Queues with Impatient Customers

M/M/N+M (G)

Palm/Erlang-A

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

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

Supporting Material (in Website)

Gans, Koole, and M.: “Telephone Call Centers: Tutorial, Review and Research Prospects.” *Review of State-of-the-Art Research*.

Brown, Gans, M., Sakov, Shen, Zeltyn, Zhao: "Statistical Analysis of a Telephone Call Center: A Queueing-Science Perspective." *Analysis of Arrivals, Services and Patience*.

Garnett, M. and Reiman: "Designing a Call Center with Impatient Customers." *Erlang-A, based on Garnett's MSc thesis*.

M. and Zeltyn: "The Impact of Customer Patience on Delay and Abandonment: Some Empirically-Driven Experiments with the M/M/N+G Queue." *On the relation between P(Ab) and E(Wait)*.

Zeltyn: Ph.D. thesis, on M/M/N+G.

Palm: "Intensitatsschwankungen im fernsprechverkehr,"
(In English) Ericsson Technics, 1943.

Palm: "Methods of judging the annoyance caused by congestion." Tele, 1953: Recommended.

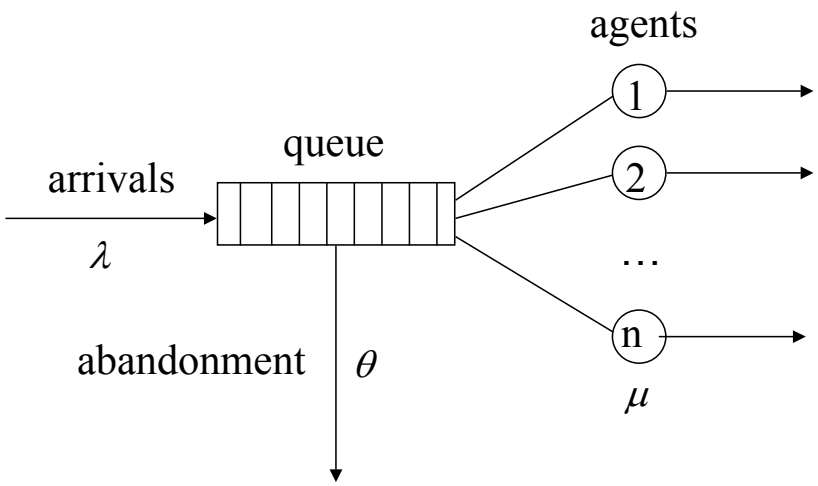
Baccelli and Hebuterne: "On queues with impatient customers." In Performance '81, ed. Gelenbe, 1981.

The Erlang-A (Palm, M/M/n+M) Model

Simplest model with abandonment, used by well-run call centers.

Building blocks:

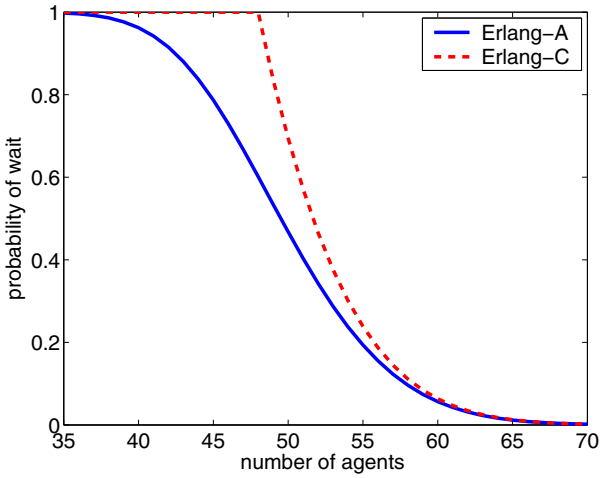
- λ – Poisson arrival rate.
- μ – Exponential service rate.
- n – number of service agents.
- θ – individual abandonment rate.



