constant workforce

- strategy : no queues



=> excess capacity

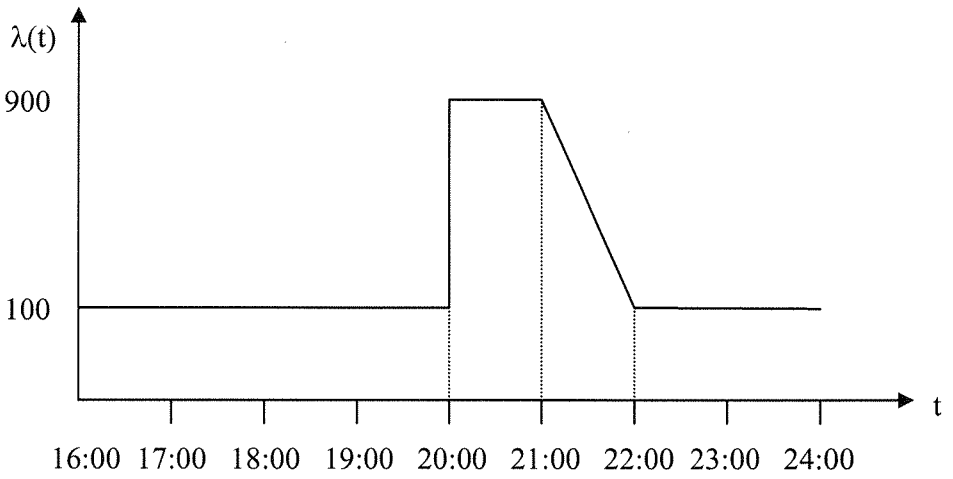
- strategy: least constant capacity that accomodates all arrivals, and leaves no queue at end.



"Pictures": Summary 1.

Homework 5 – Staffing Through Fluid Models

This question is based on the question that was presented in the recitation:
The arrival rate of customer calls to the call center is given by the following graph:



① Rates (Arrivals, Capacity)

Assume that an hour work of a service representative costs 37.5 shekels, while a minute waiting of a customer costs 1 shekel. Also, let us assume that the staffing must remain fixed during a shift.

② Cumulative Rates (Arrivals, Departures, Services)

Based on the above, answer the following questions:

a) Draw the cumulative arrivals graph.
b) Using the cumulative arrivals graph solve (using Excel's solver) the optimization problem of minimizing waiting and staffing costs. (**Hint**: You should use only the cumulative arrivals graph – there is not need to use differential-equation representations used in class).
c) Using the optimality criterion taught in class (as appears in Hall, pages 215-218) determine the optimal number of service representatives. Compare your result with your answer to (b). Compare the cost of your recommendation with the cost that was obtained in the recitation by using a different approach.
d) Based on your answer to (c), draw the queue length as a function of time.
e) Note that the above question is a special case of the one analyzed in the recitation. There we allowed to vary the staffing level <u>every hour</u>. Can the above "Cummulative Approach" be adopted to allow varying staffing levels ? (If so, describe briefly how it can be done - <u>there is no need to do any calculations</u>).

③ Δ of Cumulatives (Queues, Waiting)

24

From Data to Models: (Predictable vs. Stochastic Queues)

Fix a day of given category (say Monday = M, as distinguished from Sat.)
Consider data of many M's.
What do we see ?

- Unusual M's, that are outliers.
  Examples: Transportation : storms,...
  Hospital: : military operation, season,...)

  Such M's are accommodated by emergency procedures:
  redirect drivers, outlaw driving; recruit help.
  $\Rightarrow$ Support via scenario analysis, but carefully.

- Usual M's, that are "average".
  In such M's, queues can be classified into:
  - Predictable:
    queues form systematically at nearly the same time of most M's + avg. queue similar over days + wiggles around avg. are small relative to queue size.

    e.g., rush-hour (overloaded / oversaturated)

    Model: hypothetical avg. arrival process served by an avg. server
    Fluid approx / Deterministic queue :macroscopic
    Diffusion approx = refinements :mesoscopic

  - Unpredictable:
    queues of moderate size, from possibly at all times, due to (unpredictable) mismatch between demand/supply
    $\Rightarrow$ Stochastic models :microscopic

  Newell says, and I agree:
  Most Queueing theory devoted to unpredictable queues,
  but most (significant) queues can be classified as predictable.

A Service Center in RUSH HOUR

amazon.com

We're sorry!

Predictable ? yes / no ?
for whom ?

Our store is closed temporarily for scheduled maintenance. If you enter your e-mail address, we'll notify you as soon as we reopen. Again, our apologies for the inconvenience.

Thanks for your patience,

Your friends at Amazon.com

Please enter your e-mail address: [ ] Submit

Scales (Fig. 2.1 in Newell's book: Transportation)

	Horizon	Max. count/queue	Phenom
(a)	5 min	100 cars/5–10	(stochastic) instantaneous queues
(b)	1 hr	1000 cars/200	rush-hour queues
(c)	1 day = 24 hr	10,000 / ?	identify rush hours
(d)	1 week	60,000 / –	daily variation (add histogram)
(e)	1 year		seasonal variation
(f)	1 decade		↑ trend

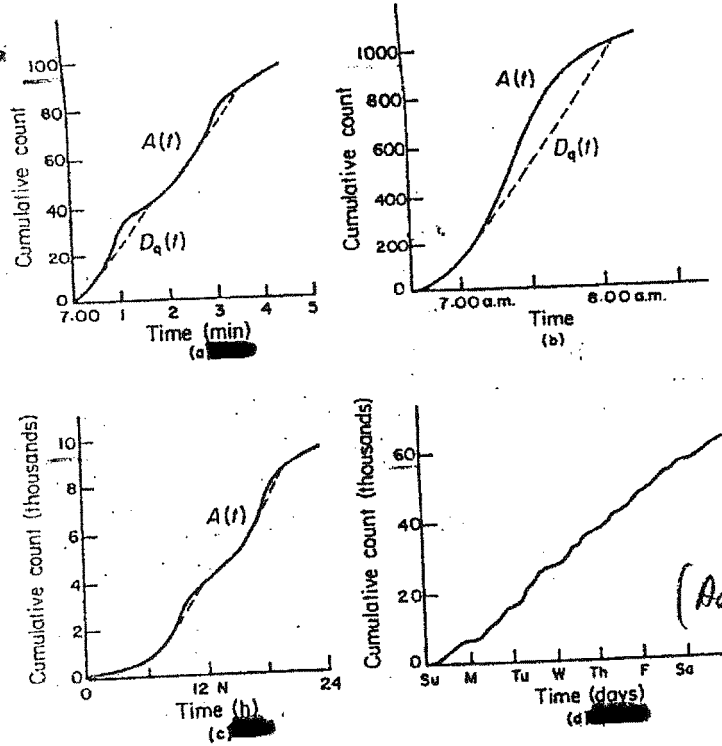
Scales in Tele-service

Horizon	Decision	e.g.
year	strategic	add centers / permanent workforce
month	tactical	temporary workforce
day	operational	staffing (Q-theory)
hour	regulatory	shop-floor decisions

Newell, G. F.
Applications of Queueing Th.
2nd Ed., 1982

Transportation
Scales, Phenomenon

max. count.



Instantaneous queues (Stochastic queues)

rush-hour queues

identify rush-hours (predictable queues)

daily variations

(Add histograms)

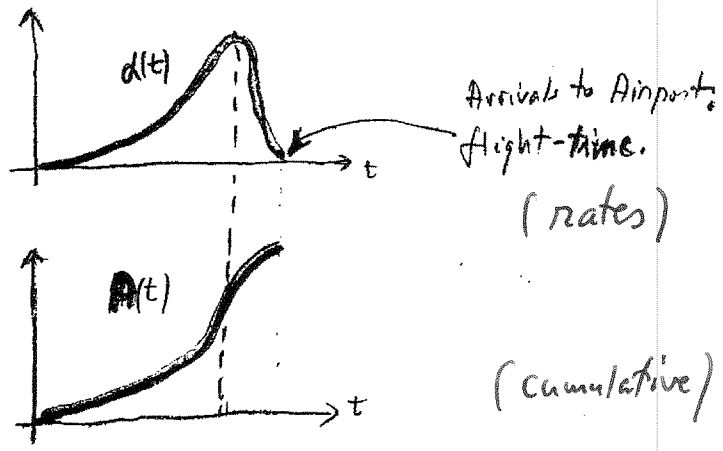
Figure 2.1 *Cumulative arrivals on various time scales*

Can add: 1 year
1 decade
to detect trends, seasonal variations, etc.

28

- **Cumulative** data (vs. rates)

$$A(t) = \int_0^t \alpha(u)\,du$$

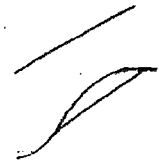


(rates) (1)

(cumulative) (2)

- Newell says:
Most Q-theory is (a),
but Most Q-applications is (b).

Test
- Averaging out many (a)'s ⇒
- " " " (b)'s ⇒

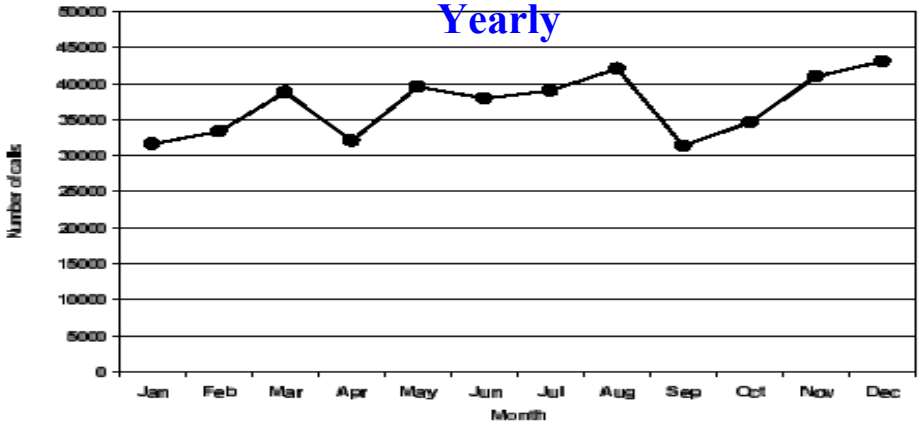


Better look at Queues

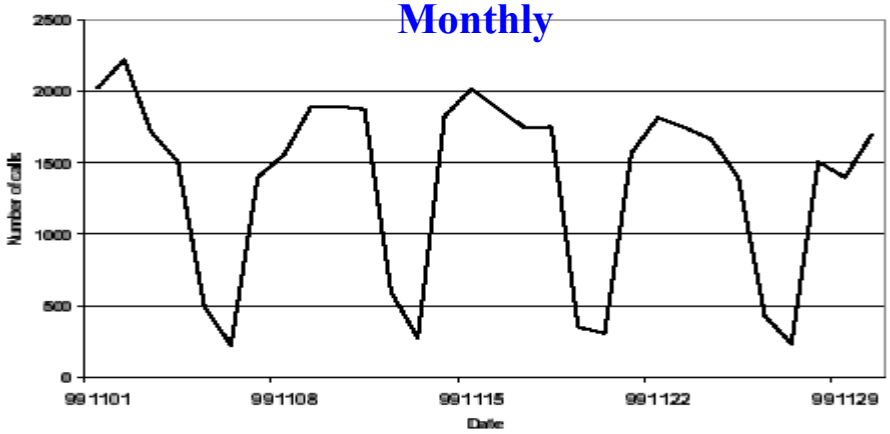
(congestion: queues, waiting) (3)

Scales: Arrival Process, 1999

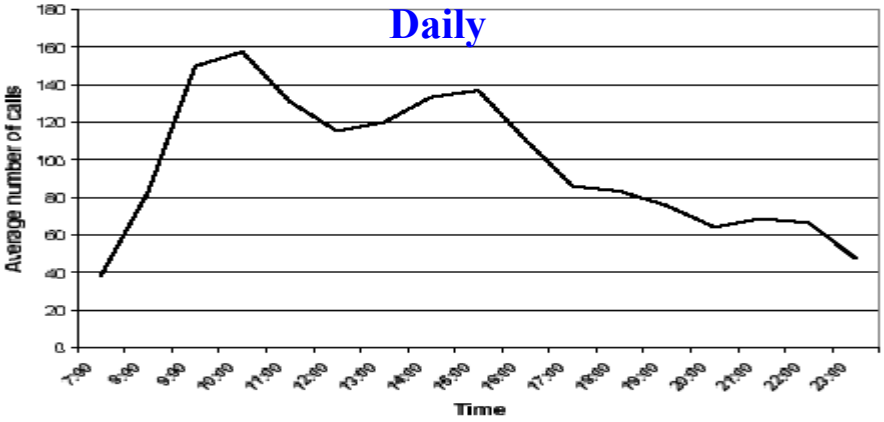
Yearly



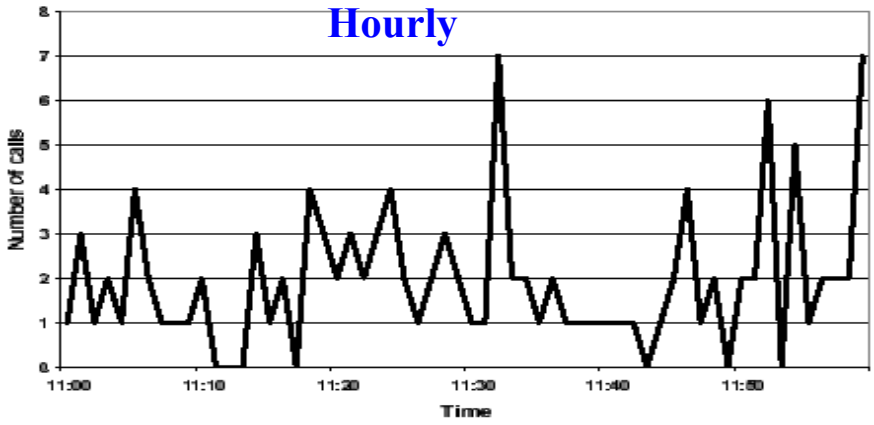
Monthly



Daily



Hourly



Arrival Process, in 1976

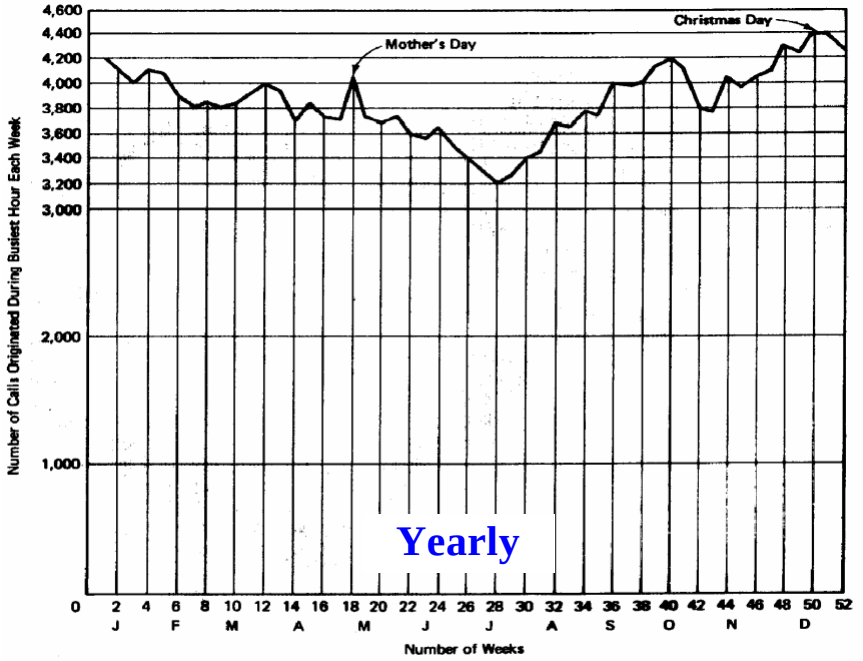


Figure 1 Typical distribution of calls during the busiest hour for each week during a year.

(E. S. Buffa, M. J. Cosgrove, and B. J. Luce, "An Integrated Work Shift Scheduling System")

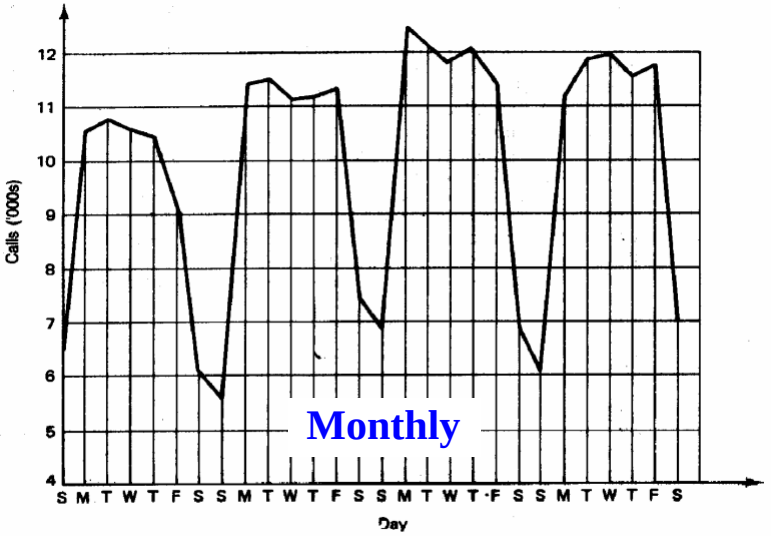


Figure 2 Daily call load for Long Beach, January 1972.

