Approximations / Hierarchical Modelling of Stochastic Networks

Hierarchy

Micro models	ex. Queueing networks
Macro	Fluid nets
Meso	Diffusion nets

Approximations (strong : Hungarian)

F - approx : 1^{st} order, via FSLLN

D - approx : 2^{nd} order, FCLT

Framework DCP (Oblique Reflection)

SCALING and CENTERING of DATA

Here: Visual + intuitive convergence.

(Rigorous: Weak convergence, or convergence of stochastic processes *in distribution*).

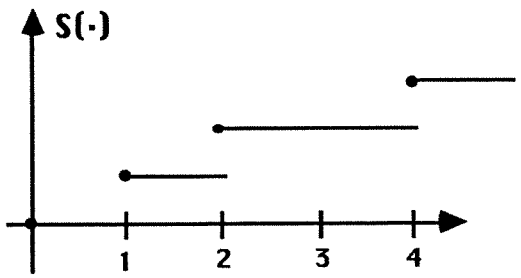
Bernoulli process:

$$\Delta_n = 2 \quad wp\ 1/2 \qquad n = 1, 2, \ldots, \ iid$$
$$= 0 \quad wp\ 1/2\ .$$

Define

$$S_n = \Delta_1 + \Delta_2 + \cdots + \Delta_n\,, \quad n \geq 1\,,$$

$$S(t) = \sum_{k=1}^{\lfloor t \rfloor} \Delta_k \quad (= S_{\lfloor t \rfloor})\,, \quad t \geq 0:$$



- Fit $\{S(t), 0 \leq t \leq n\}$, n large, into a 1×1 frame (a computer screen).

  What to expect?

FLUID and DIFFUSION Approximations (Donsker's Theorem)

$$\Delta_n = \begin{cases} 2 & 1/2 \\ wp & \\ 0 & 1/2 \end{cases} \quad ; \quad S_n = \sum_{k=1}^{n} \Delta_k \quad , \quad S(t) = \sum_{k=1}^{\lfloor t \rfloor} \Delta_k, \quad t \geq 0 .$$

Fluid approximation: $S(t) \approx t, \ t \geq 0$; supported by

$$\textbf{FSLLN} \ \frac{1}{n} S(nt) \xrightarrow{wp1} t, \quad t \geq 0 .$$

Diffusion refinement: $S(t) \approx t + B(t), \quad t \geq 0,$ supported by

$$\textbf{FCLT} \ \sqrt{n}\left[\frac{1}{n} S(nt) - t\right] \Longrightarrow B(t), \quad t \geq 0 ,$$

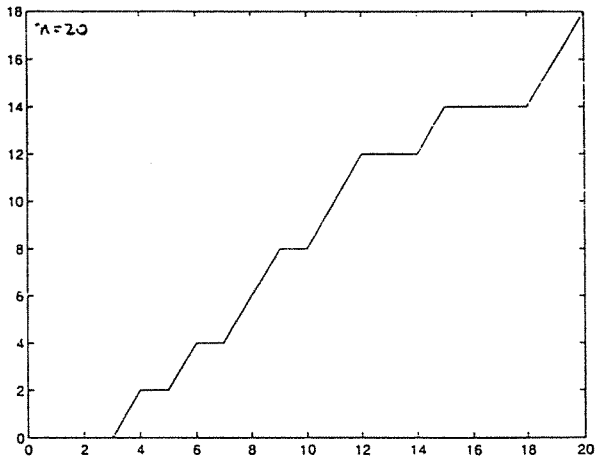
where $B = \{B(t), \ t \geq 0\}$ is Levy with normal (Gaussian) marginals, indeed SBM.

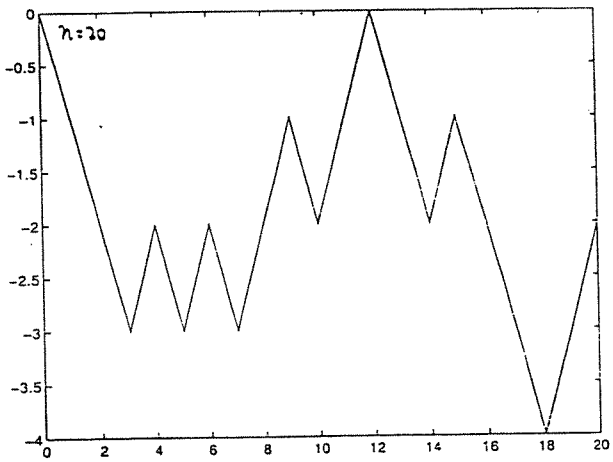
□

Invariance Principle: $\{\Delta_k\}$ mean μ, std. σ, then $\boxed{S(t) \approx \mu t + \sigma B(t), \ t \geq 0.}$

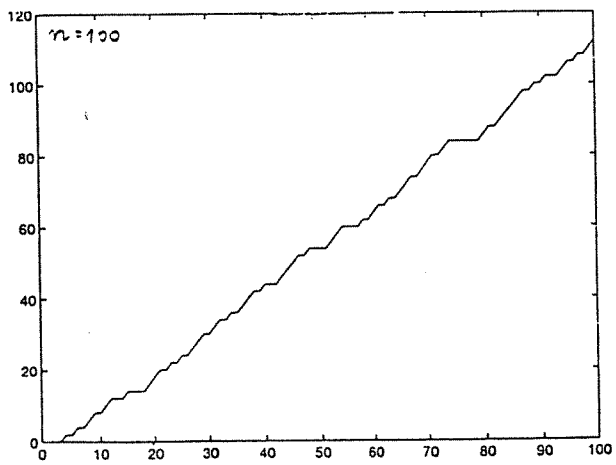
n

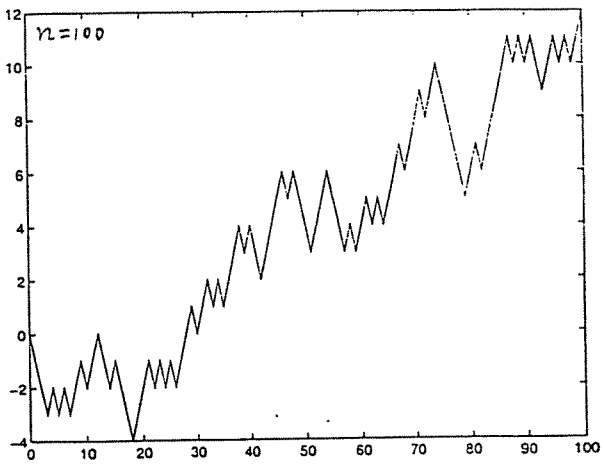
20





100



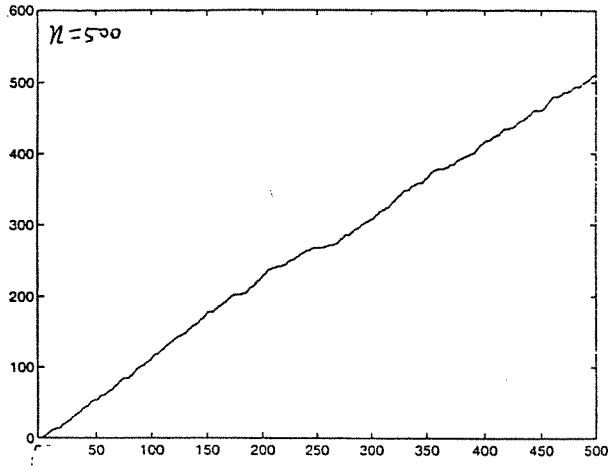


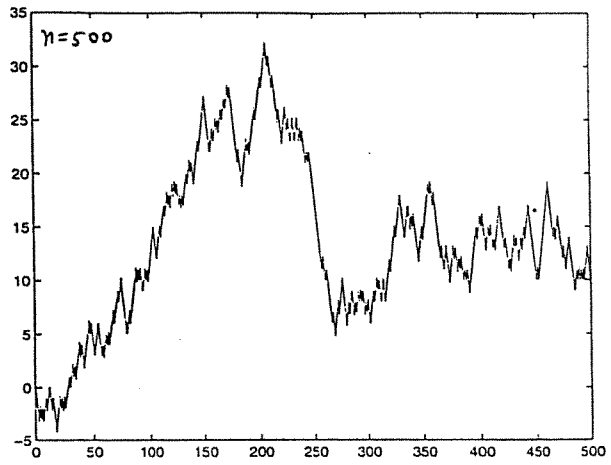
3

Fluid approximation: Deterministic Functional
Strong Law of Large Numbers (FSLLN)

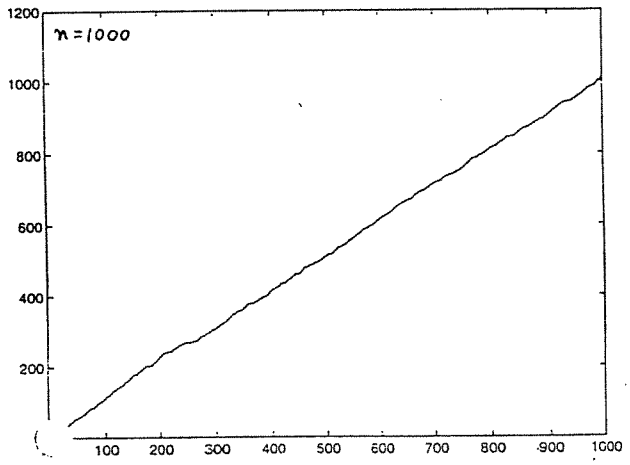
Diffusion deviation: Stochastic Functional
Central Limit Theorem (FCLT)

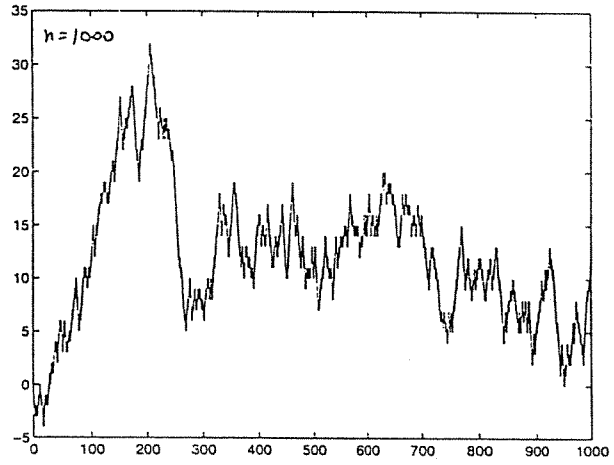
500



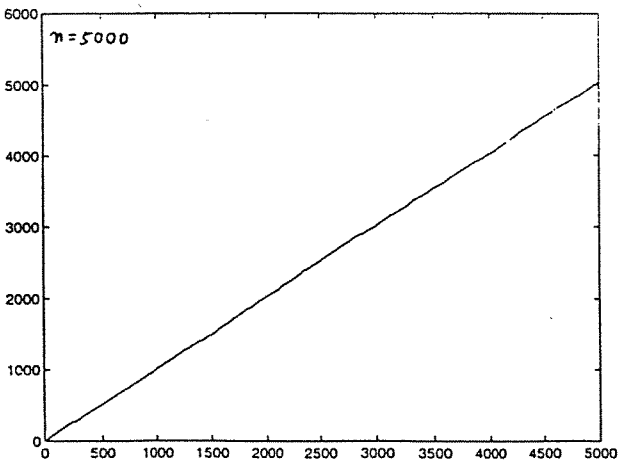


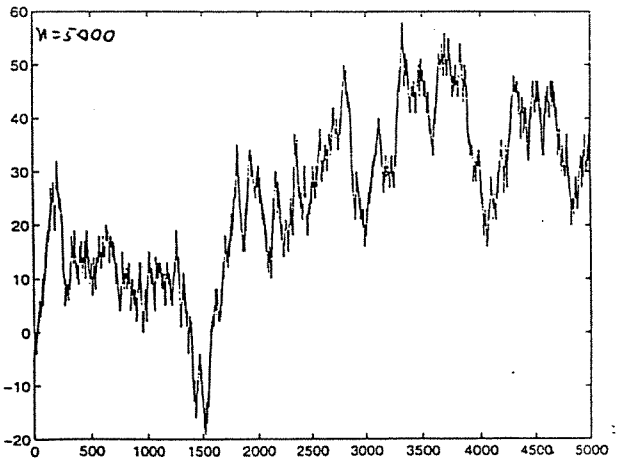
1000





5000





Interpolating $S(0), S(1), \ldots, S(n)$

Interpolating deviations from trend

4

SLLN (Strong Law of Large Numbers)

$$\frac{1}{n}S(n) \xrightarrow{wp1} E\Delta_1 = 1,\ n \uparrow \infty\ .$$

1. **Rescaling** time: $\{S(nt),\ 0 \le t \le 1\}$;

   at $t = 1$: $ES(n) \approx n$;

2. **Aggregating** space: $\{\frac{1}{n}S(nt),\ t \ge 0\}$

   $\xrightarrow{wp1} \{t,\ t \ge 0\},\ n \uparrow \infty.$

(But recall the Cauchy distribution)

CLT (Central Limit Theorem)

$$\sqrt{n}[\frac{1}{n}S(n) - 1] \stackrel{d}{\Longrightarrow} N(0,1),\ n \uparrow \infty\ .$$

3. **Centering** $\frac{1}{n}S(nt) - t$: deviations from trend

4. **Amplifying** $\sqrt{n}[\frac{1}{n}S(nt) - t] \stackrel{d}{\Longrightarrow} N(0,t),\ n \uparrow \infty.$

Obtained:

$$S(nt) \approx n[t + \tfrac{1}{\sqrt{n}}N(0,t)] = nt + \sqrt{n}N(0,t)$$
$$\stackrel{d}{=} nt + N(0,nt)$$

Expect $S(t) \stackrel{d}{\approx} t + B(t),\ \ t \geq 0,$

where $B = \{B(t),\ t \geq 0\}$ is a **Levy process** with normal Gaussian marginals, namely a **Brownian Motion** $BM(0,1) = SBM$ (= Standard Brownian Motion).

Invariance Principle (Donsker's Theorem)

Let $\Delta_1, \Delta_2, \ldots$ be *iid* with mean $= \mu$ and variance $= \sigma^2$ (*general* distribution).

For $S = \{S(t),\ t \geq 0\}$ as before, we have

$$S(t) \stackrel{d}{\approx} \mu \cdot t + \sigma B(t), \quad t \geq 0\,,$$

where $B = \{B(t),\ t \geq 0\}$ is an SBM.

Equivalently, $S \stackrel{d}{\approx} BM(\mu, \sigma^2)$.

Terminology: $S(t)$

$\bar{S}(t) = \mu t$ — *fluid approximation* (deterministic);

$\hat{S}(t) = \sigma B(t)$ — *diffusion refinement* (stochastic).

- **FSLLN** (Functional SLLN)

$$\bar{S}^n(t) = \frac{1}{n} S(nt) \longrightarrow \bar{S}(t), \quad t \geq 0.$$

Convergence wp1 (strong) to a deterministic fluid limit.

- **FCLT** (Donsker's Theorem)

$$\hat{S}^n(t) = \sqrt{n}\left[\frac{1}{n} S(nt) - \mu t\right] \Longrightarrow \hat{S}(t), \quad t \geq 0.$$

Convergence in distribution (weak) to a stochastic diffusion limit.

- **Strong Approximation** (Kolmos, Major, Tusnady)

Assume moments. Then $\exists$ prob. space supporting $B = SBM$ and $\tilde{S}$ such that

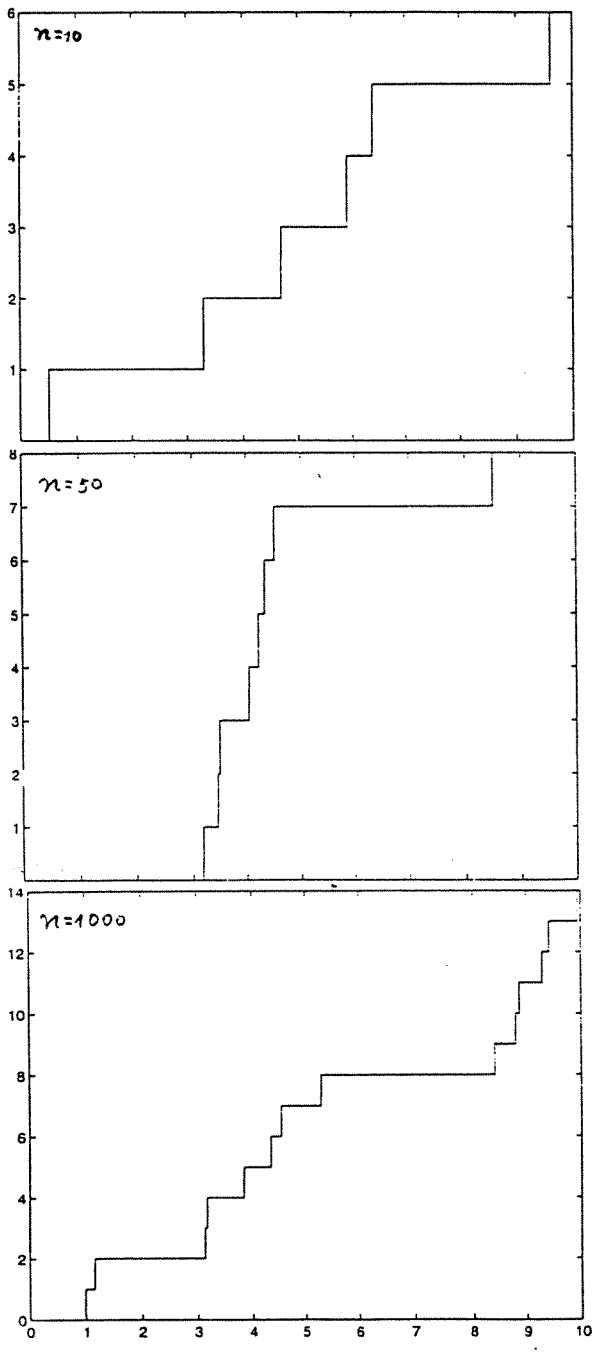
1. $\tilde{S} \stackrel{d}{=} S$
2. $\sup_{0 \leq t \leq \tau} \left|\tilde{S}(t) - (\mu t + \sigma B(t)\right| = o\left(\sqrt{\tau}\right) \ , \quad \tau \uparrow \infty$

Note: This implies both FSLLN and FCLT (but requires "too many" moments).