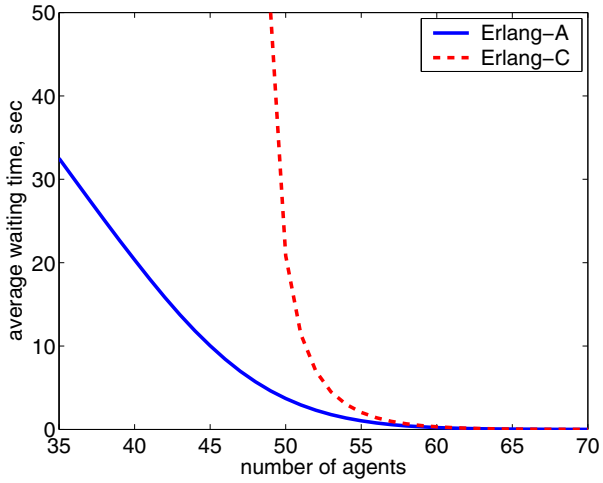
Erlang-A vs. Erlang-C

48 calls per min, 1 min average service time,
2 min average patience

probability of wait
vs. number of agents



average wait
vs. number of agents



If 50 agents:

	M/M/n	M/M/n+M	M/M/n, $\lambda \downarrow 3.1\%$
Fraction abandoning	–	3.1%	-
Average waiting time	20.8 sec	3.7 sec	8.8 sec
Waiting time's 90-th percentile	58.1 sec	12.5 sec	28.2 sec
Average queue length	17	3	7
Agents' utilization	96%	93%	93%

"The fittest survive" and wait less - much less.
Abandonment reduces workload when needed – at high-congestion periods.

Operational Performance Measures

The most popular performance measure is $P\{W \leq T; \mathrm{Sr}\}$ or even worse $P\{W \leq T \,|\, \mathrm{Sr}\}$.

We recommend either:

- $P\{W \leq T; \mathrm{Sr}\}$ - fraction of well-served;
- $P\{\mathrm{Ab}\}$ - fraction of poorly-served.

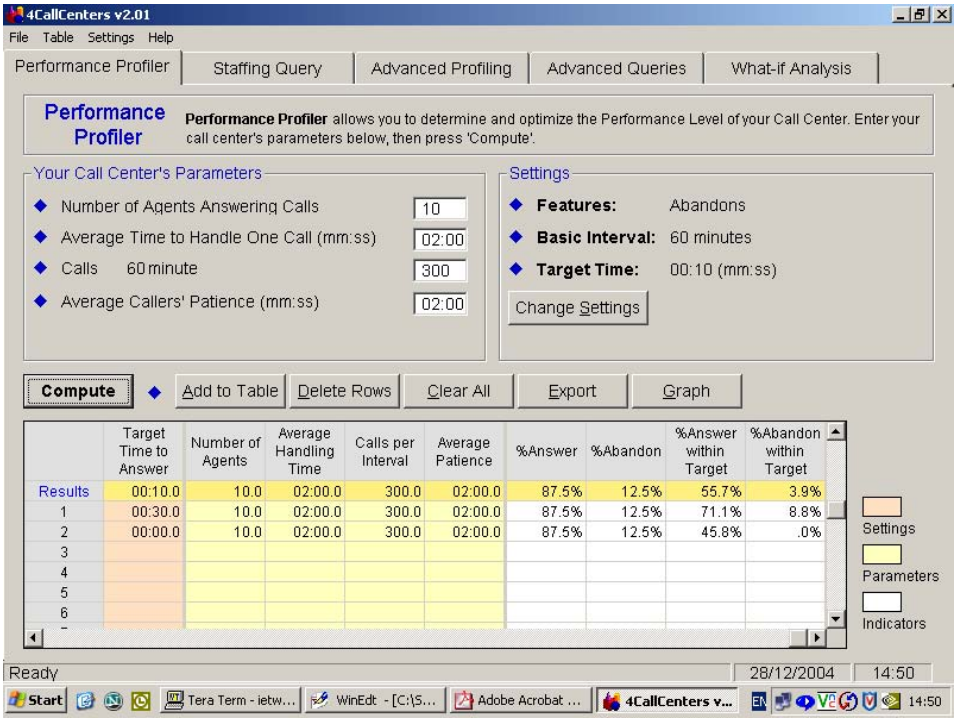
or a four-dimensional refinement:

- $P\{W \leq T; \mathrm{Sr}\}$ - fraction of well-served;
- $P\{W > T; \mathrm{Sr}\}$ - fraction of served, with a potential for improvement (say, a higher priority on next visit);
- $P\{W > \epsilon; \mathrm{Ab}\}$ - fraction of poorly-served;
- $P\{W \leq \epsilon; \mathrm{Ab}\}$ - fraction of those whose service-level is undetermined.

