Laws of Congestion

- **The Law for (The) *Causes* of Operational Queues**
  - Scarce Resources
  - Synchronization Gaps (in DS-PERT Networks)
  - Linear-effects of scarcity and log-effects of synchronization
- **The Laws of *Conservation***
  - Little's Law for Customers, Service-providers and Managers: $L = \lambda \cdot W$
  - Little's Law for the Offered Load (Utilization Profiles): $\rho = \frac{\lambda \cdot E[S]}{N}$
- **Laws of Completely *Random Arrivals***
  - Levy/Watanabe Axioms of Randomness
  - The Law of Poisson-Counting (Law of Rare Events)
  - The Law of Independent Memoryless (Exponential) Inter-arrivals
  - The Brownian-Law of Rescaling & Centering High-rate Arrivals
  - The Law of "Time-Changing" Time-homogeneous Arrivals
  - The Law of Accelerating Time-inhomogenous Arrivals
    (or, Smoothing out Stochastic-Variability around Predictable-Variability)
  - The Laws of Decomposition-Superposition
- **Laws of *Sampling***
  - Random Sampling: Wolff's PASTA = Poisson Arrivals See Time Averages
  - Biased Sampling: Costs of Randomness; (Coefficient of Variation; Form Factor)
- **Laws of *Human Service* Durations**
  - What is Service Duration?
  - The Theoretical Law of Phase-Type Durations
  - Empirical Laws of Exponential or Log-Normal Service Durations
  - The Law of Consistent Incentives: "Abandoning" Service-providers
- **Laws for *Service Systems with Abandonment***
  - The Law of the "Fittest-survive" (and Wait Less – Much Less);
  - The Linear Law of Abandonment-rates for Casual/Uninformed Customers;
  - Palm's Law of Irritation (Survival-functions and Hazard-rates);
  - (The) Impatience/"Loyalty" Index;
  - The Law of Information-shocks
    (or The Phases of Patience: Optimism, Facing Reality, Accepting Reality)
    (or The Phases of Patience: Customers' Heterogeneity);
  - The Adaptivity/Learning Cycle (Anticipation, Experience, Perception,...).

- **The Two-moment Law for *Average Congestion*, in Efficiency-Driven Systems**
  - Congestion Index (Efficiency vs. Quality, in the face of Stochastic Variability.)

$$\frac{E[W_q]}{E[S]} = \frac{E[L_q]}{N \cdot \rho} \approx \frac{\rho}{(1-\rho)} \cdot \frac{C_a^2 + C_s^2}{2} \times \frac{1}{N}$$

  - Khintchine-Pollaczek (Exact in $M/G/1$; $\rho = P\{W_q > 0\}$, "but only in numerator")
  - Allen-Cunneen Approximation, for "not-too-many" E-Driven Servers (GI/GI/N)

$$E_{GI/GI/N}[W_q] \approx E_{M/M/N}[W_q] \cdot \frac{C_a^2 + C_s^2}{2} = E[S] \times \frac{P_{M/M/N}\{W_q > 0\}}{(1-\rho)} \cdot \frac{C_a^2 + C_s^2}{2} \times \frac{1}{N}$$

- **The Invariance *Exponential* Law for Long Delays**
  - Kingman's Exponential Law for the Distribution of Delay
  - "80:20 Rules": Tails of The Delay-Distribution in Efficiency Driven Operations
- **The Law of "Simplicity"**: Simple Theoretical Models describe Ideal Robust Realities.
- **QED Q's** (= Quality and Efficiency Driven Queues).

Little's Law for the Offered Load (Erlangs)

Copy of Summary Interval - Order PK

Printed: 7/18/97 10:06:25 AM

Date: 7/7/97

Split/Skill: Order PK

Time	Avg Speed Ans	Avg Aban Time	ACD Calls	Avg ACD Time	Avg ACW Time	Aban Calls	% ACD Time	% Ans Calls	Avg Pos Staff	Calls Per Pos	%Serv Lev	%Aux Time	%ACW Time	%ACD Time
Totals	:00:02	:00:28	10456	:03:47	:00:25	46	53	98	70	149		8		
12:00 AM*	:00:00	:00:00	26	:04:31	:00:02	1	76	51	7	4	51	2	16	61
12:30 AM*	:00:03	:04:10	14	:07:27	:00:33	1	89	52	5	3	48	1	26	63
1:00 AM*	:00:00		9	:04:54	:11:29	0	91	90	1	7	90	0	26	65
5:30 AM*			0			0	0		0	0		33	0	0
6:00 AM*	:00:00		12	:03:21	:00:19	0	21	100	7	2	100	9	2	19
6:30 AM*	:00:00		27	:02:51	:00:20	0	32	100	14	2	100	5	3	29
7:00 AM*	:00:00		62	:03:34	:00:15	0	38	100	21	3	100	13	4	34
7:30 AM*	:00:00		93	:03:11	:00:34	0	36	100	30	3	100	7	4	32
8:00 AM*	:00:00		120	:03:37	:00:40	0	39	100	47	3	100	8	6	33
8:30 AM*	:00:00		193	:03:04	:00:14	0	44	100	61	3	100	10	7	37
9:00 AM*	:00:01		293	:03:25	:00:25	0	54	99	75	4	97	9	7	47
9:30 AM*	:00:02	:00:06	381	:03:45	:00:22	2	60	97	91	4	93	8	8	52
→ 10:00 AM*	:00:02	:00:01	416	:03:49	:00:26	1	63	97	94	4	96	5	8	55 ←
10:30 AM*	:00:00		349	:03:35	:00:33	0	52	99	96	4	99	6	8	44
11:00 AM*	:00:00		352	:03:50	:00:27	0	51	100	102	3	100	7	6	45
11:30 AM*	:00:00		349	:03:44	:00:18	0	49	100	97	4	100	8	5	45
12:00 PM*	:00:01		354	:03:59	:00:18	0	52	95	95	4	95	8	5	47
12:30 PM*	:00:00		336	:03:38	:00:21	0	52	99	97	3	99	9	6	46
1:00 PM*	:00:00		347	:03:53	:00:32	0	51	99	98	4	99	11	8	44
1:30 PM*	:00:00		368	:03:52	:00:14	0	56	99	99	4	99	11	7	50
2:00 PM*	:00:01		393	:03:55	:00:17	0	51	100	106	4	100	10	5	46
2:30 PM*	:00:00		403	:03:58	:00:13	0	54	100	112	4	100	10	4	50
3:00 PM*	:00:00	:00:04	410	:04:02	:00:16	1	57	98	110	4	98	8	5	51
3:30 PM*	:00:00		347	:03:59	:00:14	0	60	100	100	3	100	7	5	45
4:00 PM*	:00:00		382	:03:48	:01:37	0	54	100	98	4	100	6	7	47
4:30 PM*	:00:00		379	:03:41	:00:19	0	55	99	97	4	99	8	5	50
5:00 PM*	:00:00		411	:03:53	:00:19	0	53	100	109	4	100	9	5	48
5:30 PM*	:00:01		387	:03:58	:00:19	0	58	99	96	4	99	10	6	51
6:00 PM*	:00:01	:00:21	371	:03:28	:00:25	1	53	98	91	4	98	9	6	47
6:30 PM*	:00:00		280	:03:26	:00:13	0	41	100	90	3	100	8	4	37
7:00 PM*	:00:00		269	:03:24	:00:17	0	42	100	78	3	100	9	5	38

$$\rho = \frac{\lambda \cdot \bar{S}}{N}$$

$$= \frac{416 \times (3{:}49 + 0{:}28)}{94}$$

$$= \ldots$$

ACD: PEC

58

Congestion Index : $\frac{EW_q}{ES} = \frac{\bar{L}_q}{N\rho}$ ← observable   משוב

מעריב 25/II/92

כך דופקים את האזרח

משרד הרישוי חולון

מחכים שעתיים בשביל 19 שניות

מאת אבי פלד

[illegible] למשרד הרישוי בחולון [illegible] מקונה ישראלי [illegible] מבחן קילומטרז' [illegible] נמשכת 45 דקות. בשעה 08.09 נמצאים באולם ההמתנה כ-60 איש. באשנב המודיעין יושב פקיד [illegible] לבדיקת ההיסטוריה של מכונית שאני רוצה לקנות?