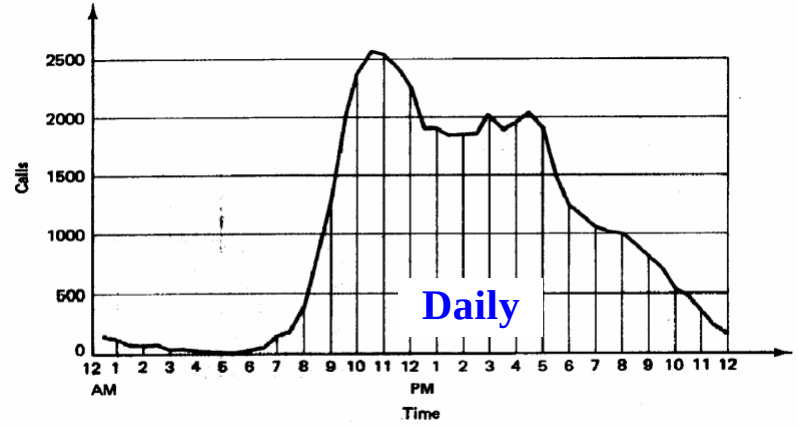


Figure 3 Typical half-hourly call distribution (Bundy D A).

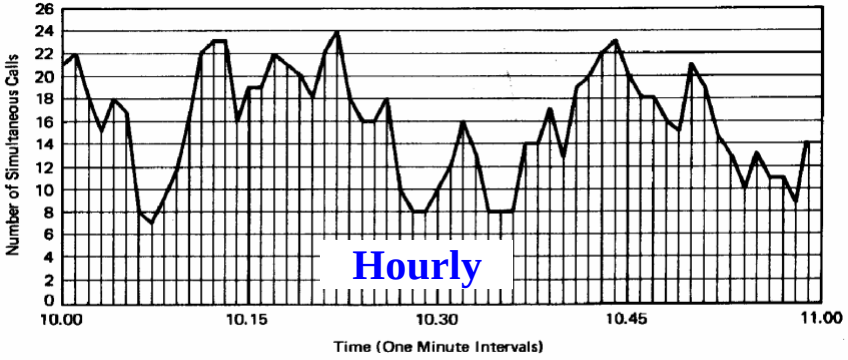
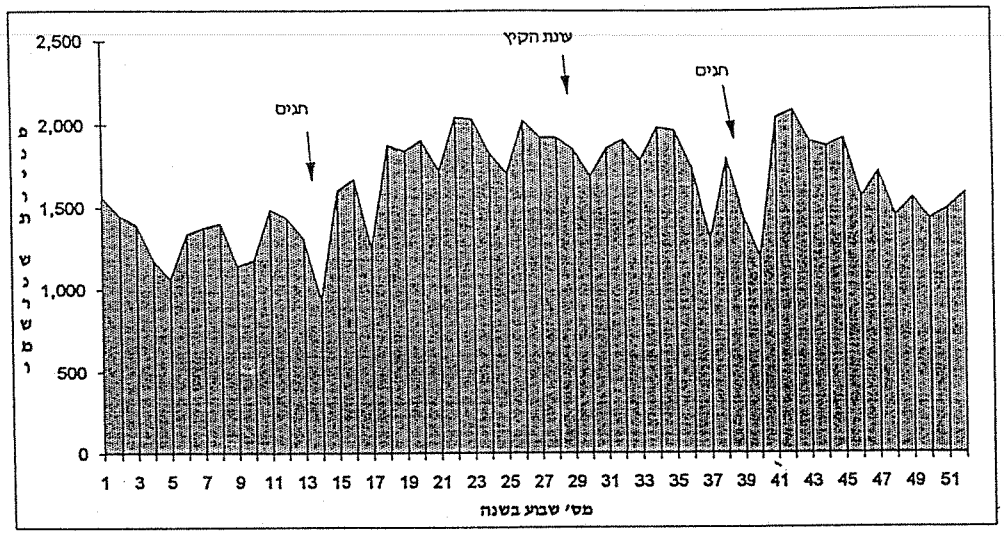


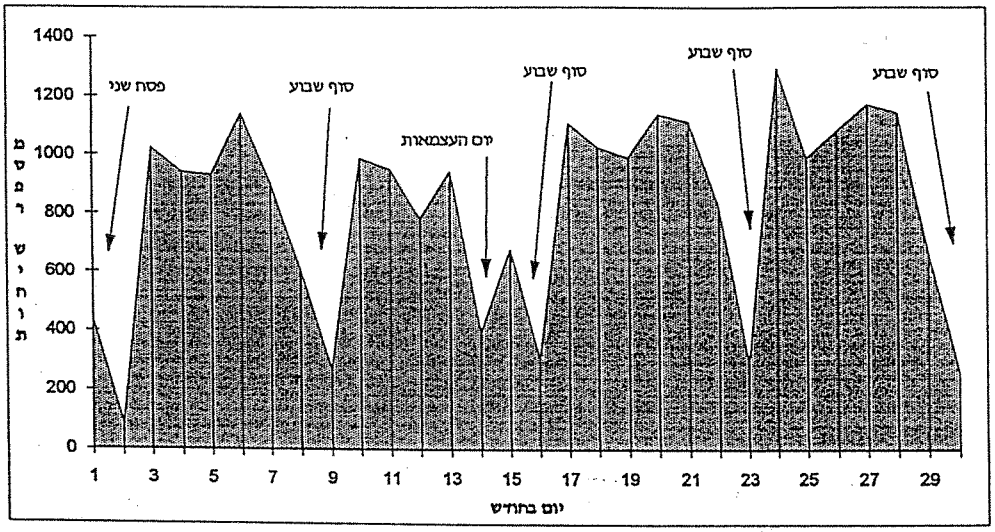
Figure 4 Typical intrahour distribution of calls, 10:00–11:00 A.M.

מופע פניות במשך שנת 1993



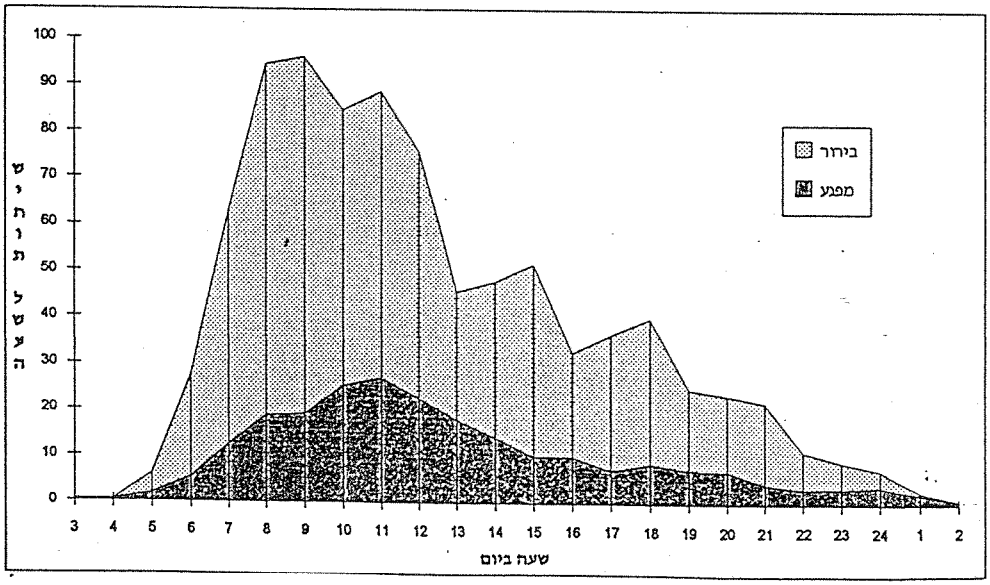
מקור הנתונים: פניות שהוקלדו במחשב המוקד העירוני, תל אביב, בשנת 1993.

מופע שיחות בחודש אפריל 1994



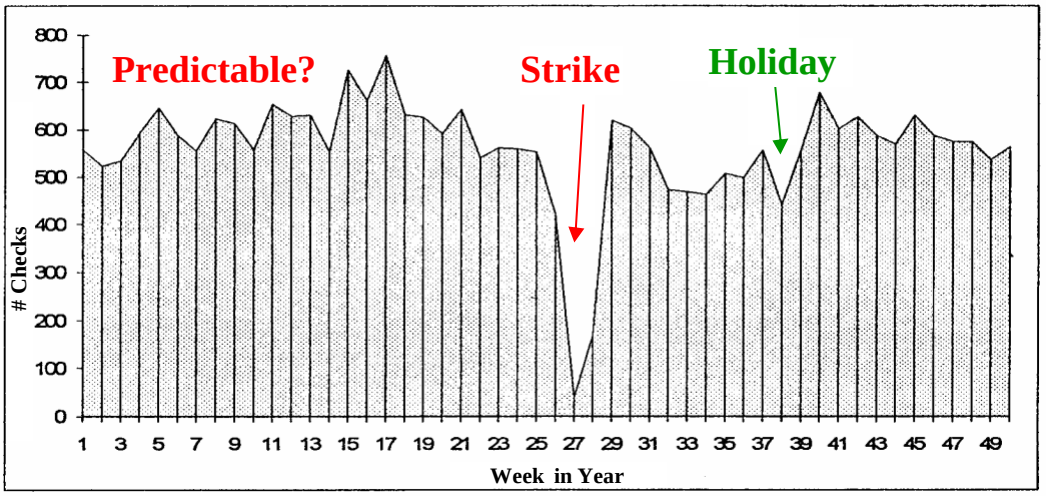
מקור הנתונים: שיחות שנרשמו במערכת ה- ACD במוקד העירוני, תל אביב, בחודש אפריל 1994.

מופע שיחות לפי שעה ביום

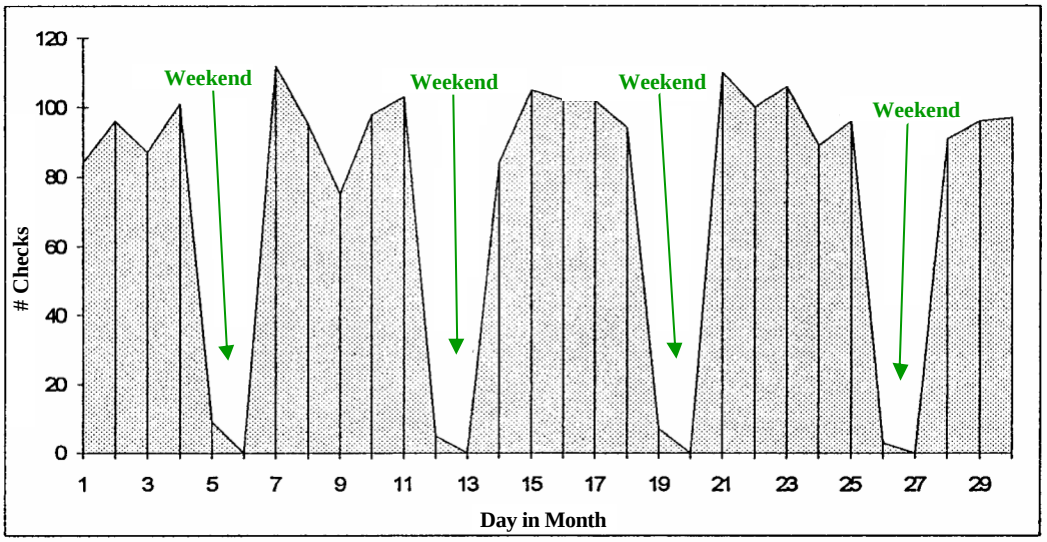


Custom Inspections at an Airport

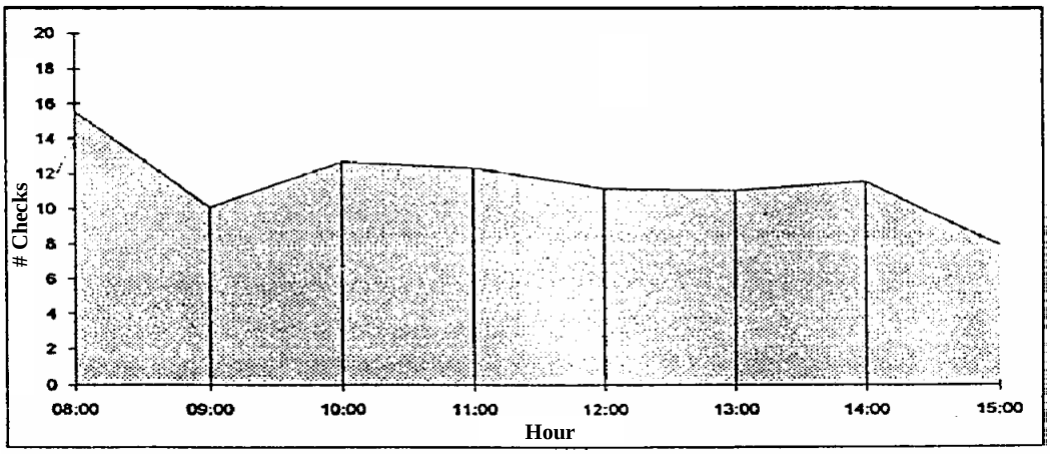
Number of Checks Made During 1993:



Number of Checks Made in November 1993:



Average Number of Checks During the Day:



Source: Ben-Gurion Airport Custom Inspectors Division

Teaching Note

Service Engineering

April 30, 1997
Last Revised 11/11/2003

A Deterministic Model of a Service Station (Fluid View)

Primitives

$Z(0)$	initial content
$\alpha(t)$	input rate
$\mu(t)$	*potential* service rate

(⇒ single resource)

Model: (Think cumulants)

$$\text{Inflow:} \quad A(t) = \int_0^t \alpha(u)du, \quad t \geq 0\,;$$

$$\text{Potential Outflow:} \quad M(t) = \int_0^t \mu(u)du, \quad t \geq 0\,.$$

- We could start with primitives A, M, in which case they need not be continuous; for example, they could be counting processes.

Netflow: $X(t) = Z(0) + A(t) - M(t), \quad t \geq 0.$

Introduce $Y(t) =$ cumulative potential lost during $[0, t]$.

⇒ Outflow: $D = M - Y$ (**A** arrivals; **D** departures)

⇒ Balance:

$$\begin{aligned} Z(t) &= Z(0) + A(t) - D(t) \\ &= Z(0) + A(t) - [M(t) - Y(t)] \\ &= X(t) + Y(t), \qquad t \geq 0\,. \end{aligned}$$

Model $Z = X + Y$

Feasible $Z \geq 0,\ Y \uparrow 0 \quad (Y(0) = 0);$

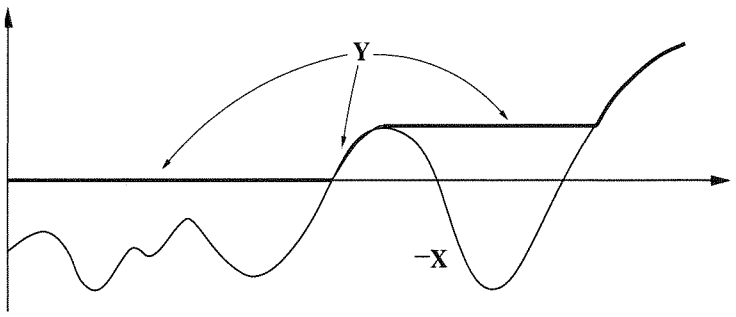
Efficient Y least (hence, Y unique);

Existence: $Y = \overline{(-X)^+}$ $(Y = -\underline{X}$, when $Z(0) = 0)$;

$\underline{X}(t) = \inf_{0 \leq u \leq t} X(u)$, which is called the **lower envelope** of X.

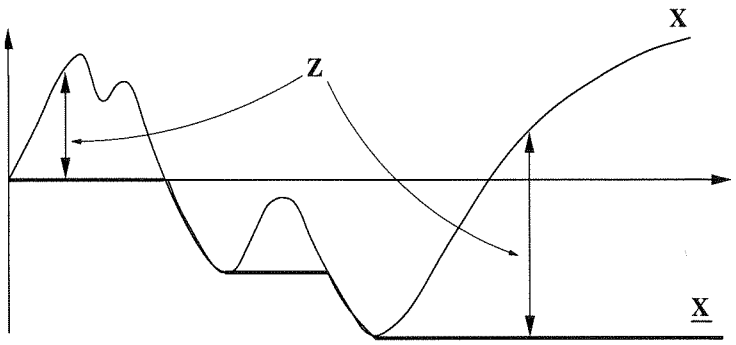
"Proof"

Least $Y \uparrow 0$
s.t. $Y \geq -X$



When $Z(0) = 0$:

$Z = X - \underline{X}$,
$\underline{X}$ = lower envelope.



Equivalent characterization via complementarity: (LCP/DCP)

$$Y \text{ least } <=> ZdY = 0, \text{ i.e. } Y \text{ increases at } t \text{ only when } Z(t) = 0\,.$$

In words: potential lost due to idleness.

Claim (Skorohod) Given $X \in$ RCLL (**R**ight **C**ontinuous **L**eft **L**imit),

there exists a unique (Y, Z) such that

$$\begin{aligned} Z &= X + Y, \\ Z &\geq 0, \quad Y \uparrow 0, \\ ZdY &= 0\,. \end{aligned}$$

Proof Existence by checking $Y = \overline{(-X)^+}$ $(= -\underline{X \wedge 0})$.

Uniqueness by Lyapunov-function argument:

(Note: if minimality is established, then uniqueness is automatic.)

If (Y_i, Z_i), $i = 1, 2$, are two solutions, then consider

$$\eta = \frac{1}{2}(Y_1 - Y_2)^2.$$

Assume, for simplicity, continuous Y_i's, in which case differentiate:

$$\begin{aligned} d\eta = (Y_1 - Y_2)(dY_1 - dY_2) &= (Z_1 - Z_2)(dY_1 - dY_2) \\ &= -Z_1 dY_2 - Z_2 dY_1 \leq 0 \, . \end{aligned}$$

Deduce that η decreases, but also

$$\begin{aligned} \eta(0) = 0 &\Rightarrow \eta \equiv 0 \\ &\Rightarrow Y_1 \equiv Y_2. \end{aligned}$$

Outflow $D(t) = M(t) - Y(t) = \int_0^t \delta(u)du$, where $\delta(u)$ = outflow rate,

$$\Rightarrow \quad Y(t) = \int_0^t [\mu(u) - \delta(u)]du \, .$$

In terms of rates: $dY \geq 0$ implies $\delta \leq \mu$.

