Class 5

Modelling a Service Station (I): Empirical/Deterministic Models; Fluid/Flow Models/Approximations of Predictable Queues

- 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 required 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)
- Phases of Congestion: under-, over- and critical-loading
- Scales (Transportation, Telephone (1976, 1993, 1999))
- Simple Important Models: EOQ, Aggregate Planning
- Queues with Abandonment and Retrials (=Call Centers; Time- and State-dependent Q's).
- Bottleneck analysis in a (feed-forward) Fluid Network, via National Cranberry
- Addendum
- (Skorohod's Deterministic Fluid Model (of a service station): teaching note)

Recitation 5: Fluid models, with application to staffing.

HW 5: "Fluid Models".

Service Engineering

Class 5

Fluid/Flow Models;
Models/Apparoximations, Empirical/Deterministic

- Introduction
- Scenario Analysis: Empirical Models + Simulation.
- Transportation: Predictable Variability.
- Fluid/Empirical models of Predictable Queues.
- Four "pictures": rates, queues, outflows, cumulative graphs.
- Phases of Congestion.
- Examples: Peak load vs. peak congestion; EOQ; Aggregate Planning.
- From Data to Models; Scales.
- Queueing Science.
- A fluid model of call centers with abandonment and retrials.
- Bottleneck Analysis, via National Cranberry Cooperative.
- Summary of the Fluid Paradigm.

Keywords: Blackboard Lecture

- Classes 1-4 = Introduction to Introduction:
  On Services, Measurements, Models: Empirical, Stochastic.
  Today, our first model of a Service Stations: Fluid Models.
- Fluid Model vs. Approximation
- Model: Fluid/Flow, Deterministic/Empirical; eg. EOQ.
- Conceptualize: busy highway around a large airport at night.
- Types of queues: Perpetural, Predictable, Stochastic.
- On Variability: Predictable vs. Stochastic (Natural/Artificial).
- Scenario Analysis vs. Averaging, Steady-State.
- Descriptive Model (Inside the Black Box), via 4 "pictures": rates, queues, outflows, cummulants.
- Mathematical Model (Black Box), via differential equations.
- Resolution/Units (Scales).
- Applications:
  - Phenomena:
    Peaks (load vs. congestion); Calmness after a storm;
  - Managerial Support:
    Staffing (Recitation); Bottlenecks (Cranberries)
- Bottlenecks.

Types of Queues

- תור תמיד **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.
- תור [illegible] **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.
- תור סטוכסטי **Stochastic Queues**: number-arrivals exceeds servers' capacity during stochastic (random) periods.
  - **Examples**: supermarkets, telephone services, bank-branches, emergency-departments(.Routine, MCE)
  - **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.

Crowded airports

Landing flap

Apr 4th 2007
From The Economist print edition

Rex

A tussle over Heathrow threatens a longstanding monopoly

TO DEATH and taxes, one can now add jostling queues of frustrated travellers at Heathrow as one of life's unhappy certainties. Stephen Nelson, the chief executive of BAA, which owns the airport, does little to inspire confidence that those passing through his domain this Easter weekend will avoid the fate of the thousands stranded in tents by fog before Christmas or trapped in twisting lines by a security scare in the summer. In the *Financial Times* on April 2nd he wrote of the difficulties of managing "huge passenger demand on our creaking transport infrastructure", and gave warning that "the elements can upset the best laid plans".

Blaming the heavens for chaos that has yet to ensue may be good public relations but Mr Nelson's real worries have a more earthly origin. On March 30th two regulators released reports on his firm, one threatening to cut its profits and the other to break it up. First the Civil Aviation Authority (CAA), which oversees airport fees, said it was thinking of reducing the returns that BAA is allowed to earn from Heathrow and Gatwick airports. Separately the Office of Fair Trading (OFT) asked the Competition Commission to investigate BAA's market dominance. As well as Heathrow, Europe's main gateway on the transatlantic air route, BAA owns its two principal London competitors, Gatwick and Stansted, and several other airports.

The "Fluid View" or Flow Models of Service Networks

Service Engineering (Science, Management)

December, 2006

1 Predictable Variability in Time-Varying Services

Time-varying demand and time-varying capacity are common-place in service operations. Sometimes, *predictable* variability (eg. peak demand of about 1250 calls on Mondays between 10:00-10:30, on a regular basis) dominates stochastic variability (i.e. random fluctuations around the 1250 demand level). In such cases, it is useful to model the service system as a deterministic *fluid model*, which transportation engineers standardly practice. We shall study such fluid models, which will provide us with our first mathematical model of a service-station.

A common practice in many service operations, notably call centers and hospitals, is to time-vary staffing in response to time-varying demand. We shall be using fluid-models to help determine time-varying staffing levels that adhere to some pre-determined criterion. One such criterion is "minimize costs of staffing plus the cost of poor service-quality", as will be described in our fluid-classes.

Another criterion, which is more subtle, strives for *time-stable* performance in the face of *time-varying* demand. We shall accommodate this criterion in the future (in the context of what will be called "the square-root rule" for staffing). For now, let me just say that the analysis of this criterion helped me also understand a phenomenon that has frustrated me over many years, which I summarize as "The Right Answer for the Wrong Reasons", namely: how come so many call centers enjoy a rather acceptable and often good performance, despite the fact that their managers noticeably lack any "stochastic" understanding (in other words, they are using a "Fluid-View" of their systems).

2 Fluid/Flow Models of Service Networks

We have discussed why it is natural to view a service network as a queueing network. Prevalent models of the latter are *stochastic* (random), in that they acknowledge *uncertainty* as being a central characteristic. It turned out, however, that viewing a queueing network through a "deterministic eye", animating it as a *fluid network*, is often appropriate and useful. For example, the Fluid View often suffices for bottleneck (capacity) analysis (the "Can we do it?" step, which is the first step in analyzing a dynamic stochastic network); for motivating congestion laws (eg. Little's Law, or "Why peak congestion lags behind peak load"); and for devising (first-cut) staffing levels (which are sometime last-cut as well).

Some illuminating "Fluid" quotes:

- "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 postoffice is the shape of the input flow rate: about 70% of all letter mail enters a post office within 4-hour period". (From Oliver and Samuel, a classical 1962 OR paper).
- " . . . 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 non-stationary Poisson arrivals pattern can be accurately approximated with a deterministic model". (Hall's textbook, pages 187-8). Note: the statement is based on a Poisson model, in which mean = variance.

There is a rich body of literature on Fluid Models. It originates in many sources, it takes many forms, and it is 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, mentioned previously). 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 clear in Hall's book.)

Some main advantages of fluid-models, as I perceive them, are:

- They are simple (intuitive) to formulate, fit (empirically) and analyze (elementary). (See the Homework on Empirical Models.)
- They cover a broad spectrum of features, relatively effortlessly.
- Often, they are all that is needed, for example in analyzing capacity, bottlenecks or utilization profiles (as in National Cranberries Cooperative and HW2).
- They provide useful approximations that support both performance analysis and control. (The approximations are formalized as first-order deterministic fluid limits, via Functional (Strong) Laws of Large Numbers.)

Fluid models are intimately related to Empirical Models, which are created *directly* from measurements. As such, they constitute a natural first step in modeling a service network. Indeed, refining a fluid model of a service-station with the outcomes of Work (Time and Motion) Studies (classical Industrial Engineering), captured in terms of say histograms, gives rise to a (stochastic) model of that service station.

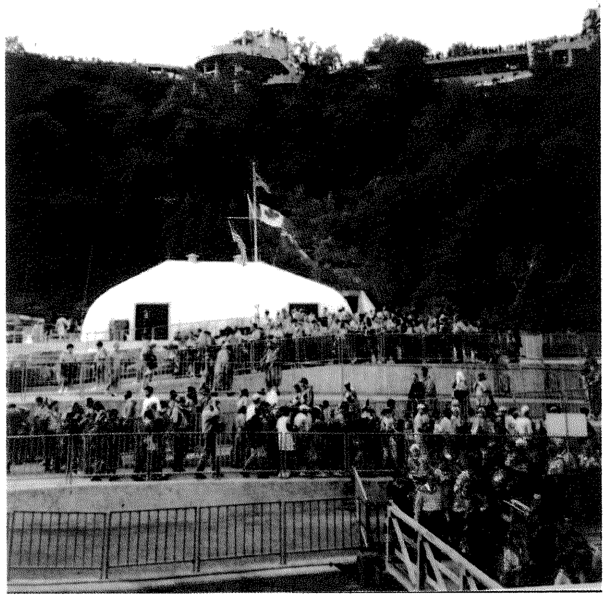
Contents

- Scenario Analysis: Empirical Models + Simulation.
- Transportation: Predictable Variability.
- Fluid/Empirical models of Predictable Queues.
- Four "pictures": rates, queues, outflows, cumulative graphs.
- Phases of Congestion.
- Examples: Peak load vs. peak congestion; EOQ; Aggregate Planning.
- From Data to Models; Scales.
- Queueing Science.
- A fluid model of call centers with abandonment and retrials.
- Bottleneck Analysis, via National Cranberry Cooperative.
- Summary of the Fluid Paradigm.

Conceptual Fluid Model

Customers/units are modeled by **fluid (continuous) flow**.

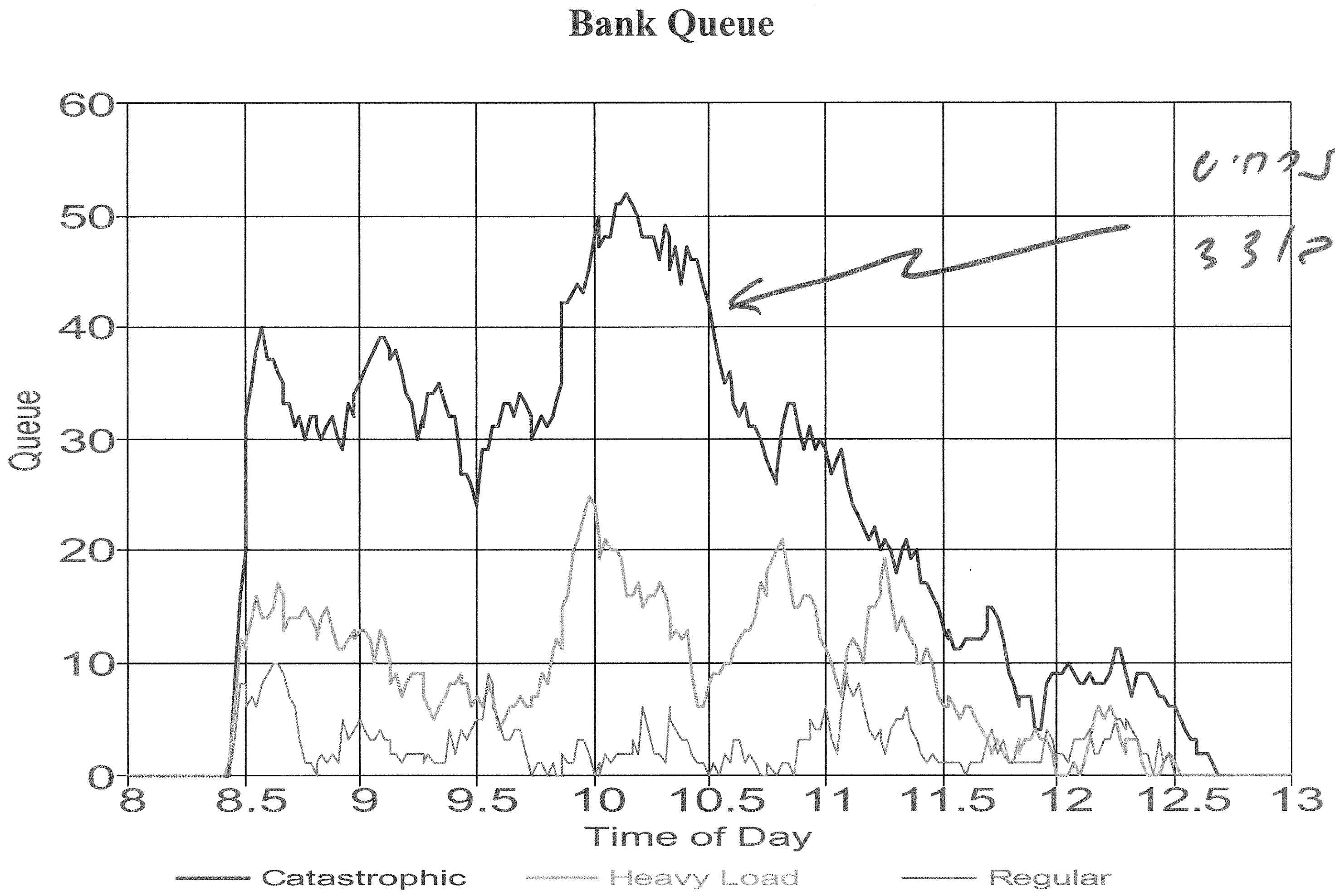
Labor-day Queueing at Niagara Falls



- Appropriate when **predictable variability** prevalent;
- Useful **first-order** models/approximations, often **suffice**;
- Rigorously justifiable via Functional Strong Laws of Large Numbers.