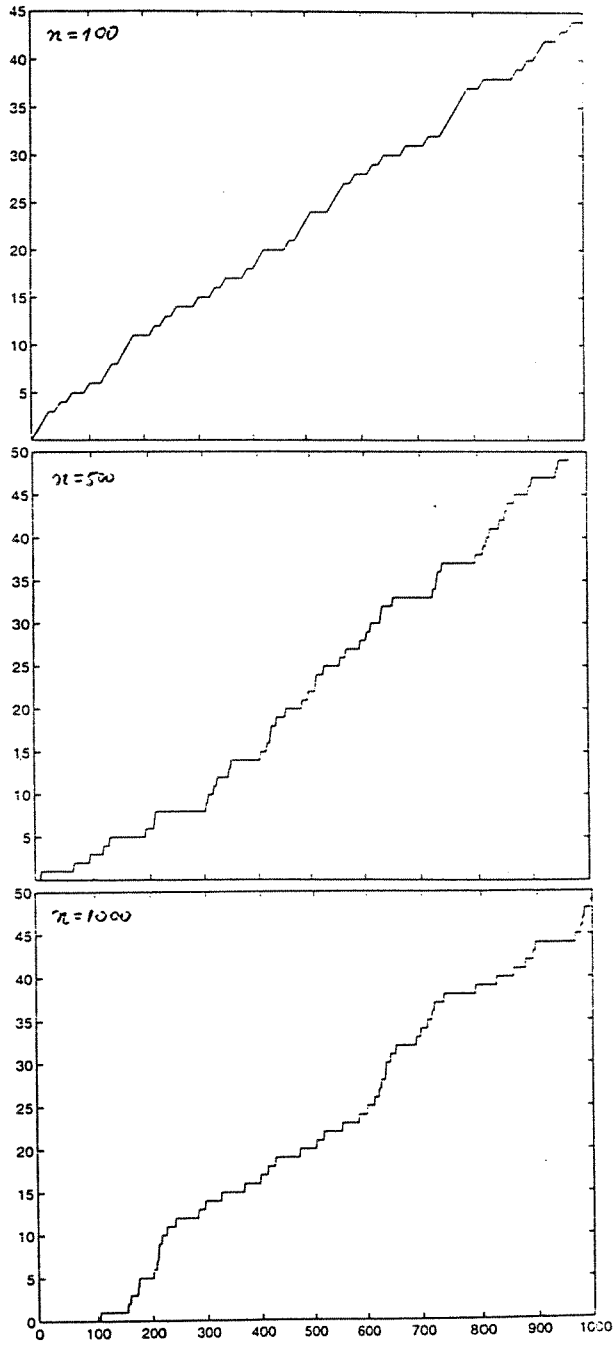
Bernoulli ⇒ Poisson

$$\forall n,\ iid \qquad \Delta_k^n = \begin{cases} 1 & p^n = \frac{\lambda}{n} \\ wp & \\ 0 & q^n = 1 - p^n \end{cases}, \quad S^n(t) = \sum_{k=1}^{\lfloor t \rfloor} \Delta_k^n, \quad t \geq 0.$$

Rare Events: $S^n(nt) \sim$ Binomial $(\lfloor nt \rfloor, \frac{\lambda}{n}) \Longrightarrow$ Poisson (λt), $t \geq 0$.



$S^n(nt)$, $0 \leq t \leq 10$; $\lambda = 1$.



$S^n(t)$, $0 \leq t \leq n$; $\lambda = 50$.

Stochastic-Process Limits

An Introduction to Stochastic-Process Limits And their Application to Queues

Ward Whitt

AT&T Labs - Research
The Shannon Laboratory
Florham Park, New Jersey

Draft
June 13, 2001

Marginals

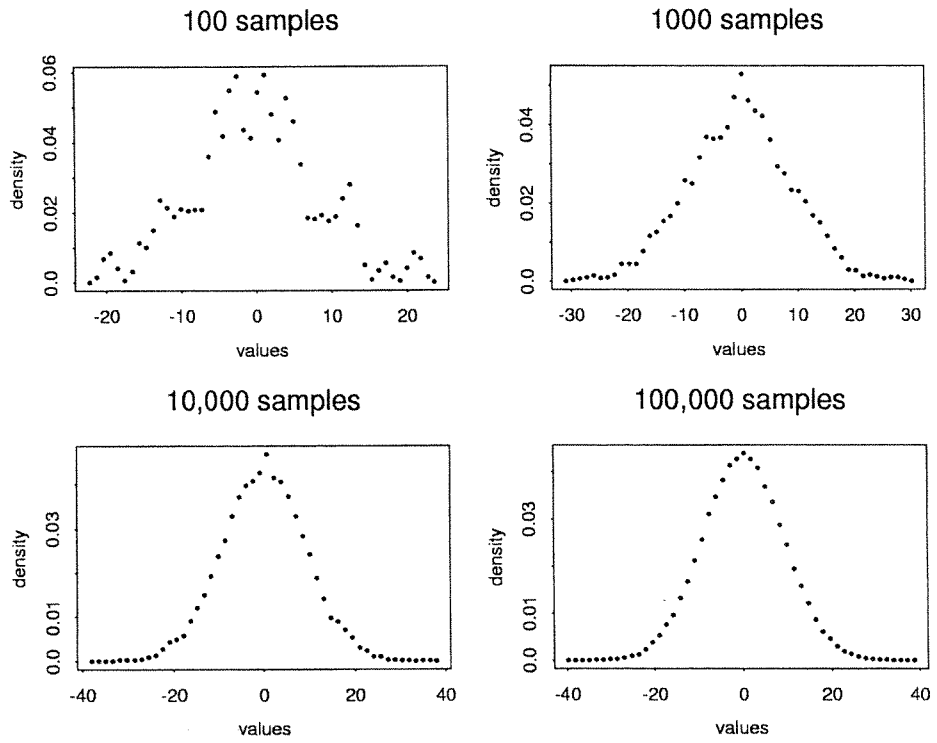


Figure 1.7: Estimates of the probability density of the final position of the random walk, obtained from 10^j independent samples of the centered partial sum $S_{1000} - 500$ for $j = 2, \ldots, 5$, for the case in which the steps U_k are uniformly distributed in the interval $[0, 1]$, based on the nonparametric density estimator *density* from *S*.

density estimates converge to a normal pdf as $n \to \infty$.

It is not our purpose to delve deeply into statistical issues, but it is worth remarking that we obtain new interesting plots, like the random walk plots, when we do. Our brief examination of the distribution of the final position of the random walk suggests looking for a more precise statistical test to determine whether or not the final position of the random walk is indeed approximately normally distributed. To evaluate whether some data can be regarded as an independent sample from any specified probability distribution, it is natural to carefully investigate how the empirical distribution of a sample from that probability distribution tends to differ from the underlying probability distribution itself.

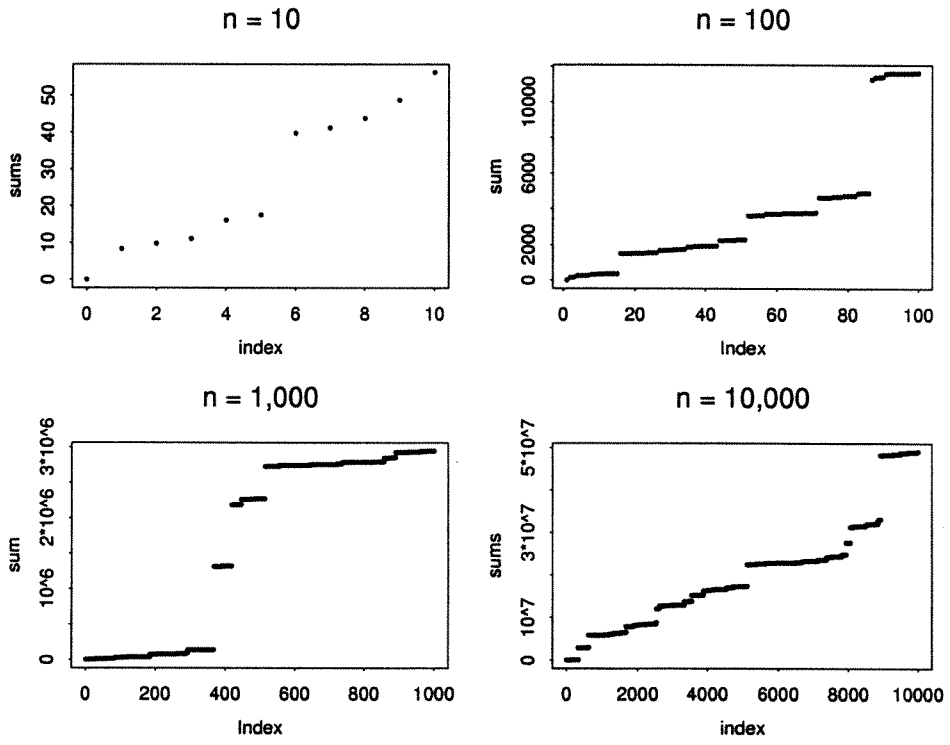


Figure 1.19: Possible realizations of the first 10^j steps of the uncentered random walk $\{S_k : k \geq 0\}$ with steps distributed as $U_k^{-1/p}$ in case (iii) of (3.5) for $p = 1/2$ and $j = 1, \ldots, 4$.

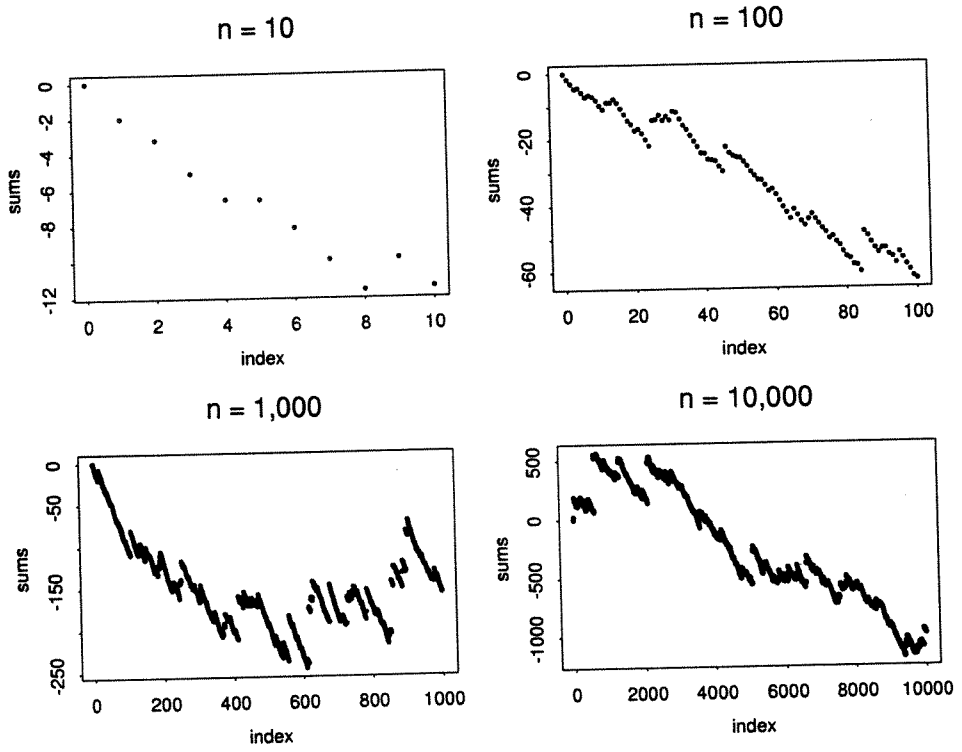


Figure 1.20: Possible realizations of the first 10^j steps of the centered random walk $\{S_k - 3k : k \geq 0\}$ associated with the Pareto steps $U_k^{-1/p}$ for $p = 3/2$, having mean 3 and infinite variance, for the cases $j = 1, \ldots, 4$.

1.20. As before, the centering causes the plotter to automatically blow up the picture. However, now the slight departures from linearity for large n in Figure 1.18 are magnified. Now, just as in Figure 1.19, we see jumps in the plot!

Once again, probability theory offers an explanation. Just as the SLLN ceases to apply when the IID summands have infinite mean, so does the (classical) CLT cease to apply when the IID summands have finite mean but infinite variance. Such a case occurs with the Pareto(p) summands in case (iii) in (3.5) when $1 < p \leq 2$. Thus, consistent with what we see in Figure 1.18, the SLLN holds, but the CLT does not, for the Pareto(p) random variable $U^{-1/p}$ in case (*iii*) when $p = 3/2$.

We have arrived at another critical point, where an important intellectual step is needed. We need to recognize that, *even though the sample paths are*

Teaching Note

Service Engineering

June 26, 1997
Revised 6/2/2000

Dynamic Randomness: The Poisson Process

Hall, Chapter 3: The *Arrival* Process

(with the framework of Levy Processes)

Review:

- Introduction to Services and Queues (Service Nets = Queueing Nets).

  Why queues? Scarce resources, coordination gaps/design constraints.

  There is a "future for queues" in service systems.

  → Poisson
  → Brownian
  → Gamma

- Measurements; Empirical Models I; Scenario Analysis.

  Little's Law $L = \lambda W$ & $H = \lambda G$;

  Capacity analysis (via first order data);

  There is a "future for models" in service systems.

- The Processing Network Paradigm (ReEngineerng = BPR)
  via Dynamic Stochastic Project Networks (DSP-nets).

  - Can we do it? Bottleneck analysis $\leftrightarrow$ the fluid view;
  - How long will it take? Typically stochastic networks;
  - Can we do better? Parametric/sensitivity/what-if.

- A Deterministic Service Station; Empirical Models II

  Skorohod's model (cumulatives a necessity): $z = x + y$
  $z \geq 0,\ y \uparrow 0$
  $\underline{z dy = 0}$

  Modelling scope: Lindley's equation
  Workload

  Towards Stochastic Service Station:

  arrivals' epochs;

  service durations.

Today: Model for Completely Random Arrivals.

Approximating Stochastic Networks

$Z = f(X)$ X Point Process, levy, ...

The mapping f continuous / Lipshitz

$\Rightarrow$ $Z \approx f(X^\bullet)$, $X^\bullet \approx X$ simpler

Fluid approx. $X^\bullet \in BV$ Bounded Variation

Diffusion approx. $X^\bullet \in BM$ Brownian Motion

Both

Long-run $Z \approx f(X^n)$, $n \uparrow \infty$ horizon

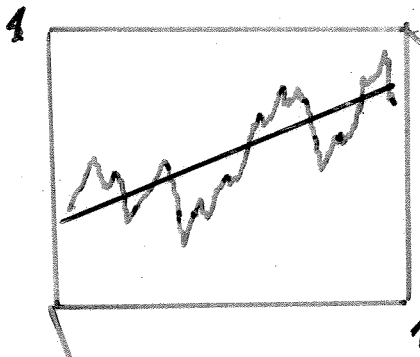
or strong approx., weak converg.