Now, either

$\delta = \mu$ or

$\delta < \mu \Leftrightarrow dY > 0,$

$\Rightarrow Z = 0$ (since $ZdY = 0$),

$\Rightarrow d(X + Y) = 0$ (consider a neighbourhood and differentiate),

$\Rightarrow (\alpha - \mu) + (\mu - \delta) = \alpha - \delta = 0.$

Thus (Hall, pg. 190, Def. 6.6),

$$\delta(t) = \begin{cases} \mu(t) & \text{when } Z(t) > 0, \\ \alpha(t) & \text{when } Z(t) = 0 \, . \end{cases}$$

Note that the above is *not* a direct definition of δ, since it uses Z, which is defined in terms of δ.

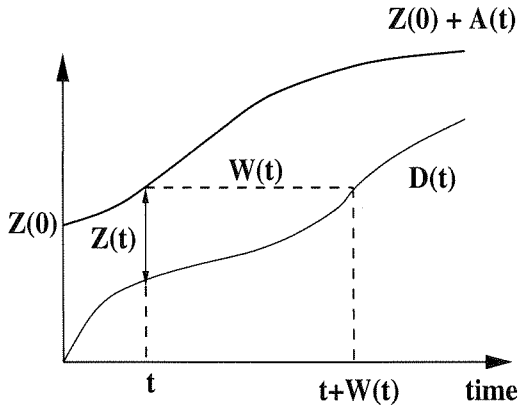
How to calculate **Delay**?

Define

$$\begin{aligned} W(t) &= \text{work-load at time } t \\ &\quad (= \text{time to process all that is present at time } t) \\ &= \text{under FCFS, virtual waiting time.} \end{aligned}$$

Defining relation for W:

$$D(t + W(t)) = Z(0) + A(t)$$



Hence, $Z(t + W(t)) = Z(0) + A(t + W(t)) - A(t)$.

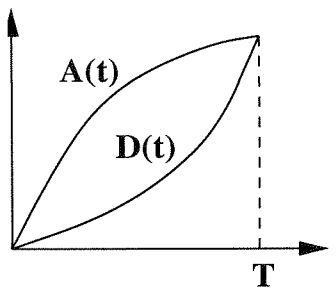
MOP's over a finite horizon T:

Averages

Inflow: $\bar{\alpha} = \frac{1}{T}\int_0^T \alpha(t)dt$;

Outflow: $\bar{\delta} = \frac{1}{T}\int_0^T \delta(t)dt$;

Throughput: λ, defined when $\bar{\alpha} = \bar{\delta}$ as their common value.



eg. $\lambda = \frac{1}{T}A(T) = \frac{1}{T}D(T)$.

Queue length (Inventory): $\bar{Z} = \frac{1}{T}\int_0^T Z(t)dt = \frac{1}{T}\times$ Area.

Delay: $\bar{W} = \frac{1}{A(T)}\int_0^T W(t)dA(t) \quad \left(= \frac{\int_0^T W(t)\alpha(t)dt}{\int_0^T \alpha(t)dt}\right)$.

$\uparrow$

Rieman-Stiltjes

35

Intuition:

- Discrete arrivals $\Rightarrow \bar{W} = \frac{1}{A(T)} \sum_{n=1}^{A(T)} W_n$ (as in Hall, Chap. 2);
- Absolutely continuous: $\alpha(t)dt$ arrivals during $(t, t+dt)$, each suffering a delay of $W(t)$.

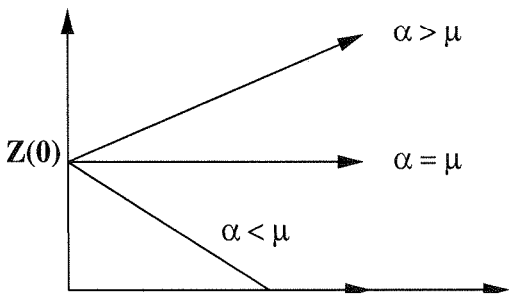
Little's Conservation Law: $\bar{Z} = \lambda \cdot \bar{W}$.

Cumulative lost potential $Y(T)$.

Efficiency $\varepsilon(T) = 1 - \frac{Y(T)}{M(T)} =$

$$\text{actual} \searrow \quad = \frac{D(T)}{M(T)} \quad \left(= \frac{\int_0^T \delta(t)dt}{\int_0^T \mu(t)dt}, \text{ when applicable} \right).$$
$$\text{potential} \nearrow$$

Example *constant rates* $\alpha(t) \equiv \alpha$, $\mu(t) \equiv \mu$.
(linear model)

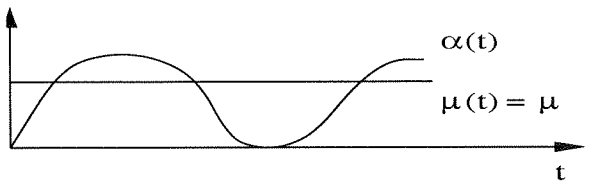


overloaded (supercritical)	$(\rho > 1)$
critically loaded (critical)	$(\rho = 1)$
underloaded (subcritical)	$(\rho < 1)$

Definition: $\rho = \alpha/\mu$ **traffic (flow) intensity.**

Natural *extension*: piecewise constant rates, as in National Cranberry (HBS case): later

Example *periodic rates* e.g.



(If α has a period $T_\alpha = 8$, μ has a period $T_\mu = 3$, take period $T = T_\alpha \cdot T_\mu = 24$.)

36

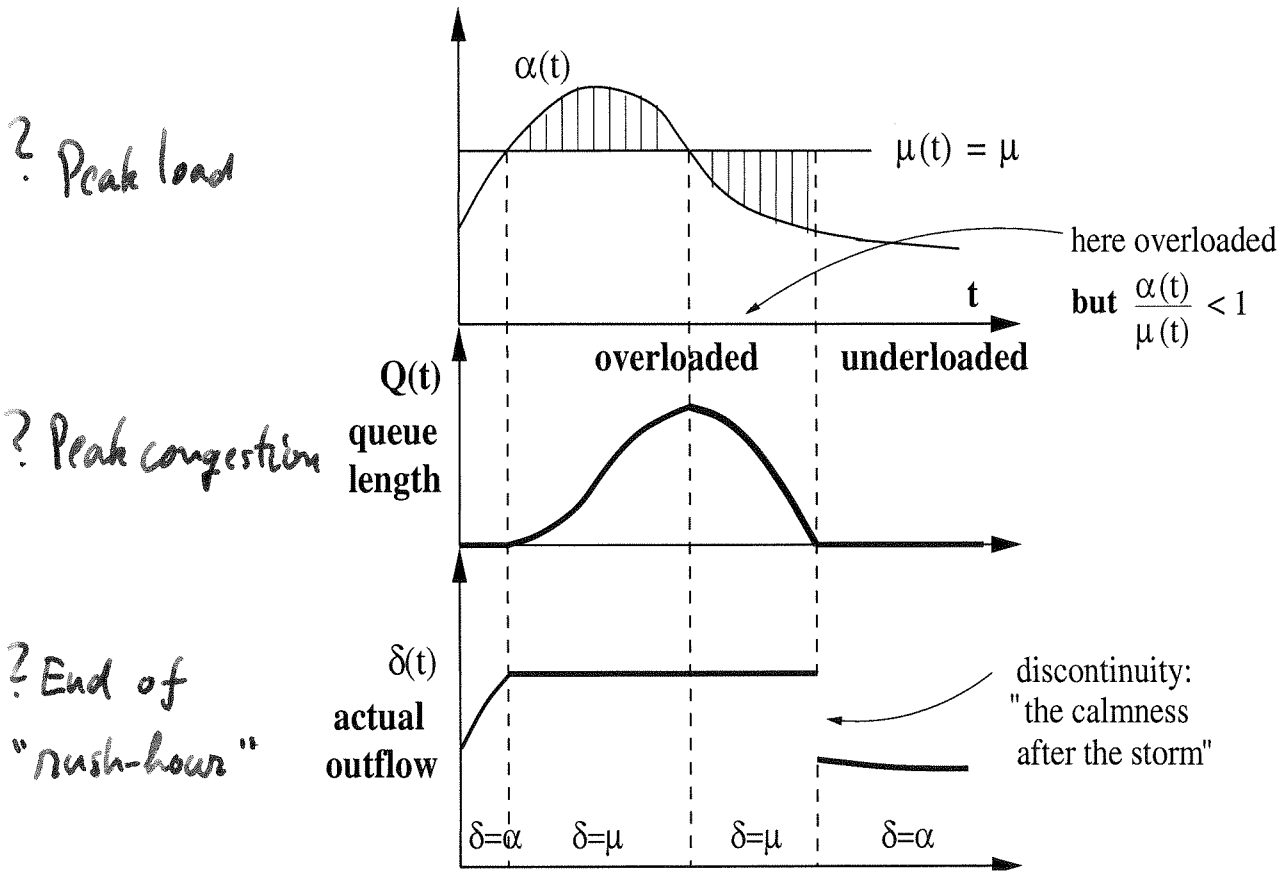
Long-run: $\bar{\alpha} = \frac{1}{T}\int_0^T \alpha(t)dt$; $\bar{\mu} = \frac{1}{T}\int_0^T \mu(t)dt$;

$\rho = \bar{\alpha}/\bar{\mu}$ (Heyman-Whitt).

Short-run: Phase-transitions (different from Hall, pg. 189–190, that has stagnant → growth → decline → stagnant).

Short-Run Phase Transitions

Overloaded at t	:	$Z(t) > 0$;		
Underloaded	:	$Z(t) = 0$	and $\delta(t) < \mu(t)$	(excess capacity, $dY(t) > 0$);
Critically loaded	:	$Z(t) = 0$	and $\delta(t) = \mu(t)$	(balanced capacity, $dY(t) = 0$).

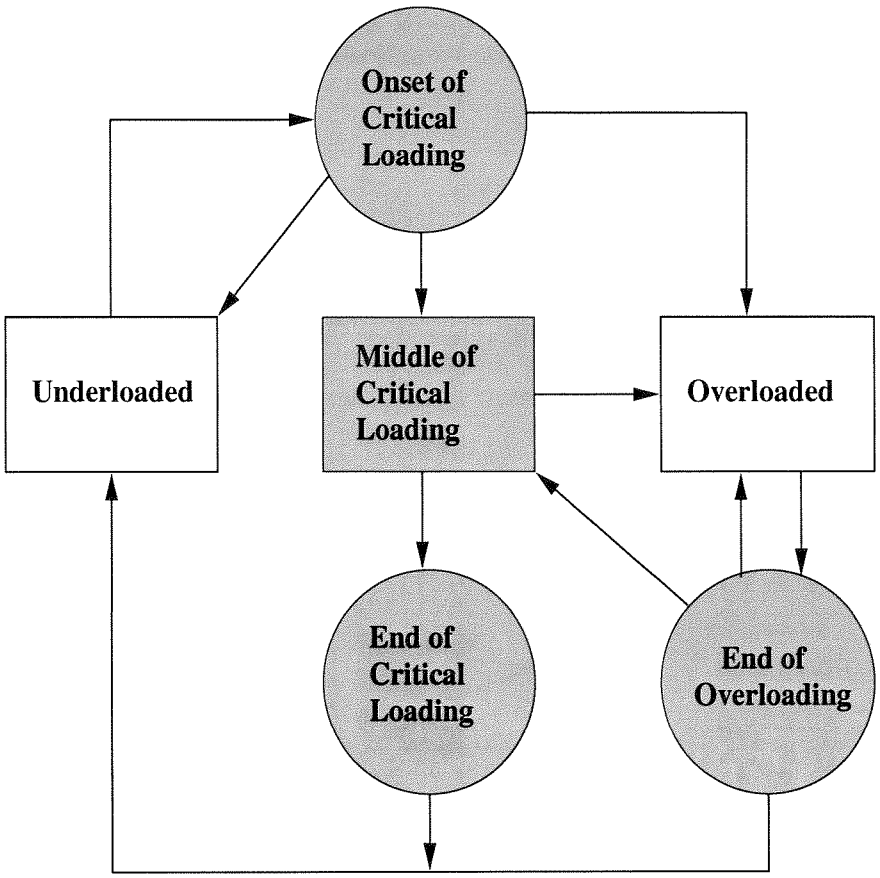


The analogue of ρ, traffic intensity, is here (assume $Z(0) = 0$):

$$\rho(t) = \sup_{0 \le s \le t} \frac{\int_s^t \alpha(u)du}{\int_s^t \mu(u)du} \qquad \begin{cases} > 1 & \text{overloaded} \\ = 1 & \text{critically loaded} \\ < 1 & \text{underloaded} \end{cases}$$

37

For finer approximations, we must acknowledge more phases, as depicted in the following figure.



Phase transition diagram for the asymptotic regions.
(Massey & Mandelbaum.)

References:

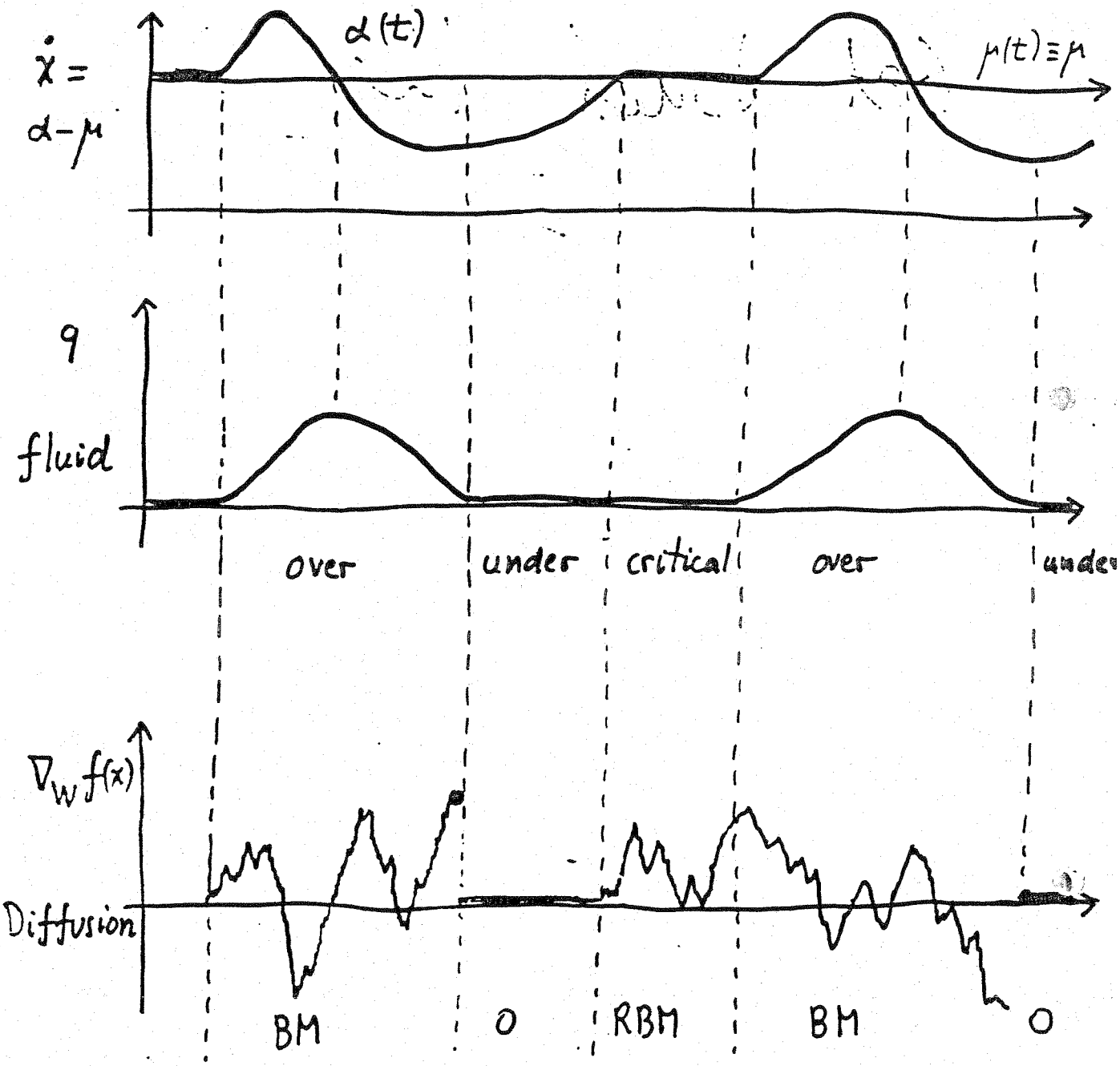
- Hall, R.W., "*Queueing Methods for Service and Manufacturing*", Prentice Hall, 1991.
- Harrison, J.M., "*Brownian Motion and Stochastic Flow Systems*", Wiley, 1985.
- Mandelbaum, A. and Massey, William, A., "Strong approximations for time-dependent queues", *Math. of Operations Research,* 20, 33-64, 1995.

Finer approximations based on stochastic (Brownian) refinements (later in course)

Why? confidence intervals

Phase Transitions $\rho(t)$: <1, $=1$, >1

31



M_1: queue of size $\frac{1}{\sqrt{\varepsilon}} = \sqrt{n}$ depletes during $\sqrt{n}$

but accelerated by n

Dynamic acceleration: slow down $\pm\sqrt{n}$ around jumps

Why Models? here is one answer:

Mathematical Framework

Reflection Mapping (Regulator) $\qquad X \to X - \underline{X \wedge 0}$

$$(X \to X - \underline{X}\,, \quad \text{when } X(0) = 0)\,.$$

Fundamental:

- Flow analysis (Fluid Models);
- Economics;
- Stochastic Processes;
  - Skorohod (needed cumulant Y!);
  - Queueing Models (later);
- Approximations.

Idea of Approximations: $\quad Z = f(X), \quad f$ continuous (Lipshitz).

Hence, $\quad X \approx \tilde{X}$ implies $Z \approx \tilde{Z} = f(\tilde{X})$

$X \approx \bar{X}$ fluid $\Rightarrow \bar{Z} = f(\bar{X})$ fluid approximations.