"תעודת זהות, רשיון רכב או ייפוי כוח, באשנב שלוש" – ענה הפקיד המיומן.

אני לא מכיר את בעל הרכב ואני רוצה לבדוק אם הרכב מעוקל.

"תעודת זהות, רשיון רכב או ייפוי כוח, באשנב שלוש", ביקלם את התשובה באותה אדישות.

אבל...

"תעודת זהות, רשיון רכב או ייפוי כוח, באשנב שלוש", חזר הפקיד תקליט. צריך להבין, באשנב שלוש אין סיכוי לקבל מידע על רכב, אם אין לך את רשיון הרכב.

בשעה 08.17 נפל המחשב במשרד הרישוי. אחרי שבע דקות נשמעה במערכת הכריזה ההודעה הבאה: "עקב תקלה במחשב יופסק מתן השירות. לא ידוע מתי יתחדש. ניתן לקבל הארכת תוקף לרשיונות. אתכם הסליחה". הפקידות המשועממות קפצו על ההזדמנות [illegible] לה חולב עם אמא או עם חברות.

בתור לאשנב שלוש, שבו צעיר שתפס מקום כבר ב-07.30 בבוקר, [illegible] הוא הלך [illegible] עצמו) מקרב הממתינים, הודיע שהשירות לא יתחדש עד הצהריים. אחד "מבין במחשבים", הסביר ש"צ'וק אחד התבלבל".

אבל ב-09.18 החל המחשב לגלות סימני חיים. הבשורה המשמחת התפשטה במהירות והמונים הצטרפו לתור שבסופו מנפיקים כפל רשיון נהיגה.

לאחר 42 דקות הגעתי, חשבון פשוט גילה שטיפול באדם נמשך בממוצע כארבע דקות. הצגתי תעודת עיתונאי וביקשתי כפל רשיון נהיגה. בתוך 19 שניות, על השעון, עזבתי את המקום עם המסמך המבוקש. שיפור ניכר ביחס לממוצע.

ובעניין בירור ההיסטוריה של המכונית: מלשכת דובר משרד התחבורה נמסר, שאין מוסרים פרטים על רכב אלא ברשות בעליו, מחשש לפגיעה בצנעת הפרט.

סקירה 19 שנ'

גישה

תור התחלה 60 איש

יחס

מחשב נפל

והמטרה הנה עדיין לא מספרת את "איכות המתנה" ו"אי־גבולות", משמעות ואותו ה[illegible]

- חוסר ודאות | [illegible]

- [illegible] בקריה [illegible]

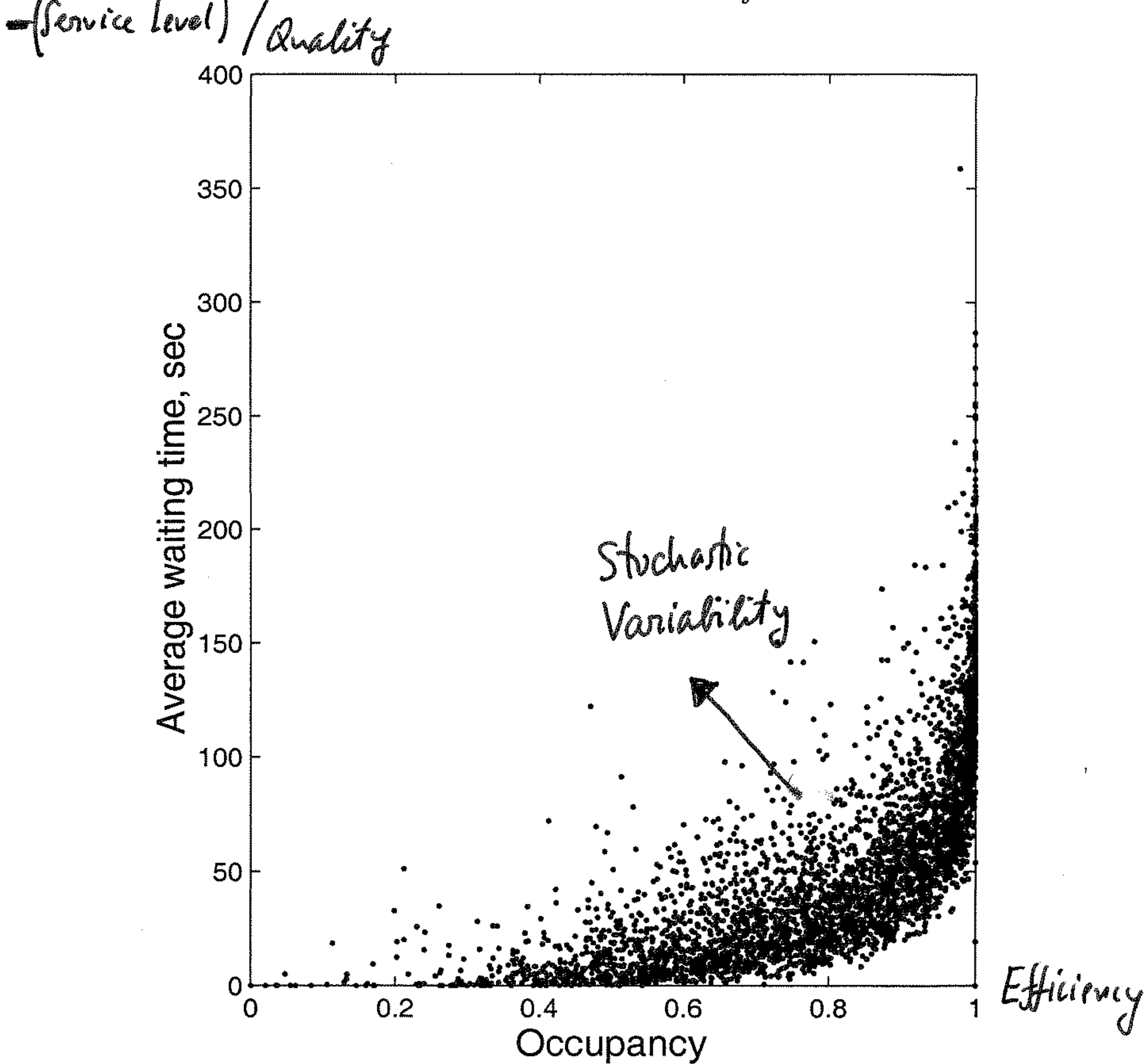
6

2

The Efficiency - Quality Tradeoff

Congestion Curves
(Empirical Proof of Khinchine-Pollatcheck Formula)