Short-run $Z \approx f(X^\varepsilon)$, $\varepsilon \downarrow 0$ neighbourhood

acceleration, perturbation

$\Rightarrow$ Five-level Hierarchical Framework

Hierarchical Framework



Long-run, aggregation

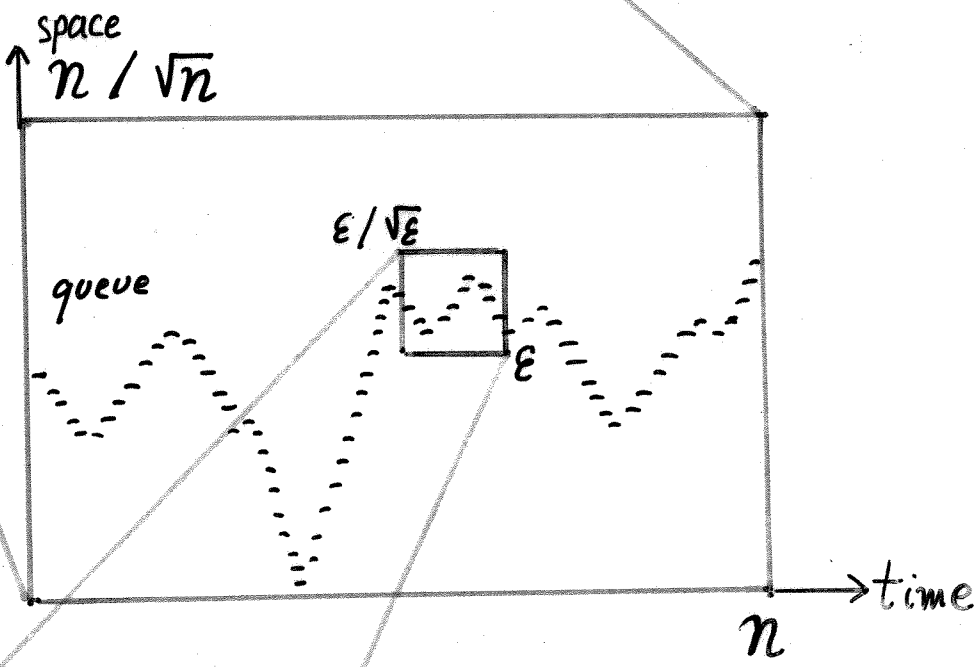
$\frac{1}{n} Z(nt)$

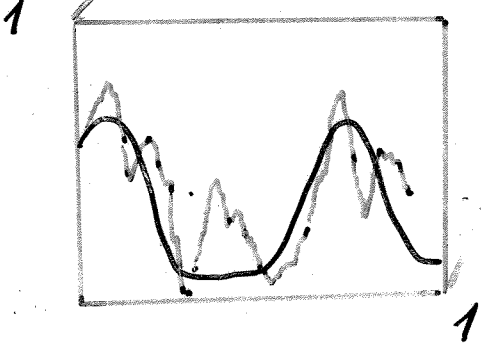
$q = f(x)$, $R = f(x + W)$

$Z = f(X)$

↓ ↑

Jackson	Poisson
NJN	Renewal
Dam	Levy
$(M_t/M_t/1)^N$	Periodic





Short-run, acceleration

$\frac{1}{n} Z^n(t)$

$q = f(x)$, $R = f(x) + \nabla_W \cdot f(x)$

Hierarchical Modelling: a Framework

Ex. Q - length, open periodic Q-net $Q = f(X)$

Strategic compress long-range evolution of network

into short-range horizon of observer

$$\bar{\bar{Q}} = f(\bar{\bar{X}})$$ f macro

bottlenecks, stability

$$\hat{\hat{Q}} = f(\hat{\hat{X}})$$ d meso

sojourn times

Tactical $Q = f(X)$ q-net micro

$$\hat{Q} = f(\hat{X})$$ d-net meso

envelopes

$$\bar{Q} = f(\bar{X})$$ f-net macro

rush-hour

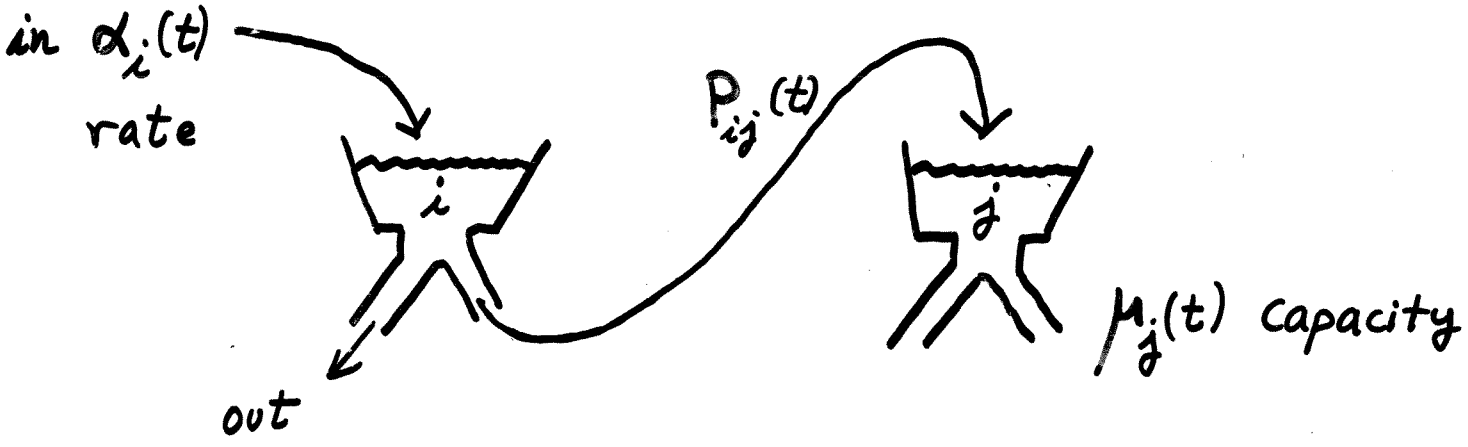
Operational amplify short-range horizon of observer

to long-range evolution of a virtual net

Rescaling → limit theorems (state-space collapse)

Multiple scales

A Fluid Network : Animating f



Outflow rates $\quad \delta_j(t) \leq \mu_j(t)$

Efficient $\quad$ " $\quad < \mu_j(t) \Rightarrow z_j(t) = 0$

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

Content $\quad z = z(0) + \int d - \int \delta = f(x)$

The Mapping ↑ $\quad$ ↑ data (α, μ, P)

Reflection Regulator DCP(x)

$$f \begin{cases} z(t) = x(t) + \int_0^t dy(s)\,[I - P(s)] \,, & t \geq 0 \\ z \geq 0 \,, \; y \uparrow 0 \,, \; z \cdot dy = 0 \end{cases}$$

↑ eff.

where $\quad x = z(0) + \int(\alpha + \mu P) - \int \mu$ $\quad$ netflow

$y = \int \mu - \int \delta$ $\quad$ cum. lost capacity

$j : X \to Z$

Outflow $\quad \delta_j(t)$

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

Eg. constant rates $\alpha(t) \equiv \alpha, \mu, P$: Linear X

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

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

General (eg. periodic)

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

$\rho_j(t) > 1 \quad \Leftrightarrow \quad z_j(t) > 0$ super-critical

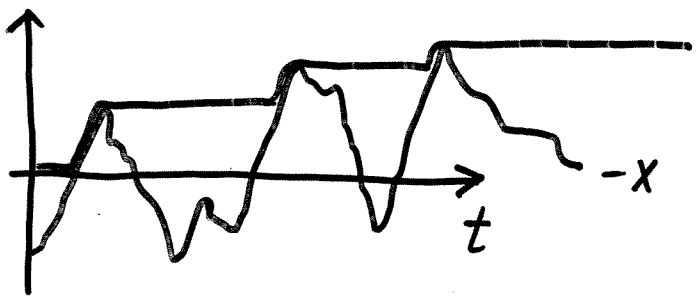
$= 1 \quad \Leftrightarrow \quad z_j(t) = 0,\ \delta_j(t) = \mu_j(t)$ critical

$< 1 \quad \Leftrightarrow \quad z_j(t) = 0,\ \delta_j(t) < \mu_j(t)$ sub-critical

Phases Time-dependent

Geometric Interpretation: Oblique Reflection

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

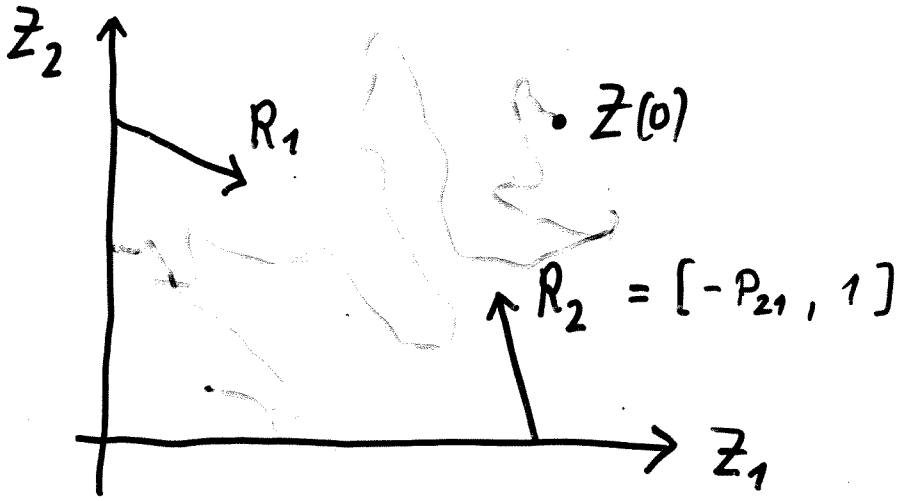


$Y = \overline{(-X)}^{+}$

$= -\underline{X}$ when $X(0) = 0$

(Skorohod)

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

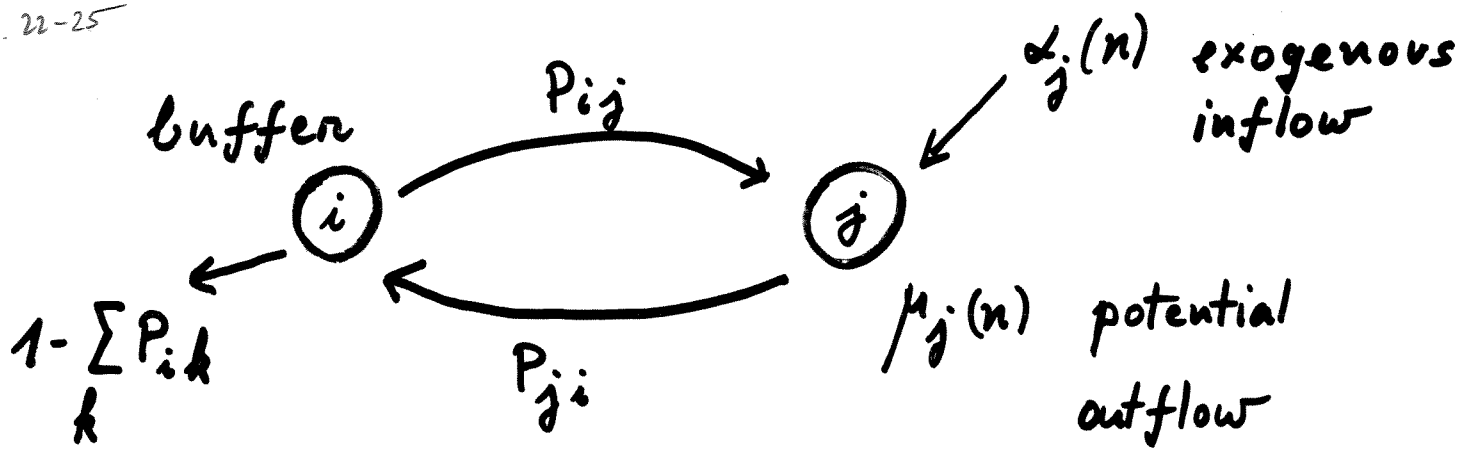


(Harrison + Reiman)

General $\quad f : X \rightarrow Z \qquad$ Lipshitz

(The detailed version Pg. 22-25)

A Flow Network in Discrete Time



At end of periods $n = 1, 2, \ldots$

Content $$z_j(n) = z_j(n-1) + \alpha_j(n) - \delta_j(n) + \sum_i \delta_i(n) P_{ij}$$

Actual outflow $\delta_j(n) \le \mu_j(n)$: efficient ?

Data $z(0) = (z_j(0))$, $\alpha = \{\alpha_j(n)\}$, μ, $P = [P_{ij}]$

Balance eq. $z(n) = z(n-1) + \alpha(n) + \delta(n)[P - I]$

Feasible $z \ge 0$, $0 \le \delta \le \mu$; efficient ?

Open spec rad $(P) < 1$; Closed P stoch. + $\alpha \equiv 0$

Efficient Operation

Content $z(n) = z(n-1) + \alpha(n) + \delta(n)[P-I]$

lost potential $\eta_j(n) = \mu_j(n) - \delta_j(n) \geq 0$

$$z(n) = z(n-1) + \underbrace{\{\alpha(n) + \mu(n)[P-I]\}}_{\xi(n)\ \text{data}} + \underbrace{\eta(n)}_{\text{control}}[I-P]$$

$$\begin{cases} z(n) = z(n-1) + \xi(n) + \eta(n)[I-P] \\ z \geq 0,\ \eta \geq 0 \qquad \text{feasible} \end{cases}$$

Fact: $\exists$ least feasible η^*

potential lost only when buffer empty

$\eta_j^*(n) > 0 \quad \Rightarrow \quad z_j^*(n) = 0$

Complementarity $z_j^*(n) \cdot \eta_j^*(n) = 0$ least η

Geometric Interpretation : Discrete Time

$$Z(n) = Z(n-1) + \xi(n) + \underbrace{\eta(n)\,[I-P]}\,, \quad n = 1,2,\ldots$$

Sequentially

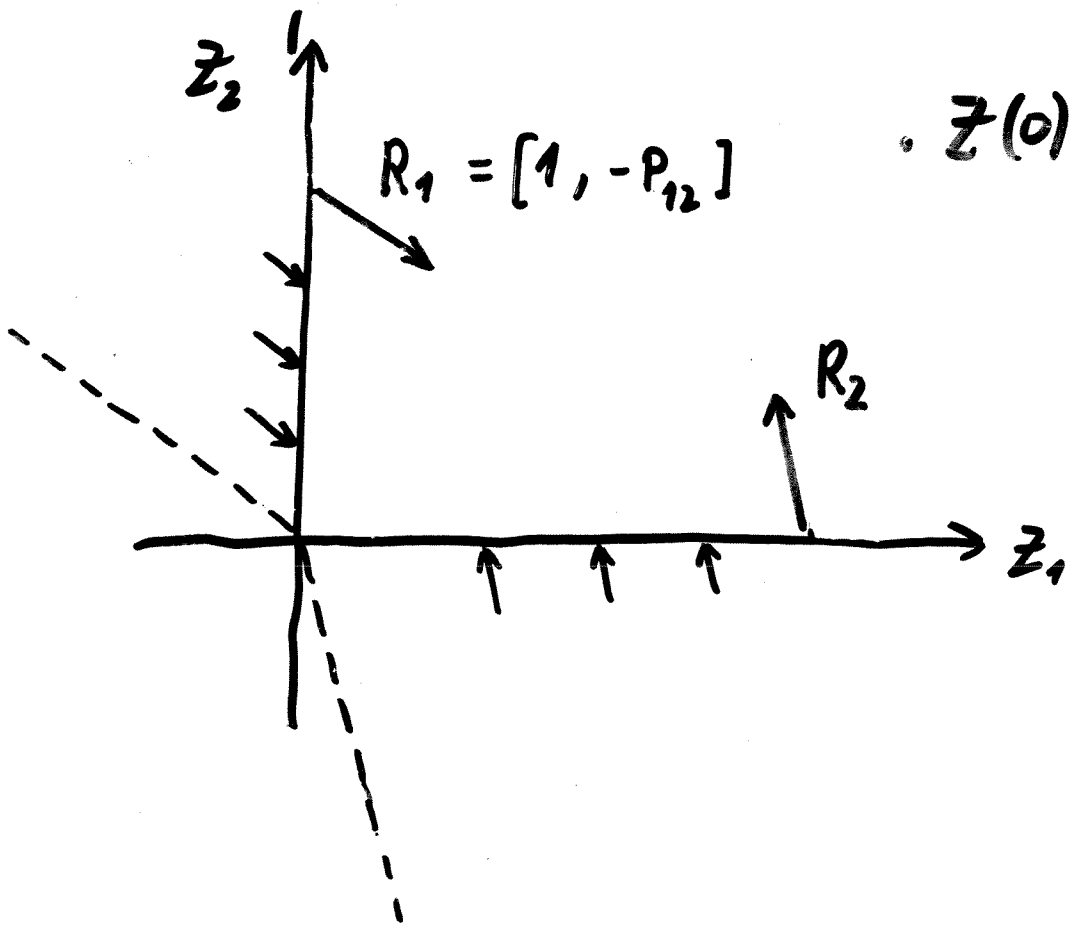
$Z(0), Z(1), Z(2), \ldots$

$\eta_1(n) R_1 + \eta_2(n) R_2 + \ldots$

$\uparrow$

2^{nd} now

$\eta(1)$ $\eta(2)$ ... least (no $*$)



Indeed $z_j(n) \cdot \eta_j(n) = 0$ LCP

Continuous Time

$$\begin{cases} z(n) - z(n-1) = \xi(n) + \eta(n)[I-P] & , n = 1, 2, \dots \\ z \geq 0 \,,\ \eta \geq 0 & \text{feasible} \\ z_j \eta_j = 0 & \text{least} \end{cases}$$

cum. $$z(t) = z(0) + \underbrace{\sum_{n \leq t} \xi(n)}_{X(t)\ \text{data}} + \underbrace{\sum_{n \leq t} \eta(n)}_{Y(t)\ \text{control}} [I-P]$$

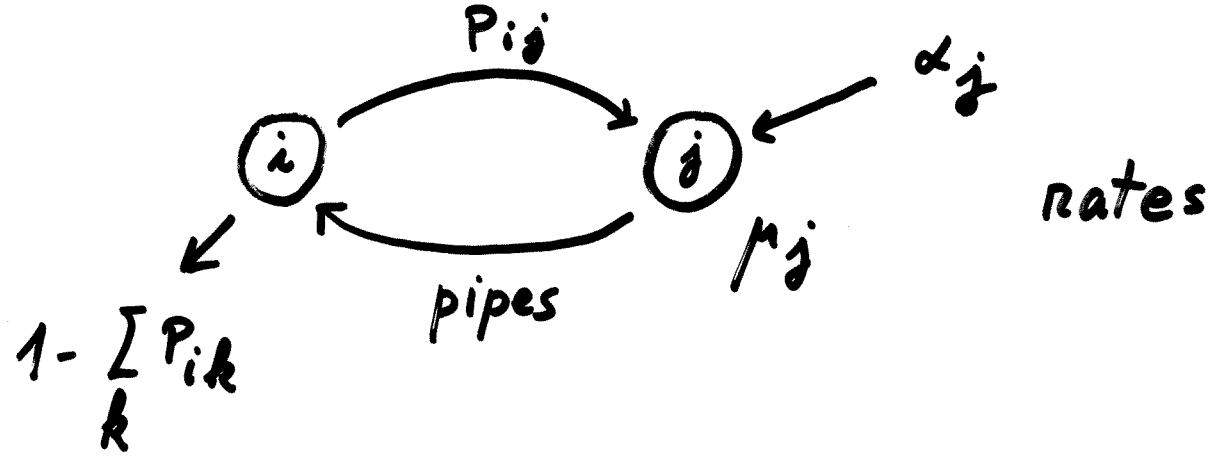
DCP Given $X = \{X(t), t \geq 0\} \in RCLL$

find $Y, Z \ni$

$$\begin{cases} Z(t) = X(t) + Y(t)[I-P] & , t \geq 0 \\ Z \geq 0,\ Y \uparrow 0 & \text{feasible} \\ Z_j \, dY_j = 0 & \text{least} \quad (dY_j(t) > 0 \Rightarrow Z_j(t) = 0) \end{cases}$$

DCP animated by a flow network $\Leftrightarrow X \in BV$ 24

Linear X: Fluid Network



Data			
	$Z_j(0)$	initial content	$Z(0)$
	α_j	exogenous inflow rates	α
	μ_j	potential outflow rates	μ
	P_{ij}	fractions	P

Most efficient operation via DCP with data

$$X(t) = Z(0) + \theta t \quad , \quad \theta = \alpha + \mu[P - I]$$

Equilibrium inflow rates : solve traffic eq.