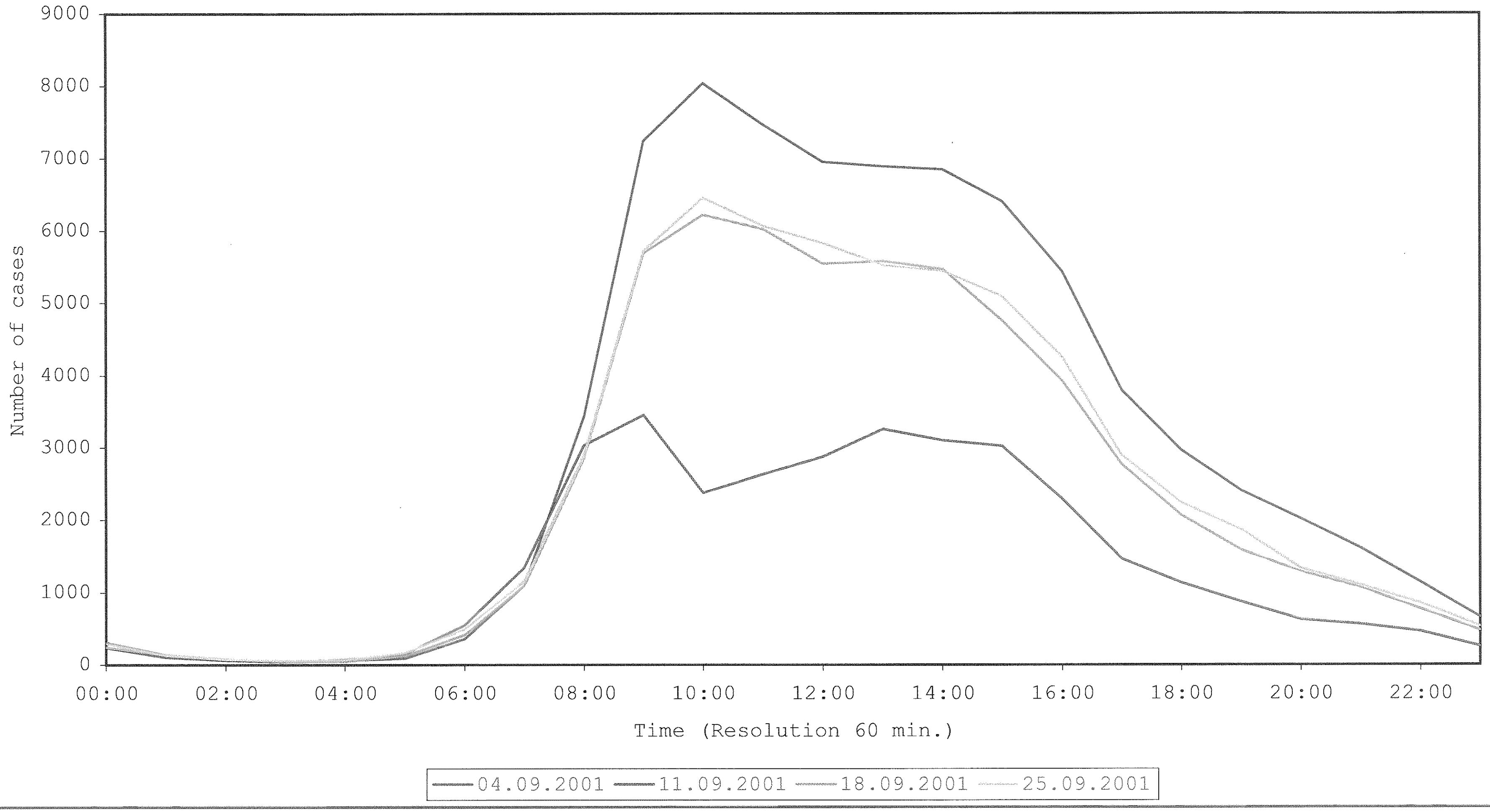
4

Empirical Fluid Model: Queue-Length at a Bank Queue Catastrophic/Heavy/Regular Day (s)

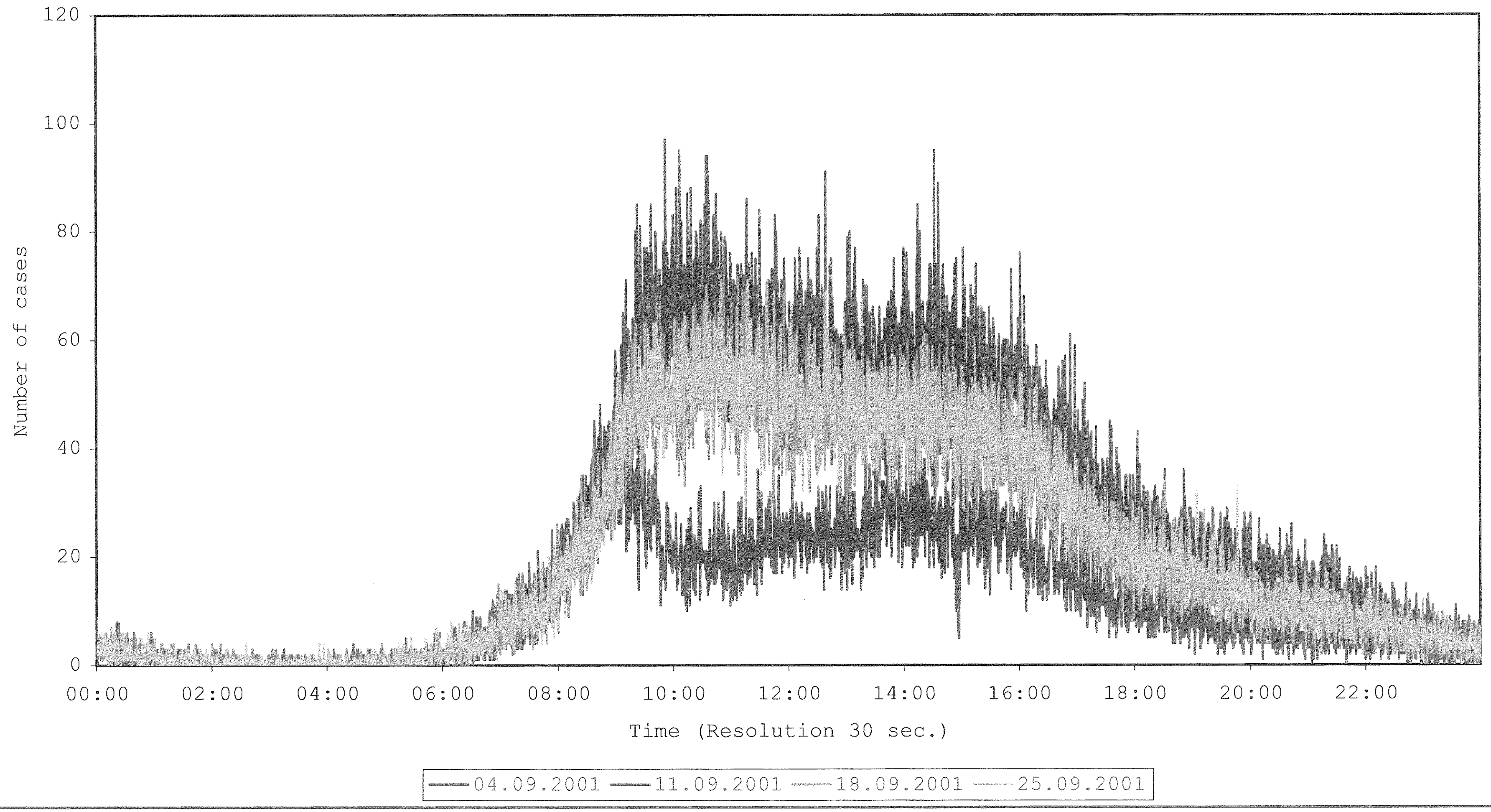


לעומת ממוצע של תרחישים דומים

Arrivals to queue
September 2001



Arrivals to queue
September 2001



⇒ Predictable

Forecasting =

(next class : arrivals)

Predicting Emergency Department Status

Houyuan Jiang‡, Lam Phuong Lam†, Bowie Owens†, David Sier† and Mark Westcott†

† CSIRO Mathematical and Information Sciences, Private Bag 10,
South Clayton MDC, Victoria 3169, Australia

‡ The Judge Institute of Management, University of Cambridge,
Trumpington Street, Cambridge CB2 1AG, UK

Abstract

Many acute hospitals in Australia experience frequent episodes of ambulance bypass. An important part of managing bypass is the ability to determine the likelihood of it occurring in the near future.

We describe the implementation of a computer program designed to forecast the likelihood of bypass. The forecasting system is designed to be used in an Emergency Department. In such an operational environment, the focus of the clinicians is on treating patients, there is no time carry out any analysis of the historical data to be used for forecasting, or to determine and apply an appropriate smoothing method.

The method is designed to automate the short term prediction of patient arrivals. It uses a multi-stage data aggregation scheme to deal with problems that may arise from limited arrival observations, an analysis phase to determine the existence of trends and seasonality, and an optimisation phase to determine the most appropriate smoothing method and the optimal parameters for this method.

The arrival forecasts for future time periods are used in conjunction with a simple demand modelling method based on a revised stationary independent period by period approximation queueing algorithm to determine the staff levels needed to service the likely arrivals and then determines a probability of bypass based on a comparison of required and available resources.

1 Introduction

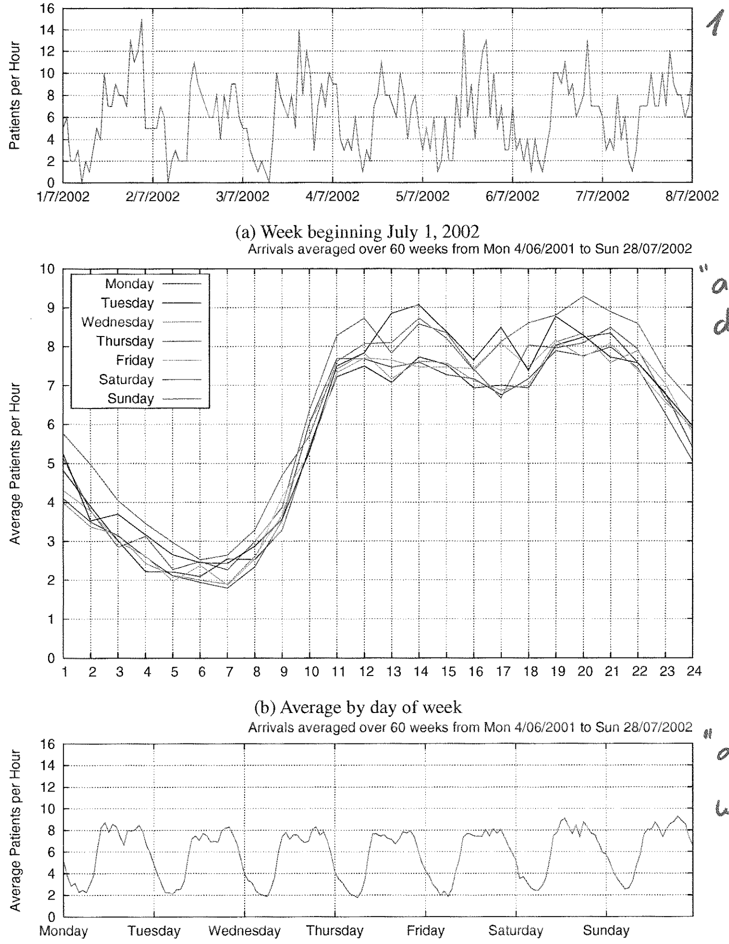
This paper describes a system designed to be part of the process for managing Emergency Department (ED) bypass. An ED is on bypass when it has to turn away ambulances, typically because all cubicles are full and there is no opportunity to move patients to other beds in the hospital, or because the clinicians on duty are fully occupied dealing with critical patients who require individual care.