$X \approx \bar{X} + \hat{X}$ diffusion $\Rightarrow \hat{Z} = f(\bar{X} + \hat{X})$ diffusion refinements.

Reference: Harrison, Chapter 2 (which covers also finite buffers, and two-node networks).

(One more)

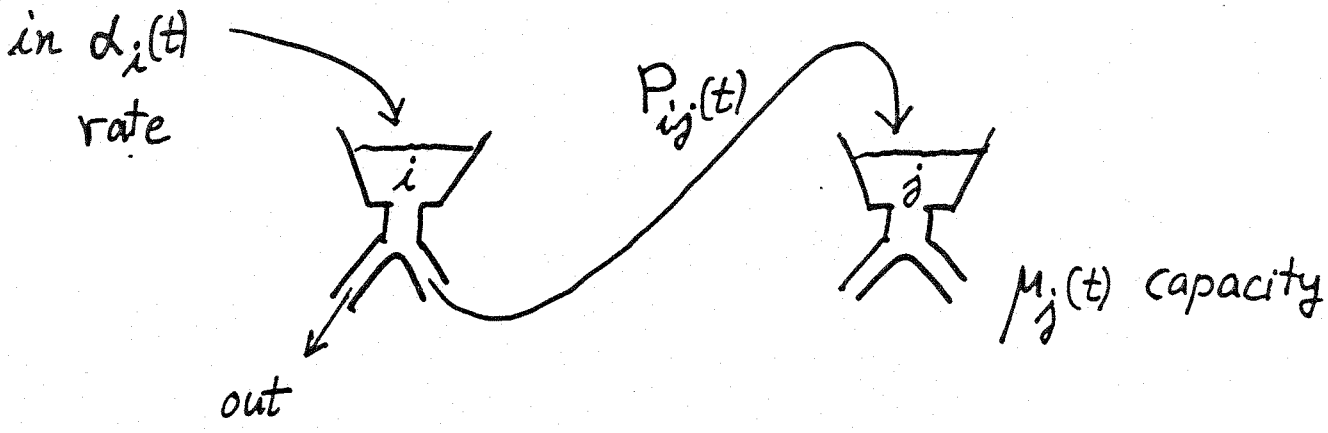
Summary of "Pictures" : 4 in total.

1. Rates (⇒ peak load)
2. Queues (⇒ congestion)
3. Outflows (⇒ end of rush-hour)
4. Cummulants (Integrals)

24

A Fluid Network

Analogue of Skorohod's Model.



Outflow rates $\delta_j(t) \leq \mu_j(t)$ efficient!

$< \mu_j(t) \Rightarrow Z_j(t) = 0$

Inflow $\lambda_j = \alpha_j + \sum_i \delta_i P_{ij}$

Content $Z = Z(0) + \int \lambda - \int \delta = f(X)$

The Mapping ↑ ↖ data (α, μ, P)

Reflection Regulator DCP(X)

$$f \begin{cases} Z(t) = X(t) + \int_0^t dY(s)\,[I - P(s)], & t \geq 0 \\ Z \geq 0,\ Y \uparrow 0,\ \int Z\,dY = 0 \end{cases}$$

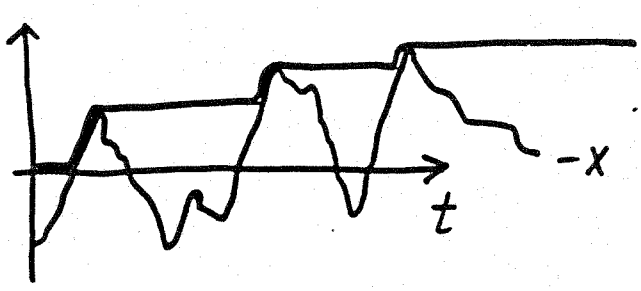
with $X = Z(0) + \int(\alpha + \mu P) - \int \mu$ netflow

$Y = \int \mu - \int \delta$ cum. lost capacity

25

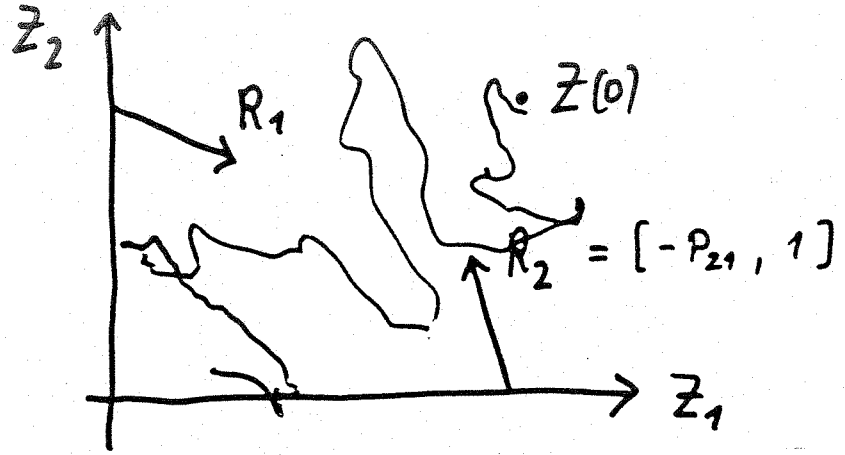
Geometric Interpretation: Oblique Reflection

Single buffer $\quad Z = X + Y \geq 0, \; Y \uparrow 0, \; \int Z\,dY = 0$



Skorohod

Two buffers $\quad Z = X + Y\,[I - P] = X + Y_1 R_1 + Y_2 R_2$



Harrison + Reiman

General $\quad f : X \to Z$ $\quad$ Lipshitz

$f: X \to Z$

26 Outflow $\delta_j(t)$

Inflow $\lambda_j(t) = \alpha_j(t) + \sum_i \delta_i(t) P_{ij}(t)$

Example: constant rates $\alpha(t) \equiv \alpha$, μ, P ~~linear~~ relaxation-pers

$\Rightarrow$ Equilibrium $\boxed{\lambda_j = \alpha_j + \sum_i (\lambda_i \wedge \mu_i) P_{ij}}$ Traffic

Traffic intensity $\rho_j = \frac{\lambda_j}{\mu_j}$ $\begin{cases} > 1 & \text{bottleneck} \\ = 1 & \text{critical} \\ < 1 & \text{"stable"} \end{cases}$

General (ex. periodic)

~~$\frac{\alpha_j(t)}{\mu_j(t)} \leq$~~ $\rho_j(t) = \sup_{0 \leq s \leq t} \frac{\int_s^t \lambda_j(u)\,du}{\int_s^t \mu_j(u)\,du}$

$\rho_j(t) > 1 \iff z_j(t) > 0$ over loaded

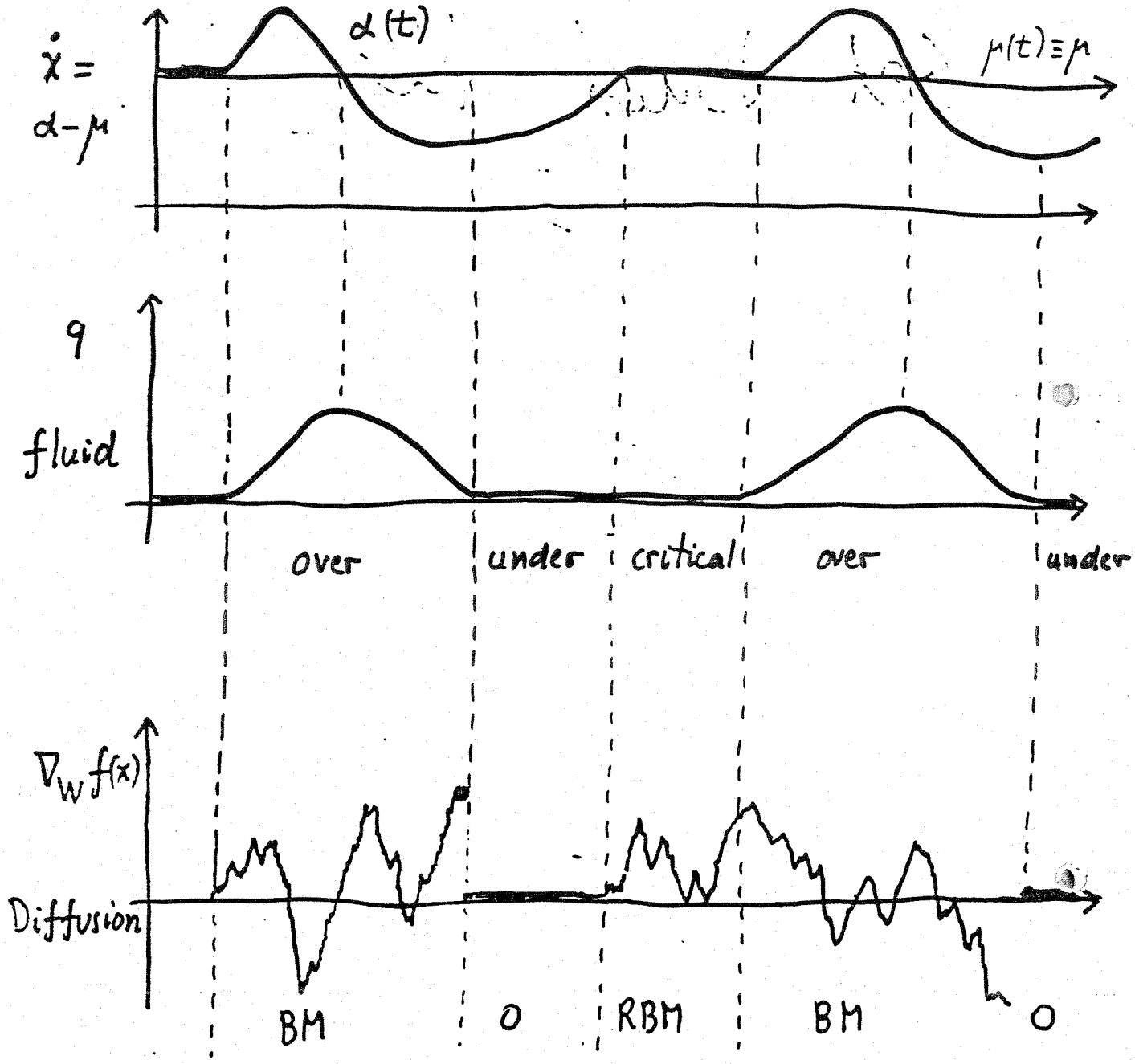
$= 1 \iff z_j(t) = 0,\ \delta_j(t) = \mu_j(t)$ critical

$< 1 \iff z_j(t) = 0,\ \delta_j(t) < \mu_j(t)$ under loaded

Phases Time depend

Phase Transitions $\rho(t)$: <1, $=1$, >1

31



M_1: queue of size $\frac{1}{\sqrt{\varepsilon}} = \sqrt{n}$ depletes during $\sqrt{n}$

but accelerated by n

Dynamic acceleration: slow down $\pm\sqrt{n}$ around jumps

Predictable Queues

Fluid Models and Diffusion Approximations

for Time-Varying Queues with Abandonment and Retrials

with

Bill Massey

Marty Reiman

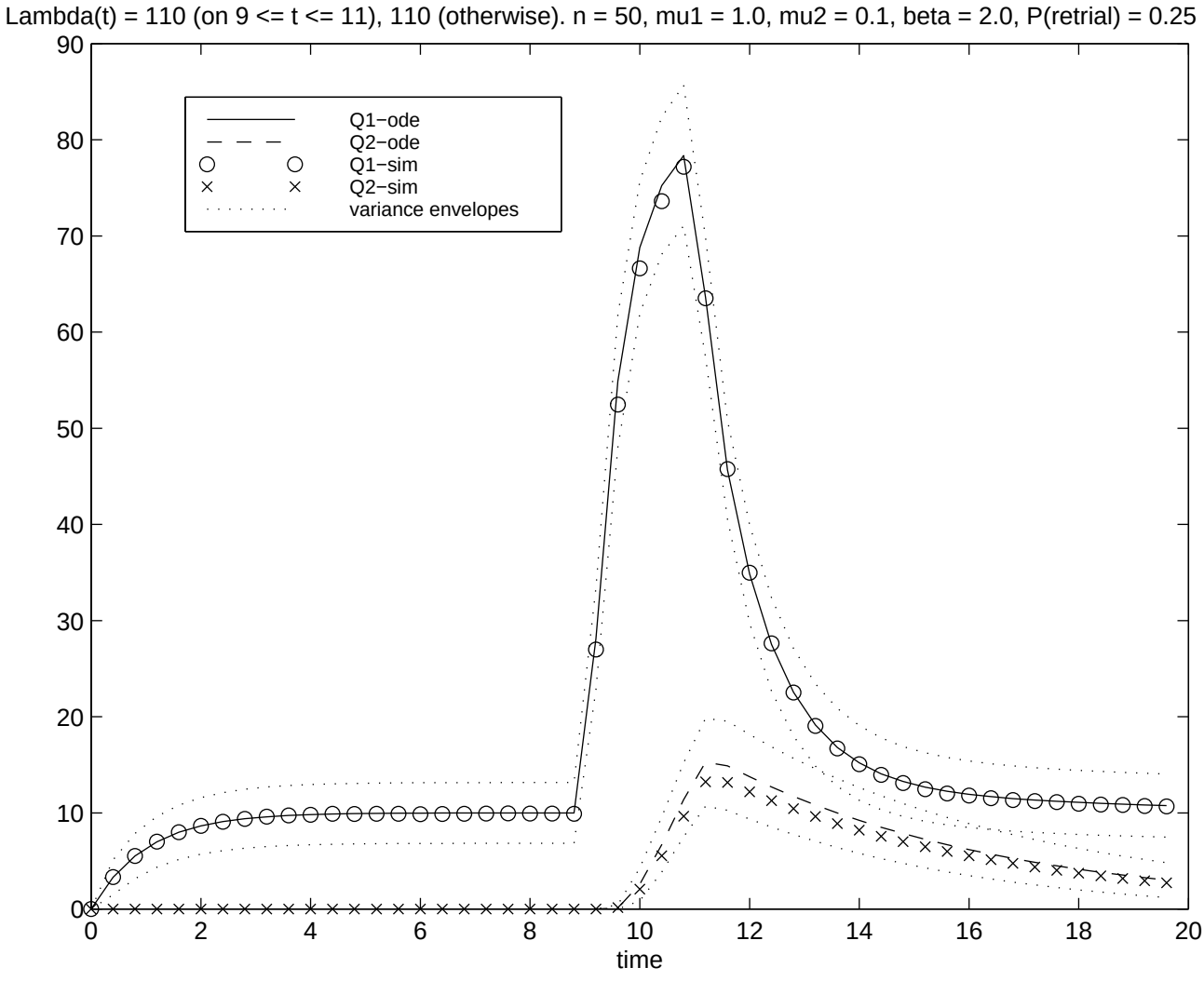
Brian Rider

Sasha Stolyar

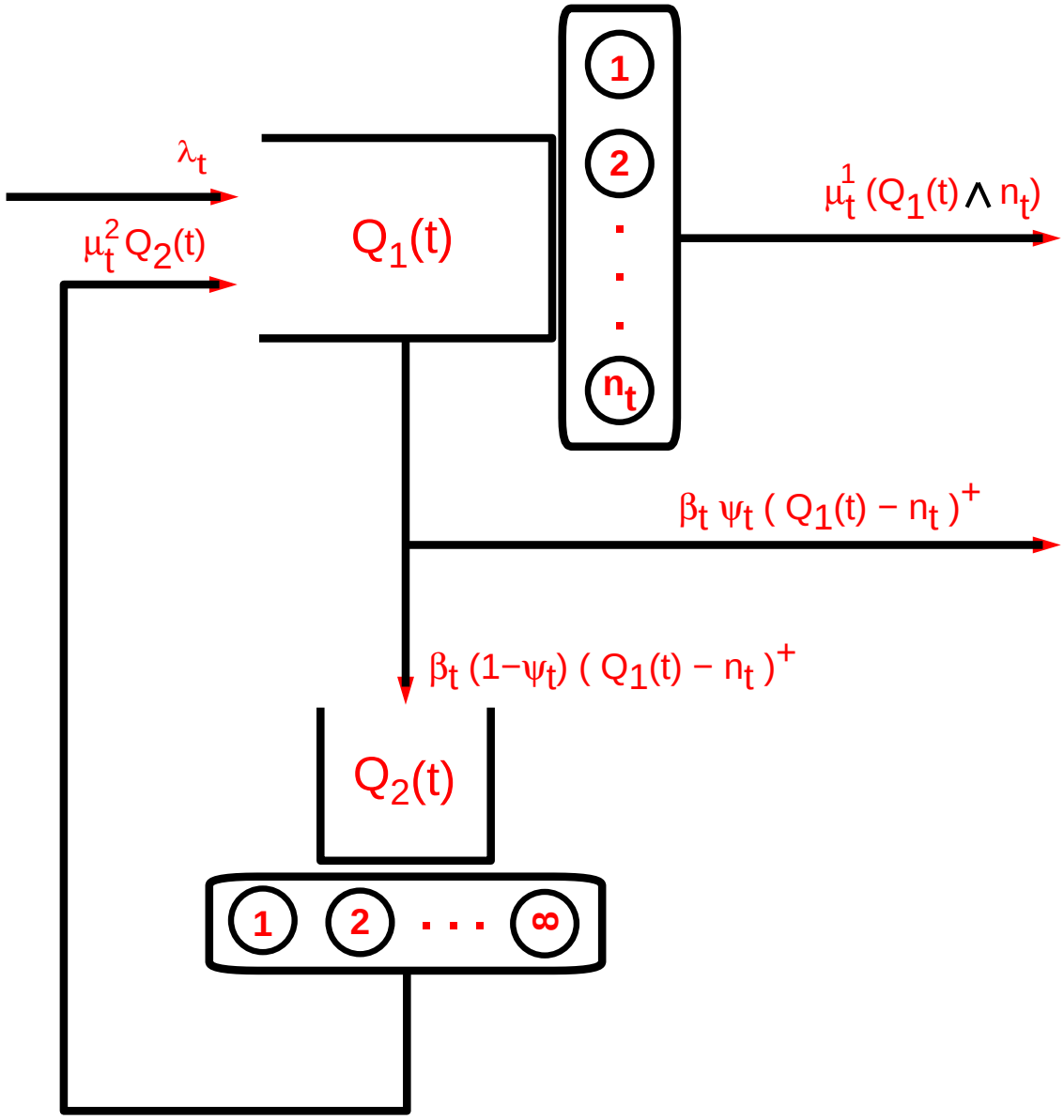
Sudden Rush Hour

n = 50 servers; $\mu = 1$

λ_t = 110 for $9 \leq t \leq 11$, $\lambda_t = 10$ otherwise



Call Center: A Multiserver Queue with Abandonment and Retrials



Primitives (Time-Varying Predictably)

λ_t exogenous arrival rate

e.g., continuously changing, sudden peak

μ_t^1 service rate

e.g., change in nature of work or fatigue

n_t number of servers

e.g., in response to predictably varying workload

β_t abandonment rate while waiting

e.g., in response to IVR discouragement

at predictable overloading

ψ_t probability of no retrial

$1/\mu_t^2$ average time to retry

Large system: $\eta \uparrow \infty$ scaling parameter. Now define

$Q^\eta(\cdot)$ via $\lambda_t \longrightarrow \eta\lambda_t$

$n_t \longrightarrow \eta n_t$

What do we get, as $\eta \uparrow \infty$?