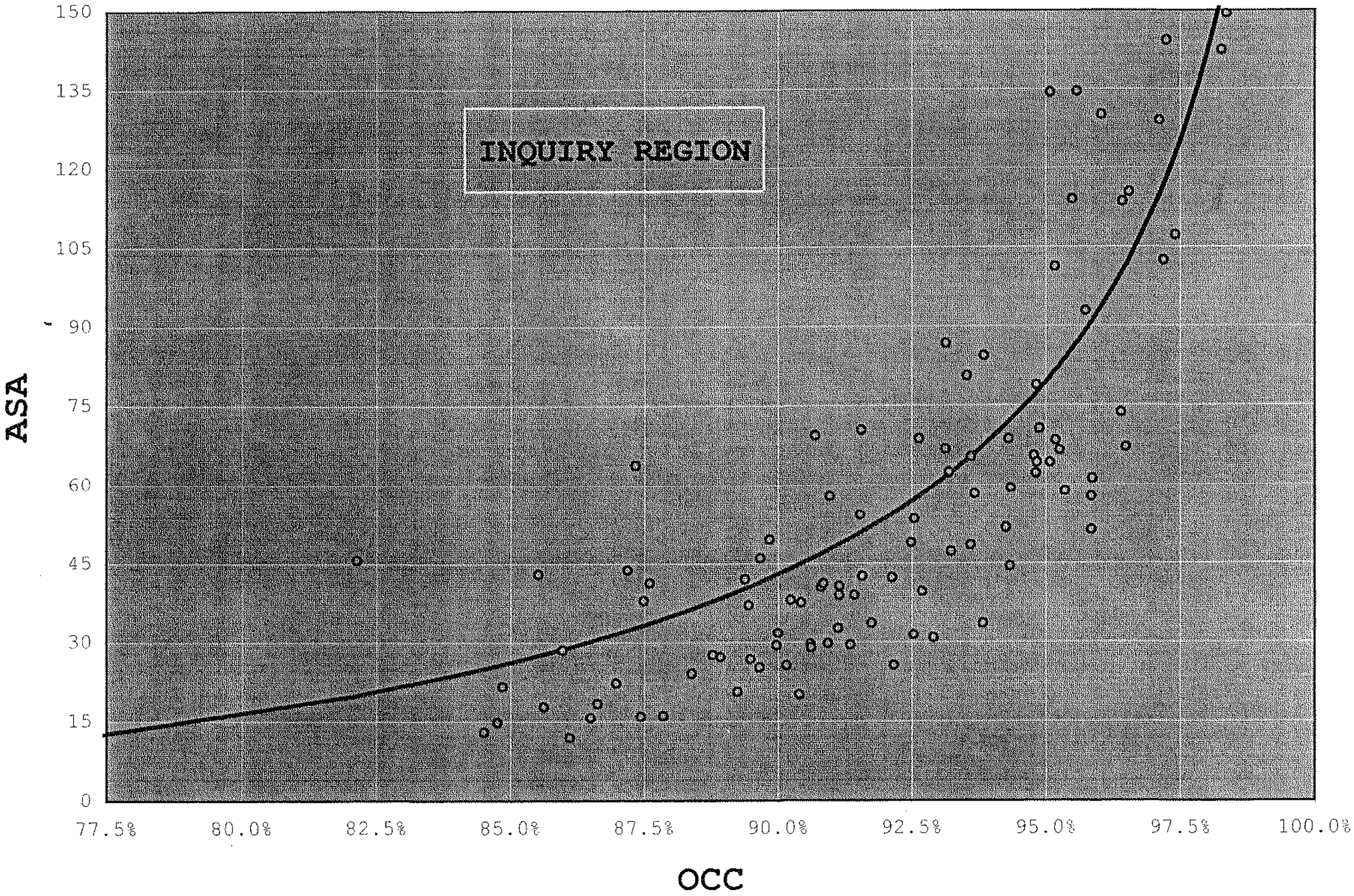
Service Level vs. Availability



The 2nd Law:

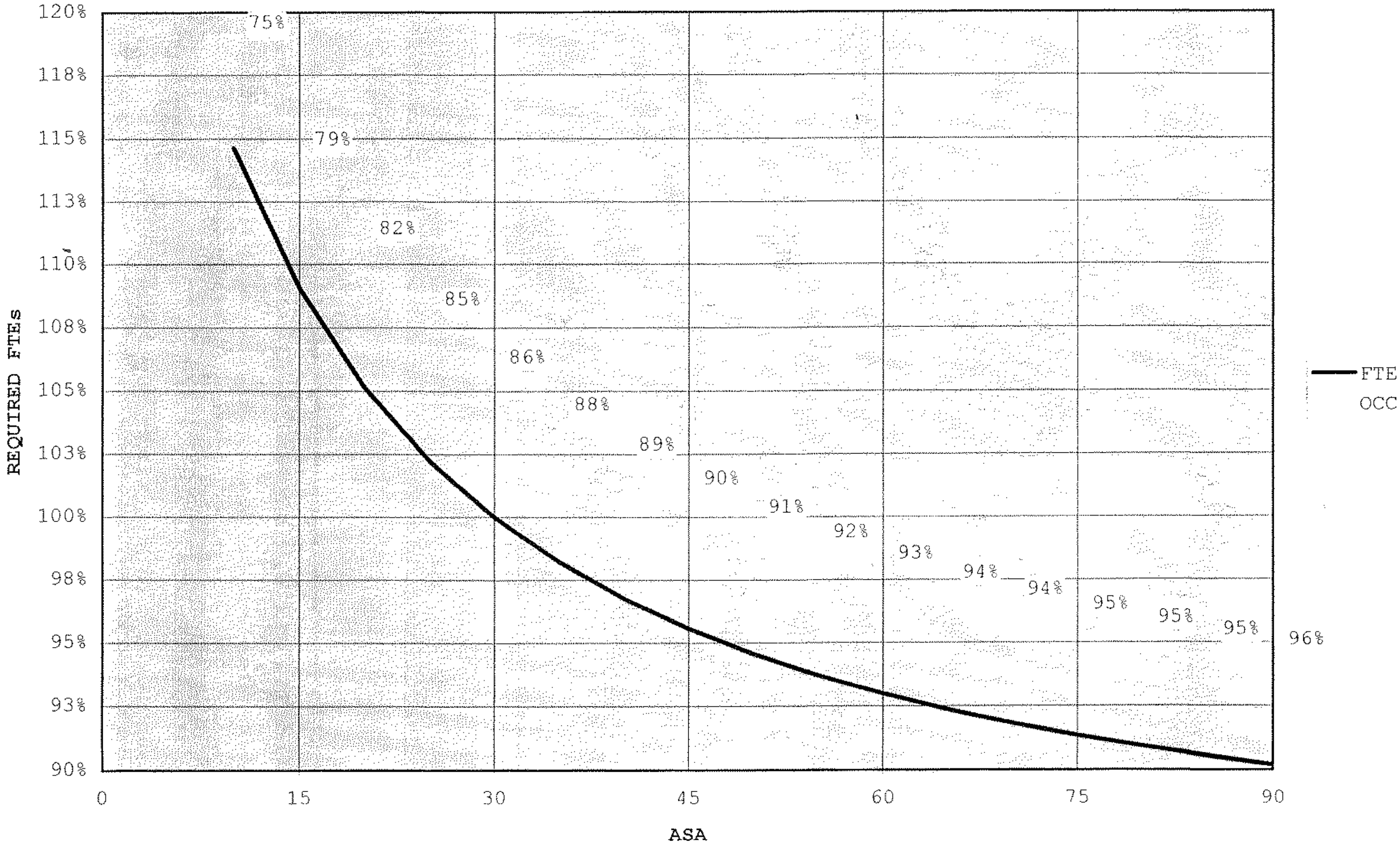
Congestion Index: $\frac{E(W_q)}{E(S)} \approx \frac{1}{N} \frac{\rho}{1-\rho} \frac{C_a^2 + C_s^2}{2}$ (N = number of servers)