Bypass management is part of the more general bed management and admission–discharge procedures in a hospital. However, a very important part of determining the likelihood of bypass occurring in the near future, typically the next 1, 4 or 8 hours, is the ability to predict the probable patient arrivals, and then, given the current workload and staff levels, the probability that there will be sufficient resources to deal with these arrivals.

Here, we consider the implementation of a multi-stage forecasting method [1] to predict patient arrivals, and a demand management queueing method [2], to assess the likelihood of ED bypass.

The prototype computer program implementing the method has been designed to run on a hospital intranet and to extract patient arrival data from hospital patient admission and ED databases. The program incorporates a range of exponential smoothing procedures. A user can specify the particular smoothing procedure for a data set or to configure the program to automatically determine the best procedure from those available and then use that method.

For the results presented here, we configured the program to automatically find the best smoothing method since this is the way it is likely to be used in an ED where the staff are more concerned with treating patients than configuring forecast smoothing parameters.



(a) Week beginning July 1, 2002

(b) Average by day of week

(c) Average weekly

1 week

"average day"

"average week"

Figure 1: Hourly patient arrivals, June 2001 to July 2002

For the optimisation we assume no a priori knowledge of the patient arrival patterns. The process involves simply fitting each of the nine different methods listed in Table 1 to the data, using the mean square fitting error, calculated using (3), as the objective function. The smoothing parameters for each method are all in (0, 1) and the parameter solution space is defined by a set of values obtained from an appropriately fine uniform discretization of this interval. The optimal values for each method are then obtained from a search of all possible combinations of the parameter values.

Predictable Variability

השתנות בר - חזוי

From the data aggregated at a daily level, repeat the procedure to extract data for each hour of the day to form 24 time series (12am–1am, 1am–2am, . . ., 11pm–12am). Apply the selected smoothing method, or the optimisation algorithm, to each time series and generate forecasting data for those future times of day within the requested forecast horizon. The forecast data generated for each time of day are scaled uniformly in each day in order to match the forecasts generated from the previously scaled daily data.

Output: Display the historical and forecasted data for each of the sets of aggregated observations constructed during the initialisation phase.

The generalisation of these stages is straightforward. For example, if the data was aggregated to a four-weekly (monthly) level, then the first scaling step would be to extract the observations from the weekly data to form four time series, corresponding to the first, second, third and fourth week of each month. Historical data at timescales of less than one day are scaled to the daily forecasts. For example, observations at a half-hourly timescale are used to form 48 time series for scaling to the day forecasts.

4.3 Output from the multi-stage method

Figures 2 and 3 show some of the results obtained from using the multi-stage forecasting method to predict ED arrivals using the 60 weeks of patient arrival data described in Section3. The forecasted data were generated from an optimisation that used the multi-stage forecasting method to minimise the residuals of (3) across all the smoothing methods in Table 1.

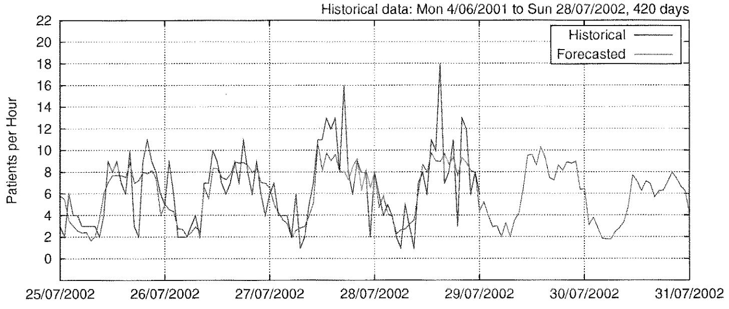


Figure 2: **Hourly** historical and forecasted data 25/7/2002–31/7/2002

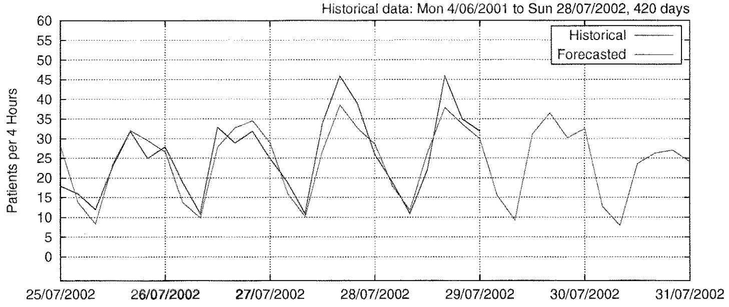
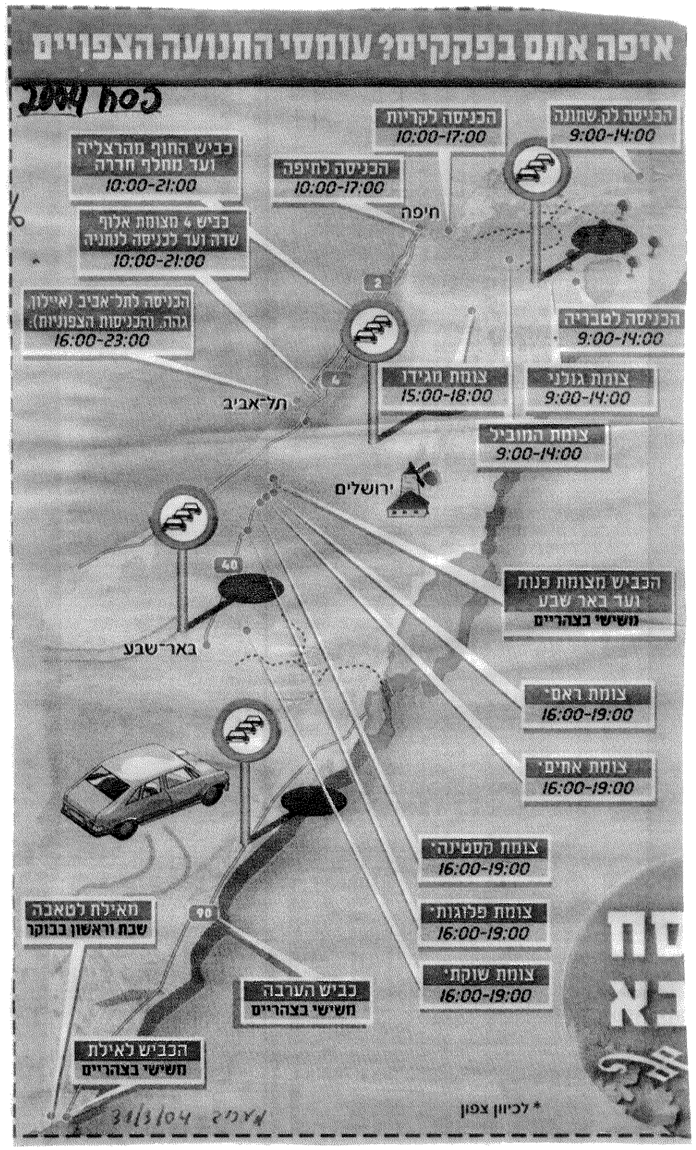


Figure 3: **Four-hourly** historical and forecasted data 25/7/2002–31/7/2002

↓ Resolution ⇒ easier (more accurate) to predict



Q-Science: Predictable Variability

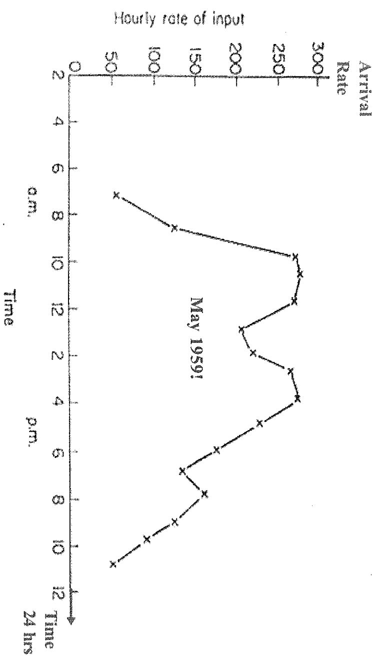
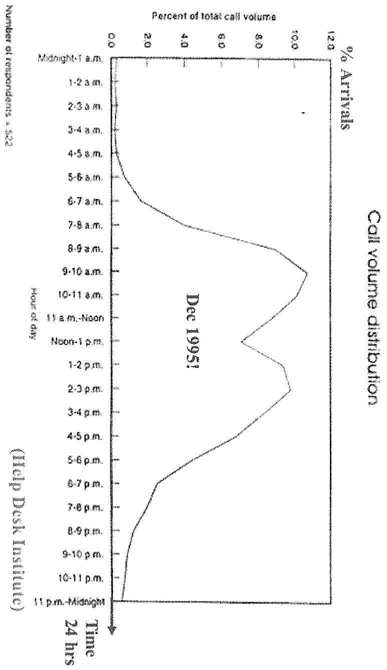


Fig. 15.1 The variation in the hourly input rates of reservations calls during a typical day (in May 1959)

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



1995 Help Desk and Customer Support Practices Report

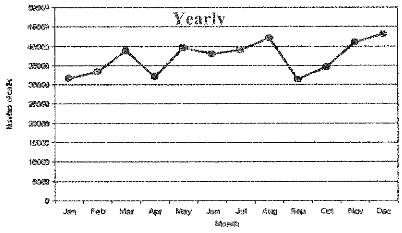
Number of respondents = 522

(Help Desk Institute)

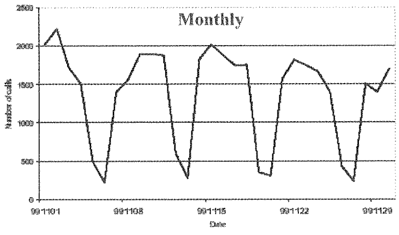
Arrivals to Service

Arrivals to a Call Center (1999): Time Scale

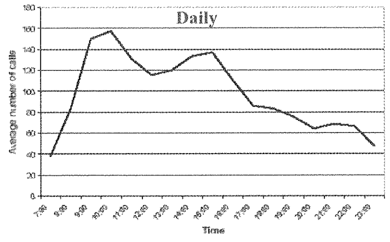
Strategic



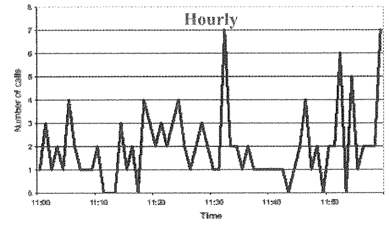
Tactical



Operational



Stochastic



Arrivals Process, in 1976

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

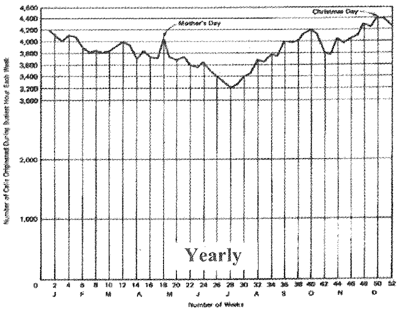


Figure 1 Typical distribution of calls during the busiest hour for each week during a year.

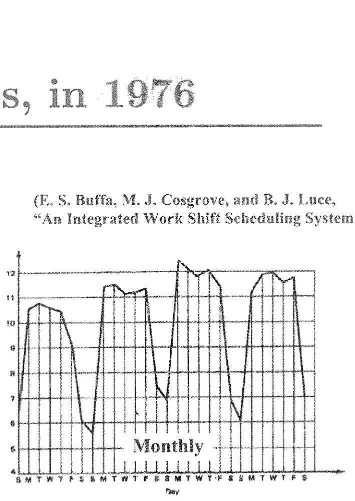


Figure 2 Daily call load for Long Beach, January 1972.

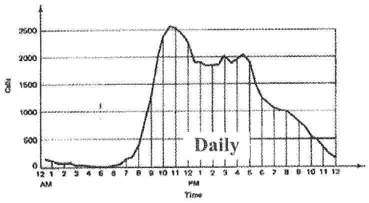


Figure 3 Typical half-hourly call distribution (Sundy D A).