Additional important performance measures:

- Delay probability $\mathrm{P}\{W > 0\}$;
- Average wait $\mathrm{E}[W]$;
- ASA (Average Speed of Answer) – used extensively in call centers; usually defined as $\mathrm{E}[W|\mathrm{Sr}]$;
- Agents' occupancy $\rho = \dfrac{\lambda \cdot (1 - \mathrm{P}\{\mathrm{Ab}\})}{n\mu}$.
- Average queue-length $\mathrm{E}[Q]$.

Operational Performance Measures: calculation via 4CallCenters

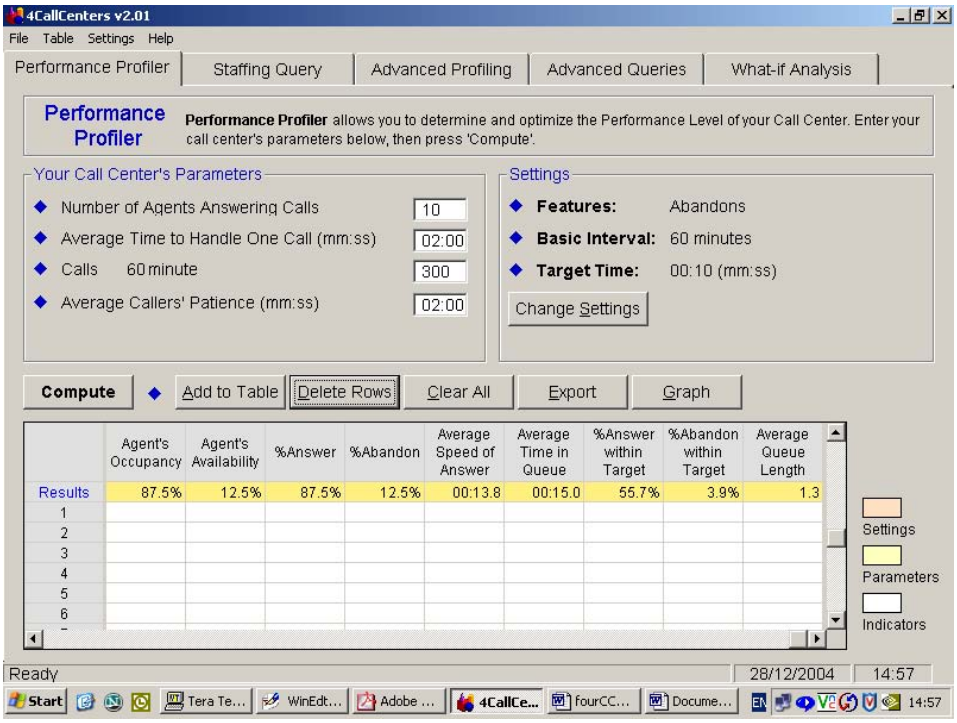


Erlang-A parameters:
$\lambda = 300$ calls per hour, $1/\mu = 2$ min, $n = 10$, $1/\theta = 2$ min.

Target times $T = 30$ sec, $\epsilon = 10$ sec.

- $\mathrm{P}\{W \le T; \mathrm{Sr}\} = 71.1\%$;
- $\mathrm{P}\{W > T; \mathrm{Sr}\} = 87.5\% - 71.1\% = 16.4\%$;
- $\mathrm{P}\{W > \epsilon; \mathrm{Ab}\} = 12.5\% - 3.9\% = 8.6\%$;
- $\mathrm{P}\{W \le \epsilon; \mathrm{Ab}\} = 3.9\%$.
- Delay probability $\mathrm{P}\{W > 0\} = 100\% - 45.8\% = 54.2\%$.

Additional performance measures



- Average Time in Queue = E$[W]$ = 15 sec;
- ASA = E$[W|\mathrm{Sr}]$ = 13.8 sec;
- Agents' Occupancy $\rho = 87.5\%$.
- Average Queue Length E$[Q] = 1.3$.

P{Ab}/E[W] relation

Theoretical: In Erlang-A (and other queueing models with exp(θ) patience):

$$\mathrm{P}\{\mathrm{Ab}\} \;=\; \theta \cdot \mathrm{E}[W]\,.$$

Proof. Balance equation:

$$\theta \cdot \mathrm{E}[Q] \;=\; \lambda \cdot \mathrm{P}\{\mathrm{Ab}\}\,. \tag{1}$$

Little's formula:

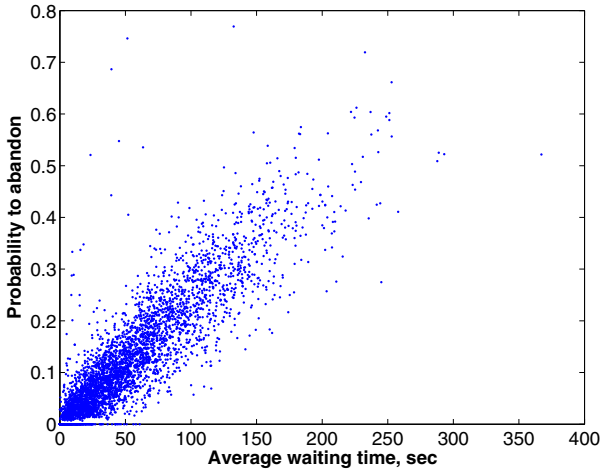
$$\mathrm{E}[Q] \;=\; \lambda \cdot \mathrm{E}[W]\,. \tag{2}$$

Substitute (2) into (1). ■

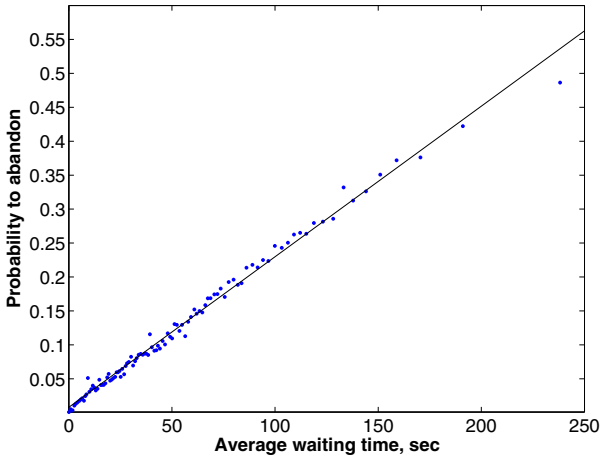
Empirical relations

Israeli bank: yearly data

hourly data



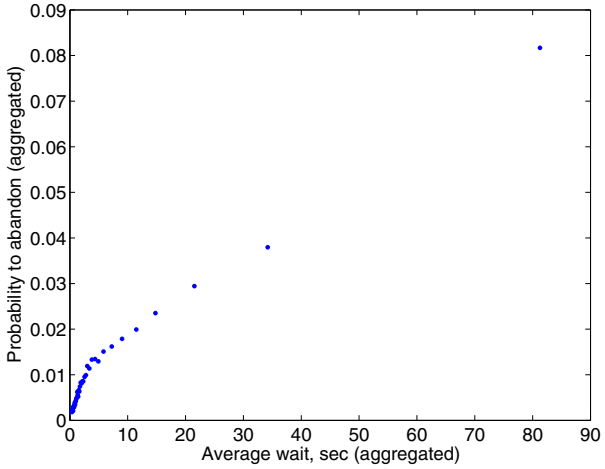
aggregated



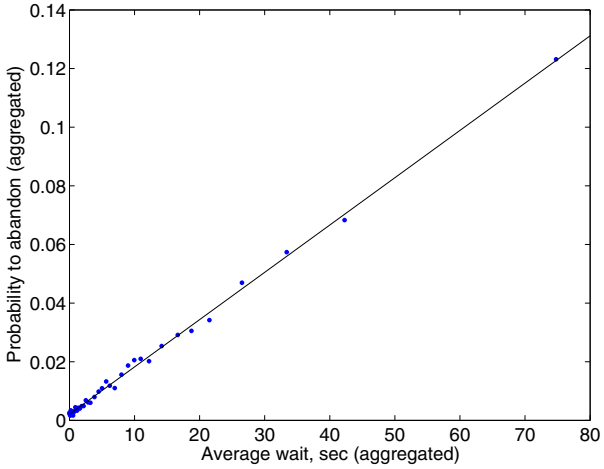
The graphs are based on 4158 hour intervals.