Fluid Model

Replacing random processes by their rates yields

$$Q^{(0)}(t) = (Q_1^{(0)}(t),\ Q_2^{(0)}(t))$$

Solution to nonlinear differential balance equations

$$\begin{aligned}
\frac{d}{dt}\, Q_1^{(0)}(t) &= \lambda_t - \mu_t^1\ (Q_1^{(0)}(t) \wedge n_t) \\
&\quad + \mu_t^2\ Q_2^{(0)}(t) - \beta_t\ (Q_1^{(0)}(t) - n_t)^+ \\
\frac{d}{dt}\, Q_2^{(0)}(t) &= \beta_1(1-\psi_t)(Q_1^{(0)}(t) - n_t)^+ \\
&\quad - \mu_t^2\ Q_2^{(0)}(t)
\end{aligned}$$

Justification: **Functional Strong Law of Large Numbers**,
with $\lambda_t \to \eta\lambda_t,\ n_t \to \eta n_t$.

As $\eta \uparrow \infty$,

$$\frac{1}{\eta}\, Q^{\eta}(t) \to Q^{(0)}(t)\,, \quad \text{uniformly on compacts, a.s.}$$

given convergence at $t = 0$

Diffusion Refinement

$$Q^\eta(t) \stackrel{d}{=} \eta\, Q^{(0)}(t) + \sqrt{\eta}\, Q^{(1)}(t) + o(\sqrt{\eta}\,)$$

Justification: **Functional Central Limit Theorem**

$$\sqrt{\eta}\,\left[\frac{1}{\eta}\, Q^\eta(t) - Q^{(0)}(t)\right] \stackrel{d}{\to} Q^{(1)}(t), \quad \text{in } D[0,\infty)\,,$$

given convergence at $t = 0$.

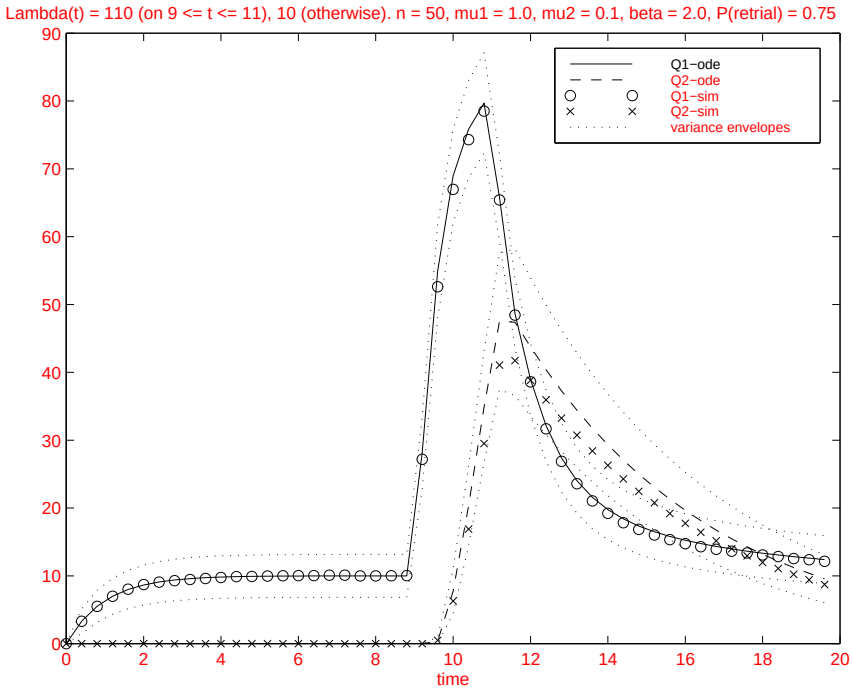
$Q^{(1)}$ solution to stochastic differential equation.

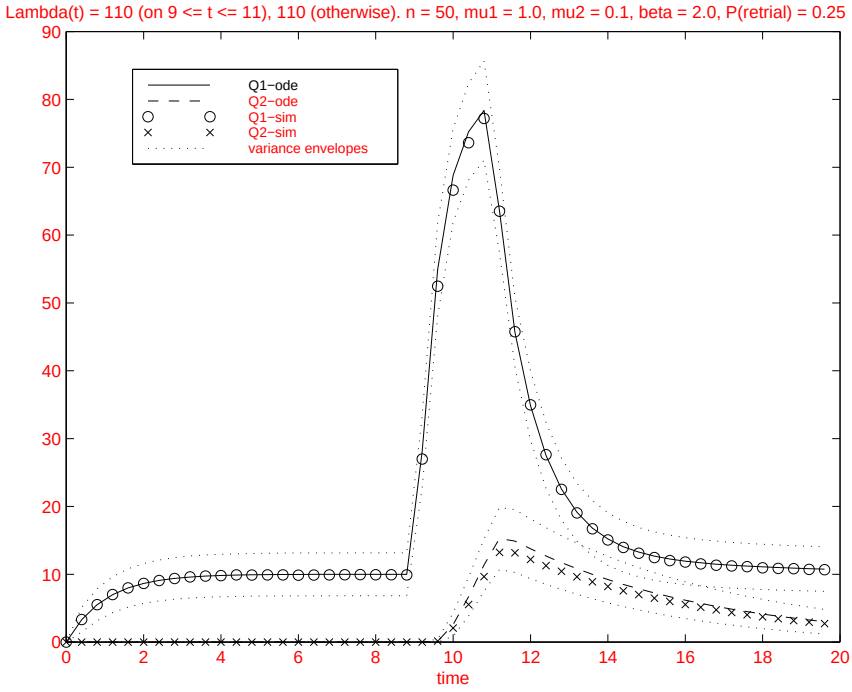
If the set of critical times $\{t \geq 0 : Q_1^{(0)}(t) = n_t\}$ has Lebesque measure zero, then $Q^{(1)}$ is a Gaussian process. In this case, one can deduce ordinary differential equations for

$$EQ_i^{(1)}(t)\,, \quad \text{Var}\, Q_i^{(1)}(t) : \quad \text{confidence envelopes}$$

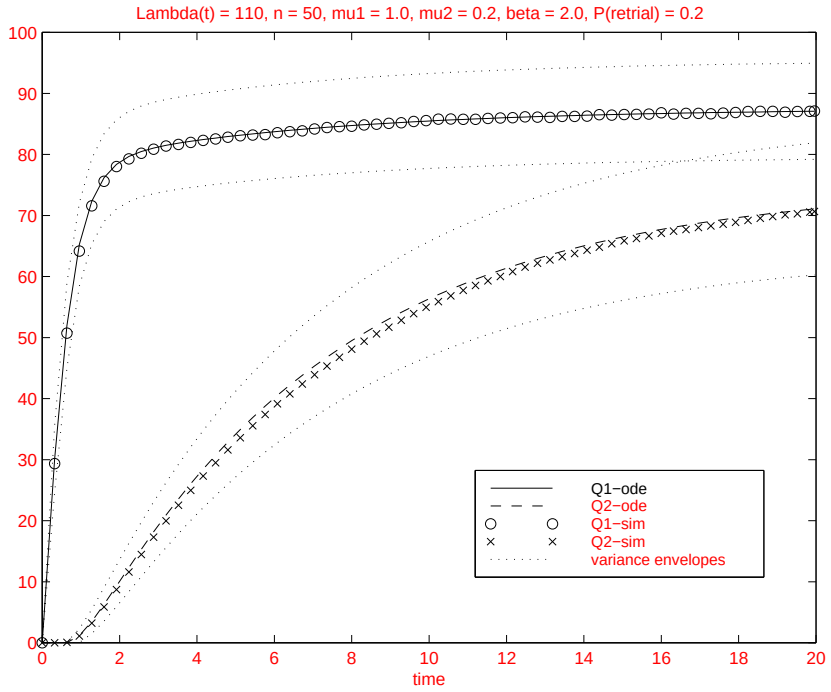
These ode's are easily solved numerically (in a spreadsheet, via forward differences).

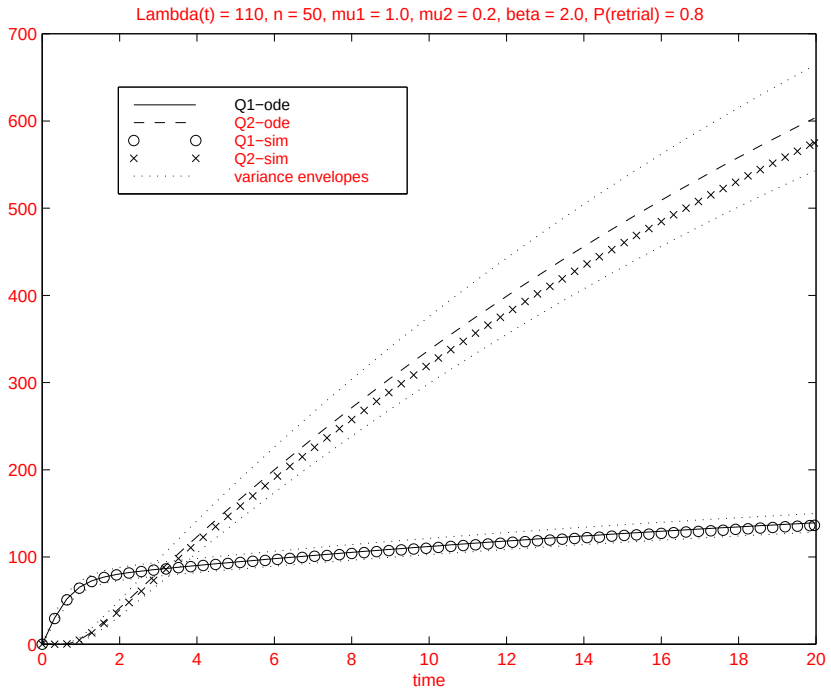
What if $Pr\{$Retrial$\}$ increases to 0.75 from 0.25 ?





Starting Empty and Approaching Stationarity





3. Numerical Examples

Our numerical examples cover the case of time-varying behavior only for the external arrival rate λ_t. We make $\mu^1 = 1$, $\mu^2 = 0.2$, and $Q_1(0) = Q_2(0) = 0$ but let n, β, and ψ range over a variety of different constants.

The first two examples, see Figure 2, that we consider actually have the arrival rate λ equal to a constant 110, with $n = 50$, $\beta = 2.0$, and $\psi = 0.2$ and 0.8. This is an overloaded system, see [8], i.e. $Q_1^{(0)}(t) > n$ for large enough t, and equations (1) and (2) indicate that $Q_1^{(0)}(t) \to q_1$ and $Q_1^{(0)}(t) \to q_2$ as $t \to \infty$. Setting $\frac{d}{dt}Q_1^{(0)}(t) = \frac{d}{dt}Q_1^{(0)}(t) = 0$ as $t \to \infty$, then q_1 and q_2 solve the linear equations

$$\lambda + \mu^2 q_2 - \mu^1 n - \beta(q_1 - n) = 0 \tag{12}$$

and

$$\beta(1 - \psi)(q_1 - n) - \mu^2 q_2 = 0. \tag{13}$$

These equations can be easily solved to yield

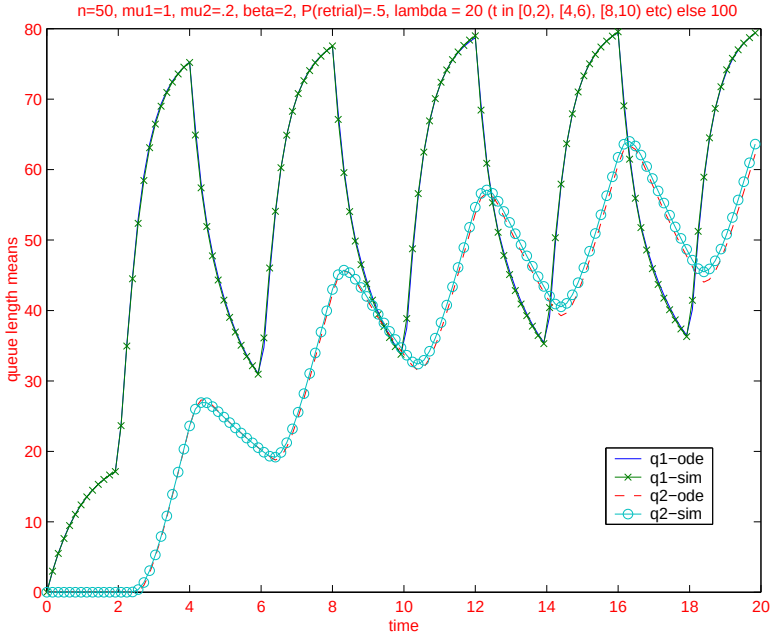
$$q_1 = n + \frac{\lambda - \mu^1 n}{\beta\psi} \quad \text{and} \quad q_2 = \frac{\beta(1-\psi)}{\mu^2}\frac{\lambda - \mu^1 n}{\beta\psi}. \tag{14}$$

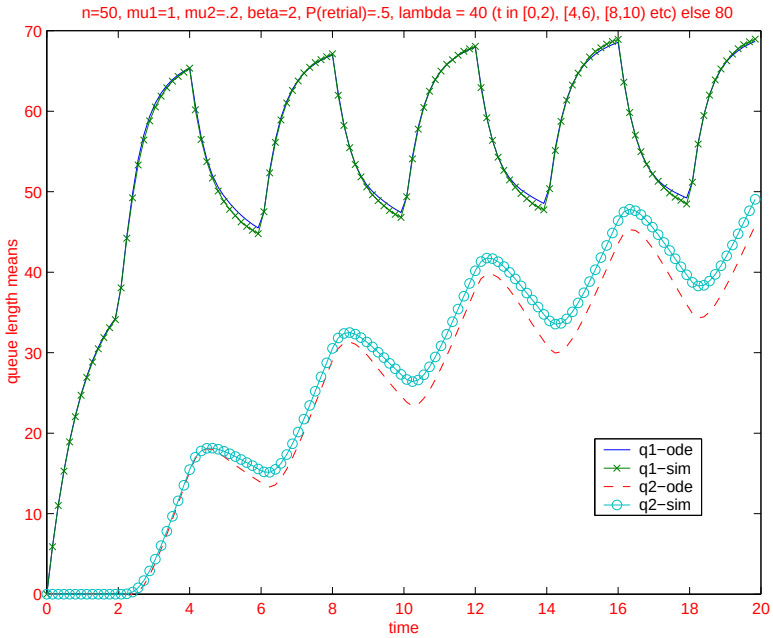
Substituting in $\psi = 0.2$ and the other parameters indicated above yields $q_1 = 200$, $q_2 = 1200$. This case corresponds to the graph of the left in Figure 2 and indicates that this system is still far from equilibrium at time 20. With $\psi = 0.8$ (so the probability of retrials is equal to 0.2) we obtain $q_1 = 87.5$ and $q_2 = 75$. This case corresponds to the graph on the right in Figure 2. Here it appears that $Q_1^{(0)}$ has essentially reached equilibrium by the time $t = 20$, while $Q_2^{(0)}$ has a bit more to go.

In general, the accuracy for the computation of the fluid approximation can be checked by a simple test that only requires a visual inspection of the graphs.