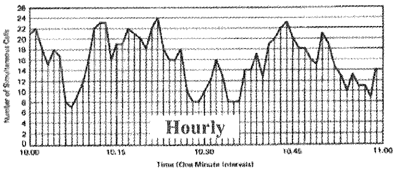
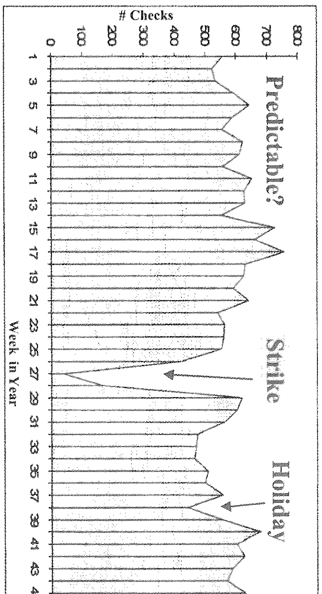


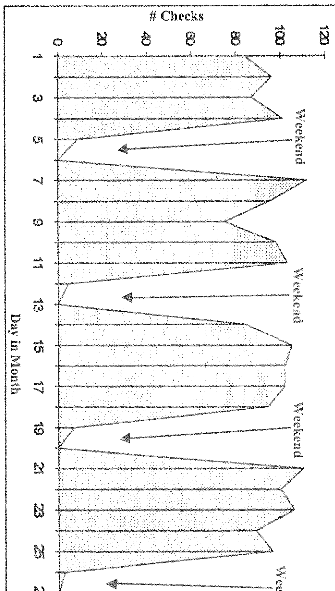
Figure 4 Typical intrahour distribution of calls, 10:00–11:00 A.M.

Custom Inspections at an Airport

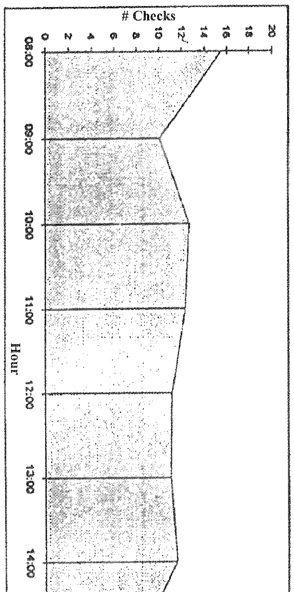
Number of Checks Made During 1993:



Number of Checks Made in November 1993:

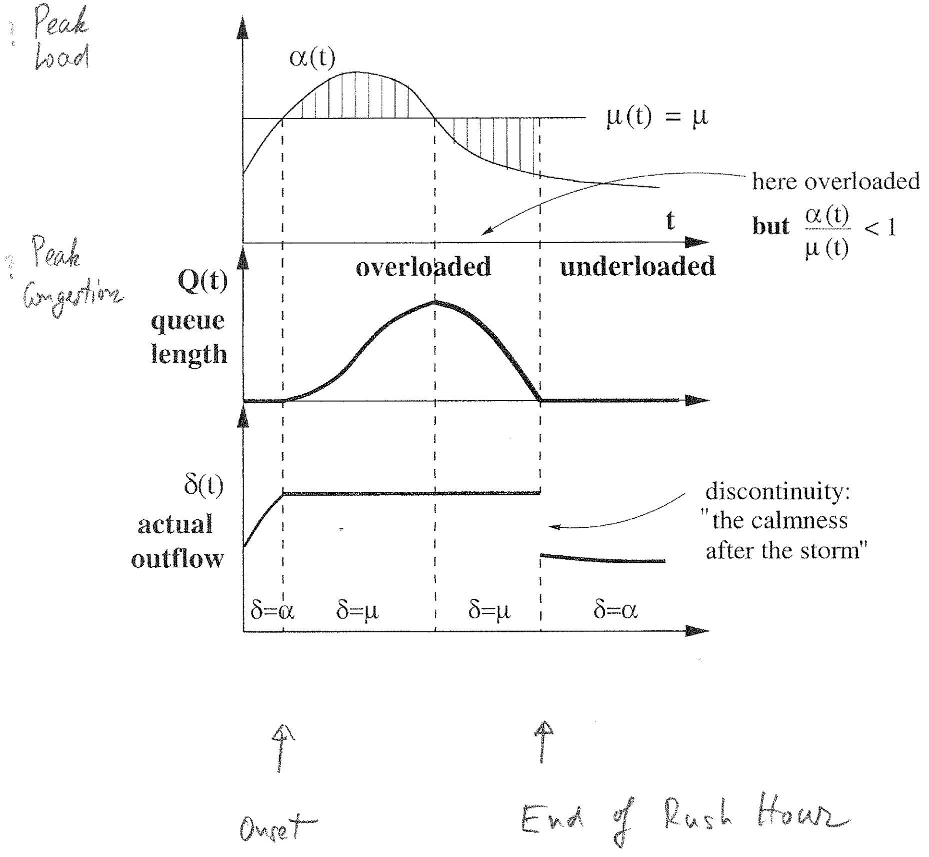


Average Number of Checks During the Day:



Source: Ben-Gurion Airport Custom Inspectors Division

Phases of Congestion (Rush Hour Analysis)



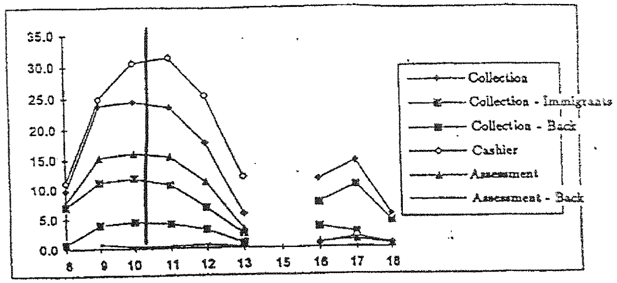
Face-to-Face Services

Phenomenon:
Peak congestion lags behind peak load

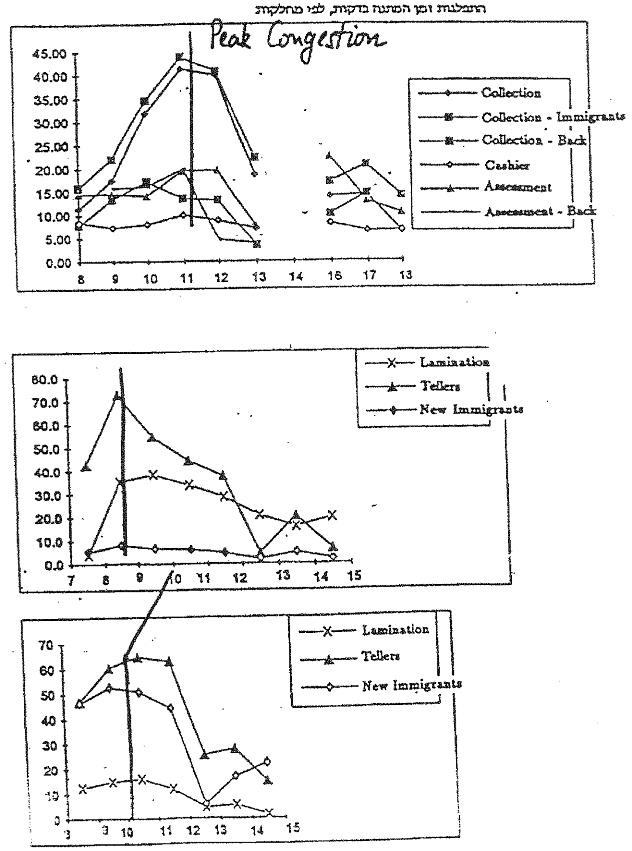
How to "explain"?
Fluid-view suffices

Peak load

Peak Congestion Lags Behind Peak Load

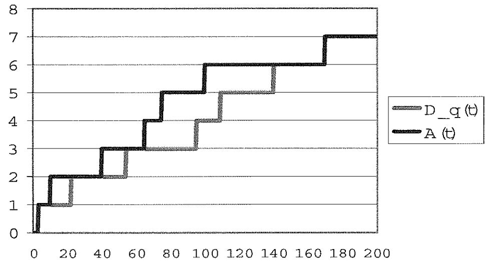


התפלגות זמן המתנה בדקות, לפי מחלקות

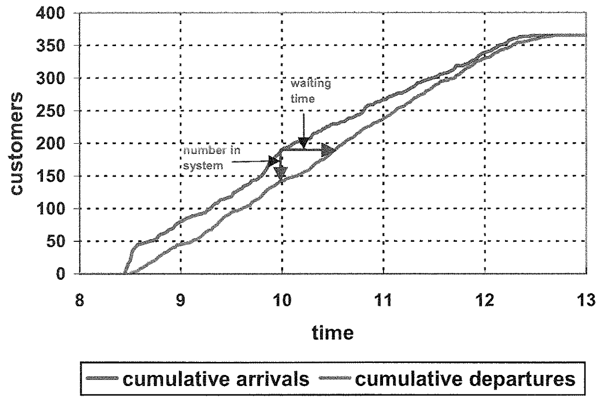


Fluid Models and Empirical Models

Recall **Empirical Models**, cumulative arrivals and departure functions.



For large systems (bird's eye) the functions look smoother.

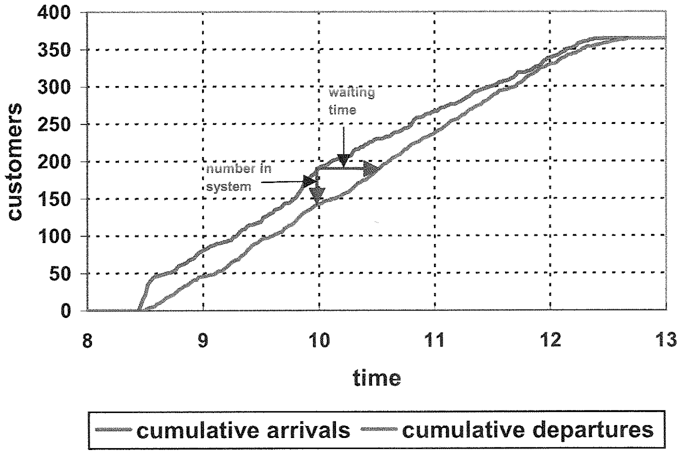


Empirical Models: Fluid, Flow

Derived directly from event-based (call-by-call) measurements. For example, an isolated service-station:

- $A(t)$ = **cumulative** # arrivals from time 0 to time t;
- $D(t)$ = **cumulative** # departures from system during $[0, t]$;
- $L(t) = A(T) - D(t) = \#$ customers in system at t.

Arrivals and Departures from a Bank Branch Face-to-Face Service



When is it possible to calculate waiting time in this way?

Phases of Congestion : via Cumulants

Hall

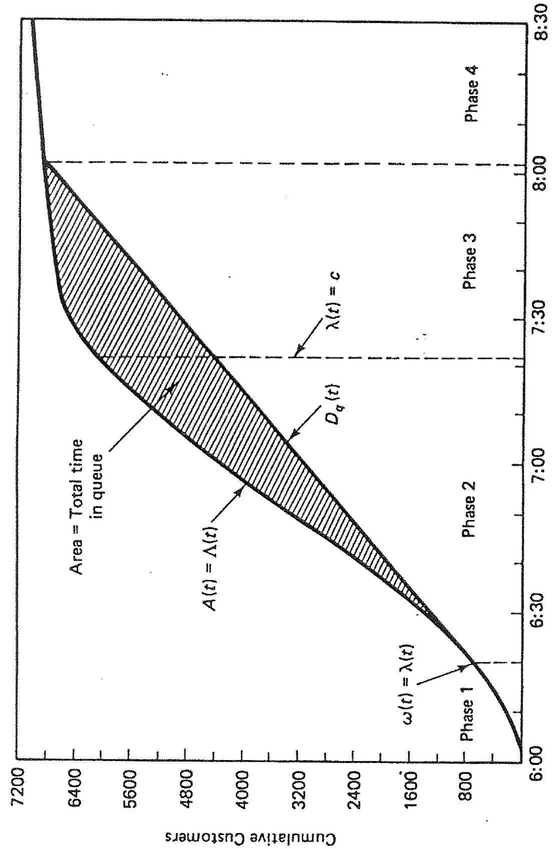


Figure 6.6 Cumulative diagram illustrating deterministic fluid model. When a queue exists, customers depart at a constant rate. Queues increase when the arrival rate exceeds the service capacity and decrease when the service capacity exceeds the arrival rate.