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

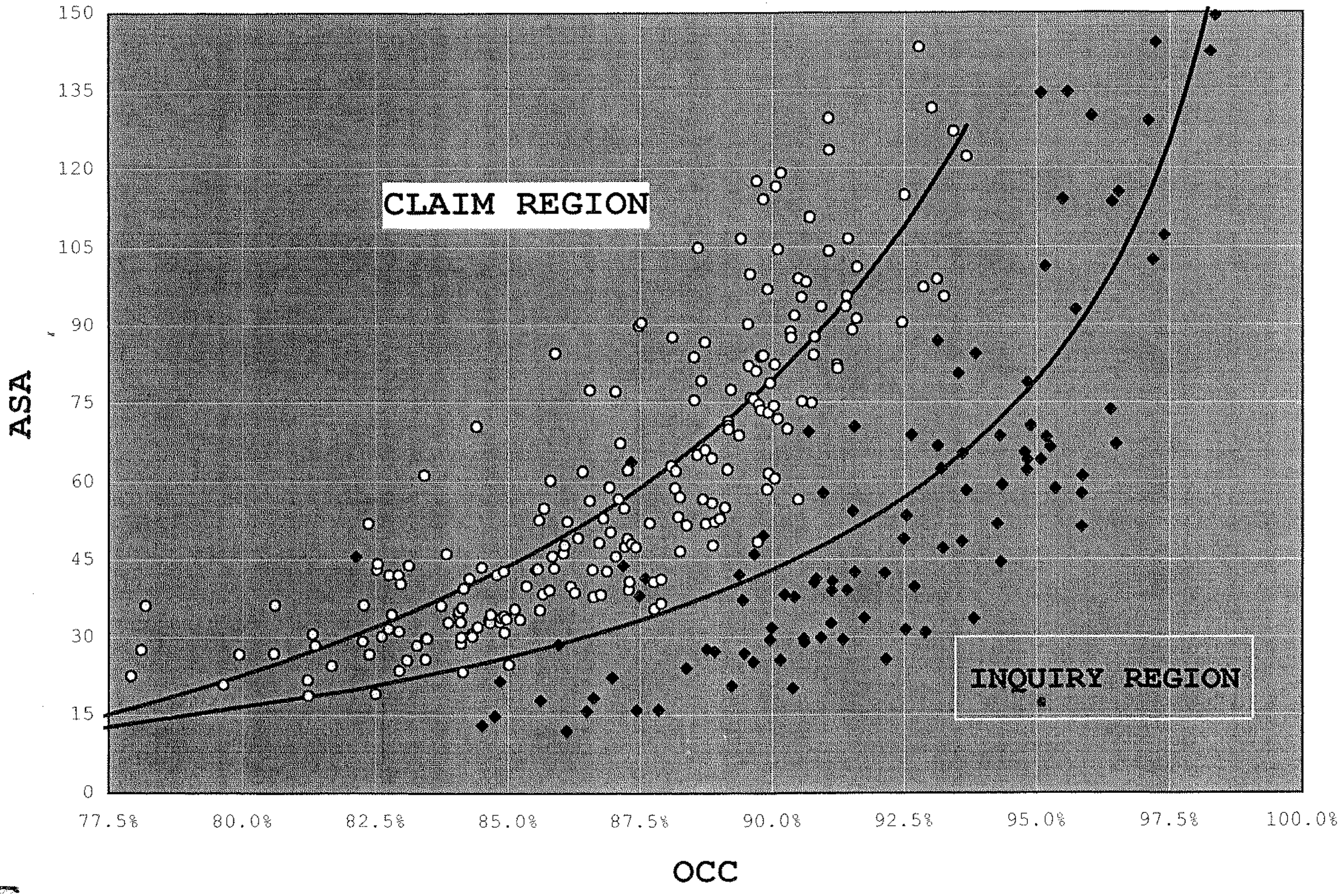


INQUIRY REGION
ASA
150
135
120
105
90
75
60
45
30
15
0
77.5%
80.0%
82.5%
85.0%
87.5%
90.0%
92.5%
95.0%
97.5%
100.0%
OCC

INQUIRY REGION



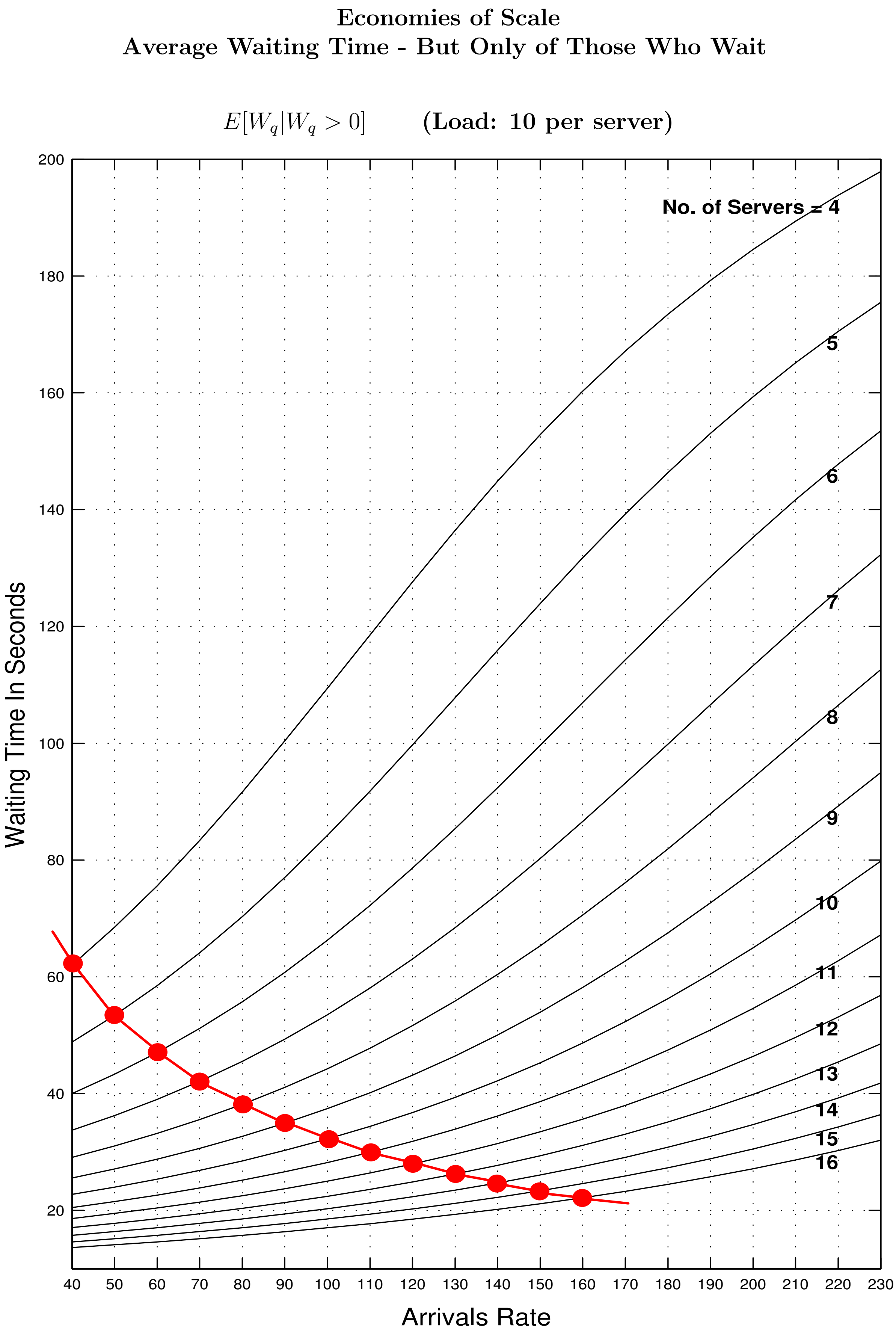
K-P / A-C Law (2 moments; performance averages)



$$\frac{\overline{W_q}}{\overline{S}} \approx \frac{1}{N} \cdot \frac{\rho}{1-\rho} \cdot \bullet$$

index — efficiency — ?