Sample Mean vs. Fluid Approximation

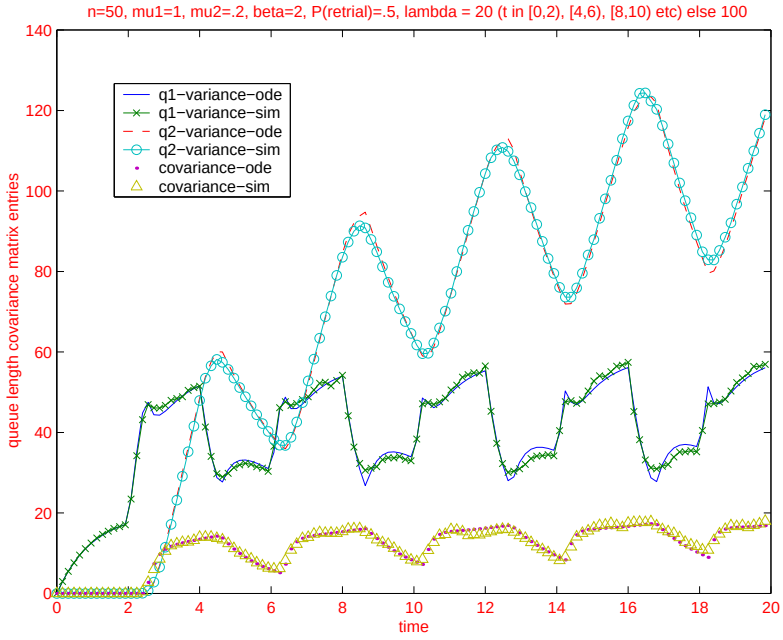
Queue Lengths (λ_t = 20 or 100)

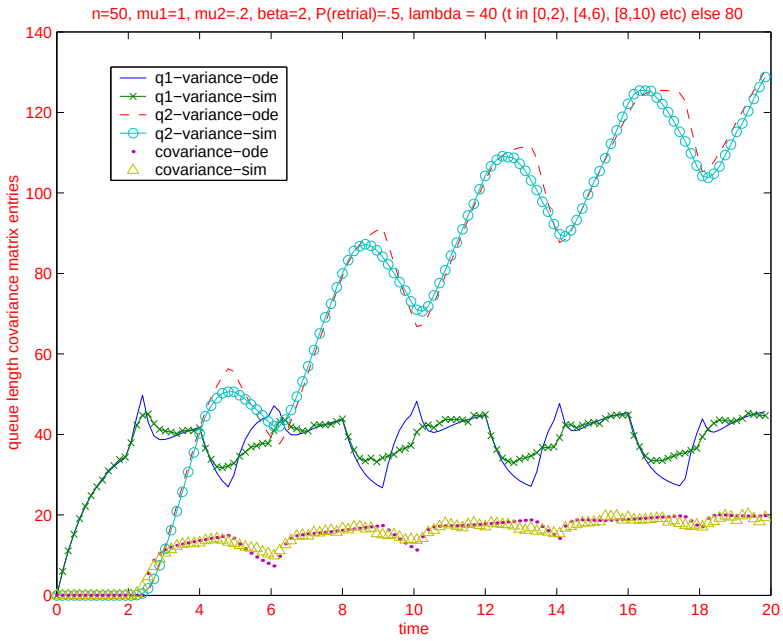




Variances and Covariances

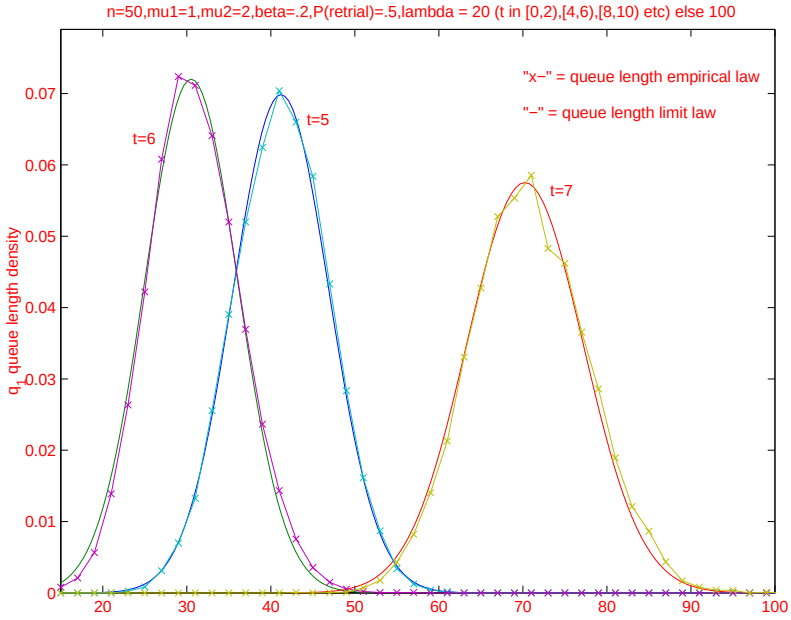
Queue Lengths

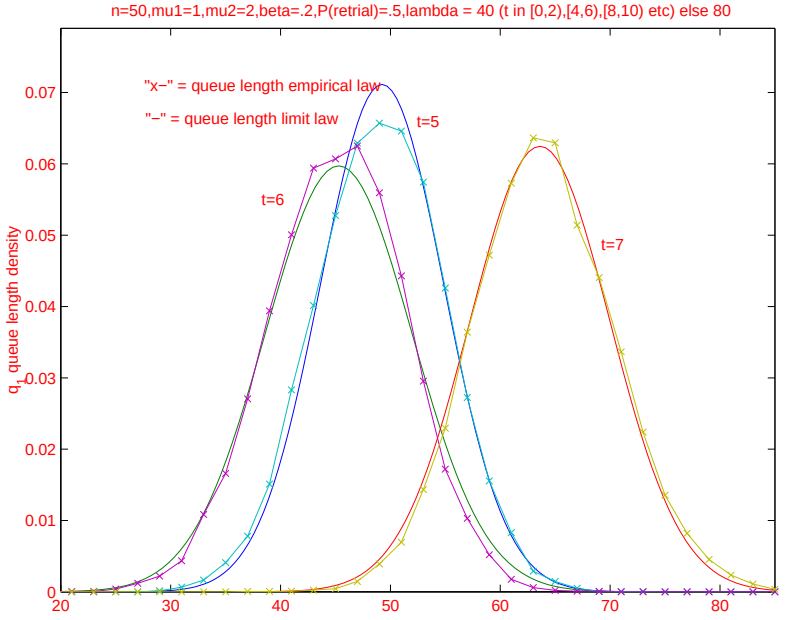




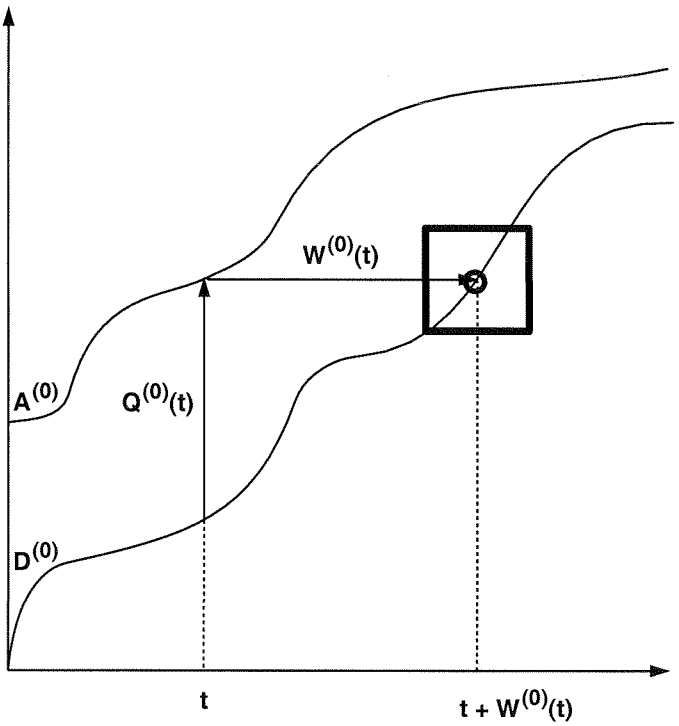
Sample Density vs. Gaussian Approximation

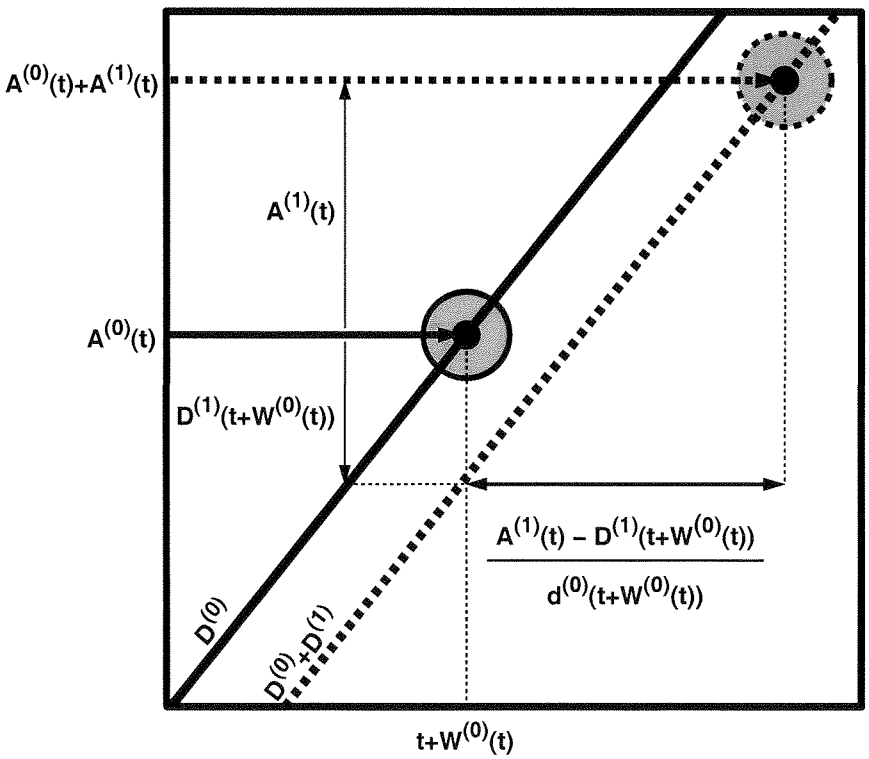
Multi-Server Queue





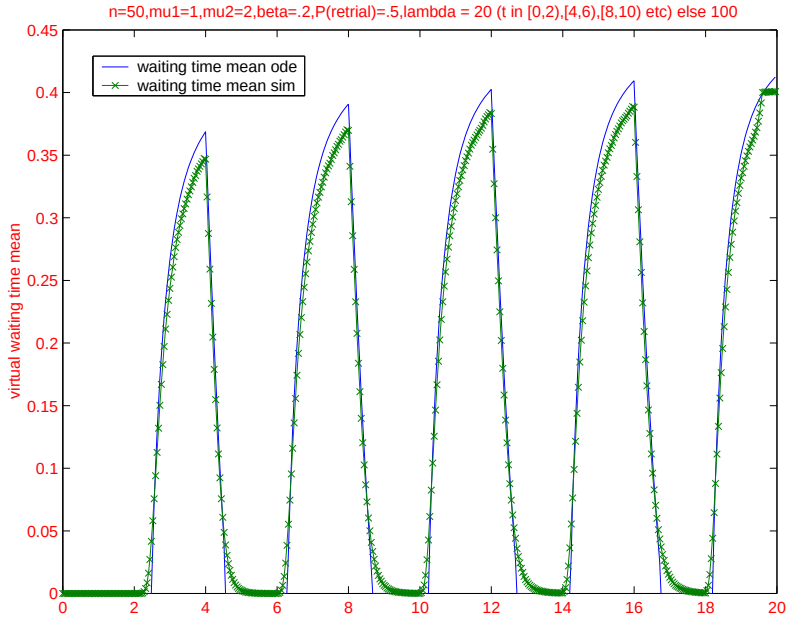
Waiting Time

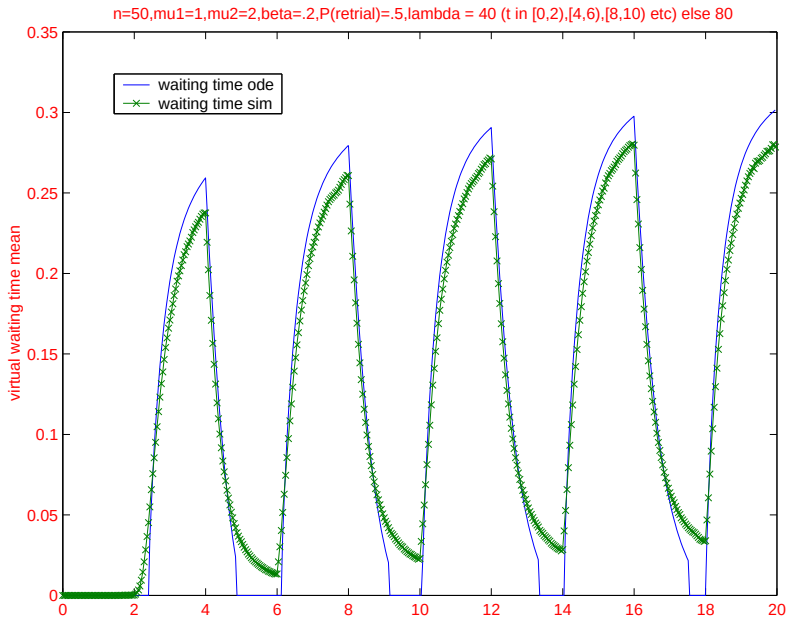




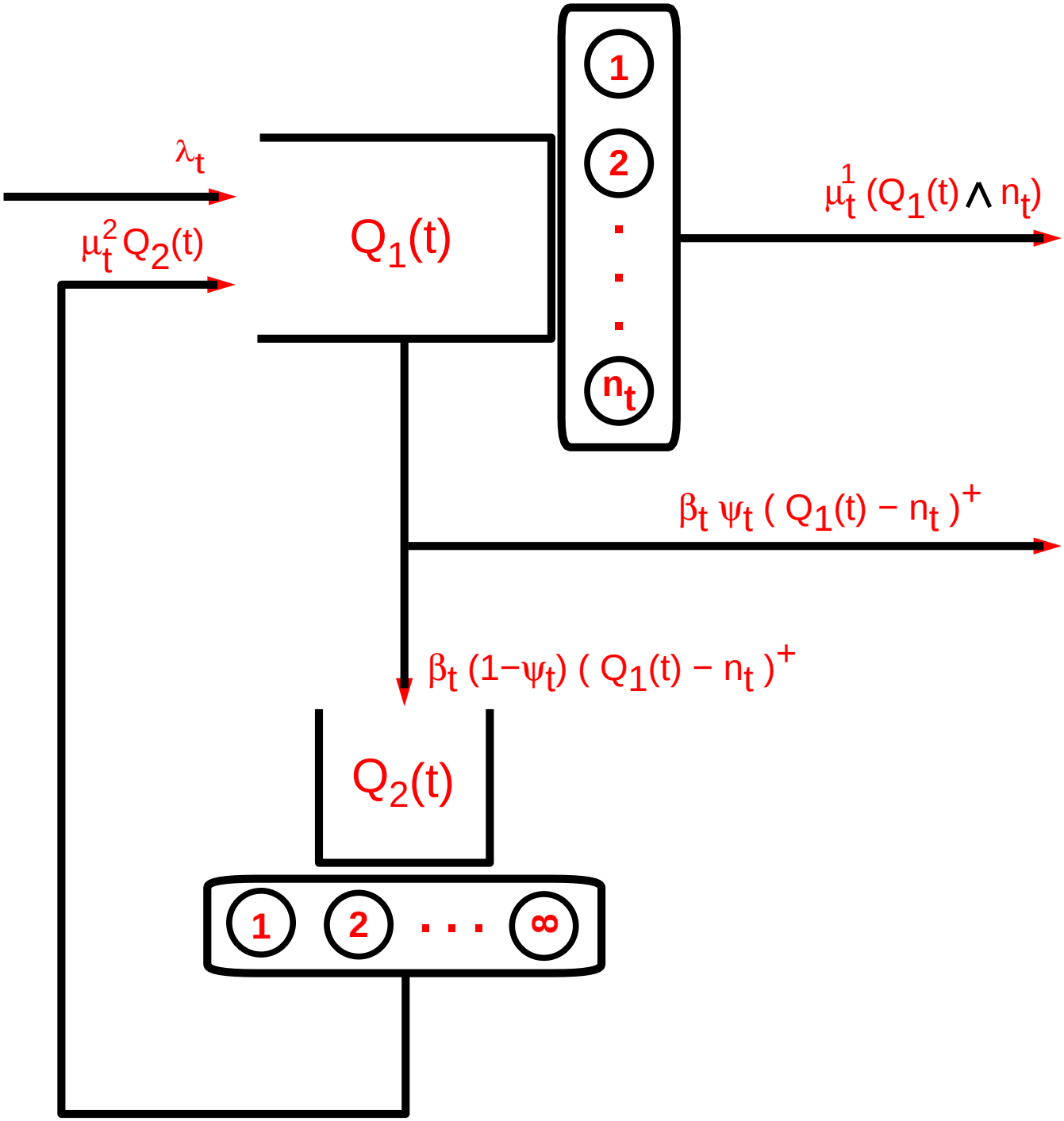
Sample Mean vs. Fluid Approximation

Virtual Waiting Time





Back to the Multiserver Queue with Abandonment and Retrials



Sample Path Construction of a Multiserver Queue with Abandonment and Retrials

$$Q_1(t) = Q_1(0) + A^a\left(\int_0^t \lambda_s ds\right)$$
$$+ A^c_{21}\left(\int_0^t Q_2(s)\mu^2_s ds\right) - A^c\left(\int_0^t (Q_1(s)\wedge n_s)\mu^1_s ds\right)$$
$$- A^b_{12}\left(\int_0^t (Q_1(s)-n_s)^+\beta_s(1-\psi_s)ds\right)$$
$$- A^b\left(\int_0^t (Q_1(s)-n_s)^+\beta_s\psi_s ds\right)$$

and

$$Q_2(t) =$$
$$Q_2(0) + A^b_{12}\left(\int_0^t (Q_1(s)-n_s)^+\beta_s(1-\psi_s)ds\right)$$
$$- A^c_{21}\left(\int_0^t Q_2(s)\mu^2_s ds\right).$$

$A_\cdot \stackrel{d}{=}$ Poisson(1), independent.

Fluid Limit for the Multiserver Queue with Abandonment and Retrials (2 O.D.E.'s)

$$\frac{d}{dt}\, Q_1^{(0)}(t) \;=\; \lambda_t + \mu_t^2\, Q_2^{(0)}(t) - \mu_t^1 \left(Q_1^{(0)}(t) \wedge n_t\right) - \beta_t \left(Q_1^{(0)}(t) - n_t\right)^+$$

and

$$\frac{d}{dt}\, Q_2^{(0)}(t) = \beta_t(1-\psi_t) \left(Q_1^{(0)}(t) - n_t\right)^+ - \mu_t^2\, Q_2^{(0)}(t)\,.$$

Can be solved numerically (forward Euler) in a spreadsheet.