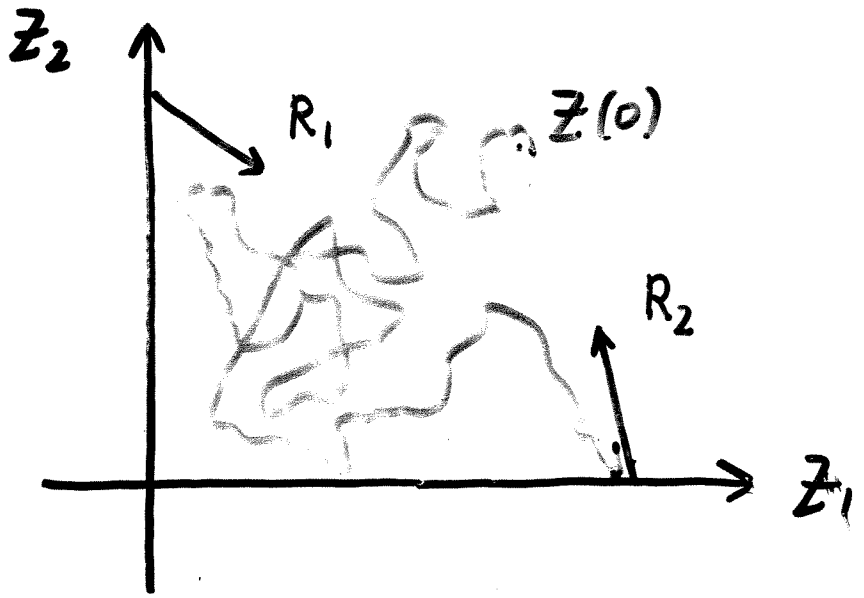
$$\lambda_j = \alpha_j + \sum_i (\lambda_i \wedge \mu_i) P_{ij} \begin{cases} < \mu_j & \text{non-bottlenecks} \\ = \mu_j & \text{balanced} \\ > \mu_j & \text{strict bottlenecks} \end{cases}$$

$$\lambda = \alpha + (\lambda \wedge \mu) P$$

Geometric Interpretation: Continuous Time

$$Z(t) = X(t) + \underbrace{Y(t)\,[I-P]}_{Y_1(t)R_1 + Y_2(t)R_2 + \dots}$$

X continuous $\Rightarrow$ Y, Z as well



$dY_1(t) > 0 \Rightarrow Z_1(t) = 0$, reflect in direction R_1

$f : X \to Z$ oblique reflection

DCP: $X \in RCLL$ Lipshitz

Teaching Note

April 30, 1997
Revised 2/2/2000

A Deterministic Model of a Service Station (Fluid View)

Primitives

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

Model: (Think cumulants)

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

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

- 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 $[0, t]$.

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

$\Rightarrow$ 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 $Z(0) = 0$, where

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

27

THE DYNAMIC COMPLEMENTARITY PROBLEM

Running Head: *DCP*

Avi Mandelbaum †

Graduate School of Business
Stanford University
Stanford, Cal. 94305, USA
and
Industrial Engineering and Management
Technion – Israel Institute of Technology
Haifa, Israel

Abstract A dynamic version of the Linear Complementarity Problem (*LCP*) is proposed. We call it the Dynamic Complementarity Problem (*DCP*). There are versions of *DCP* both in discrete and continuous time. (The latter, which constitutes a natural limit of the former, was actually conceived first; it emerged as a framework for fluid and diffusion approximations of stochastic queueing networks.) Existence and uniqueness results for discrete-time *DCP*'s are precisely analogous to their classical *LCP* counterparts. In contrast, continuous-time *DCP*'s exhibit $\mathcal{P}$-matrices for which uniqueness fails to hold. Such examples will be constructed by introducing and solving a certain system of differential inequalities. This system implicitly characterizes the matrices for which uniqueness prevails, but their explicit characterization remains a challenging open probelm.

Keywords and phrases: Linear and non-linear complementarity; $\mathcal{Q}$-matrices, $\mathcal{P}$-matrices, $\mathcal{M}$-matrices; Reflected/Regulated stochastic processes; Convex processes; Brownian networks, Diffusion and fluid approximations; Elasto-plastic models.

First Version: June 1987
Revised: October 1989 (Accepted to MOR)

AMS 1980 subject classifications (1985 Revision): Primary 90C33. Secondary 90B22, 60F15, 60F17, 60K20, 90C48.

OR/MS Index subject classifications: Primary Programming: Complementarity, Infinite dimensional; Queues: Diffusion models, Networks; Secondary Mathematics: Matrices; Networks: Stochastic.

† Research supported in part by the Stanford Business School Trust Fund, the Israeli Bat-Sheva Foundation and the Fund for Promotion of Research at the Technion.

Derivative of Reflection

Massey, Pats Ramanan, Whitt

Fluid $\frac{1}{n} X^n \Rightarrow \bar{x}$ $\qquad X^n \sim n\bar{x}$

Diffusion $\sqrt{n}\left(\frac{1}{n} X^n - \bar{x}\right) \Rightarrow \hat{x}$ $\qquad X^n \approx n\bar{x} + \sqrt{n}\,\hat{x}$

Consider $z^n = f(x^n)$

Fluid $\frac{1}{n} z^n = \frac{1}{n} f(x^n) = f\left(\frac{1}{n} x^n\right) \Rightarrow f(\bar{x}) = \bar{z}$

($=$: f homog.; $\Rightarrow$: f cont)

$$z^n \sim n f(\bar{x}) = n\bar{z}$$

Diffusion $\sqrt{n}\left(\frac{1}{n} z^n - \bar{z}\right) = \sqrt{n}\left[f\left(\frac{1}{n} x^n\right) - f(\bar{x})\right]$

$$\approx \sqrt{n}\left[f\left(\bar{x} + \frac{1}{\sqrt{n}}\hat{x}\right) - f(\bar{x})\right]$$

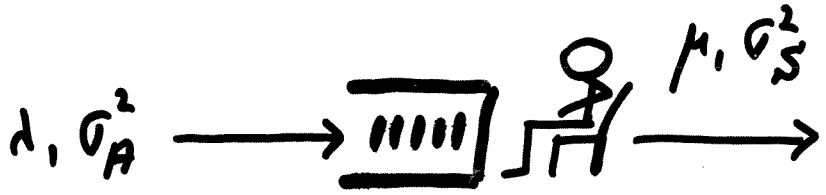
$$= \frac{1}{\varepsilon}\left[f(\bar{x} + \varepsilon\hat{x}) - f(\bar{x})\right] \Rightarrow \nabla_{\hat{x}} f(\bar{x}) = \hat{z}$$

$$z^n \approx n f(\bar{x}) + \sqrt{n}\,\nabla_{\hat{x}} f(\bar{x}) = n\bar{z} + \sqrt{n}\,\hat{z}$$

Subtlety: $\Rightarrow \nabla_{\hat{x}} f(\bar{x})$ in M_1

Common: $\bar{x} = 0$, $f(0) = 0 \Rightarrow \bar{z} = 0$, $\hat{z} = f(\hat{x})$ Critical

The GI/GI/1 Queue



Arrival process $A = \{A(t), t \geq 0\}$

Renewal process : iid interarrival times

mean $= \frac{1}{\lambda}$, var $= \sigma_A^2$ $(c_A^2 = \lambda^2 \sigma_A^2)$

Service times $S_1, S_2, \ldots$

iid services : $\frac{1}{\mu}, \sigma_S^2$ $(c_S = \mu \sigma_S)$

$$L(t) = \sum_{i=1}^{A(t)} S_i \qquad \text{offered load}$$

$$X(t) = L(t) - t \qquad \text{netflow}$$

$$V = X - \underline{X} \qquad \textbf{unfinished work}$$

(virtual waiting time)

V obtained from X via reflection : RBV

(RBM)

G/G/1 : Strong Approximations

$$S(t) = \sum_{1}^{t} S_i \approx \frac{1}{\mu} t + \sigma_S B_S(t) \qquad \text{Donsker}$$

$$A(t) \approx \lambda t + \lambda^{3/2} \sigma_A B_A(t) \qquad \text{Inverse D.}$$

$$L(t) = S[A(t)] \qquad \text{Time-change}$$

$$\approx \frac{1}{\mu}[\lambda t + \lambda^{3/2} \sigma_A B_A(t)] + \sigma_S B_S(\lambda t) \qquad \text{B-fluct.}$$

$$= \frac{\lambda}{\mu} t + \frac{\lambda^{1/2}}{\mu}\left[c_A B_A(t) + c_S \frac{1}{\sqrt{\lambda}} B_S(\lambda t)\right]$$

$$\overset{d}{=} \frac{\lambda}{\mu} t + \frac{\lambda^{1/2}}{\mu}\left(c_A^2 + c_S^2\right) B(t) \qquad \text{Self-similarity}$$

$$X(t) = L(t) - t \overset{d}{\approx} -(1-\rho) t + \sigma B(t)$$

$\rho = \frac{\lambda}{\mu} < 1$ traffic intensity ; $\sigma^2 = \frac{1}{\mu} \rho \left(c_A^2 + c_S^2\right)$

RBM $V \overset{d}{\approx} RBM\left[-(1-\rho), \sigma\right)$; $V(\infty) \overset{d}{=} exp\left(mean = \frac{\sigma^2}{2(1-\rho)}\right)$

G/G/1 $$\frac{V(\infty)}{1/\mu} \overset{d}{\approx} exp\left(mean = \frac{\rho}{1-\rho} \frac{c_A^2 + c_S^2}{2}\right)$$

Teaching Note (soon to be installed) on the G/G/1 Queue

The Congestion Index (Understanding Khinchine-Pollatschek)

$$\text{M/G/1} \;\Rightarrow\; E(W_q) = E(S)\,\frac{\rho}{1-\rho}\,\frac{1+C^2(S)}{2}$$

G/G/1:

$$\boxed{\frac{E(W_q)}{E(S)} \approx \frac{\rho}{1-\rho}\cdot\frac{C^2(A)+C^2(S)}{2}}$$

Allen-Cunneen (Hall (5.70))

- $\frac{\rho}{1-\rho}$: "utilization" availability
- $\frac{C^2(A)+C^2(S)}{2}$: stochastic variability
- pure (unitless)

Fact: Right-hand side upper bound, becoming asymptotically exact as $\rho \uparrow 1$ (Heavy traffic).

Kingman's Exponential Law of Congestion (Invariance principle)

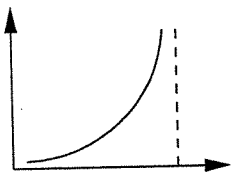
$$\frac{W_q}{E(S)} \approx \begin{cases} \exp\left(\text{mean} = \dfrac{1}{1-\rho}\cdot\dfrac{C^2(A)+C^2(S)}{2}\right) & \text{, wp } \rho, \\ 0 & \text{, wp } 1-\rho, \end{cases}$$

asymptotically exact as $\rho \uparrow 1$. (Later, at tails.)

Understanding the formula:

Assume $C^2(A) = C^2(S) = 1$ (as in M/M/1): $\frac{E(W_q)}{E(S)} = \frac{\rho}{1-\rho}$

Substitute $\rho = 0.5,\ 0.9,\ 0.95,\ 0.99.$

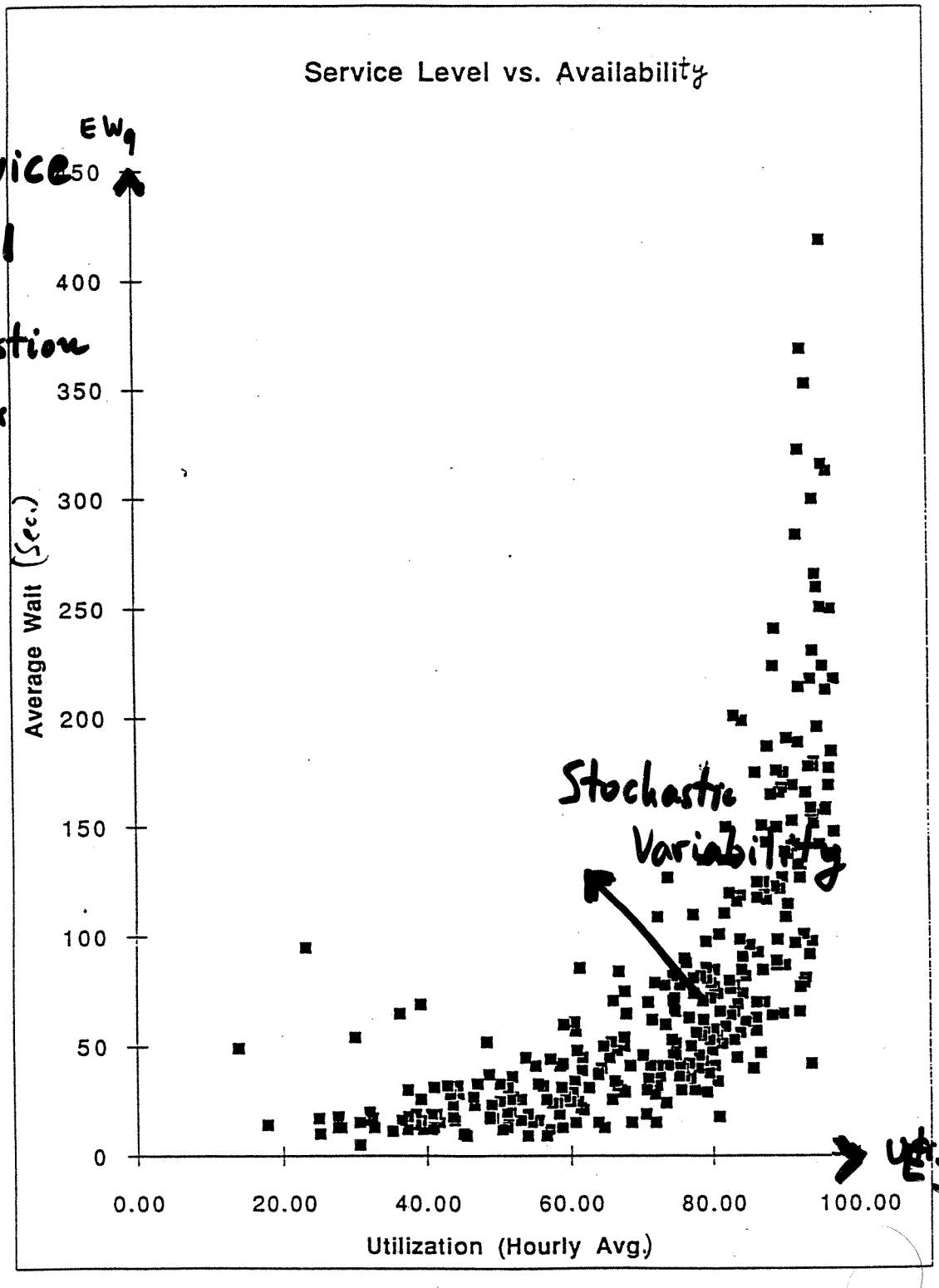


Other MOP's:

$$\begin{aligned} E(W) &= E(S) + E(W_q) \\ E(L_q) &= \lambda E(W_q) \\ E(L) &= \lambda E(W) = E(L_q) + \rho \end{aligned}$$