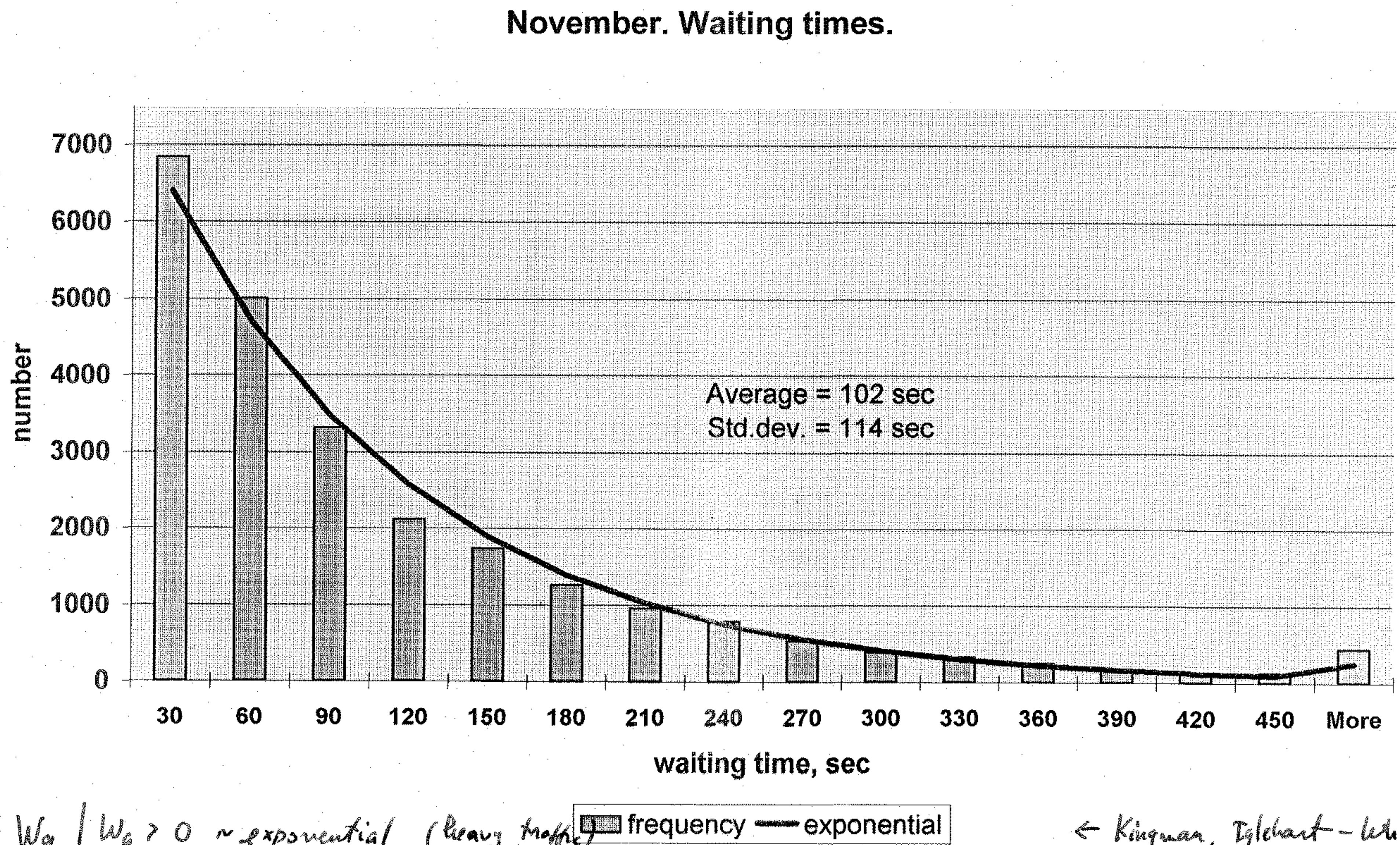
Theoretical Congestion Curves: Staffing Tools (4CallCenters)



Kingman's Exponential Invariance Law for the Distribution of Congestion waitcha :

The 3rd Law:

$$P(W_q > T \mid W_q > 0) \approx e^{-T/\bar{W}_q}$$

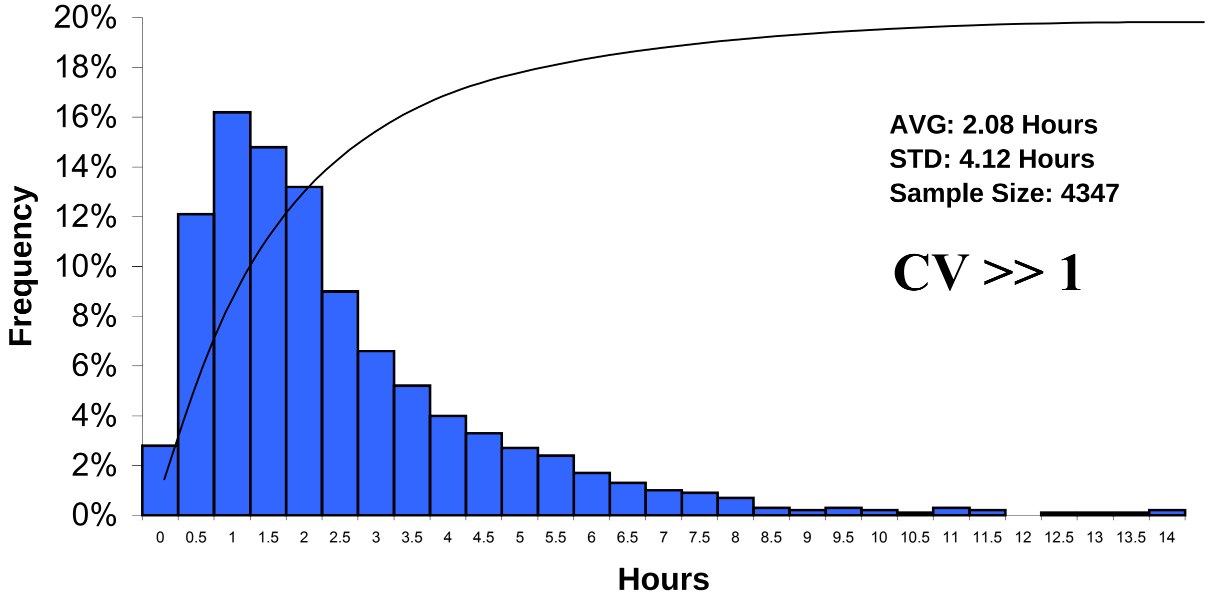


- $W_q \mid W_q > 0 \sim$ exponential (heavy traffic) ← Kingman, Iglehart–Whitt, ...
- $\exists\, \eta, \alpha \ni e^{\eta x} P(W_q > x) \to \alpha$, as $x \uparrow \infty$. (Exponential decay) ← Whitt 93, §4.2

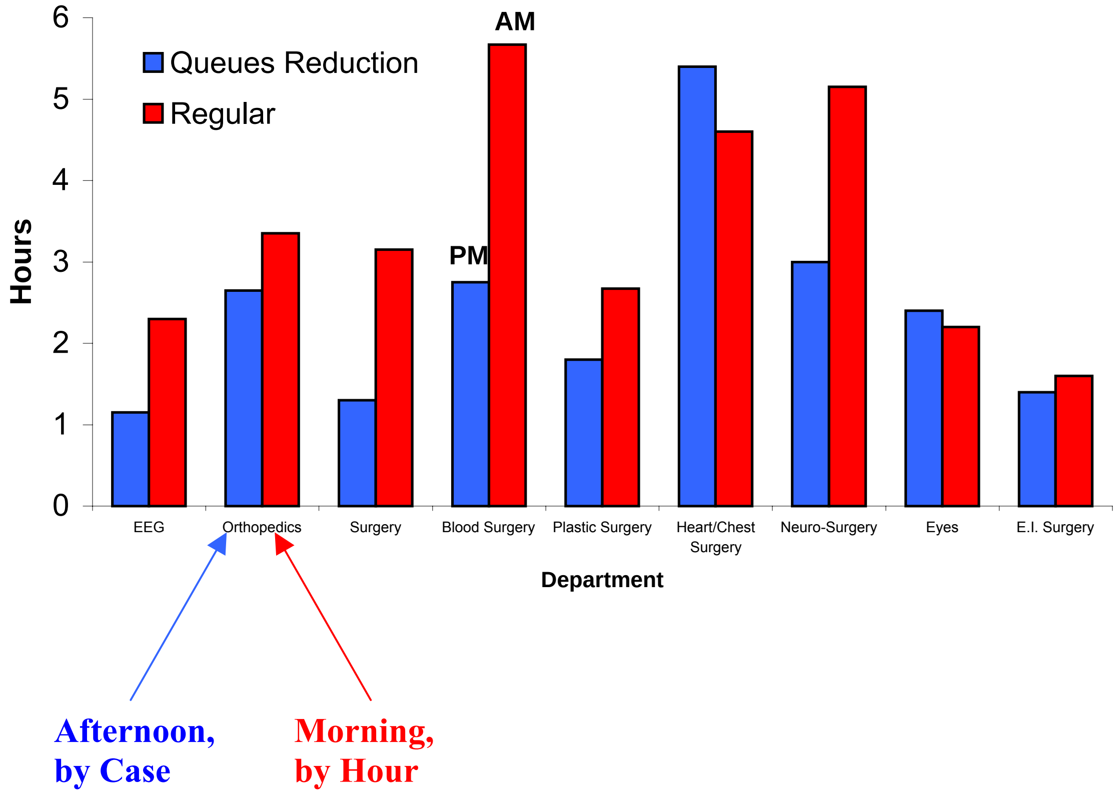
What is Service Time / Duration ?