Hall, pg. 189-90: 1. stagnant

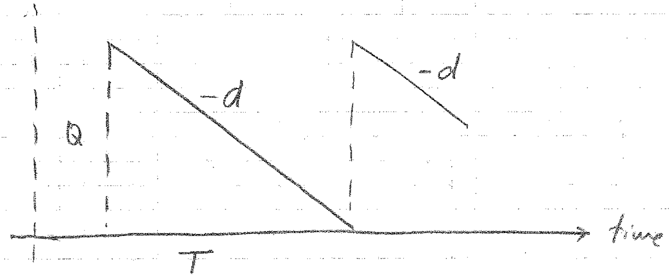
2. growth

3. decline

4. stagnant, etc.

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 $\Rightarrow Q \uparrow$)

average cost (over cycle) $= \underbrace{\frac{1}{2} Q \cdot h}_{L} + \frac{C}{T} = \frac{1}{2} Qh + \frac{Cd}{Q}$

Optimal Q^* when derivative $= 0$: $\frac{1}{2} h = \frac{Cd}{Q^2}$ ($\Rightarrow \frac{1}{2} Qh = \frac{Cd}{Q}$)

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

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

Extension: finite production rate ⋀⋀ , Q = batch size

Fluid Models: General Setup

- $\boldsymbol{A(t)}$ – **cumulative arrivals function**.
- $D(t)$ – cumulative departures function.
- $\boldsymbol{\lambda(t) = \dot{A}(t)}$ – **arrival rate**.
- $\delta(t) = \dot{D}(t)$ – processing (departure) rate.
- $\boldsymbol{c(t)}$ – **maximal potential processing rate**.
- $Q(t)$ – total amount in the system.

Queueing System as a Tub (Hall, p.188)

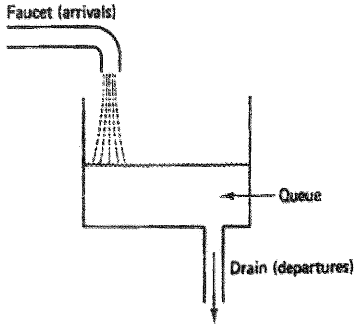


Figure 6.5 In a fluid model, the customers can be viewed as a liquid that accumulates in a tub. Queues increase when the fluid enters the tub faster than it leaves.

Mathematical Fluid Models

Differential Equations:

- $\boldsymbol{\lambda(t)}$ – **arrival rate** at time $t \in [0,T]$.
- $\boldsymbol{c(t)}$ – **maximal potential processing rate**.
- $\delta(t)$ – **effective** processing (departure) rate.
- $Q(t)$ – **total** amount in the system.

Then $Q(t)$ is <u>**a**</u> solution of

$$\dot{Q}(t) = \lambda(t) - \delta(t); \quad Q(0) = q_0, \quad t \in [0,T] .$$

In a Call Center Setting (no abandonment)

$\boldsymbol{N(t)}$ statistically-identical servers, each with service rate $\boldsymbol{\mu}$.
$\boldsymbol{c(t) = \mu N(t)}$: maximal potential processing rate.
$\delta(t) = \mu \cdot \min(N(t), Q(t))$: processing rate.

$$\dot{Q}(t) = \lambda(t) - \mu \cdot \min(N(t), Q(t)), \quad Q(0) = q_0, \quad t \in [0,T] .$$

How to actually solve? Mathematics (theory, numerical), or simply: Start with $t_0 = 0$, $Q(t_0) = q_0$.
Then, for $t_n = t_{n-1} + \Delta t$:

$$Q(t_n) \;=\; Q(t_{n-1}) + \lambda(t_{n-1}) \cdot \Delta t - \mu \min(N(t_{n-1}), Q(t_{n-1})) \cdot \Delta t .$$

Predictable Queues

Fluid Models and Diffusion Approximations

for Time-Varying Queues with Abandonment and Retrials

with

Bill Massey

Marty Reiman

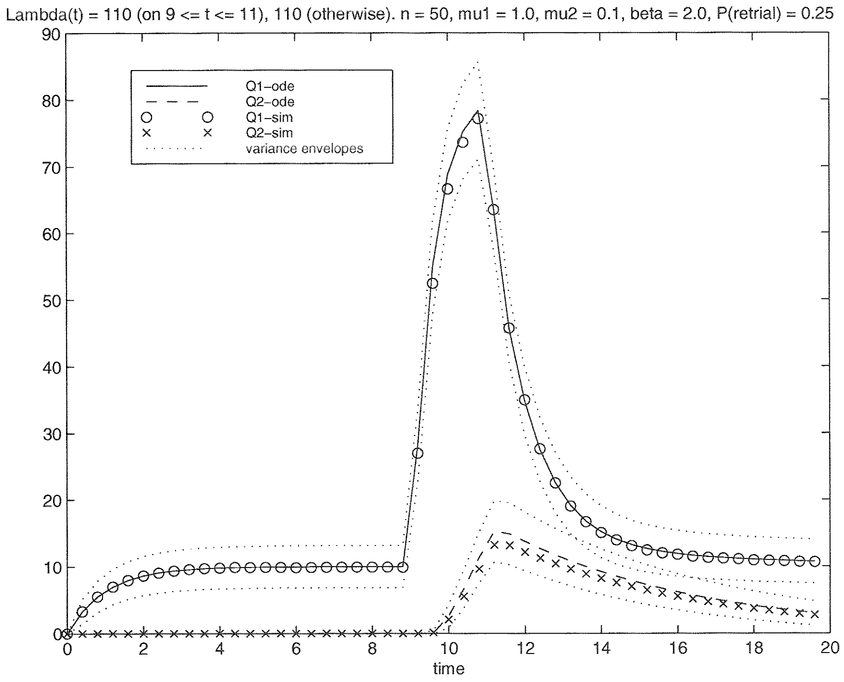
Brian Rider

Sasha Stolyar

Sudden Rush Hour

$$n \;=\; 50 \text{ servers}; \qquad \mu = 1$$

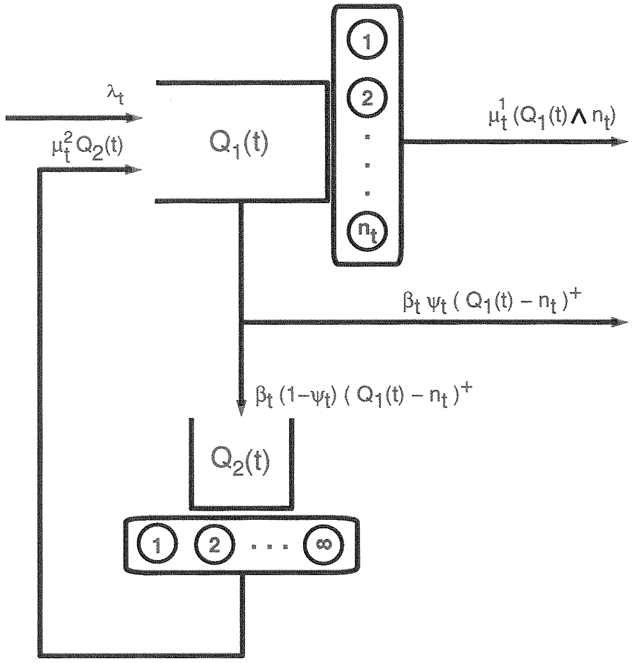
$$\lambda_t \;=\; 110 \text{ for } 9 \le t \le 11, \quad \lambda_t = 10 \text{ otherwise}$$



Time-Varying Queues with Abandonment and Retrials

Based on a series of papers with Massey, Reiman, Rider and Stolyar (all at Bell Labs, at the time).

Call Center: a Multiserver Queue with Abandonment and Retrials



λ_t
$\mu_t^2 Q_2(t)$
$Q_1(t)$
1
2
n_t
$\mu_t^1 (Q_1(t) \wedge n_t)$
$\beta_t \psi_t (Q_1(t) - n_t)^+$
$\beta_t (1-\psi_t) (Q_1(t) - n_t)^+$
$Q_2(t)$
1
2
∞

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.

$Q_1(t)$ number of customers within call center (queue+service).

β_t abandonment rate while waiting;
e.g., in response to IVR discouragement at predictable overloading.

ψ_t probability of no retrial.

μ_t^2 retrial rate;
if constant, $1/\mu^2$ – average time to retry.

$Q_2(t)$ number of customers that will retry (in orbit).

In our examples, we vary λ_t only, while other primitives are held constant.

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

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

Justification: **Functional Central Limit Theorem**

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

given convergence at $t = 0$.

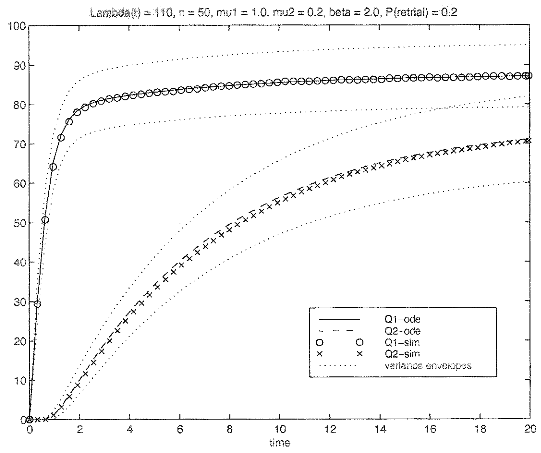
$Q^{(1)}$ solution to stochastic differential equation.

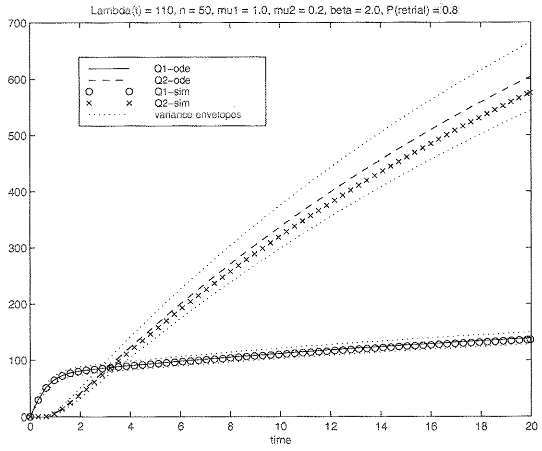
If the set of critical times $\{t \geq 0 : Q_1^{(0)}(t) = n_t\}$ has Lebesque measure zero, then $Q^{(1)}$ is a Gaussian process. In this case, one can deduce ordinary differential equations for

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

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

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

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

and

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

These equations can be easily solved to yield

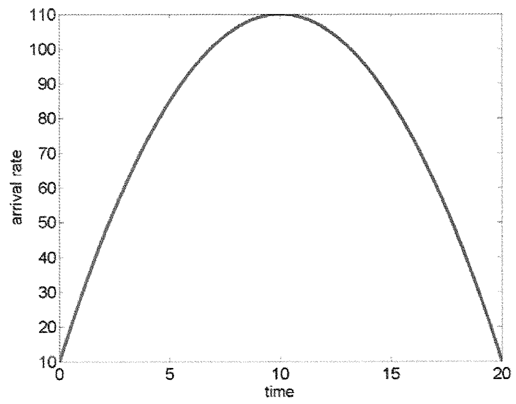
$$q_1 = n + \frac{\lambda - \mu^1 n}{\beta\psi} \quad \text{and} \quad q_2 = \frac{\beta(1-\psi)}{\mu^2}\frac{\lambda - \mu^1 n}{\beta\psi}. \qquad (14)$$

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.

Quadratic Arrival rate

Assume $\lambda(t) = 10 + 20t - t^2$.



Take P{retrial} $= 0.5$, $\beta = 0.25$ and 1.

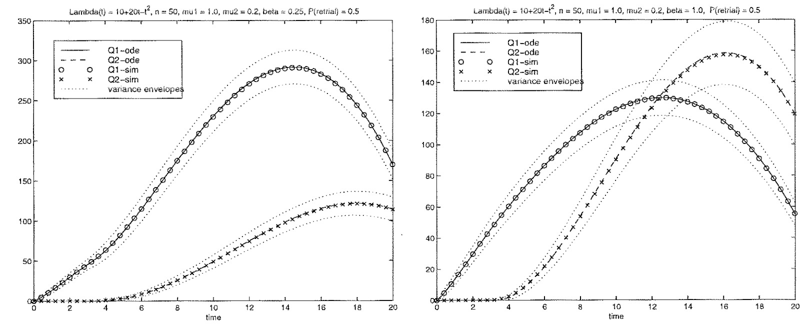


Figure 4. Numerical examples: $\beta_t = 0.25$ and 1.0.