K-T 2nd Law of Congestion

Congestion Curves

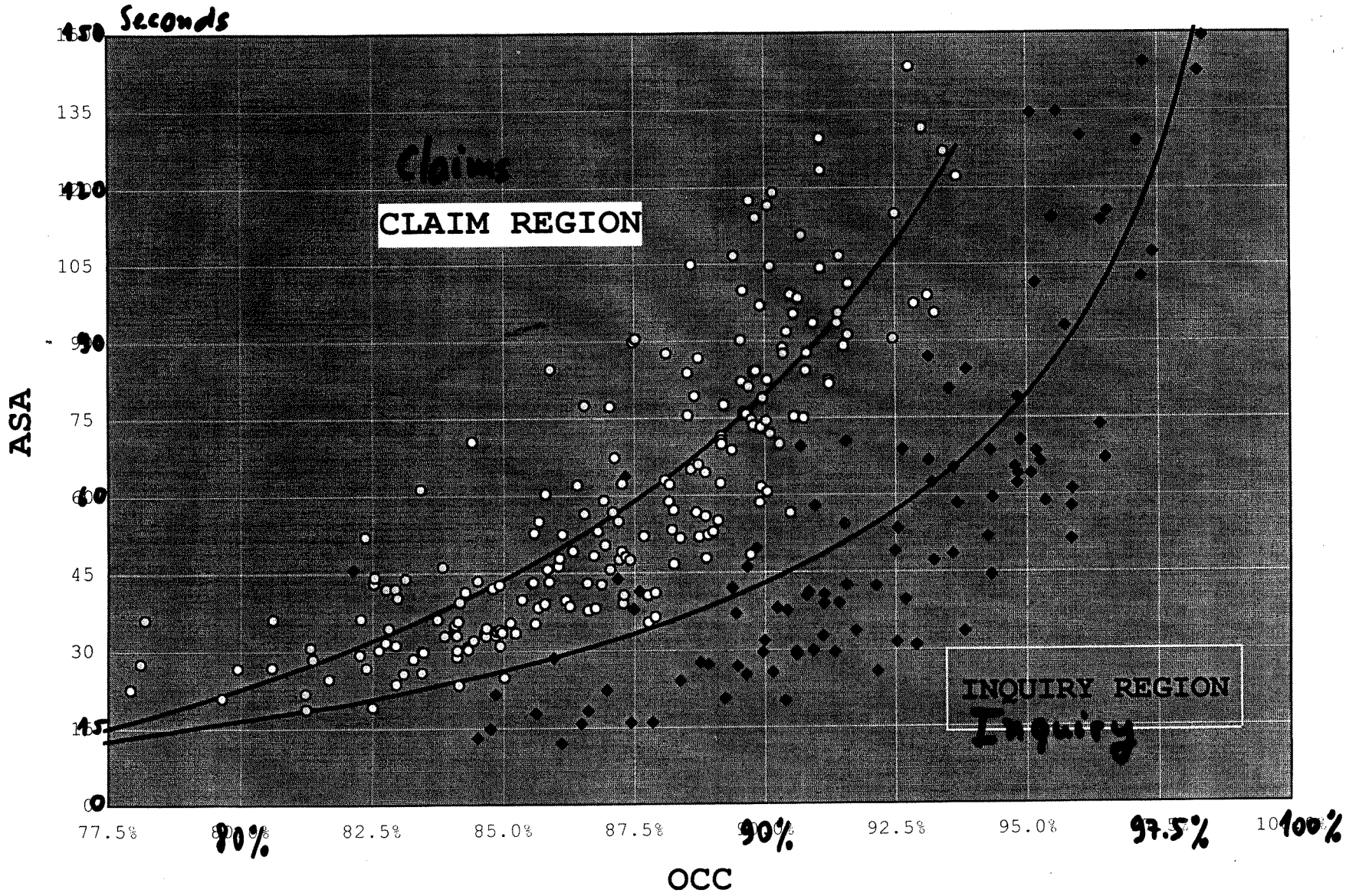


- Service Level
- Congestion Index

A **Congestion Index:** $\frac{E(W_q)}{E(s)} \approx \frac{1}{M} \cdot \frac{\rho}{1-\rho} \cdot \frac{Ca^2 + Cs^2}{2}$ (M = # servers)

$= \frac{1}{M} \cdot \frac{\rho}{1-\rho} \cdot C^2$

Health Insurance Call Center



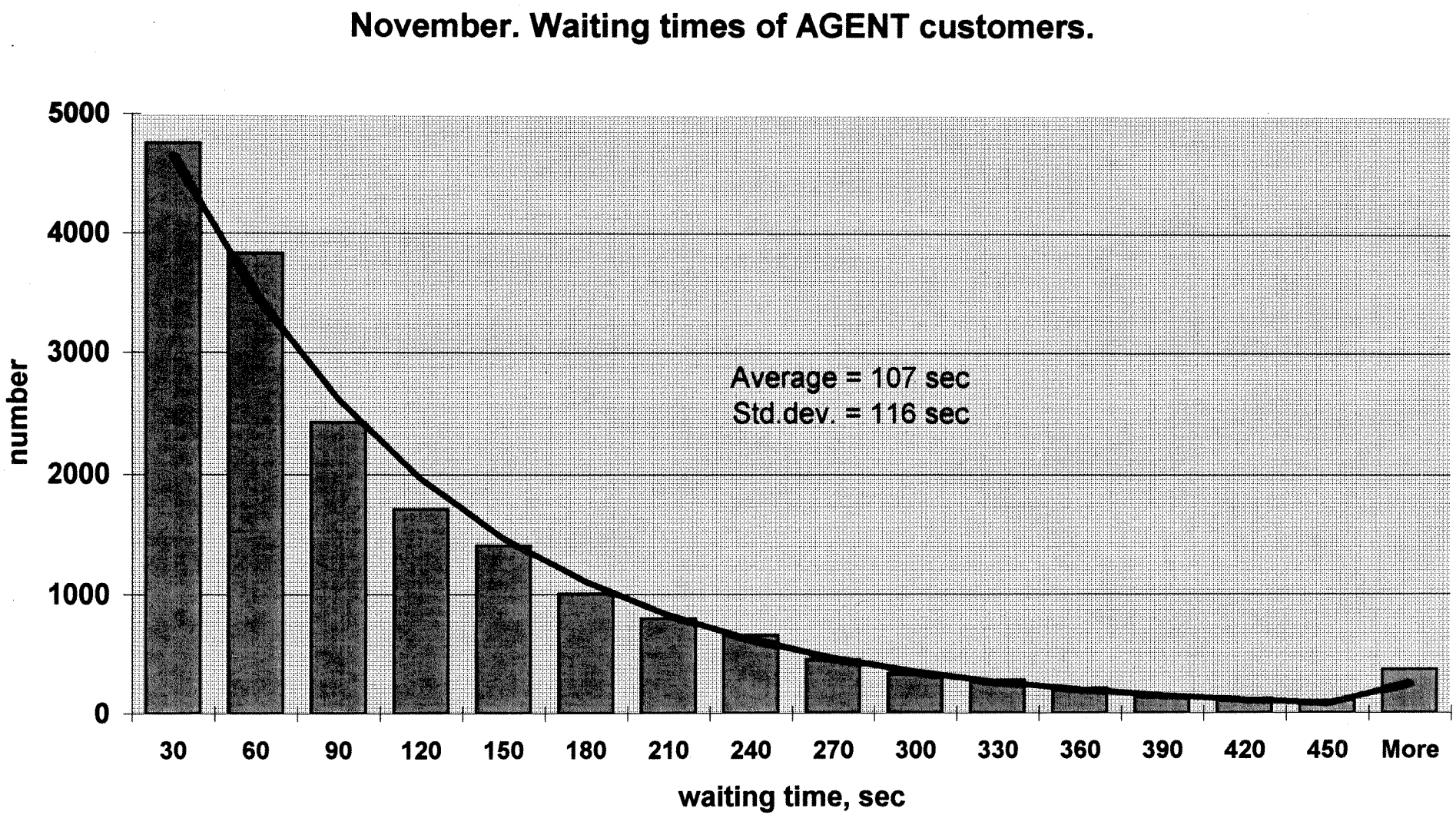
Average Speed of Answer

Occupancy

Staffing ?

Specialization ?