Operations Time In a Hospital

Operations Time Histogram:



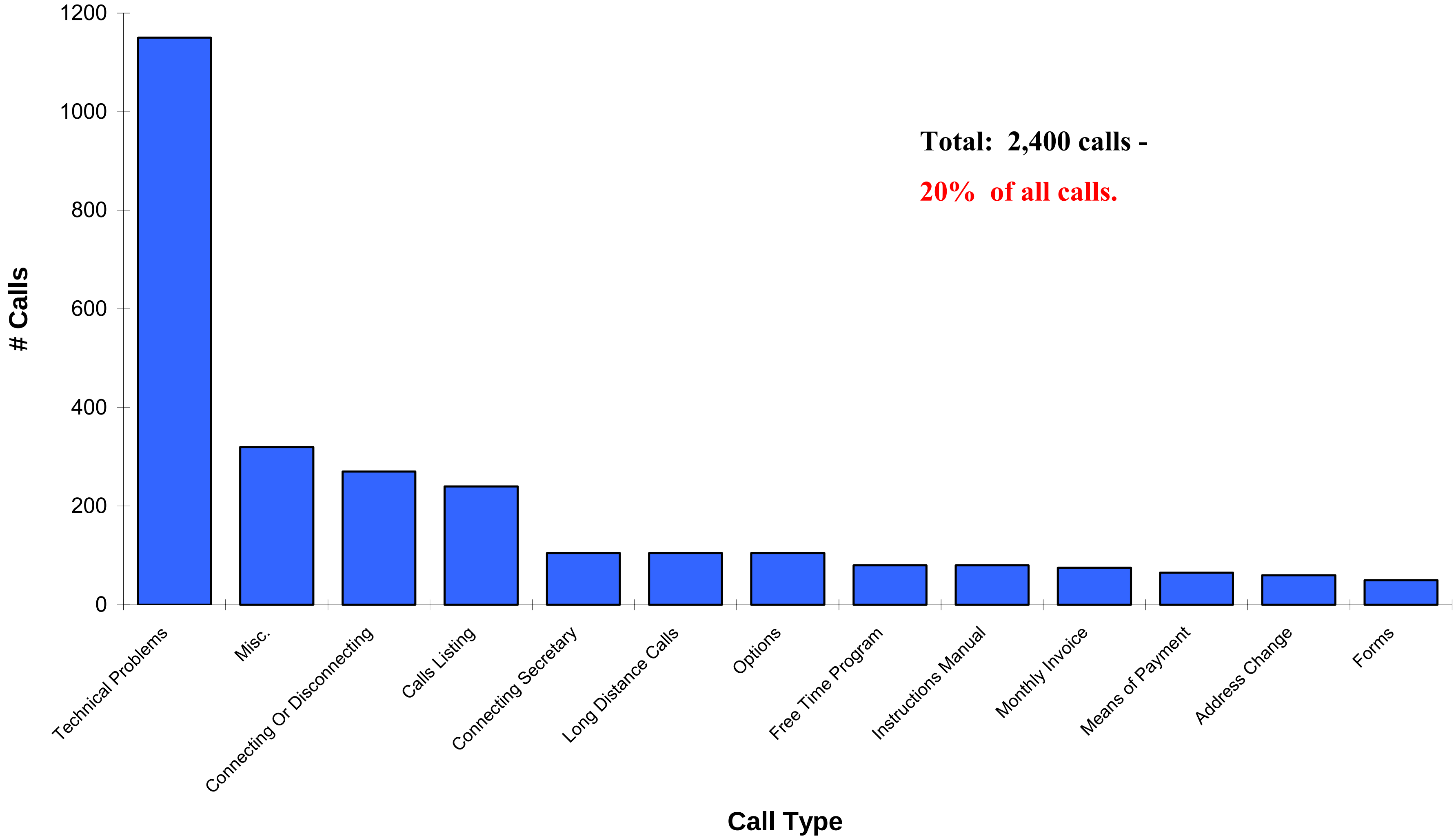
Operations Time - Morning vs. Afternoon:

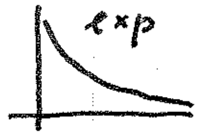


Ethical?

Even Doctors Can Manage!

What is "Service Time"?
Bank Classification of "Continued – Calls"

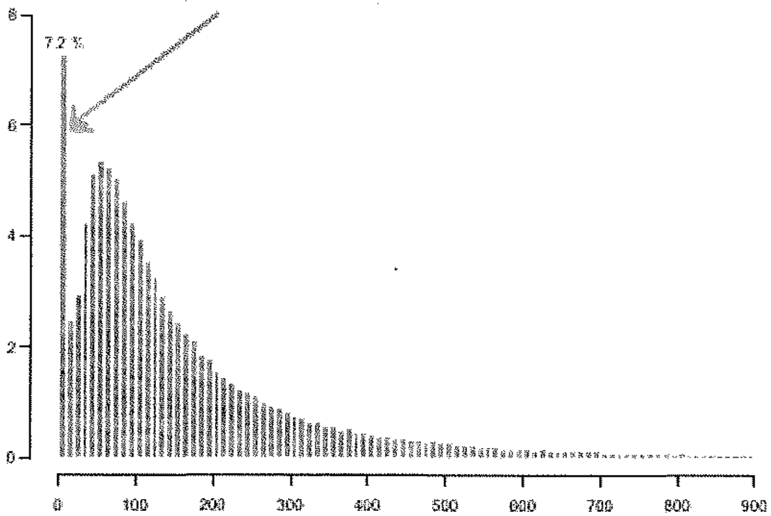


Service Times: 

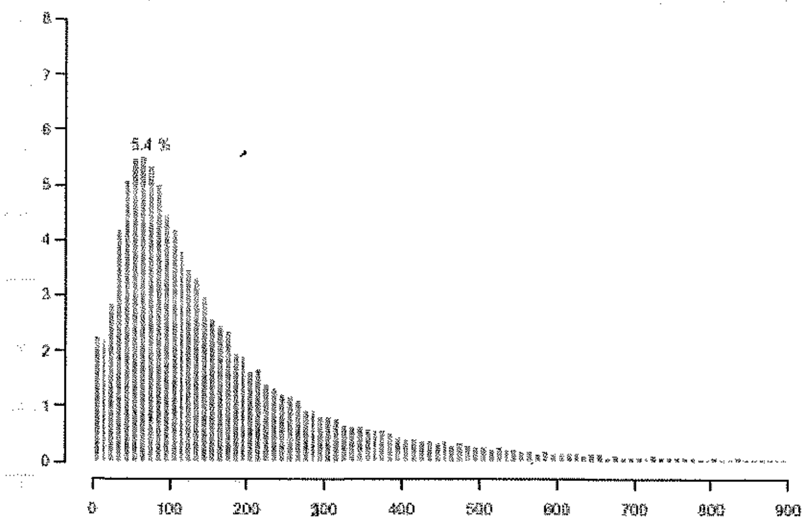
(why? when?)