Diffusion Moments for the Multiserver Queue with Abandonment and Retrials

Let $E_1(t) = E\left[Q_1^{(1)}(t)\right]$, $E_2(t) = E\left[Q_2^{(1)}(t)\right]$.

Assume the set $\left\{t \,\middle|\, Q_1^{(0)}(t) = n_t\right\}$ has Lebesque measure zero.

Then

$$\begin{aligned}\frac{d}{dt}E_1(t) &= -\left(\mu_t^1 \mathbf{1}_{\{Q_1^{(0)}(t)\le n_t\}} + \beta_t \mathbf{1}_{\{Q_1^{(0)}(t)>n_t\}}\right) E_1(t) \\ &\quad + \mu_t^2 E_2(t)\end{aligned}$$

and

$$\frac{d}{dt}E_2(t) = \beta_t(1-\psi_t)E_1(t)\mathbf{1}_{\{Q_1^{(0)}(t)\ge n_t\}} - \mu_t^2 E_2(t).$$

More Diffusion Moments

(A Grand Total of 7 O.D.E.'s)

Let $V_1(t) = \text{Var}\left[Q_1^{(1)}(t)\right]$, $V_2(t) = \text{Var}\left[Q_2^{(1)}(t)\right]$,

and $C(t) = \text{Cov}\left[Q_1^{(1)}(t), Q_1^{(1)}(t)\right]$. Then

$$\begin{aligned}
\frac{d}{dt} V_1(t) &= -2\left(\beta_t \mathbf{1}_{\{Q_1^{(0)}(t) > n_t\}} + \mu_t^1 \mathbf{1}_{\{Q_1^{(0)}(t) \le n_t\}}\right) V_1(t) \\
&\quad + \lambda_t + \beta_t \left(Q_1^{(0)}(t) - n_t\right)^+ + \mu_t^1 \left(Q_1^{(0)}(t) \wedge n_t\right) \\
&\quad + \mu_t^2 Q_2^{(0)}(t), \\
\frac{d}{dt} V_2(t) &= -2\mu_t^2 V_2(t) + \beta_t(1-\psi_t)\left(Q_1^{(0)}(t) - n_t\right)^+ \\
&\quad + \mu_t^2 Q_2^{(0)}(t) + 2\beta_t(1-\psi_t) C(t) \mathbf{1}_{\{Q_1^{(0)}(t) \ge n_t\}},
\end{aligned}$$

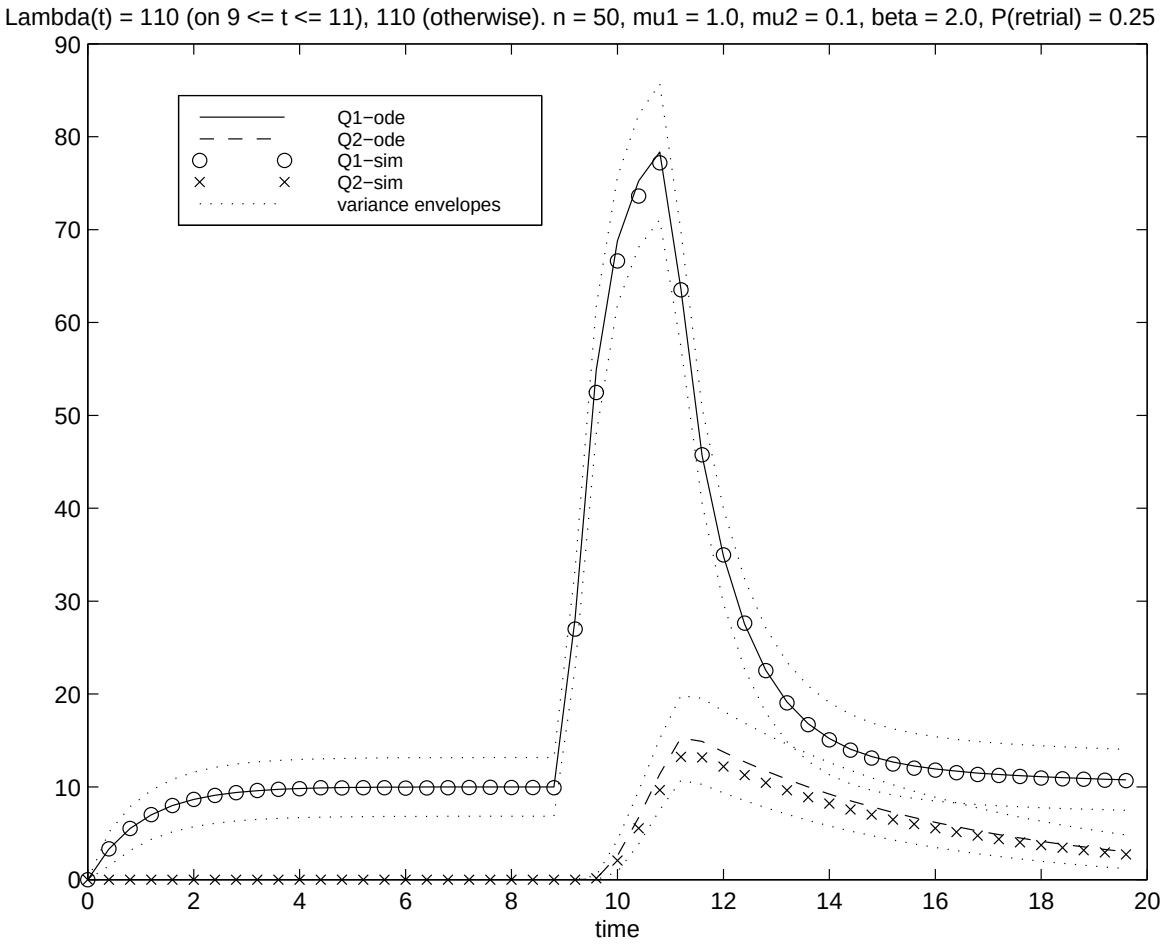
and

$$\begin{aligned}
\frac{d}{dt} C(t) &= -\left(\beta_t \mathbf{1}_{\{Q_1^{(0)}(t) \ge n_t\}} + \mu_t^1 \mathbf{1}_{\{Q_1^{(0)}(t) < n_t\}}\right) C(t) \\
&\quad + \mu_t^2 (V_2(t) - C(t)) - \beta_t(1-\psi_t)\left(Q_1^{(0)}(t) - n_t\right) \\
&\quad - \mu_t^2 Q_2^{(0)}(t)\,.
\end{aligned}$$

Example: Spiked Arrival Rate:

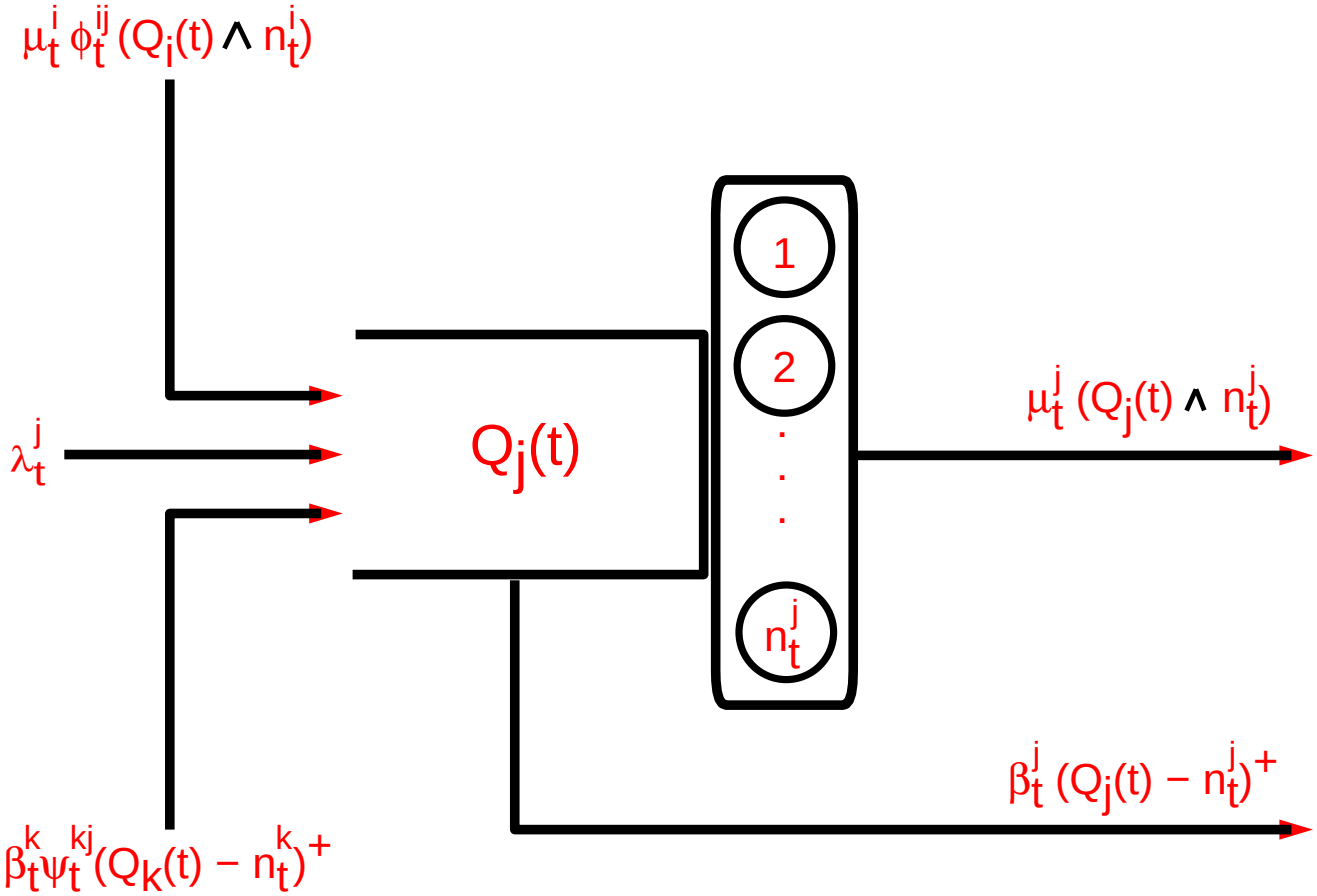
$\lambda(t) = 110$, if $9 \leq t \leq 11$ otherwise $\lambda(t) = 10$,

$\mu_1 = 1.0, \mu_2 = 0.1, \beta = 2.0, n = 50, \psi = 0.25$



Theory Generalizes to

Jackson Networks with Abandonment



Further generalizations: **Pre-Emptive Priorities**

Bottleneck Analysis

Inventory Build-up Diagrams, based on *National Cranberry*
(Recall EOQ,...) (Recall Burger-King) (in Reading Packet: *Fluid Models*)

A peak day:
- 18,000 bbl's (barrels of 100 lbs. each)
- 70% wet harvested (requires drying)
- Trucks arrive from 7:00 a.m., over 12 hours
- Processing starts at 11:00 a.m.
- Processing bottleneck: drying, at 600 bbl's *per hour*
  (Capacity = max. sustainable processing rate)
- Bin capacity for wet: 3200 bbl's
- 75 bbl's per truck (avg.)

- Draw inventory build-up diagrams of berries, arriving to RP1.

- Identify berries in bins; where are the rest? analyze it!

Q: Average wait of a truck?

- Process (bottleneck) analysis:

What if buy more bins? buy an additional dryer?

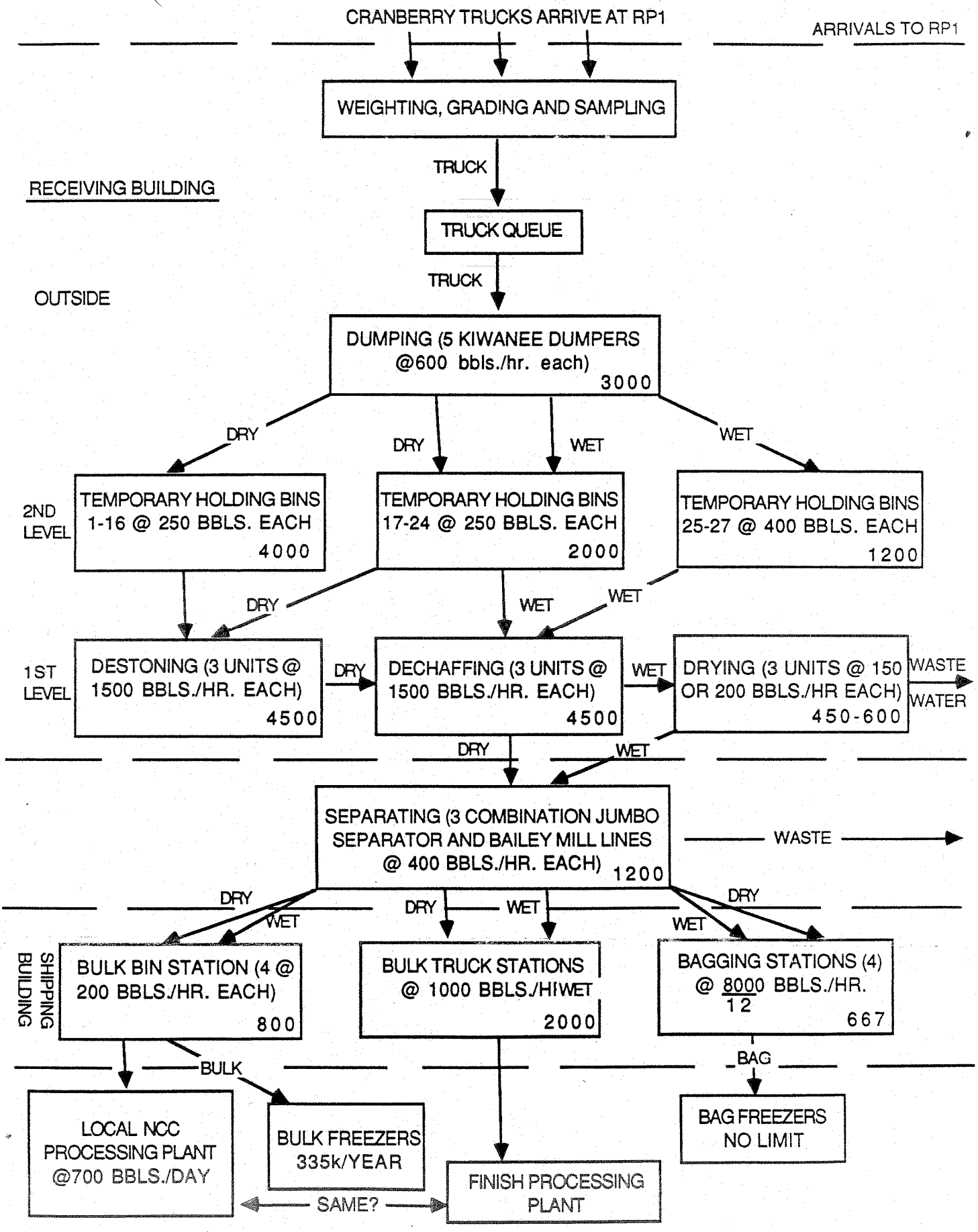
What if start processing at 7:00 a.m.?

Service analogy:

- front-office + back-office (banks, telephones)
  ↑ service ↑ production
- hospitals (operating rooms, recovery rooms)
- ports (inventory in ships; bottlenecks = unloading crews,router)
- More ?