Kingman's 3rd Law of Congestion



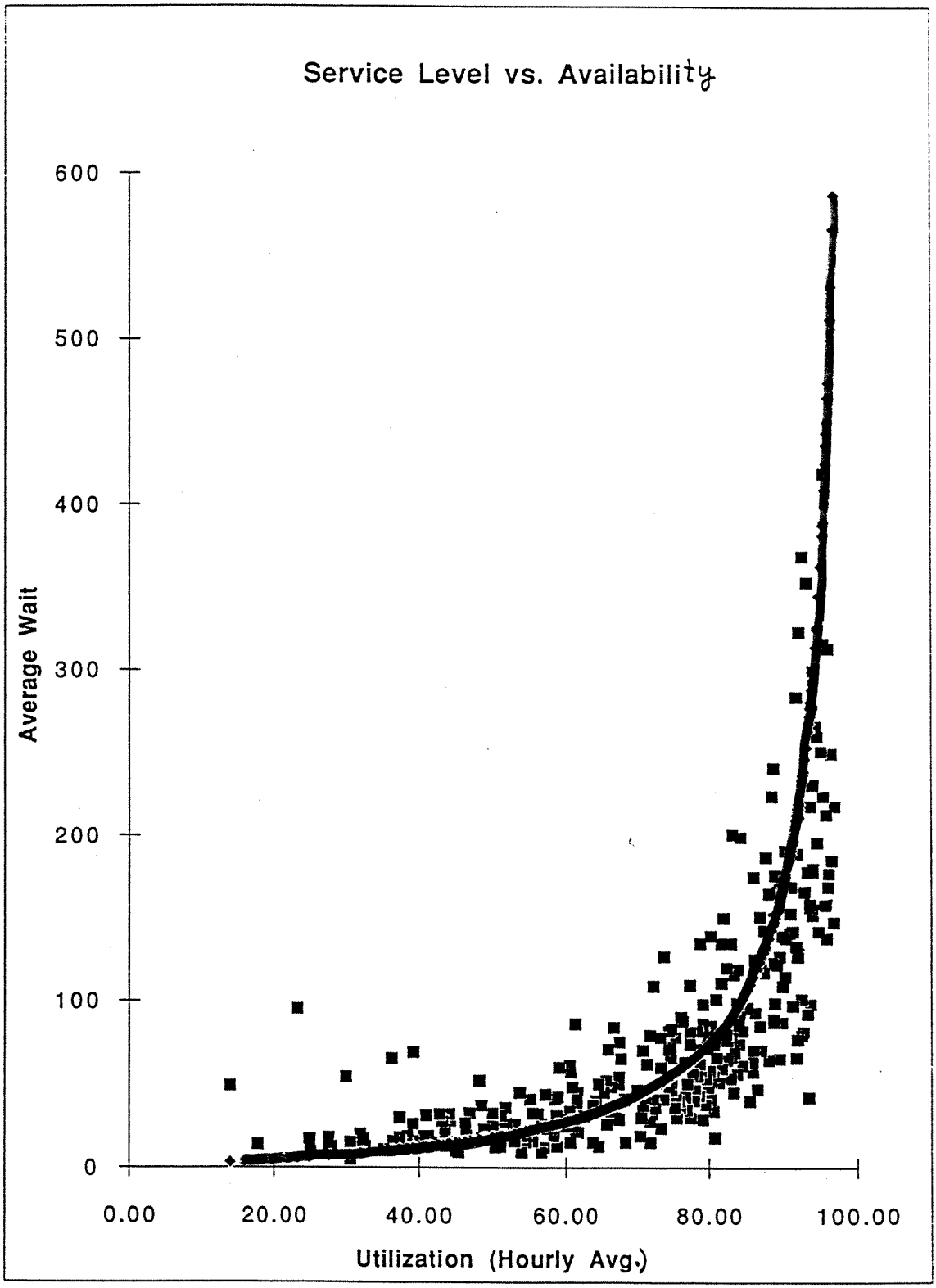
Ex. $E = \sigma$

$P(\text{Wait} > E) = e^{-1} \approx 37\%$

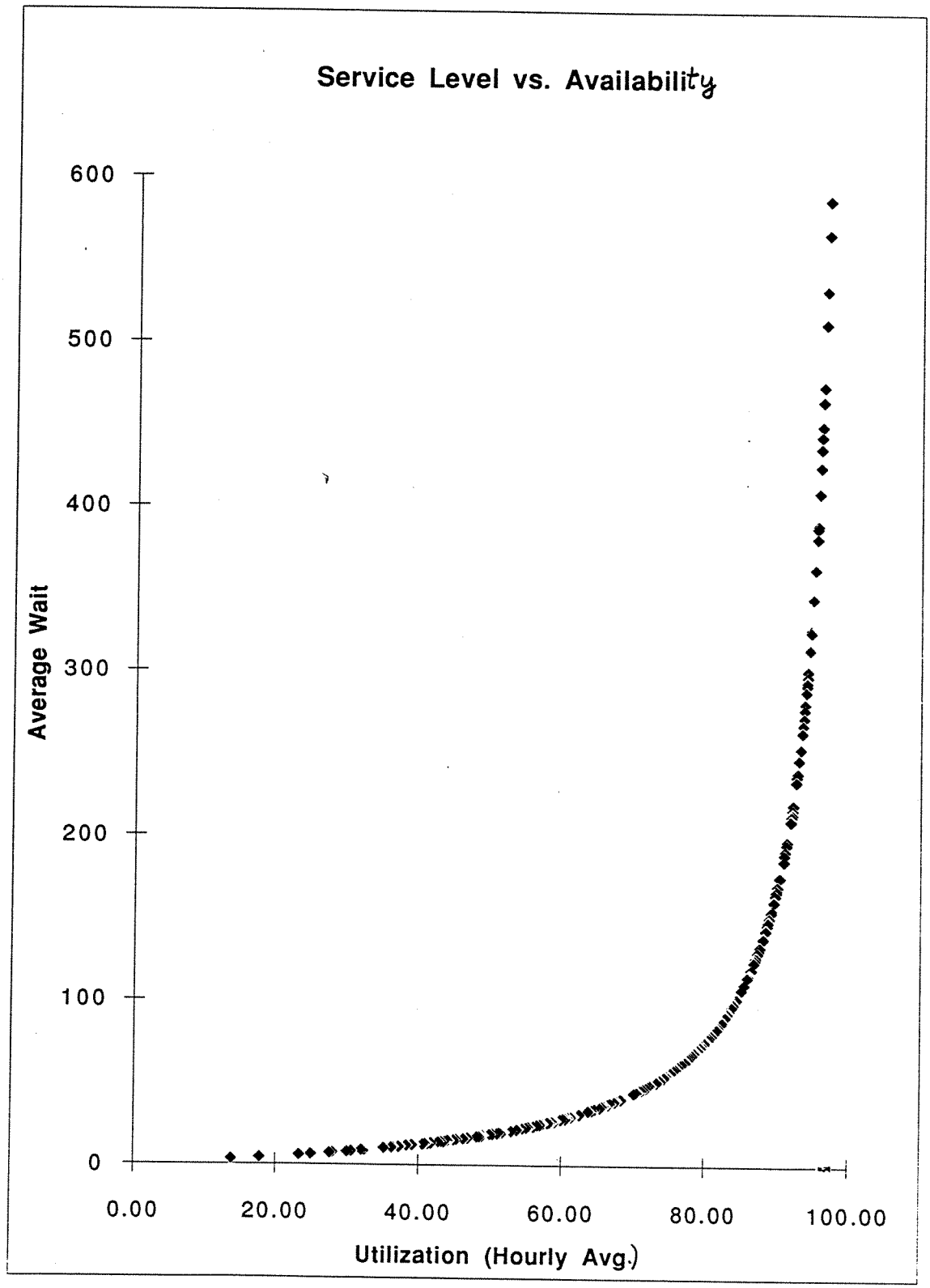
$P(\text{Wait} > 3E) \approx 5\%$

invariance

Fit $W_q \approx M \cdot \frac{\rho}{1-\rho} \cdot \boxed{?}$



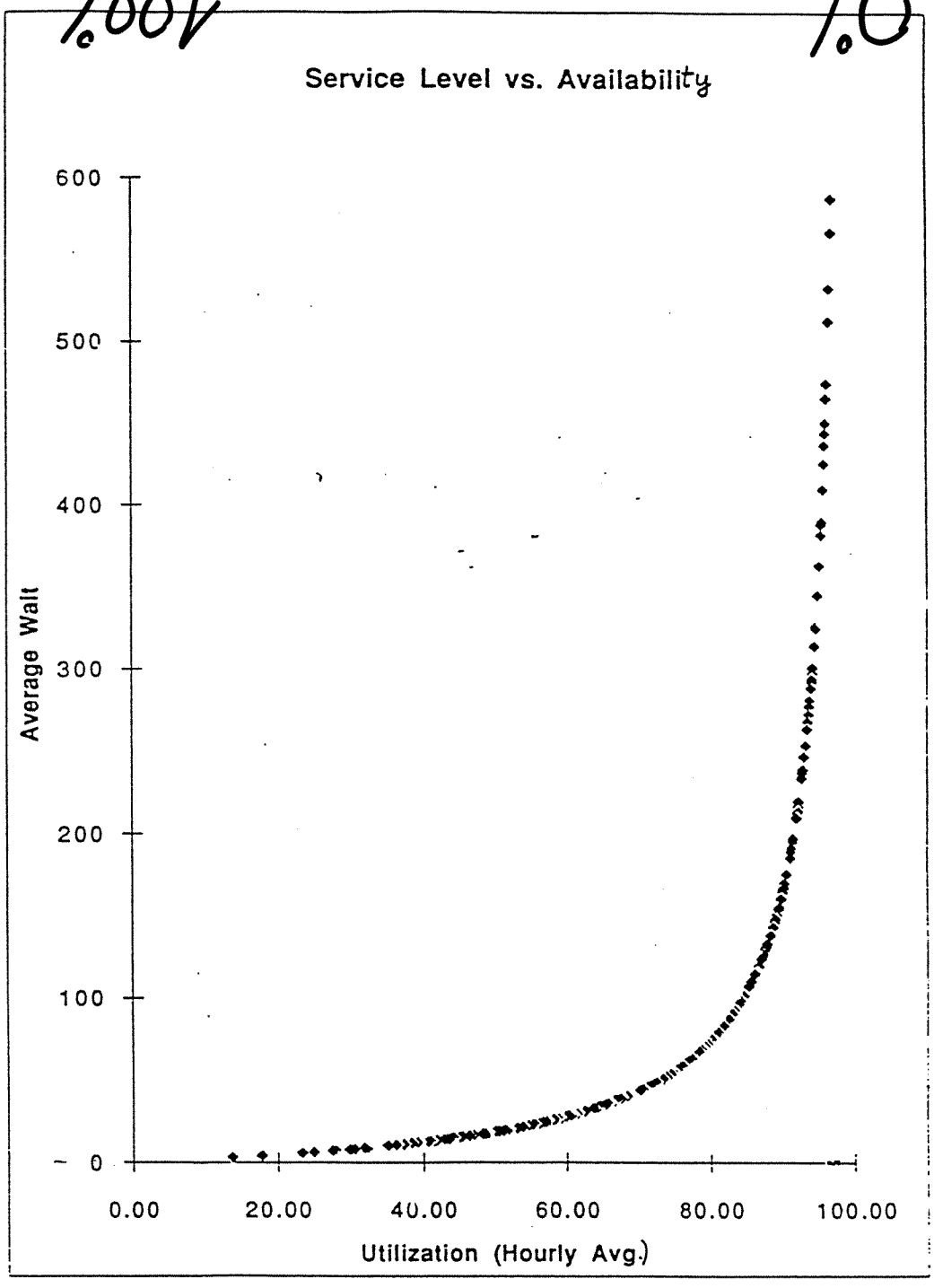
Queueing Science



$$W_q = M \times \frac{\rho}{1-\rho} \times 0.093$$

Can use

Queueing Science



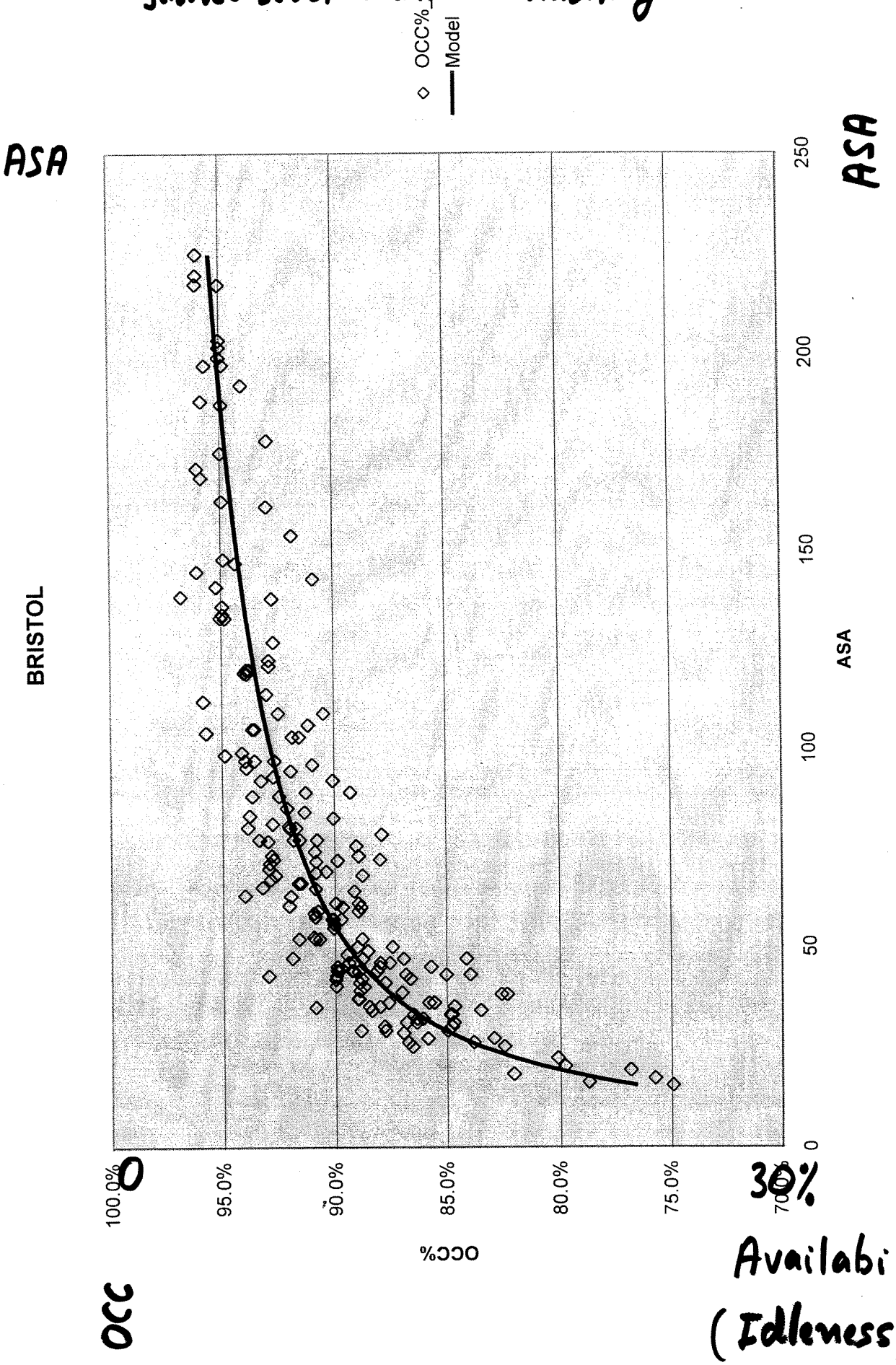
$$W_q = M \times \frac{\rho}{1-\rho} \times 0.093$$

Can use

Service level, as a function of Availability

Congestion = f (Availability) ↓

Service level ↑ with availability

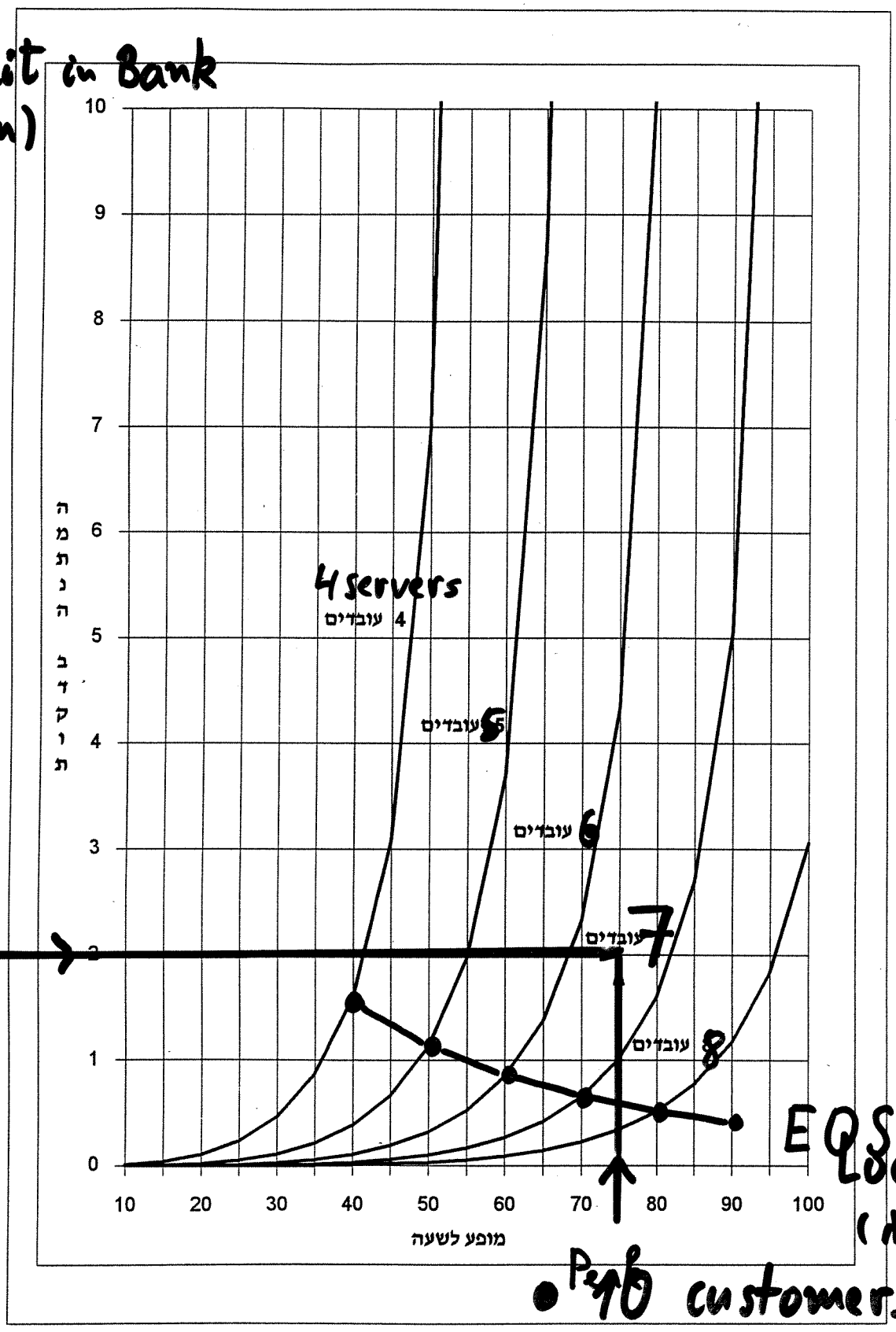


Staffing: Key performance of queue

@ given load/efficiency, service level improves w/ size

נספח 6 - אשנב כל מחלקת שרותים בנקאיים - קביעת רמת איוש לפי אמד שירות זמן

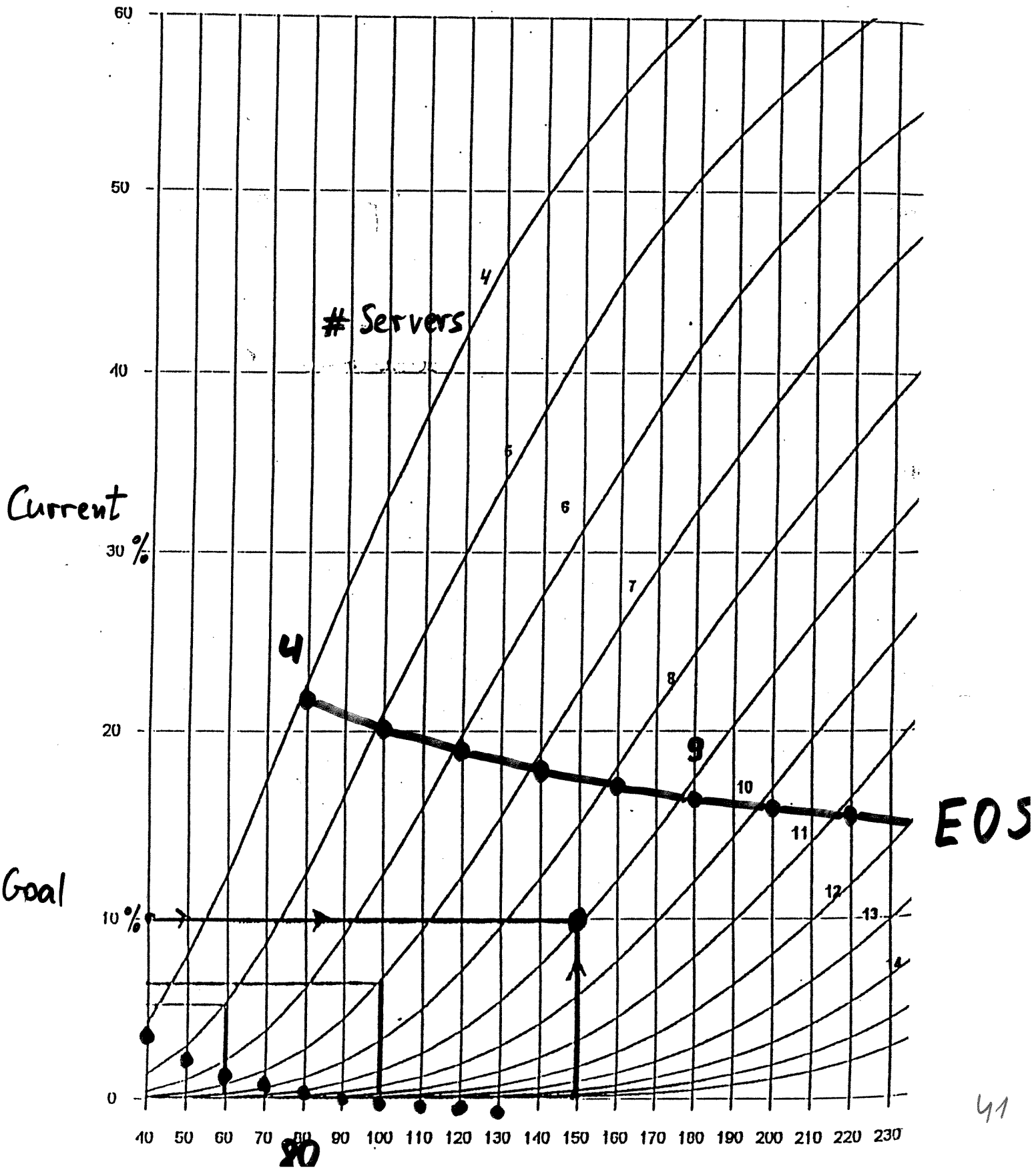
המתנה כללי



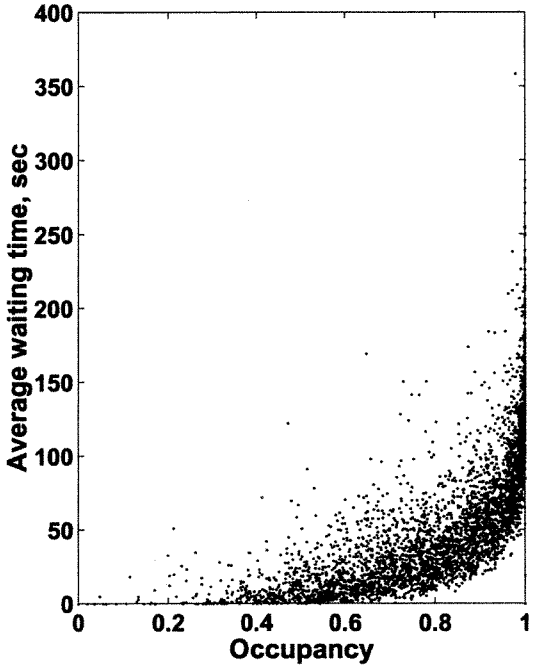
• Peak customers per hour

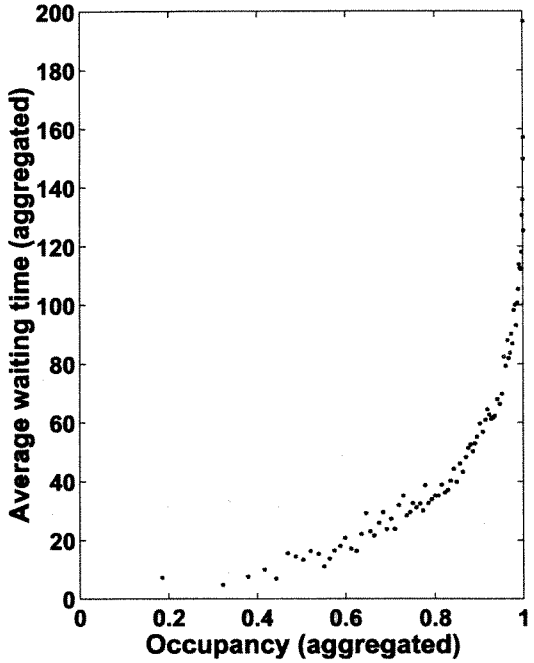
Abandonment Rate (% Arrivals = Enter)