Short Service Times

Jan-Oct



Nov-Dec



The Law of Consistent Incentives

The Fittest Survives and Waits less – Much less

Rationalized staffing $\Rightarrow$ Abandonments

Abandonments Prevail (10–40%)

Abandonments Matter! Service Level

Economics

E.g. $M/M/N$: $\lambda = 48$, $\mu = 1$, $N = 50$

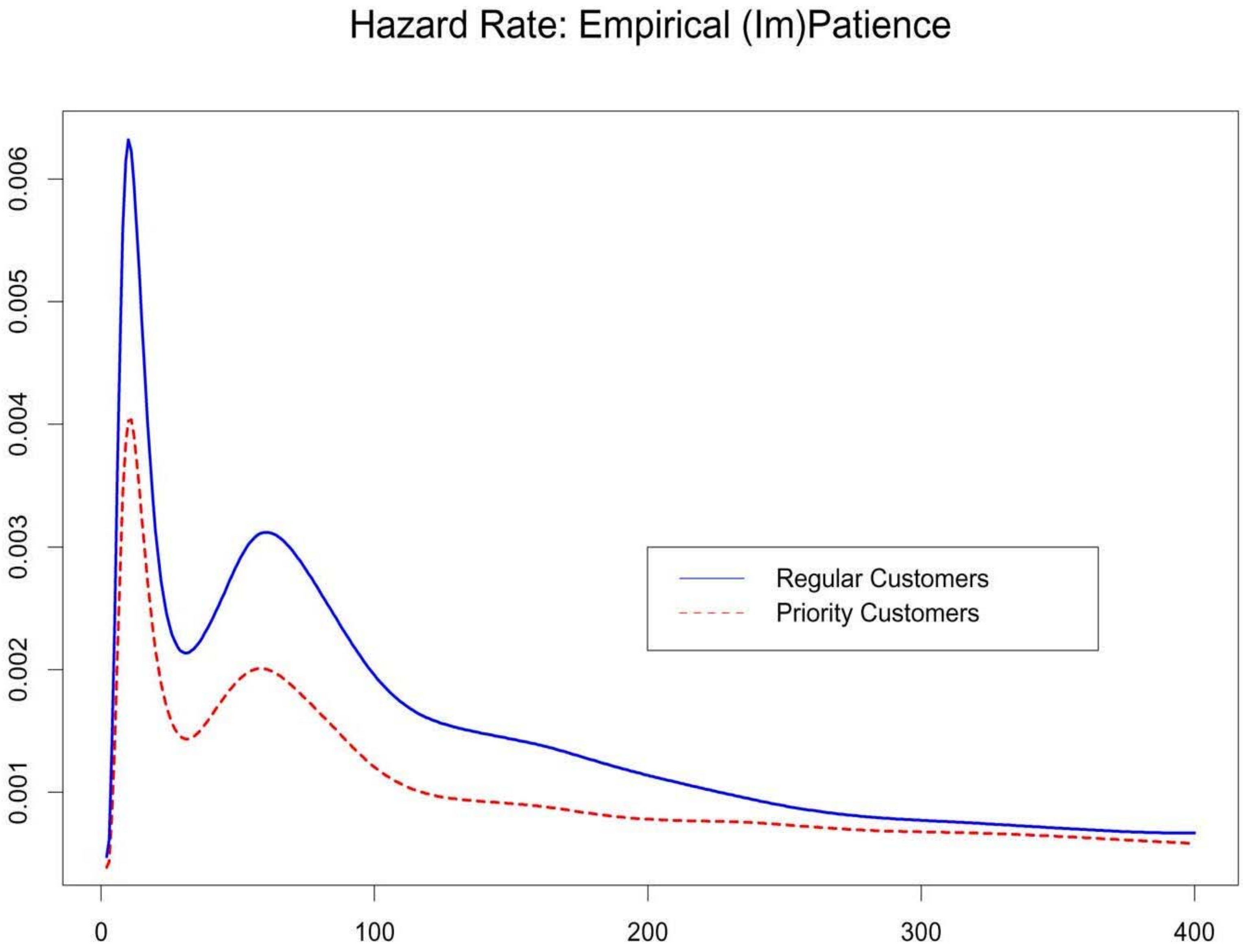
vs. $M/M/N$ + exponential patience, mean = 2 min.

	$M/M/N$	$M/M/N+M$
Fraction abandoning	–	3.1%
E[Wait]	20.8 sec.	3.7 sec.
90% percentile	58 sec.	12.5 sec.
E[Queue]	17	3
Agents' utilization	96%	93%

What if $\lambda = 50$? Robustness (vs. M/M/N with 3.1% less arrivals)

16

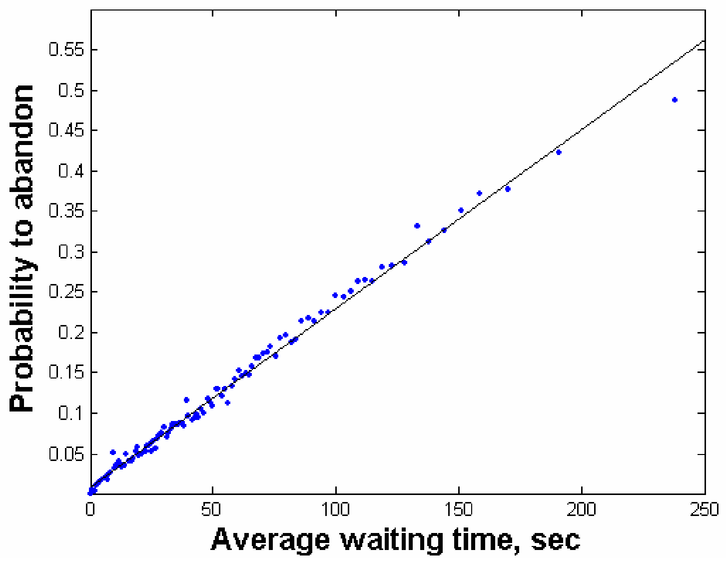
Palm's Law of Irritation: $I_t \propto h_R(t)$

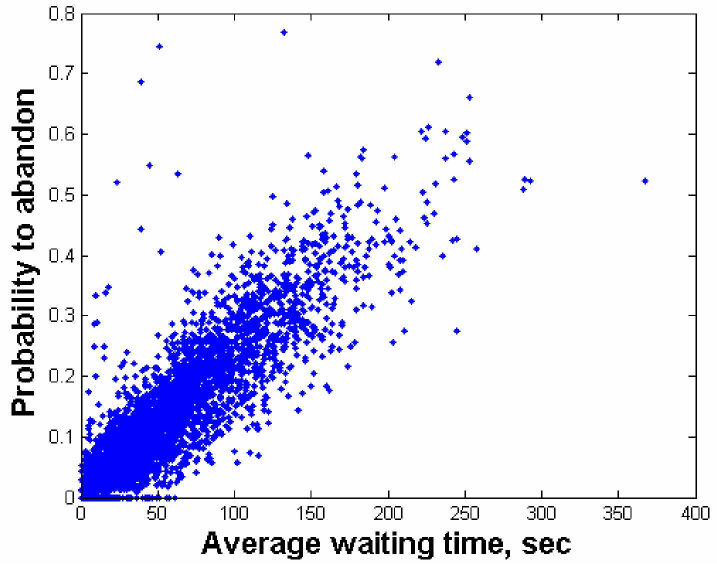


Empirically-Based Theory

Linear pattern observed: P{Abandon} = $C \bullet$ E[Wait]

Theory: Average Patience = 1/C in Erlang-A, else?





PATIENCE INDEX

- How to Define? Measure? Manage?

Statistics	Time Till	Interpretation
360K served (80%)	2 min.	**?** must = **expect**
90K abandon (20%)	1 min.	**? willing** to wait

"Time willing to wait" of served is **censored** by their "wait".

"Uncensoring" (simplified)

Willing to wait $1 + 2 \times \frac{360K}{90K} = 1 + 2 \times 4 = \mathbf{9}$ min.

Expect to wait $2 + 1 \times \frac{90K}{360K} = 2 + 1 \times \frac{1}{4} = \mathbf{2.25}$ min.

Patience Index $= \frac{\text{time willing}}{\text{time expect}} = 4 = \frac{\#\,\text{served/wait} > 0}{\#\,\text{abandon/wait} > 0}$

↑ definition ↑ measure

- Supported by ongoing research (Wharton).