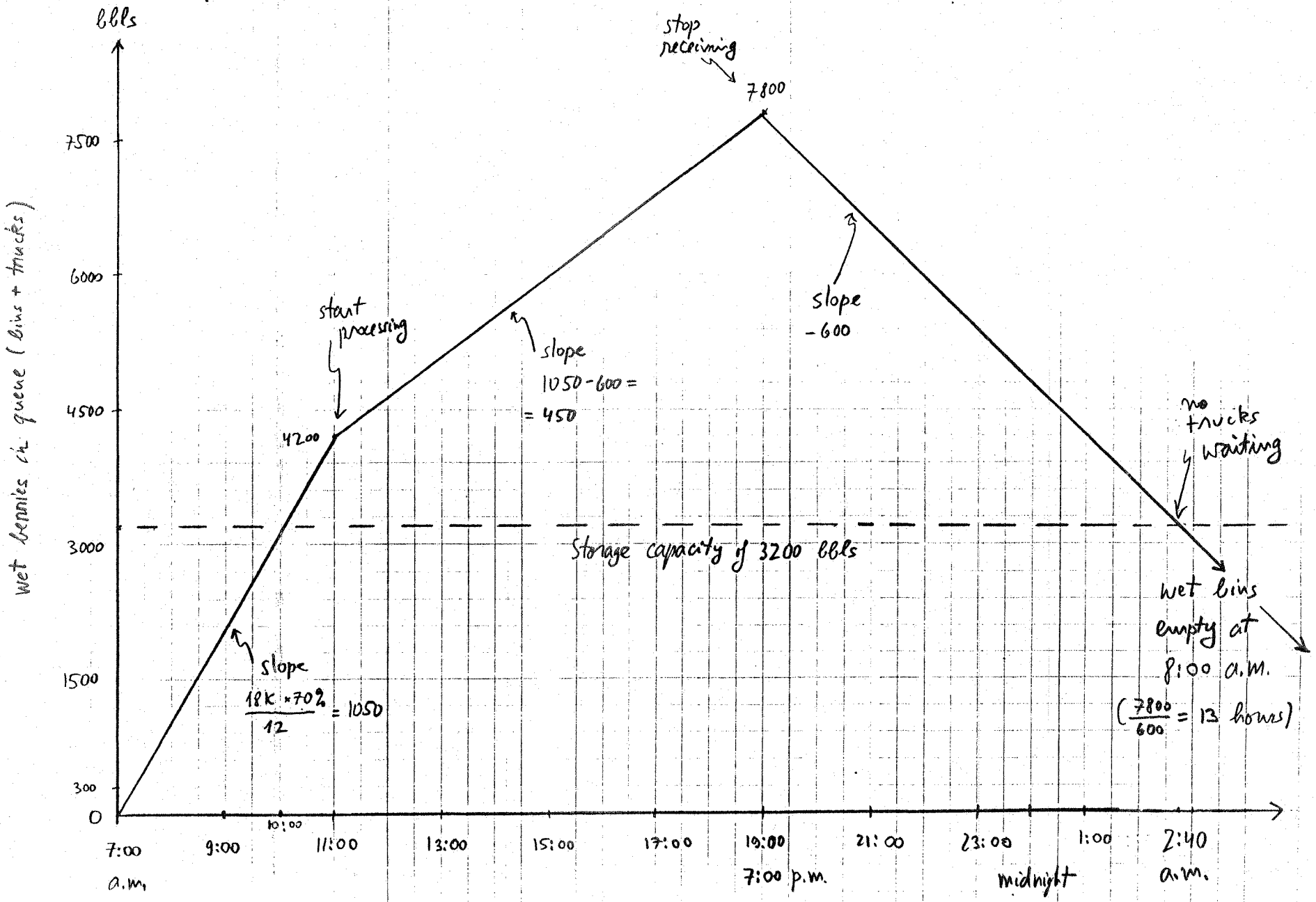
(5/13 mtg)

PROCESS FLOW DIAGRAM FOR PROCESS FRUIT AT NATIONAL CRANBERRY COOPERATIVE RP1

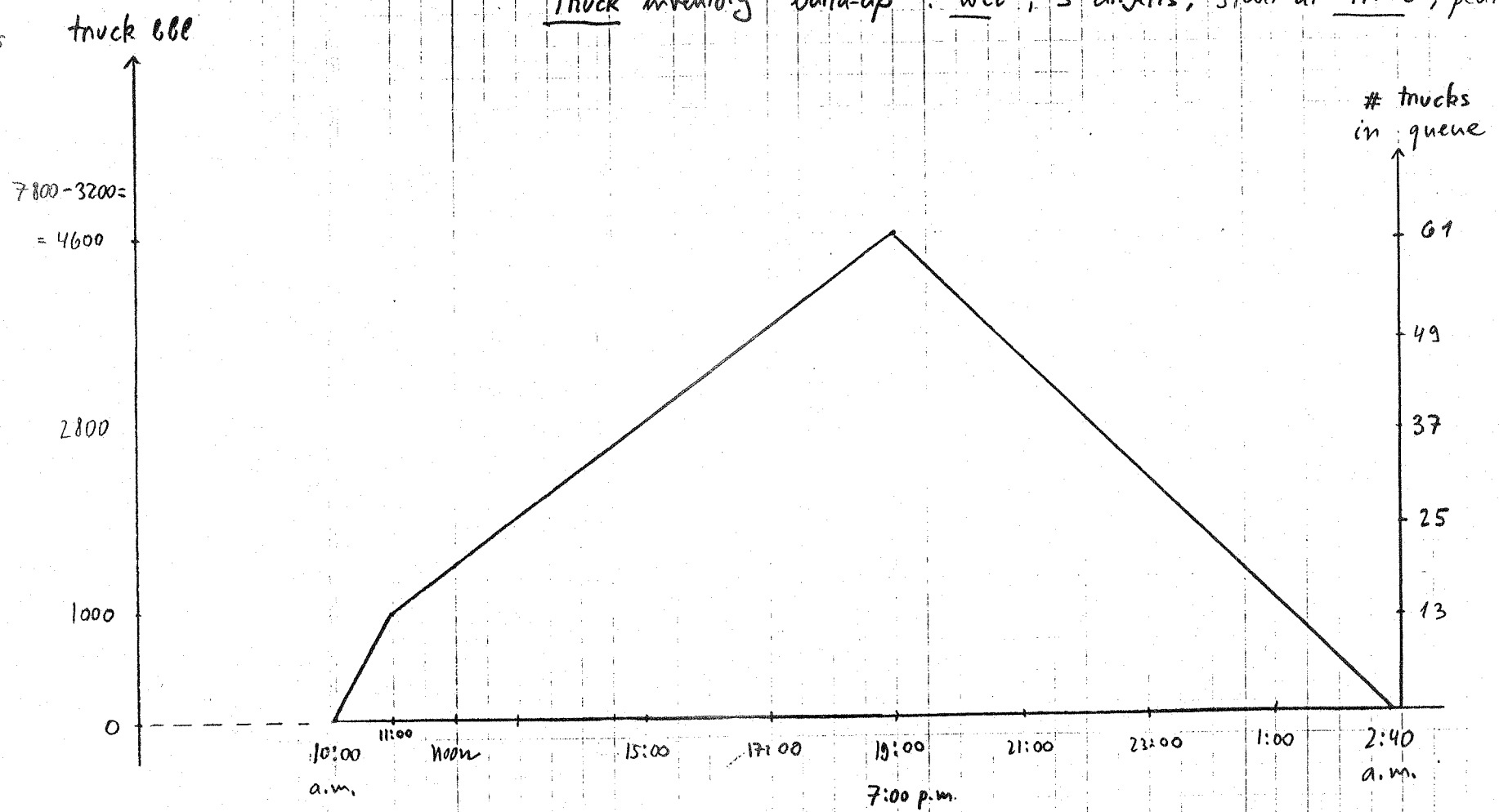


Wet
3 dryers
11:00
Total

Total inventory build-up: Wet Berries, 600 bbl/hr processing capacity,
start at 11:00, peak day 18K × 70% over 12 hours.
(bins + trucks)



3 dryers
11:00
trucks



Truck queueing analysis:

area under curve = $\frac{1}{2} \cdot 1 \cdot 1000 + \frac{1}{2} \cdot [1000 + 4600] \cdot 8 + \frac{1}{2} \cdot 4600 \cdot 7\frac{2}{3} = 40,533$ bbl·hours ; divide by 75

truck hours waiting = $40,533 \div 75$ bbl/truck = 540 truck·hours

ave. throughput rate = $[0.1 + 600 \cdot 15\frac{2}{3}] \div [16\frac{2}{3} \cdot 75] = 7.52$ trucks/hr.

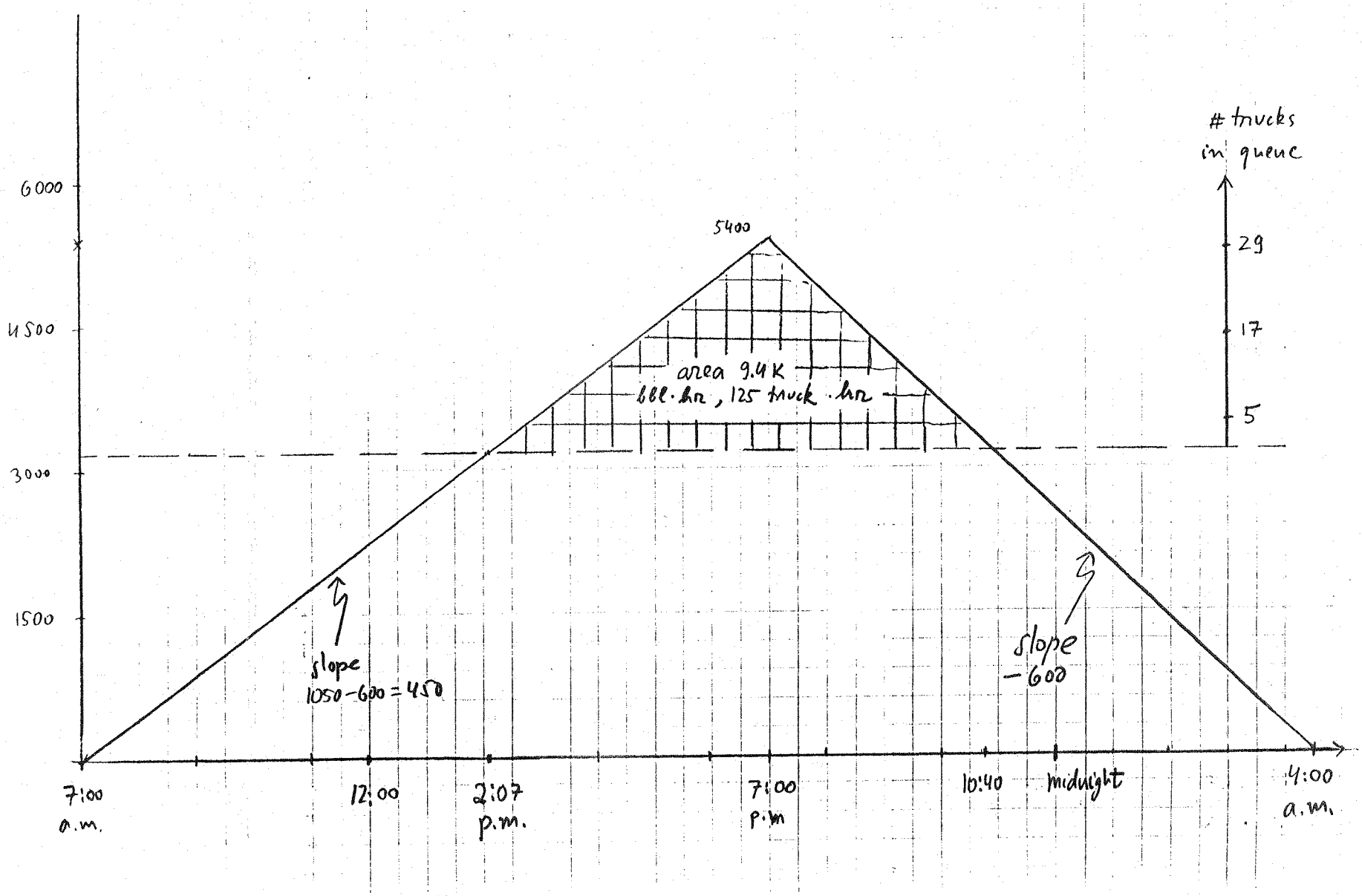
ave. WIP = $540 \div 16\frac{2}{3} = 32.4$ trucks (a "biased" average)

Given that a truck waits, it will wait on the average $32.4 \div 7.52 = 4.3$ hours. (Little)

89

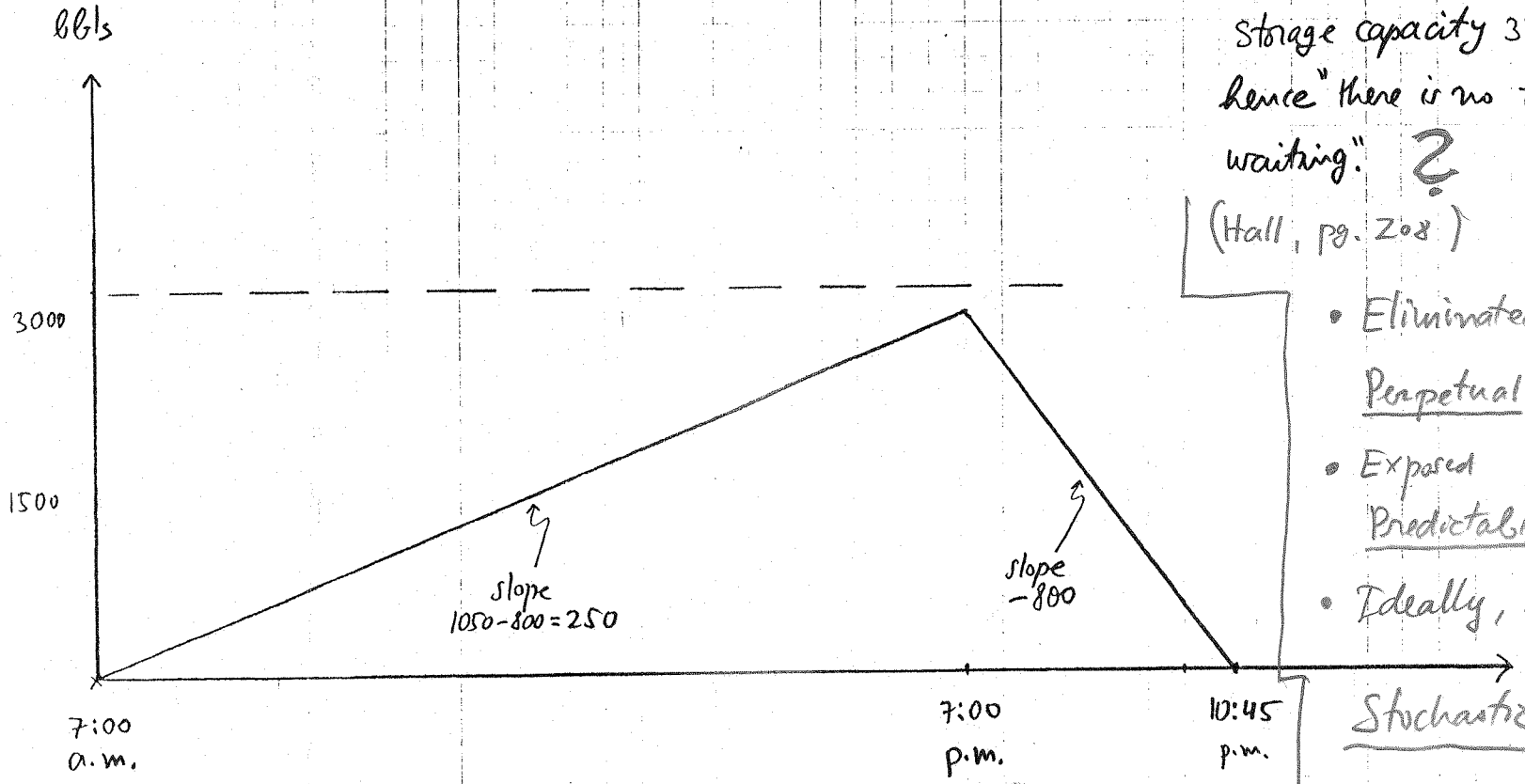
wet
3 dryers
7:00
total

Total inventory build-up: Wet Berries, 600 bbl/hr processing capacity, start at 7:00 a.m., peak day 18K × 70% over 12 hours.



Wet
4 dryers
7:00
total = bins

Total inventory build-up: Wet Berries, 800 bbl/hr processing capacity, (i.e. add 4-th dryer), start at 7:00, peak day 18K×702 over 12 hours.



Storage capacity 3200 bbls, hence "there is no truck waiting." ?

(Hall, pg. 208)

- Eliminated Perpetual Queues
- Exposed Predictable Queues
- Ideally, have only Stochastic Queues

OK as

Types of Queues

- **Perpetual Queues**: every customers waits.
  - **Examples**: public services (courts), field-services, operating rooms, ...
  - **How** to cope: reduce arrival (rates), increase service capacity, reservations (if feasible), ...
  - **Models**: fluid models.
- **Predictable Queues**: arrival rate exceeds service capacity during predictable time-periods.
  - **Examples**: Traffic jams, restaurants during peak hours, accountants at year's end, popular concerts, airports (security checks, check-in, customs) ...
  - **How** to cope: capacity (staffing) allocation, overlapping shifts during peak hours, flexible working hours, ...
  - **Models**: fluid models, stochastic models.
- **Stochastic Queues**: number-arrivals exceeds servers' capacity during stochastic (random) periods.
  - **Examples**: supermarkets, telephone services, bank-branches, emergency-departments, ...
  - **How** to cope: dynamic staffing, information (e.g. reallocate servers), standardization (reducing std.: in arrivals, via reservations; in services, via TQM) ,...
  - **Models**: stochastic queueing models.

Last page in Operations Scheduling and Control 2
Chapter 15, Scheduling handout

Unbalanced Plant

This term refers to the amount of work at each work center in a job shop. It is impossible to have a "perfectly balanced" job shop running at full capacity where the output of one work center feeds to the next one just at the time when it receives a new unit from upstream. This is because of the statistical distribution in performance times—one workstation completing a job early may have to wait for its next unit in order to start working. Thus, the workstation has idle time at that point. On the other hand, the work center may take more than the average time and delay the next workstation. The result of this "unbalance" is that jobs accumulate in various locations and are not evenly distributed throughout the system.

The Ten Commandments of Scheduling

OPT has 10 rules that are excellent for any job shop. These are shown in Exhibit S15.2.

Bottleneck Operations

A bottleneck is that operation which limits output in the production sequence. No matter how fast the other operations are, system output can be no faster than the bottleneck. Bottlenecks can occur because of equipment limitations or a shortage of material, personnel, or facilities.

Ways to Increase Output at the Bottleneck

Once a bottleneck is identified, production can be increased by a variety of possible actions:

1. Adding more of whatever resource is limited there: personnel, machines, etc.
2. Using alternate equipment or routing. For example, some of the work can be routed to other—though perhaps more costly and lesser quality—equipment.
3. Reducing setup time. If the equipment is already operating at maximum capacity, then some savings may be realized by adding jigs, handling equipment, redesign of tooling, etc. in order to speed up changeovers.
4. Running larger lot sizes. Total time at a work center consists of different kinds of time: processing time, maintenance time, setup time, and other wait time such as waiting for parts etc. Output can be increased by making fewer changeovers using larger lots and thus reducing the total amount of time spent in setups.
5. Clearing up area. Often, by doing a relayout, or removing material that may be obstructing good working conditions, output can be improved.
6. Working overtime.
7. Subcontracting.
8. Delaying the promised due date of products requiring that facility.
9. Investing in faster equipment or higher skilled personnel.

The Fluid View : Summary

- Predictable variability is dominant (Std << Mean)
- The value of the fluid-view increases with the complextity of the system from which it originates

- Legitimate models of flow systems
  - Often simple and sufficient; empirical, predictive
    - Capacity analysis
    - Inventory build-up diagrams
    - Mean-value analysis
- Approximations
  - First-order fluid approx. of stochastic systems
    - Strong Laws of Large Numbers

    (vs. Second-order diffusion approx., Central Limits)
  - Long-run
    - Long horizon, smooth-out variability (strategic)
  - Short-run
    - Short horizon, deterministic (operational)
- Technical tools
  - Lyapunov functions to establish stability (Long-run)
  - Building blocks for stochastic models (M(t)/M(t)/1)