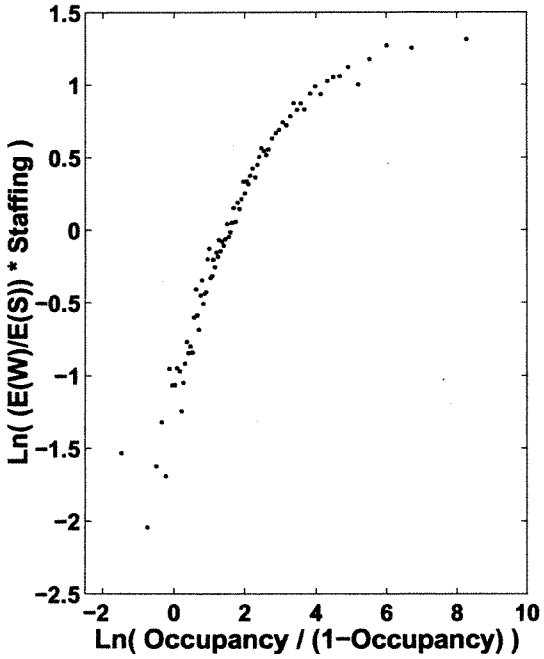
@ given load (efficiency) : 20 cust's / server / hr
service level (abandons) improves with size



Average Waiting Time vs. Agents' Occupancy (Khintchine-Pollaczek)

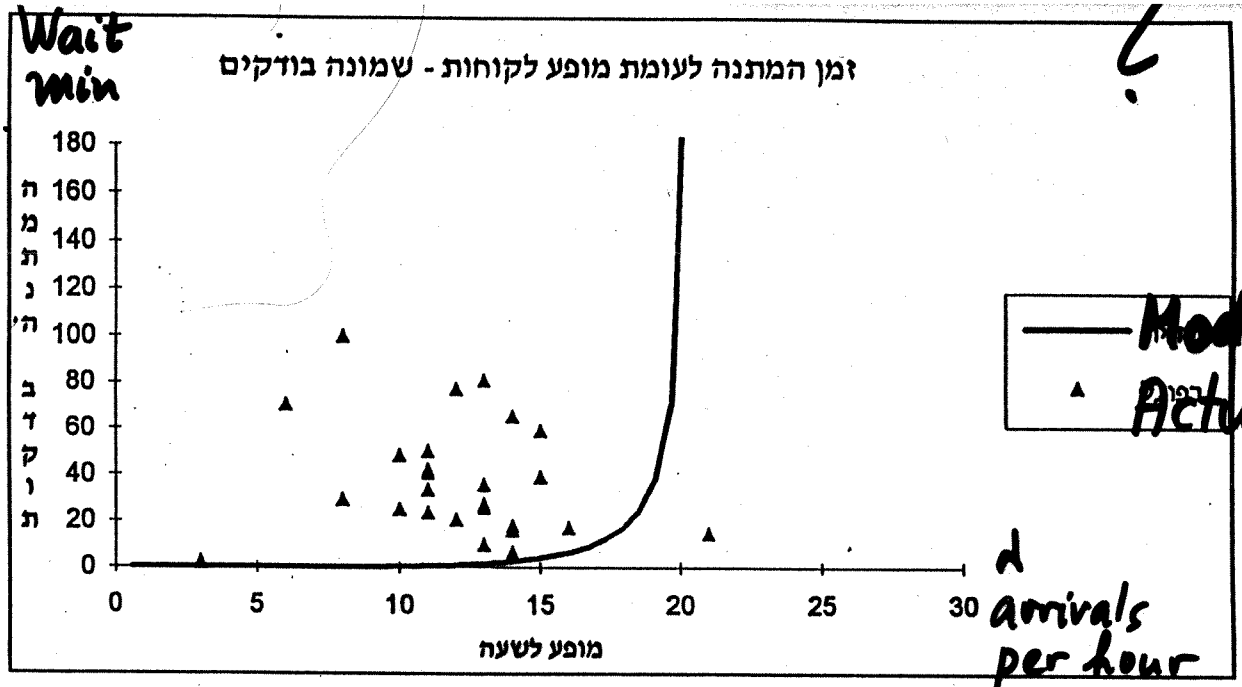






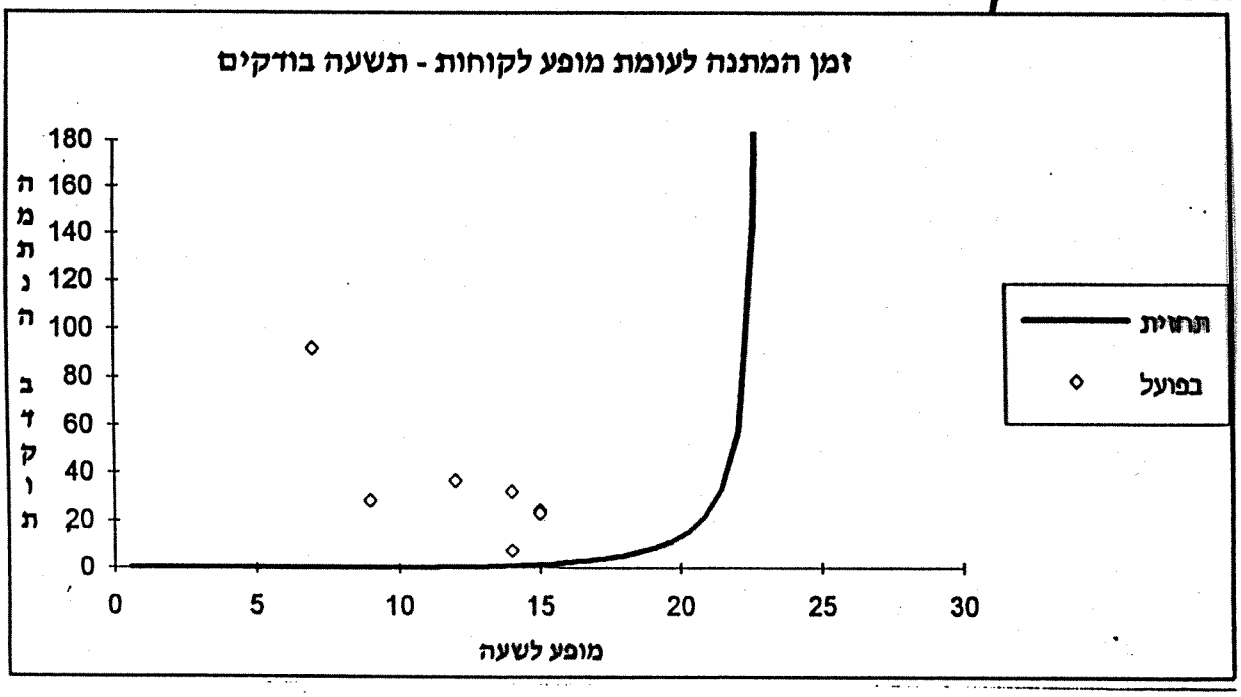
W

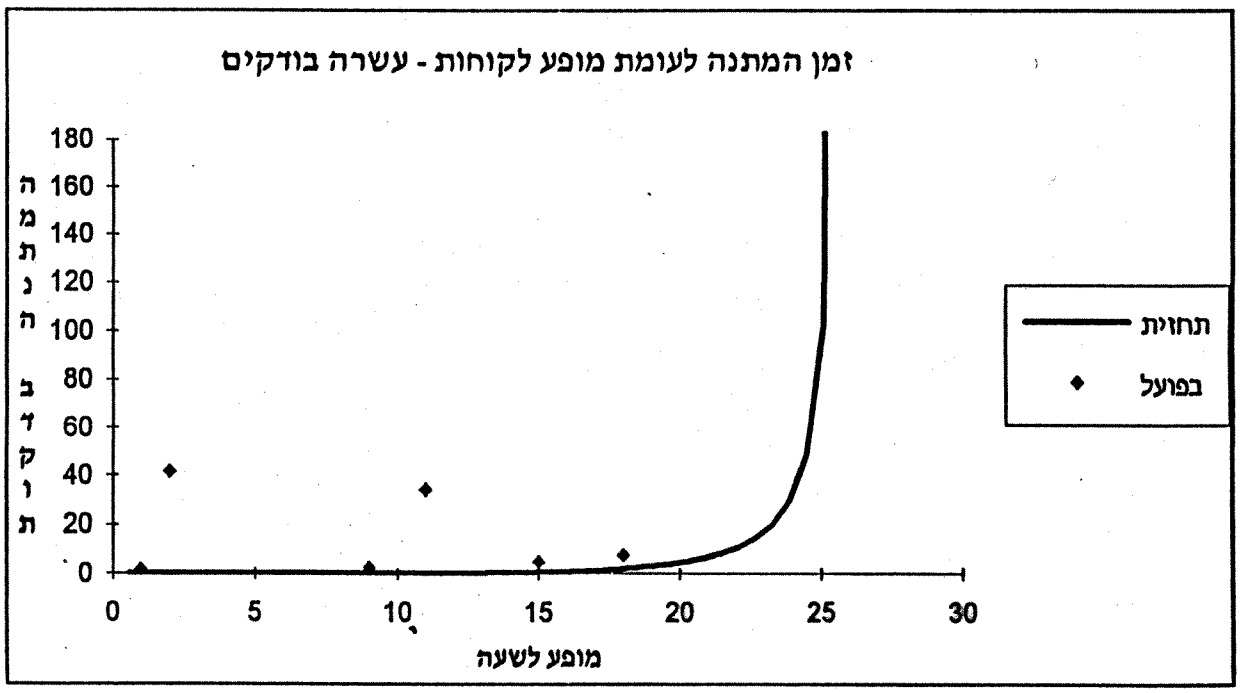
Wait
Min



Model
Actual

d
arrivals
per hour





The $M_t/M_t/1$ Queue

Arrival Time-varying Poisson

" " arrival rate $\lambda_t, t \geq 0$.

Services Think exponential(μ)

(Time-varying service rate $\mu_t, t \geq 0$.)

Stable? when periodic, $\exists$ analysis (Harrison+Lemoine,...)

Must resort to approx. : short-run

Representation $Z = \{Z_t, t \geq 0\}$ number in system

$$Z = f(X) \qquad f \text{ reflection}$$

$$X_t = \underset{\uparrow}{N_+}\left(\int_0^t \lambda_u \, du\right) - \underset{\uparrow}{N_-}\left(\int_0^t \mu_u \, du\right)$$

Std Poisson indep.

Hierarchical Modelling: Short-Run (Massey)

$M_t/M_t/1$ $\qquad X_t = N_+\left(\int_0^t \alpha_u\,du\right) - N_-\left(\int_0^t \mu_u\,du\right)$

(α_u: e.g. periodic; N_-: std. Poisson)

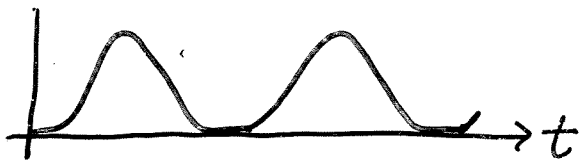
Micro $\qquad Z = f(X)$

Acceleration $\qquad Z^\varepsilon = f(X^\varepsilon)$, $\alpha_t^\varepsilon \leftrightarrow \frac{1}{\varepsilon}\alpha_t$, $\mu_t^\varepsilon \leftrightarrow \frac{1}{\varepsilon}\mu_t$, $\varepsilon \downarrow 0$

Strong Approx $\qquad X^\varepsilon \approx \frac{1}{\varepsilon}\underbrace{[\textstyle\int\alpha - \int\mu]}_{\frac{1}{\varepsilon}\chi} + W_+\left(\frac{1}{\varepsilon}\int\alpha\right) - \underbrace{W_-\left(\frac{1}{\varepsilon}\int\mu\right)}_{\frac{1}{\sqrt{\varepsilon}}W \text{ in dist.}}$

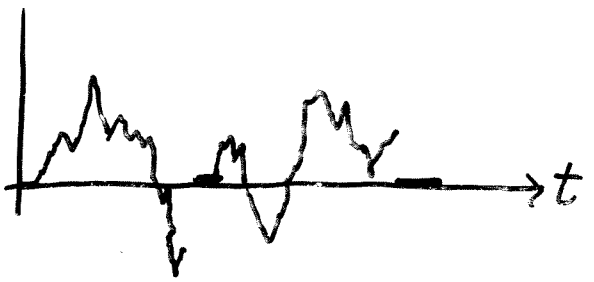
FSLLN $\qquad \varepsilon Z^\varepsilon \approx f(\chi + \sqrt{\varepsilon}W) \xrightarrow{a.s.} f(\chi) = q$ fluid

Macro $\qquad Z \sim \frac{1}{\varepsilon}q$



FCLT $\qquad \frac{1}{\sqrt{\varepsilon}}[\varepsilon Z^\varepsilon - q] \overset{M_1}{\Rightarrow} \nabla_W f(\chi) = V$ diffusion

Meso $\qquad Z \sim \frac{1}{\varepsilon}q + \frac{1}{\sqrt{\varepsilon}}V$



Phase Transitions: over → under → critical

Fluid determines time-varying phases

$q(t) > 0$ super-critical $\Leftrightarrow \rho(t) > 1$

$q(t) = 0 \quad dy > 0$ sub-critical $\Leftrightarrow \rho(t) < 1$

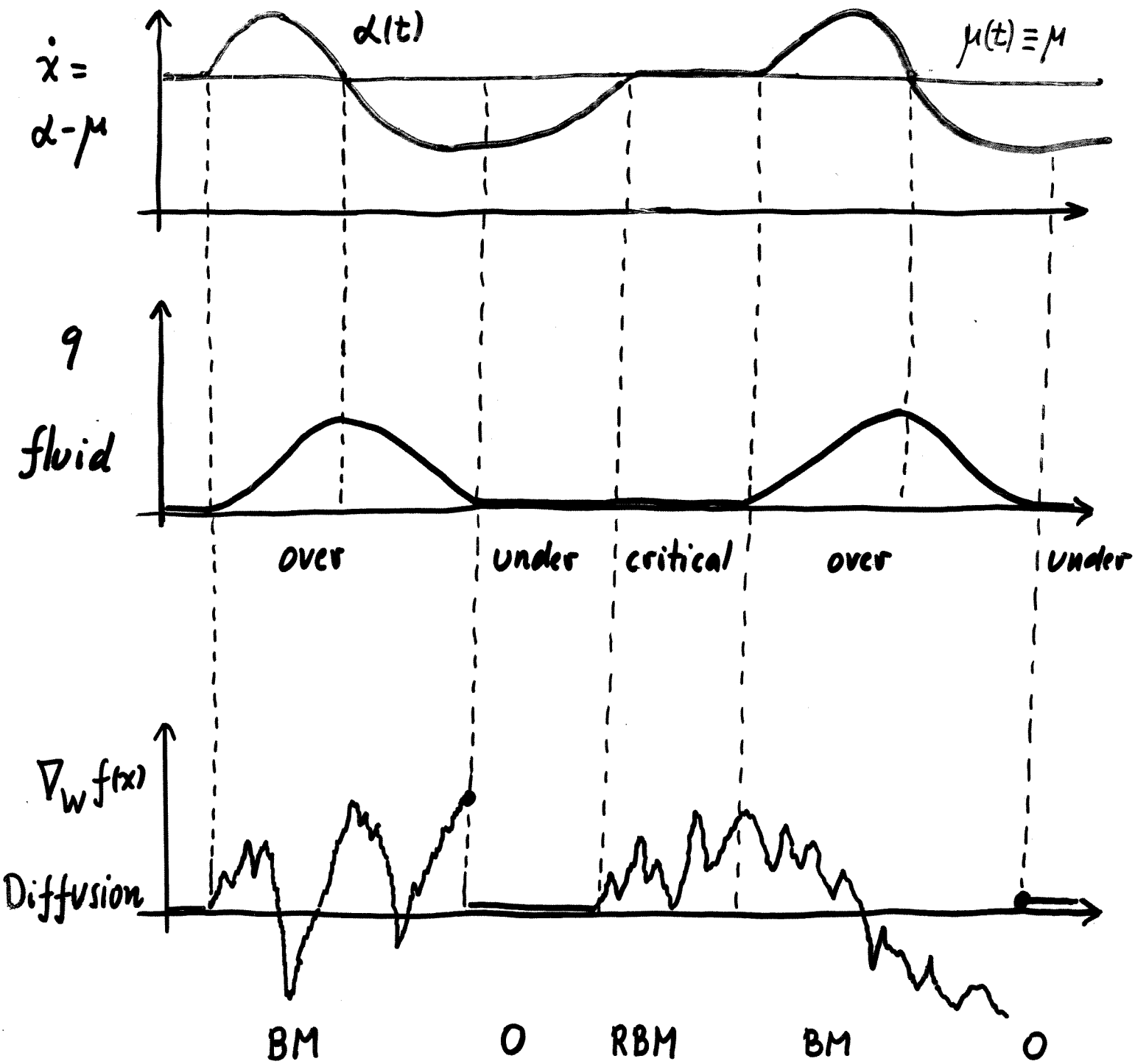
$q(t) = 0 \quad dy = 0$ critical $\Leftrightarrow \rho(t) = 1$

$$\rho(t) = \sup_{0 \le s \le t} \frac{\int_s^t \alpha_u \, du}{\int_s^t \mu_u \, du}$$

$$\left(\rho(t) \neq \frac{\alpha_t}{\mu_t} \right)$$

Diffusion approx. changes by phases

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



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.