$Q_1^{(0)}$ appears to peak roughly at the value 130 at time $t \approx 12$. Since the derivative at a local maximum is zero, then equation (1) becomes

$$\lambda_t + \mu_t^2 Q_2^{(0)}(t) \approx \mu_t^1\big(Q_1^{(0)}(t) \wedge n_t\big) + \beta_t\big(Q_1^{(0)}(t) - n_t\big)^+ \qquad (15)$$

when $t \approx 12$, as well as $Q_1^{(0)}(t) \approx Q_2^{(0)}(t) \approx 130$. The left hand side of (15) equals $106 + .2 \cdot 130 = 132$ which is roughly the value of the right hand side of (15), which is $50 + 80 = 130$.

Similarly, the graph of $Q_2^{(0)}$ appears to peak roughly at the value 155 at time $t \approx 16.5$ which also implies $Q_1^{(0)}(t) \approx 110$ and equation (2) becomes

$$\beta_t(1 - \psi_t)\big(Q_1^{(0)}(t) - n_t\big)^+ \approx \mu_t^2 Q_2^{(0)}(t). \qquad (16)$$

The left hand side of (16) is $0.5 \cdot 60 = 30$ and the right hand side of (16) is about the same or $0.2 \cdot 155 = 31$.

The reader should be convinced of the effectiveness of the fluid approximation after an examination of Figures 2 through 5. Here we compare the numerical solution (via forward Euler) of the system of ordinary differential equations for $\mathbf{Q}^{(0)}(t)$ given in (1) and (2) to a simulation of the real system. These quantities are denoted in the legends as $Q1$-ode, $Q2$-ode, $Q1$-sim, and $Q2$-sim. Throughout, the term "variance envelopes" refers to

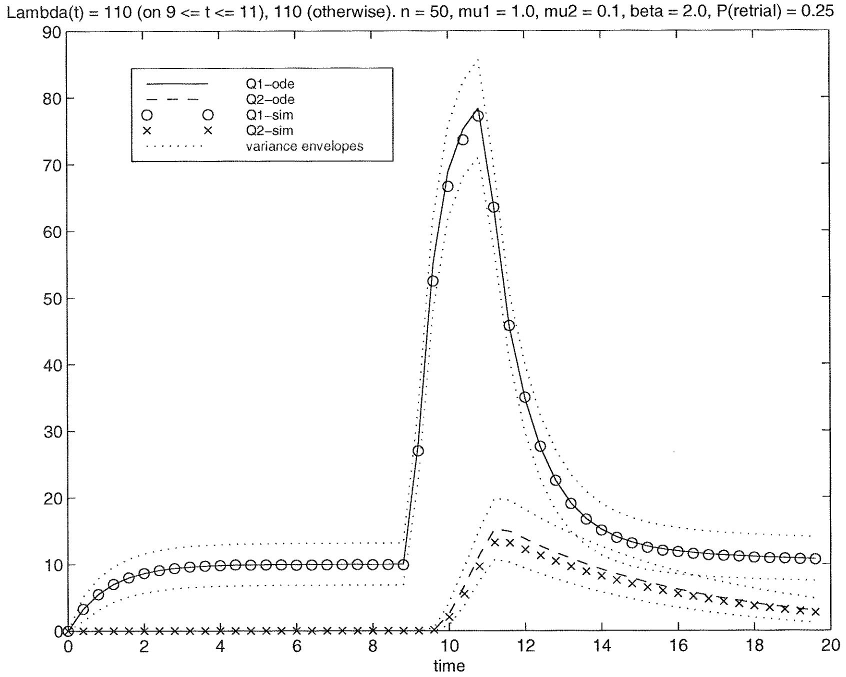
$$Q_i^{(0)}(t) \pm \sqrt{\mathsf{Var}\left[Q_i^{(1)}(t)\right]} \qquad (17)$$

for $i = 1, 2$, where $\mathsf{Var}\left[Q_1^{(1)}(t)\right]$ and $\mathsf{Var}\left[Q_2^{(1)}(t)\right]$ are the numerical solutions, again by forward Euler, of the differential equations determining the covariance matrix of the diffusion approximation $\mathbf{Q}^{(1)}$ (see Proposition 2.3). Setting $Q_1^{(1)}(0) = Q_1^{(1)}(0) = 0$ yields by

Sudden Rush Hour

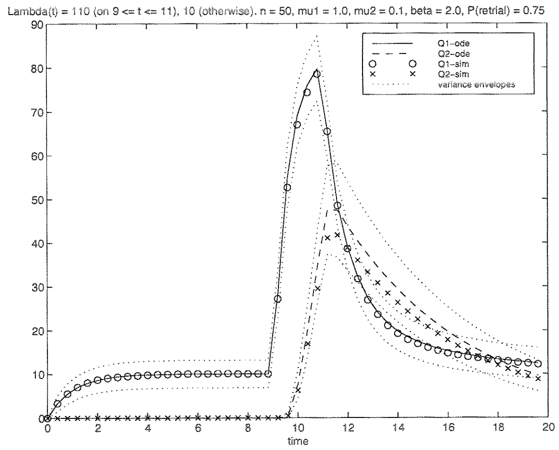
$$n = 50 \text{ servers;} \qquad \mu = 1 \ .$$

$$\lambda_t = 110 \text{ for } 9 \le t \le 11, \quad \lambda_t = 10 \text{ otherwise}$$

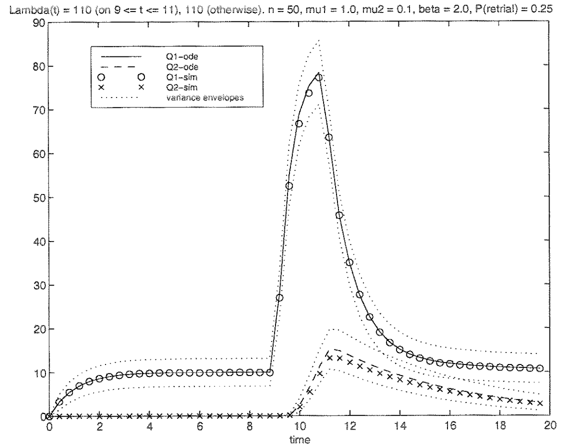


Lambda(t) = 110 (on 9 <= t <= 11), 110 (otherwise). n = 50, mu1 = 1.0, mu2 = 0.1, beta = 2.0, P(retrial) = 0.25

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



Lambda(t) = 110 (on 9 <= t <= 11), 10 (otherwise). n = 50, mu1 = 1.0, mu2 = 0.1, beta = 2.0, P(retrial) = 0.75



Lambda(t) = 110 (on 9 <= t <= 11), 110 (otherwise). n = 50, mu1 = 1.0, mu2 = 0.1, beta = 2.0, P(retrial) = 0.25

21.1

Submitted to *Manufacturing & Service Operations Management*
manuscript (Please, provide the mansucript number!)

Erlang-R: A Time-Varying Queue with Reentrant Customers, in Support of Healthcare Staffing

(Authors' names blinded for peer review)

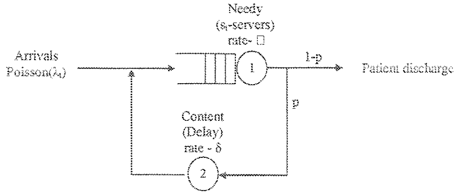
We analyze a queueing model that we call Erlang-R, where the "R" stands for Reentrant customers. Erlang-R accommodates customers who return to service several times during their sojourn within the system, and its modeling power is most pronounced in time-varying environments. Indeed, it was motivated by healthcare systems, in which offered-loads vary over time and patients often go through a repetitive service process. Erlang-R helps answer questions such as how many servers (physicians/nurses) are required in order to achieve predetermined service levels. Formally, it is merely a 2-station open queueing network which, in steady-state, evolves like an Erlang-C (M/M/s) model. In time-varying environments, on the other hand, the situation differs: here one must account for the reentrant nature of service, in order to avoid excessive staffing costs or undesirable service levels. We validate Erlang-R against an Emergency Ward (EW), operating under normal conditions as well as during a Mass Casualty Event (MCE). In both scenarios, we apply time-varying fluid and diffusion approximations: the EW is critically loaded (QED) and the MCE is overloaded (ED). In particular, for the EW we propose a time-varying square-root staffing (SRS) policy, based on the modified-offered-load, which is proved to perform well over small-to-large systems.

Key words: Healthcare; Queueing Networks; Modified Offered-Load; Time Varying Queues; Halfin-Whitt Regime; QED Regime; ED Regime; Emergency Department Staffing; Mass Casualty Events; Patient Flow

1. Introduction: The Erlang-R Model

It is natural and customary to use queueing models in support of workforce management. Most common are the Erlang-C (M/M/s), Erlang-B (M/M/s/s) and Erlang-A (M/M/s + M) models.

Figure 1 The Erlang-R Queueing Model.



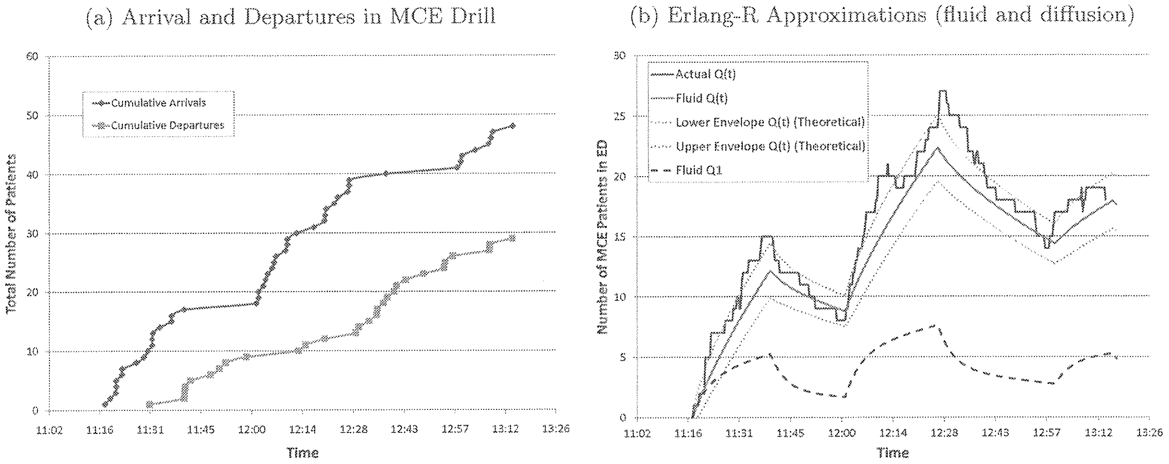
For the first, the process starts by admitting patients and referring them to an EW physician. The physician examines them in order to decide between discharge vs. hospitalization - a decision that could require a series of medical tests. Thus, the process that a patient experiences, from the physician's perspective, fits Erlang-R: a physician visit is a Needy state; and between each visit, the patient is in a Content state, which represents the delay caused by undergoing medical tests such as X-rays, blood tests or examinations by specialists. After each visit to the physician, a decision is made to release the patient from the EW (home or hospitalized), or to direct the patient to additional tests. We shall verify later, in Section 6, that the *simple* Erlang-R model captures the essence of the *complete* EW process, enough to render the model useful for staffing applications.

EWs often accommodate MCEs, and these are inherently transient (Zychlinski et al. 2012). Based on data from an MCE drill, as described in Section 7.1, we demonstrate that our time-varying Erlang-R can accurately forecast MCE census and hence support its management. Ours is a Chemical MCE, and these share treatment protocols that are especially amenable to Erlang-R modeling: every T minutes or so, each patient must be monitored and given an injection, where T depends on severity. (In our case, patients were triaged into 4 levels of severity: the most acute required treatment every 10 minutes, the second level every 30 minutes, etc.)

Our second example is the Radiology reviewing process (Lahiri and Seidmann 2009). After a mammography test, the radiologist interprets the results. In some cases, part of the information on the patient is lacking: the radiologist starts the reviewing but the case must be put on hold. One then waits for this additional information to arrive, after which the reviewing process starts again. With radiologists being the servers, this can be modeled using our Needy-Content cycle.