U.S. bank

Retail



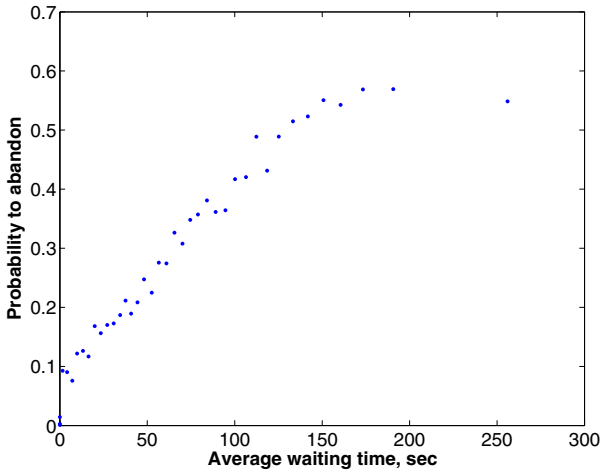
Telesales



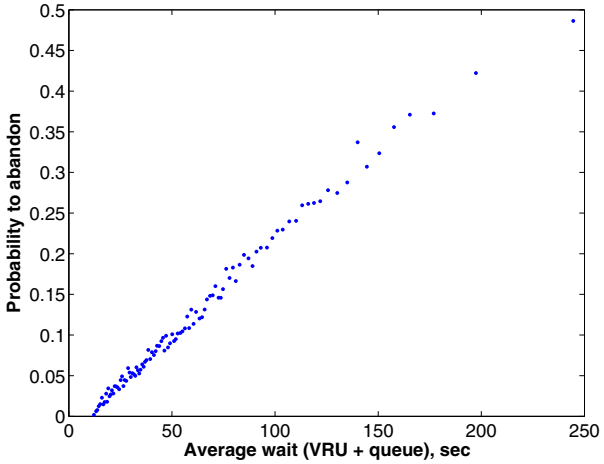
Retail – significant abandonment during first seconds of wait.

Linear patterns with non-zero intercepts

Israeli data: new customers



VRU-time included in wait



Left-hand plot $\approx$ exp patience with balking:
0 with probability p, $\exp(\theta)$ with probability $(1-p)$.

Right-hand plot $\approx$ delayed patience: $c + \exp(\theta)$, $c > 0$.

Erlang-A: parameter estimation and prediction

Estimation: inference from historical data (e.g. exp, normal) were parameters assumed fixed over time.
Prediction: forecast behavior of sample outside of original data set.

Arrivals (λ)

- Typically Poisson, time-varying rates, constant at 15/30/60 min scale;
- Significant uncertainty concerning future rates $\Rightarrow$ prediction;
- Predict separately *daily volumes* and *fraction* of arrivals per time interval.

Services (μ)

- Typically stable from day to day $\Rightarrow$ estimation;
- Can change depending on time-of-day;
- Typically, service time $\neq$ talk time.

First approach:
service time = talk time + wrap-up time (after-call work) + ...;

Second approach:

$$\text{service time} = \frac{\text{Total Working Time} - \text{Total Idle Time}}{\text{Number of Served Customers}}.$$

Number of agents (n)

- Output of WFM software given λ, μ, θ, performance goals. One gets number of FTE's (Full Time Equivalent positions).
- Agents on schedule = FTE's $\cdot$ RSF (Rostered Staff Factor) (RSF > 1). Reasons: absenteeism, unscheduled breaks, ...
- Obtaining historical data on n can be hard.

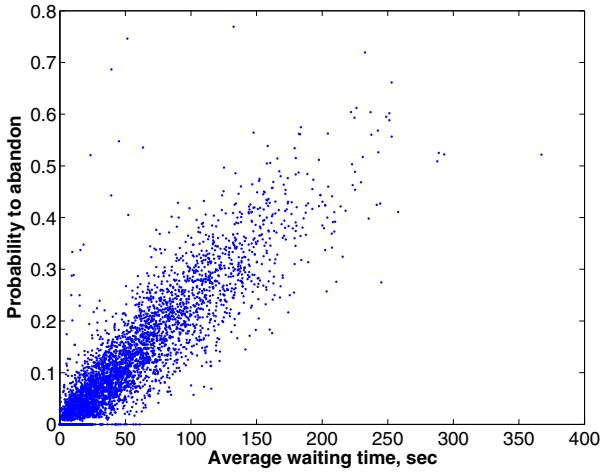
Patience (θ)

Observations are **censored**!

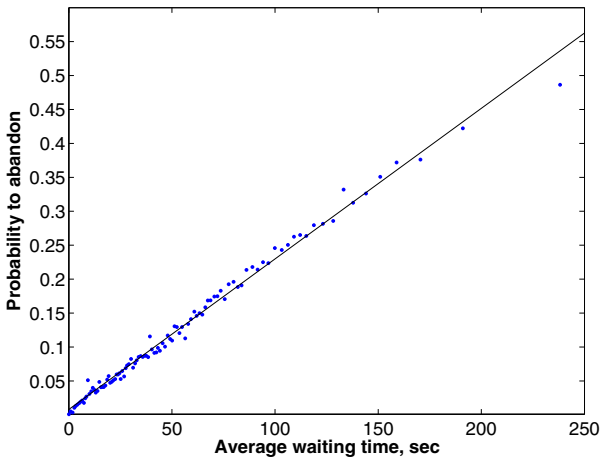
- Customer abandoned $\Rightarrow$ patience R known;
- Customer served $\Rightarrow$ offered wait V known $\Rightarrow R > V$.