Phase Transitions

$\rho(t) < 1$ $\rho(t) = 1$ $\rho(t) > 1$

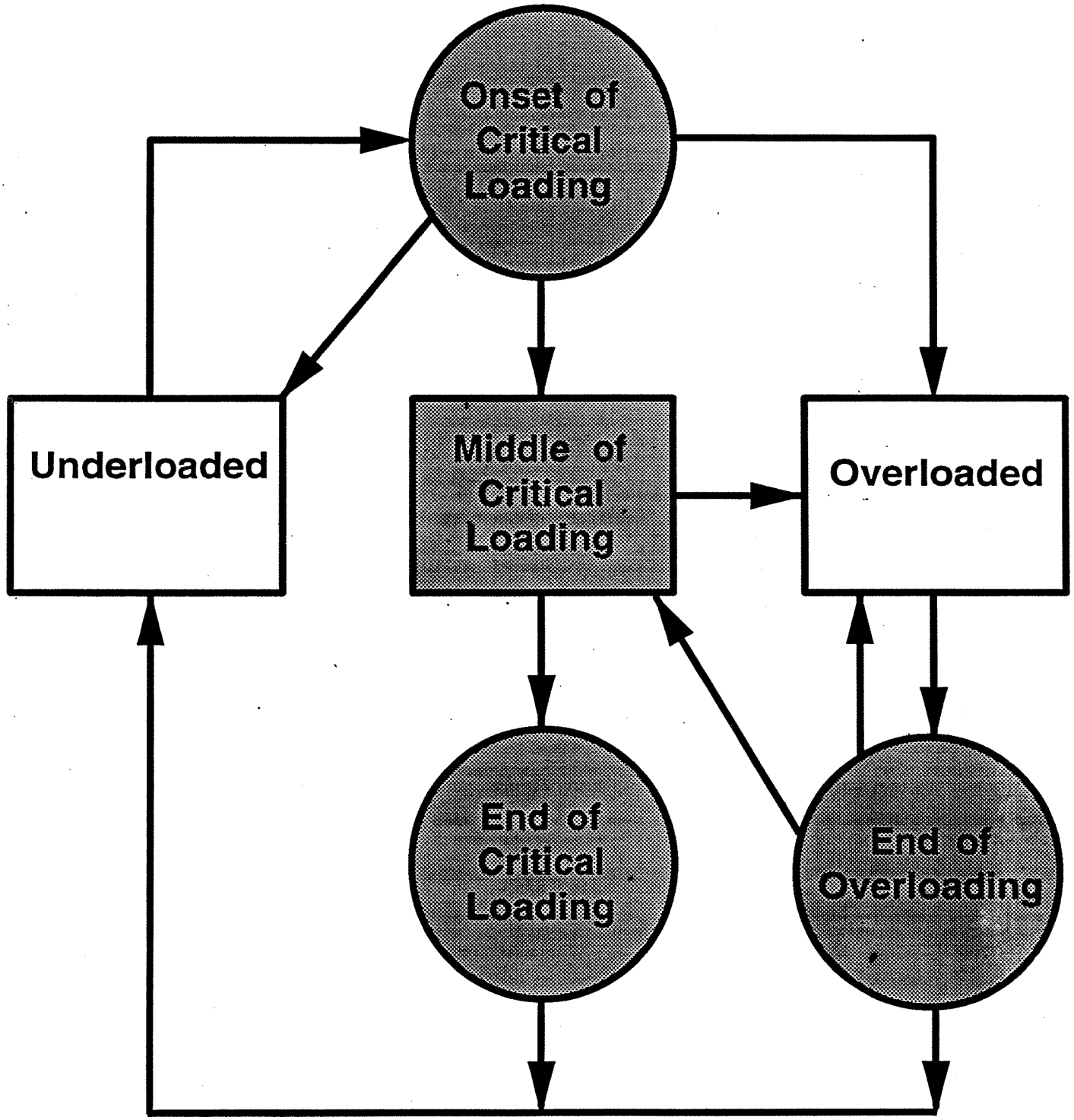


Figure 3.1: Phase transition diagram for the asymptotic regions

Refined regions for asymptotic analysis

Approximations - Examples

Middle + End of Critical Loading

$$Q \overset{d}{=} \sqrt{\varepsilon}\, V + o(\sqrt{\varepsilon}) \quad , \quad V = \nabla_w f(x)$$

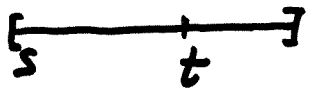
Middle : continuous, reflected BM

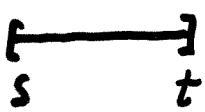
End : V left-continuous, not right-cont. wp > 0

Overloading

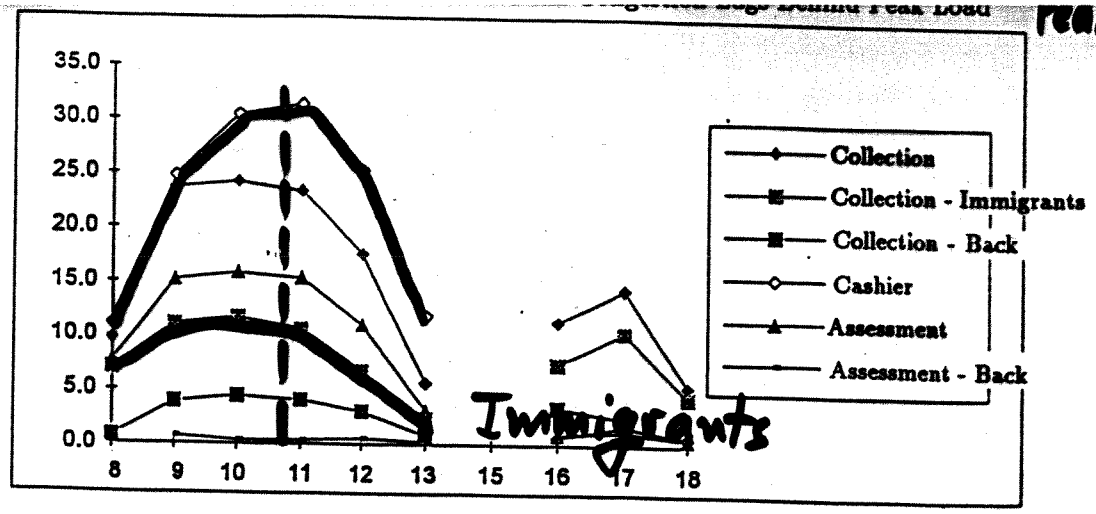
$$Q(t) \overset{d}{=} \int_{\substack{s \\ \uparrow \text{onset}}}^{t} (\alpha - \mu)\, ds + \sqrt{\varepsilon}\,[V(s) + \Delta W(s,t)] + o(\sqrt{\varepsilon})$$

Eg. Peak load < Peak congestion (queue)

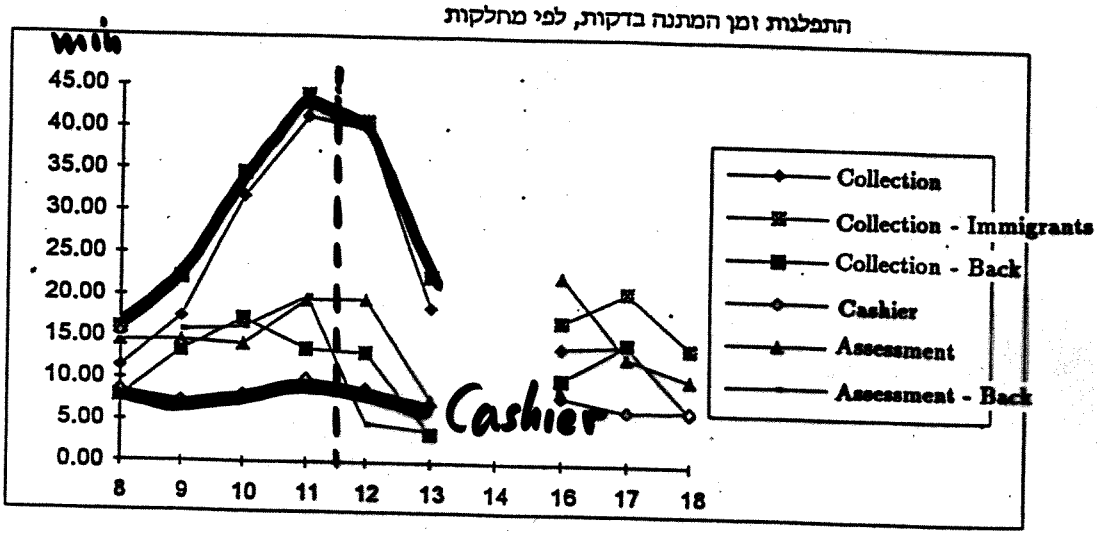
$t = \text{argmax}\ q(\cdot)$ over 

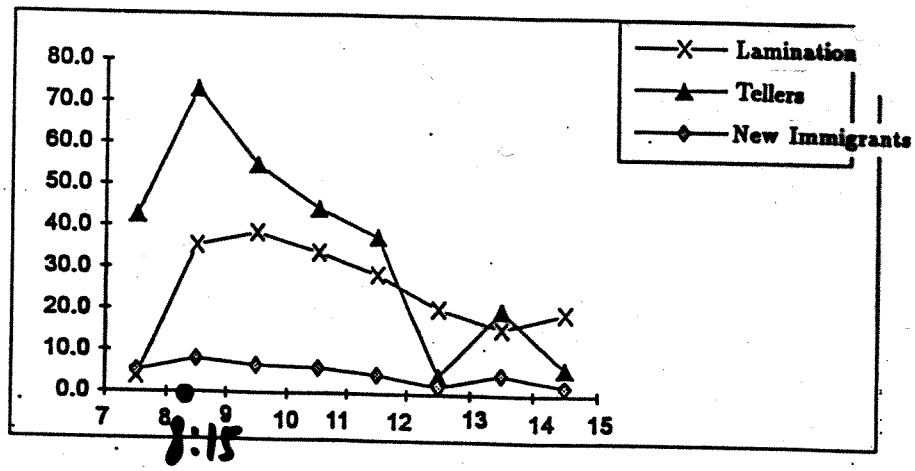
$\Rightarrow \exists\ u = \text{argmax}\ \alpha(\cdot)$ over 

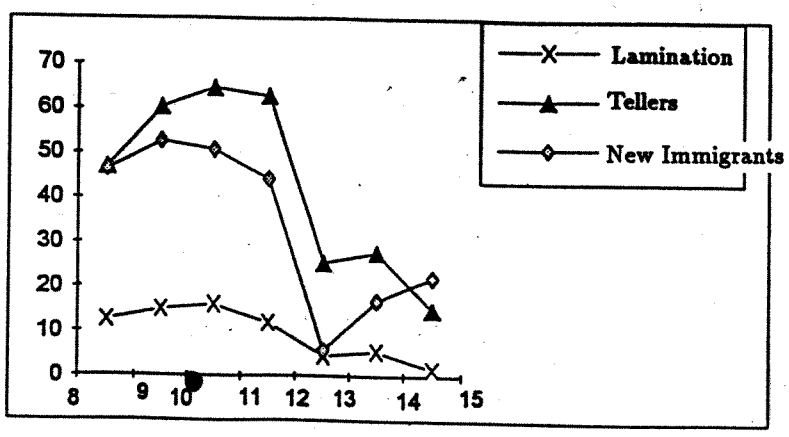
Arrival Rates: 6 Dept's



Waiting Times







Short-Run: State-Dependent

$Z : \alpha(3), \mu(3)$

$Z : \alpha^{\varepsilon}(3) = \frac{1}{\varepsilon}\alpha(\varepsilon 3), \mu^{\varepsilon}(3) = \ldots, \varepsilon \downarrow 0$ Acceleration

FSLLN $\quad \varepsilon Z^{\varepsilon} \underset{a.s.}{\to} q = f(x)$ Fluid

D.E. $\quad \dot{x} = [\alpha(q) + \mu(q)P] - \mu(q)$

FCLT $\quad \frac{1}{\sqrt{\varepsilon}}[\varepsilon Z^{\varepsilon} - q] \underset{M_1}{\Rightarrow} V = \nabla_W f(x)$ Diffusion

S.D.E $\quad dW = a(q)V\,dt + b(q)\,dB$

Phases $\quad \rho(t)$ = traffic intensity, based on q

$<1, =1, >1$

1 Station
- Repairman (Iglehart)
- Multiple Servers (Halfin, Whitt)
- Finite Population (Kogan, Krichagina, Lipster,
- Reneging (Coffman, Pukhalski, Reiman, Wright

Networks $\quad P(\text{state})$ - Routing

Teaching Note

Service Engineering

April 15, 1997
Revised July 1999, 30/1/2000

REVIEW: MARKOV JUMP-PROCESS (MJP)

MJP $X = \{X_t,\ t \geq 0\}$ on $\mathcal{S} = \{i, j, \ldots\}$ countable.
Markov property: $P_r\{X_t = j | X_r,\ r < s;\ X_s = i\} = P_{ij}(s,t), \quad \forall\, s < t,\ \forall\, i, j \in \mathcal{S}.$
Time homogeneity: $P_r\{X_{s+t} = j | X_s = i\} = P_{ij}(t), \quad \forall\, s, t,\ i, j,$ transition probabilities.

Characterization: π^0 = initial distribution and $P(t) = [P_{ij}(t)],\ t \geq 0$, stochastic.
Finite-dimensional distributions:
$P_r\{X_0 = i_0,\ X_{t_1} = i_1, \ldots, X_{t_n} = i_n\} = \pi^0(i_0) P_{i_0,i_1}(t_1) \ldots P_{i_{n-1},i_n}(t_n - t_{n-1}).$

$P(t)$: stochastic ; $P(s+t) = P(s)P(t),\ \forall\, s, t$ (Chapman Kolmogorov);
$\exists\, P(0) = I$; $\exists\, \dot{P}(0) = Q = [q_{ij}]$, infinitesimal generator $\left(\sum_{j \in \mathcal{S}} q_{ij} = 0\right)$.

Micro to Macro : $\dot{P}(t) = P(t)Q\ \ (= QP(t))$ and $P(0) = I$
Forward (Backward) equations.

Solution : $P(t) = \exp[tQ] = \sum_{n=0}^{\infty} \frac{t^n}{n!} Q^n\ ,\ t \geq 0.$

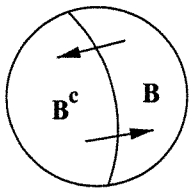
Animation: $i \xrightarrow{q_{ij}} j; \quad \forall\, i, j \in \mathcal{S}\ \exists$ exponential clock at rate q_{ij}, call it (i, j).
Given i, consider clocks $(i, j),\ j \in \mathcal{S}$; move to the "winner" when rings.
Thus: stay at $i \sim \exp(q_i = \sum_{j \neq i} q_{ij})$ and switch to j with probability $P_{ij} = q_{ij}/q_i$
$(q_{ij} = q_i P_{ij}, i \neq j; q_{ii} = -q_i)$.

Transient analysis	vs. long-run/limit	stability/steady-state
	$\exists\ \lim_{t \uparrow \infty} P_{ij}(t) = \pi_j,\ \forall\, i;$	$\pi = \pi P(t),\ \forall\, t.$

Calculation via steady-state equations: $\dot{P}(\infty) = P(\infty)Q \Rightarrow \left\{ \begin{array}{l} 0 = \pi Q \\ \sum_i \pi_i = 1,\ \pi_i \geq 0 \end{array} \right\}$
or balance equations: $\sum_{i \neq j} \pi_i q_{ij} = -\pi_j q_{jj} = \sum_{i \neq j} \pi_j q_{ji},\ \forall\, j.$

Transition rates: $\pi_i q_{ij}$ = long-run average number of switches from i to j.

Cuts: $\sum_{i \in B} \sum_{j \in B^c} \pi_i q_{ij} = \sum_{i \in B^c} \sum_{j \in B} \pi_i q_{ij},\ \forall\, B \subset \mathcal{S}.$



Hierarchical Modelling: a Framework

Strategic	F-nets	Macro	↑	Long-run,
	D-nets	Meso		Aggregation
Tactical	Q-nets	Micro		
	D-nets	Meso	↓	Short-term,
Operational	E-nets	Macro		Amplification

Anthony, R.N. "Planning and Control Systems: A Framework for Analysis", 1965

"I kept recalling the point that President James B. Conant demonstrated so vividly in *On Understanding Science*, namely The development of a framework or a conceptual scheme often has led to progress, even though the framework turns out to be wrong ... the framework will have served a useful purpose if it prepares the way for a better one."