21.1

Figure 10 **Chemical MCE Drill: Arrivals, Departures, and Erlang-R Approximations.**


(a) Arrival and Departures in MCE Drill

(b) Erlang-R Approximations (fluid and diffusion)

one to use Erlang-R for off-line *Planning* of an MCE, *Initial-Reaction* at its outset (customized to the MCE type, severity and scale), and subsequently online MCE *Control* until the event winds up. We focus as before on staffing. To this end, we use data from a Chemical MCE drill. The MCE took place in July 2010 at 11:00 a.m. and lasted till 13:15; its casualties were transported to an Israeli hospital where our data were collected. The short horizon of MCEs (here 2 hours) and the protocol of chemical events (periodic treatment of patients) renders the transient Erlang-R, with its recurrent service structure, naturally appropriate.

Our data is for the severely wounded non-trauma patients. Figure 10a depicts cumulative arrival and departure counts, collected roughly during 11:15–13:15. The arrival rate is clearly time-varying: periods with no arrivals alternate with approximately constant arrival rates, with the rates decreasing as time progresses. (Our hospital partners, experienced in managing MCEs, inform us that this piecewise-constant pattern of arrival rate is typical of MCEs: it is attributed to the fact that casualties are transported from the MCE scene by a finite number of ambulances, who traverse back and forth.) The estimated arrival rate function (customers per minute) is as follows ($1_{[a,b)}$ is an indicator function):

$$\lambda_t = 0.773 \times 1_{[0,22)}(t) + 0.884 \times 1_{[44,69)}(t) + 0.5 \times 1_{[102,117)}(t), \quad 0 \le t \le 120. \tag{16}$$

21.2

the arrival rate and the number of physicians are scaled up by η while the Needy and Content service rates remain unscaled:

$$\begin{aligned} Q_1^\eta(t) &= Q_1^\eta(0) + A_1^a\left(\int_0^t \eta\lambda_u du\right) - A_2^d\left(\int_0^t \eta p\mu\left(\frac{1}{\eta}Q_1^\eta(u)\wedge s_u\right)du\right) \\ &\quad - A_{12}\left(\int_0^t \eta(1-p)\mu\left(\frac{1}{\eta}Q_1^\eta(u)\wedge s_u\right)du\right) + A_{21}\left(\int_0^t \eta\delta\left(\frac{1}{\eta}Q_2^\eta(u)\right)du\right), \\ Q_2^\eta(t) &= Q_2^\eta(0) + A_{12}\left(\int_0^t \eta p\mu\left(\frac{1}{\eta}Q_1^\eta(u)\wedge s_u\right)du\right) - A_{21}\left(\int_0^t \eta\delta\left(\frac{1}{\eta}Q_2^\eta(u)\right)du\right). \end{aligned} \tag{11}$$

THEOREM 6. *(FSLLN) Through the scaling* (11), *we have*

$$\lim_{\eta\to\infty}\frac{Q^\eta(t)}{\eta} = Q^{(0)}(t), \quad t \ge 0,$$

where $Q^{(0)}(\cdot)$, *the* fluid approximation/model, *is the solution of the following ODE:*

$$\begin{aligned} Q_1^{(0)}(t) &= Q_1^{(0)}(0) + \int_0^t \left(\lambda_u - \mu\left(Q_1^{(0)}(u)\wedge s_u\right) + \delta Q_2^{(0)}(u)\right)du, \\ Q_2^{(0)}(t) &= Q_2^{(0)}(0) + \int_0^t \left(p\mu\left(Q_1^{(0)}(u)\wedge s_u\right) - \delta Q_2^{(0)}(u)\right)du. \end{aligned} \tag{12}$$

The convergence to $Q^{(0)}(\cdot)$ *is a.s. uniformly on compacts (u.o.c).*

The theorem follows from Theorem 2.2 in Mandelbaum et al. (1998). We continue by developing diffusion approximations for Erlang-R. These are used for calculating variances and covariances which, in turn, yield confidence intervals for the number of patients in the system.

THEOREM 7. *(FCLT) Through the scaling* (11) *and with the fluid limits* (12), *we have*

$$\lim_{\eta\to\infty}\sqrt{\eta}\left[\frac{Q^\eta(t)}{\eta} - Q^{(0)}(t)\right] \overset{d}{=} Q^{(1)}(t), \quad t \ge 0, \tag{13}$$

where $Q^{(1)}(\cdot)$, *the* diffusion model/approximation, *is the solution of an SDE (Stochastic Differential Equation), as given by* (26) *in the Internet Supplement, Section A.6. The convergence to* $Q^{(1)}(\cdot)$ *is the standard Skorohod* J_1 *convergence in* $D[0,\infty)$.

The theorem is a consequence of Theorem 2.3 in Mandelbaum et al. (1998). Our fluid and diffusion models are easiest to apply when durations of critical-loading are negligible (the zero-measure assumption in Mandelbaum et al. (2002)). They are thus natural as models for MCEs, during which overloading constantly prevails. Formally:

21.3

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.

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 ?

Flow Diagram: National Cranberry

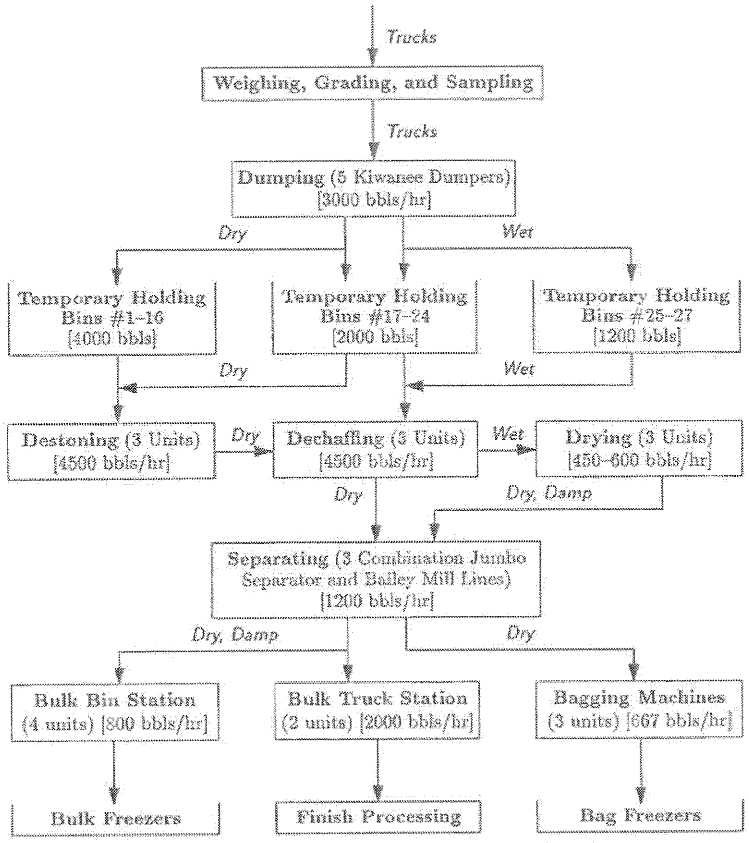
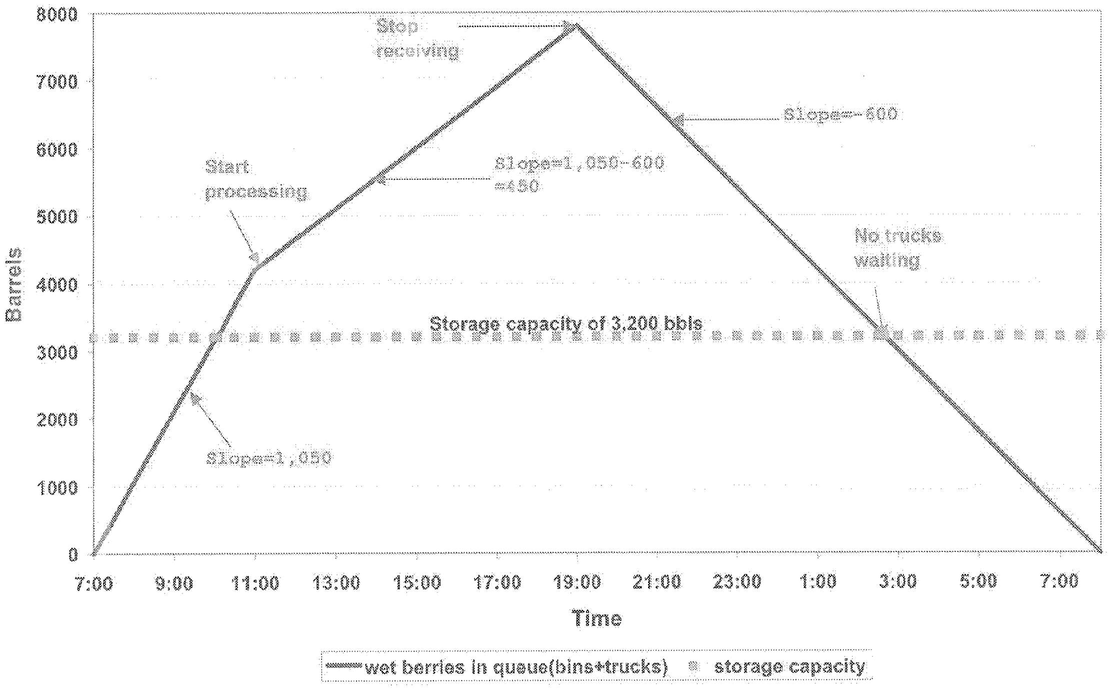

Trucks
Weighing, Grading, and Sampling
Trucks
Dumping (5 Kiwanee Dumpers) [3000 bbls/hr]
Dry
Wet
Temporary Holding Bins #1–16 [4000 bbls]
Temporary Holding Bins #17–24 [2000 bbls]
Temporary Holding Bins #25–27 [1200 bbls]
Dry
Wet
Destoning (3 Units) [4500 bbls/hr]
Dry
Dechaffing (3 Units) [4500 bbls/hr]
Wet
Drying (3 Units) [450–600 bbls/hr]
Dry
Dry, Damp
Separating (3 Combination Jumbo Separator and Bailey Mill Lines) [1200 bbls/hr]
Dry, Damp
Dry
Bulk Bin Station (4 units) [800 bbls/hr]
Bulk Truck Station (2 units) [2000 bbls/hr]
Bagging Machines (3 units) [667 bbls/hr]
Bulk Freezers
Finish Processing
Bag Freezers

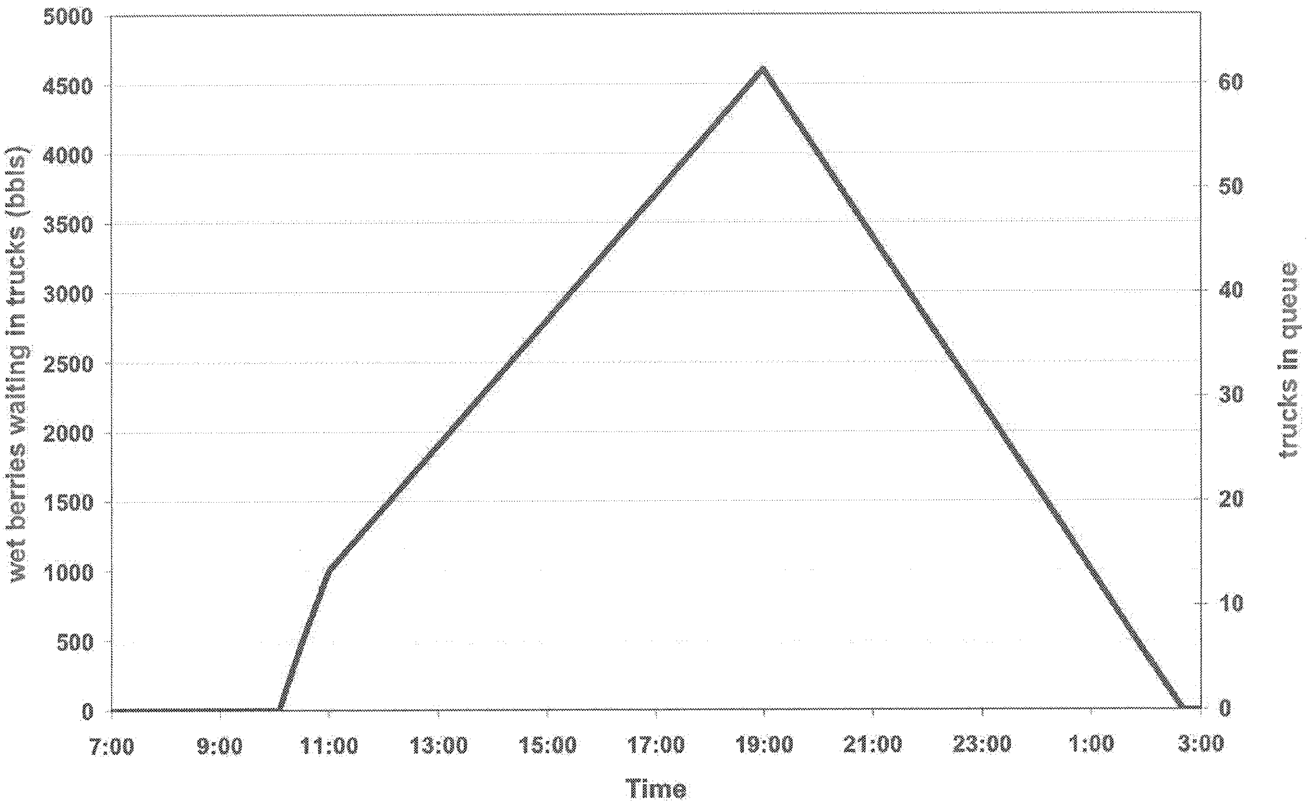
Figure 1: This process flow diagram shows the flow of berries through RP1, including truck arrivals, segregation into wet and dry berries, and the option to partially dry wet berries, creating damp berries. Capacities of work stations, shown as rectangles, are given in barrels per hour, and capacities of storage points, shown as uncovered rectangles, are given in barrels.