Patience Index

Let the means of V and R be m_V and m_R, and define

$$\text{Patience Index} \triangleq \frac{m_R}{m_V}.$$

- Call-by-call data
- Survival analysis. High-censoring might be a problem.
- Ancillary measure:

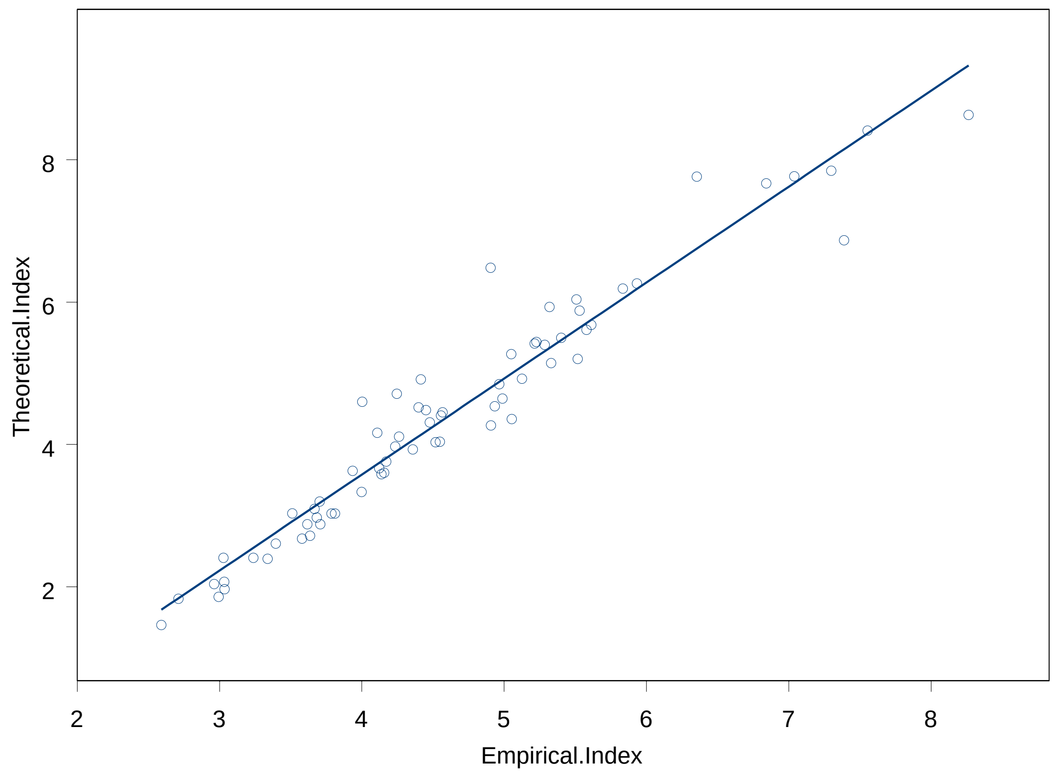
$$\text{Empirical Index} \triangleq \frac{\#\text{ served}}{\#\text{ abandoned}}.$$

▷ The usual plug-in MLE for Patience Index if V and R are independent exponential.

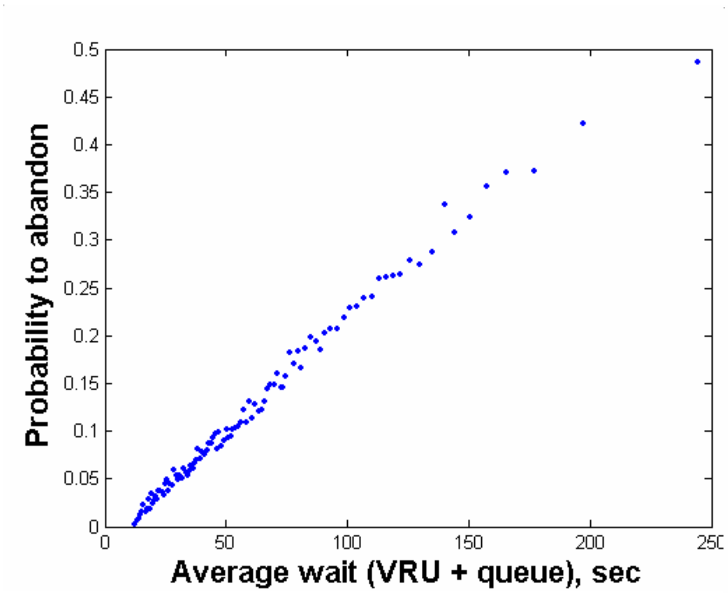
▷ Works well empirically .

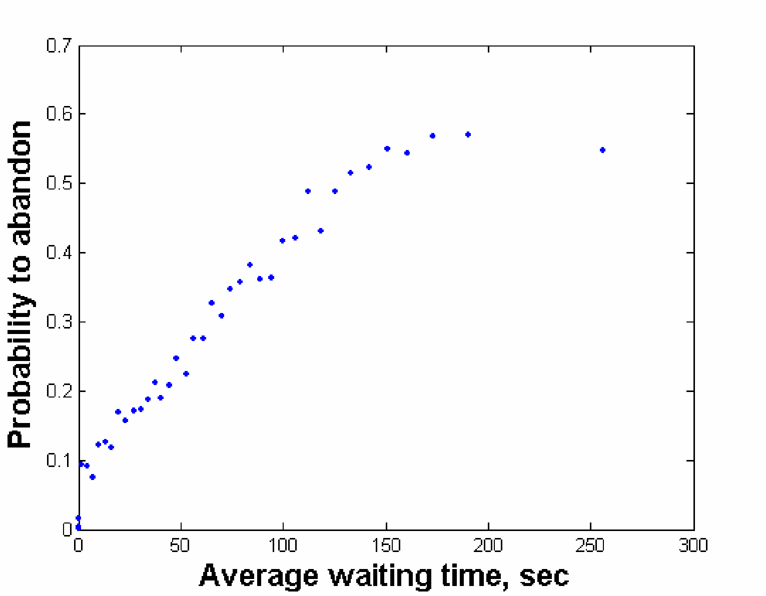
21

Figure 24: Patience Indices: empirical vs. theoretical ($R^2 = 0.94$)

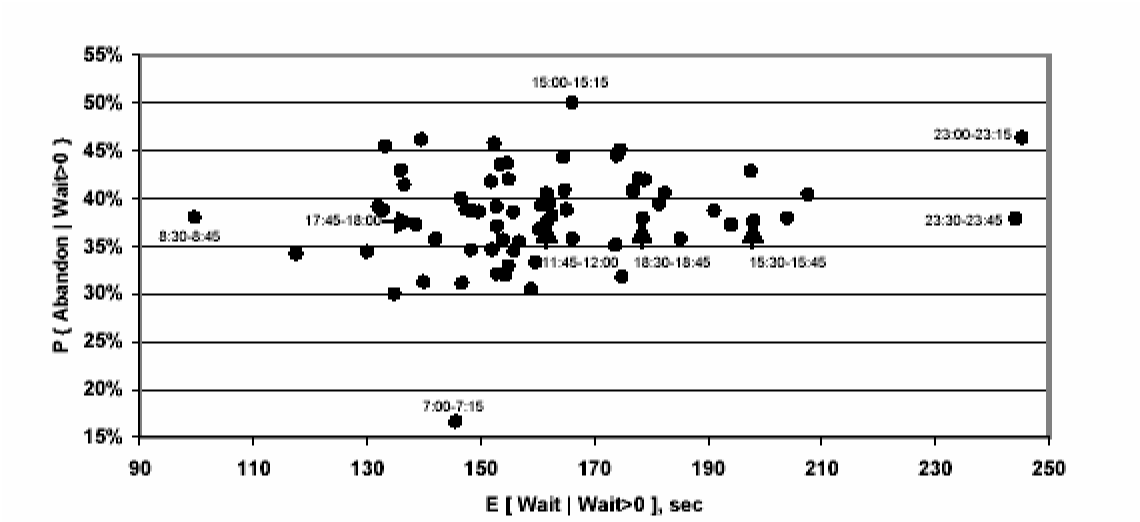


Human behavior





Delayed Abandons (IVR) Balking (New Customers)



Learning (Internet Customers)

Customer-Focused Queueing Theory

– 200 abandonment in Direct-Banking

– Not scientific

Reason to Abandon	**Actual** Abandon Time (sec)	**Perceived** Abandon Time (sec)	Perception Ratio
Fed up waiting (77%)	70	164	2.34
Not urgent (10%)	81	128	1.6
Forced to (4%)	31	35	1.1
Something came up (6%)	56	53	0.95
Expected call-back (3%)	13	25	1.9

$\Rightarrow$ Rational Abandonment from Invisible Queues (with Shimkin).

Fitting a Simple Model to

a Complex Reality

Erlang-A Formulae vs. Data Averages