Total inventory build-up: Wet berries (bins & trucks)



Processing capacity 600 bbl/hr; Start at 11:00; Peak day 18k*70% over 12 hours.

Trucks inventory build-up Wet berries



Trucks queue analysis

- Area over curve =

$$\frac{1}{2}\cdot 1000\cdot 1+\frac{1}{2}\cdot[1000+4600]\cdot 8+\frac{1}{2}\cdot 4600\cdot 7\tfrac{2}{3}=40{,}533\ bbl\cdot hours$$

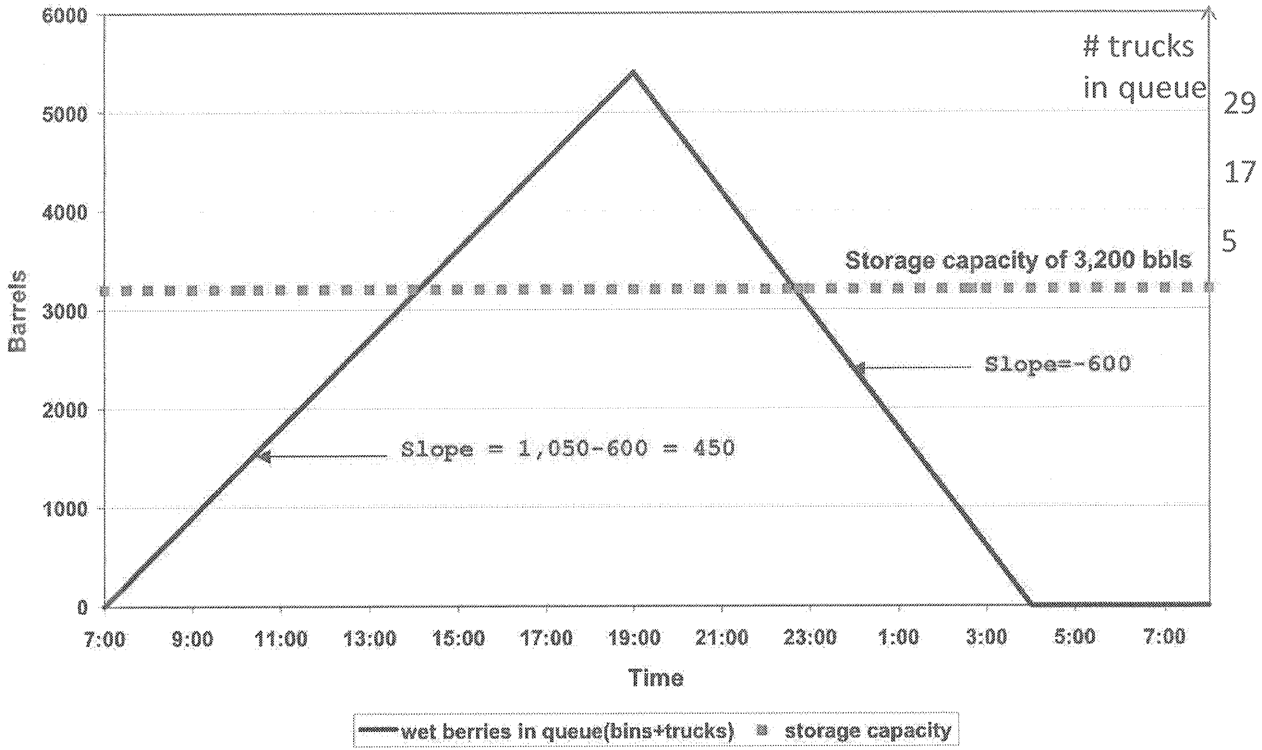
  Divide by 75.

- Truck hours waiting = 40,533/75 bbl/truck = 540 truck•hours
- Ave throughput rate =

$$[0\cdot 1+600\cdot 15\tfrac{2}{3}]/[16\tfrac{2}{3}\cdot 75]=7.52\ trucks/hour$$

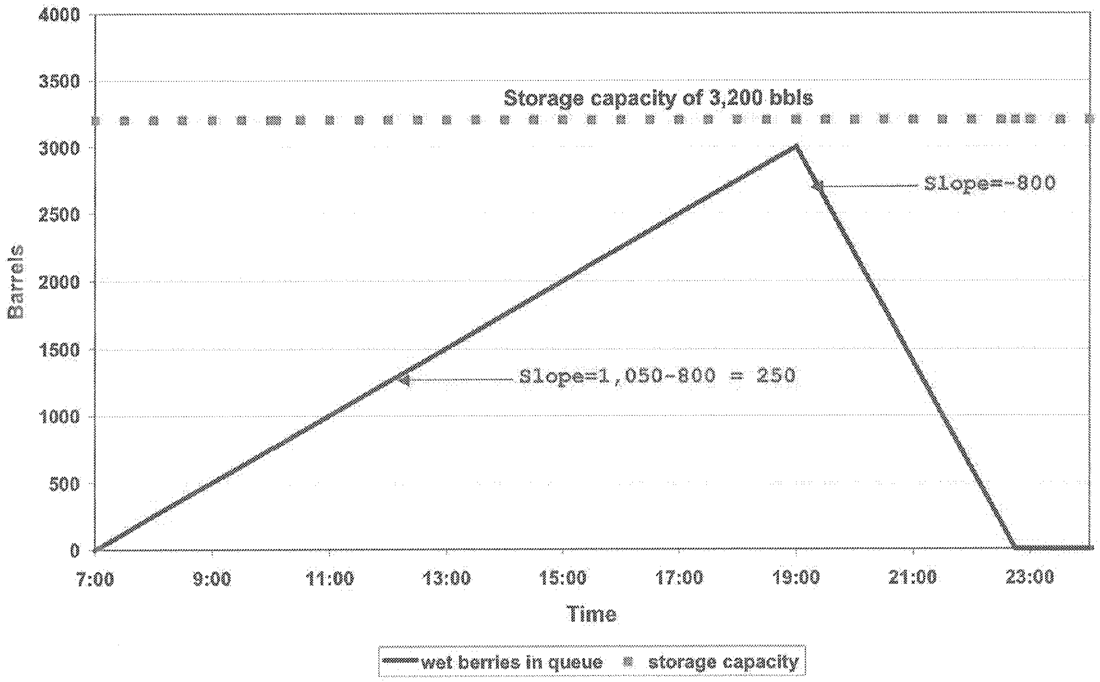
- Ave WIP = $540/16^2/_3$=32.4 trucks (a “biased” average)
- Given that a truck waits, it will wait on average 32.4/7.52 = 4.3 hours (Little’s Law)

Total inventory build-up: Wet berries



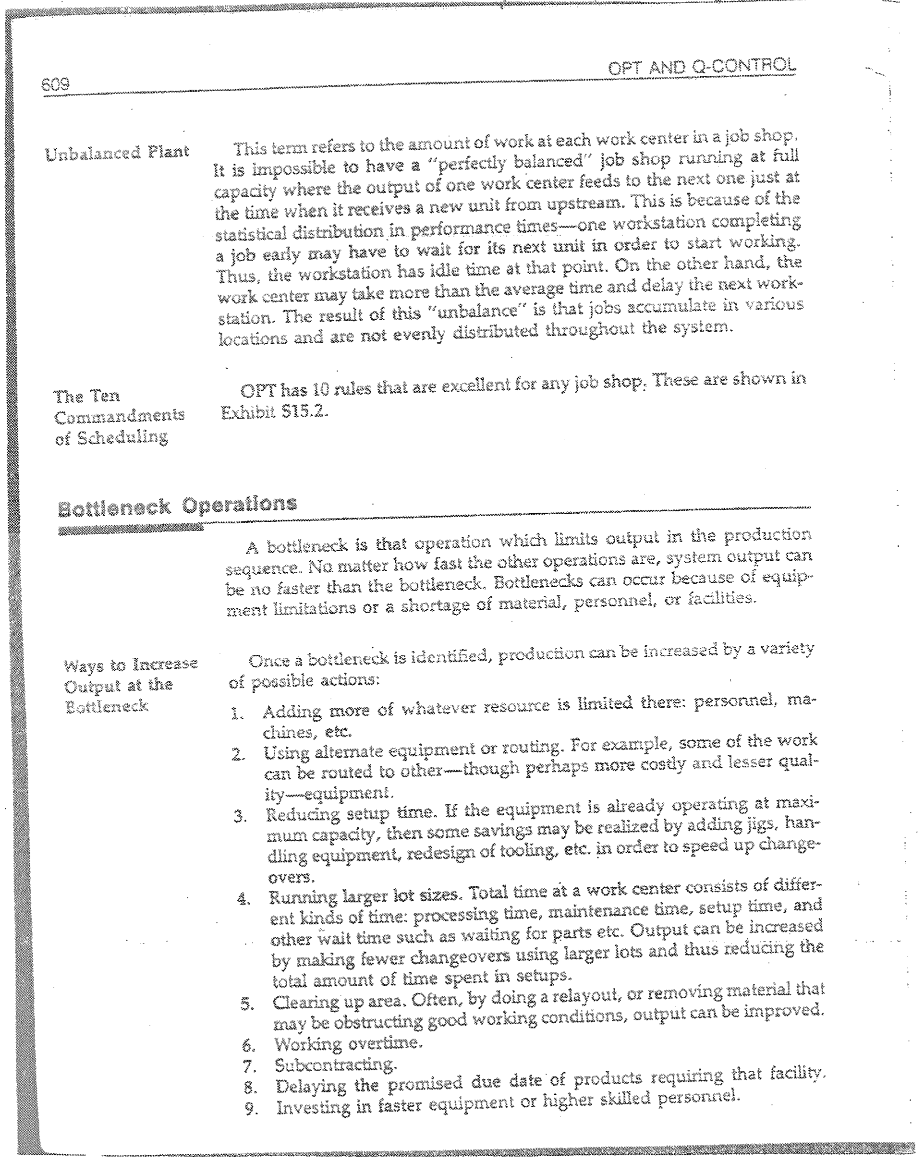
Processing capacity 600 bbl/hr; Start at 7:00; Peak day 18k*70% over 12 hours.

Total inventory build-up: Wet berries



Processing capacity 800 bbl/hr (i.e., add 4th dryner); Start at 7:00; Peak day 18k*70% over 12 hours.

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



609 OPT AND Q-CONTROL

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)

Stochastic Model of a Basic Service Station

Building blocks:

- Arrivals
- Service durations (times)
- Customers' (im)patience.

First study these building blocks one-by-one:

- Empirical analysis, which motivates
- Theoretical model(s).

Then integrate building blocks, via protocols, into Models.
The models support, for example,

- Staffing Workforce
- Routing Customers
- Scheduling Servers
- Matching Customers-Needs with Servers-Skills (SBR).