Avoiding direct "uncensoring": use $\mathrm{P}\{\mathrm{Ab}\} = \theta \cdot \mathrm{E}[W]$.

hourly data



aggregated



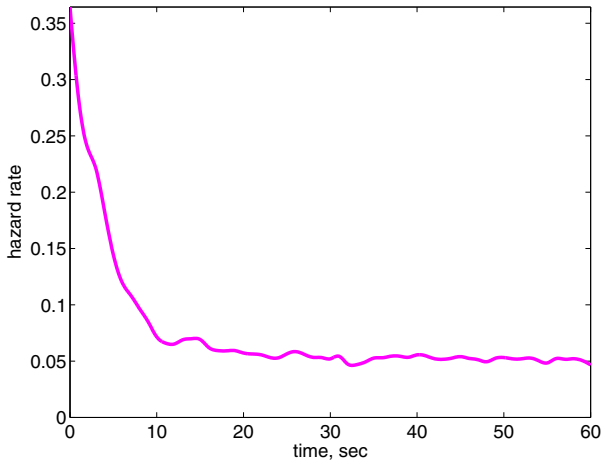
Regression $\Rightarrow$ average patience $(1/\theta) \approx 450$ sec.

Estimating patience distribution

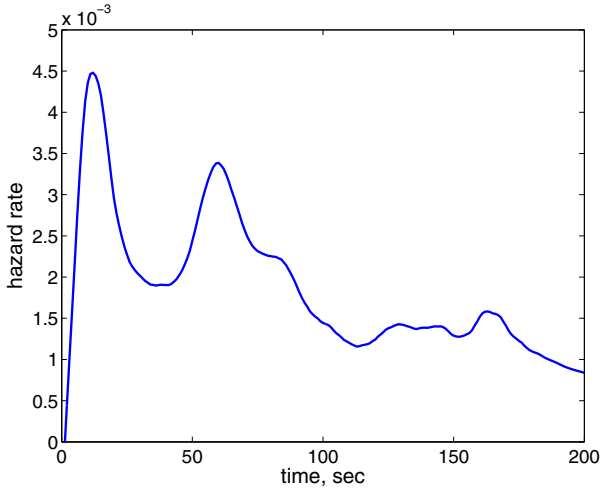
Are patience times really exponential?
To "uncensor data" use Kaplan-Meier (product-limit) estimator.
Output: estimates of hazard rate and survival function of patience.

Empirical hazard rates of patience times

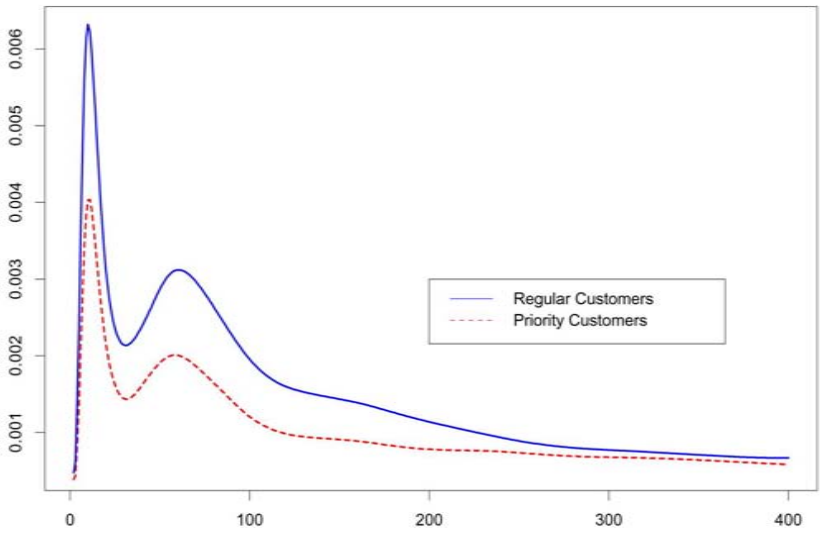
U.S. bank



Israeli bank

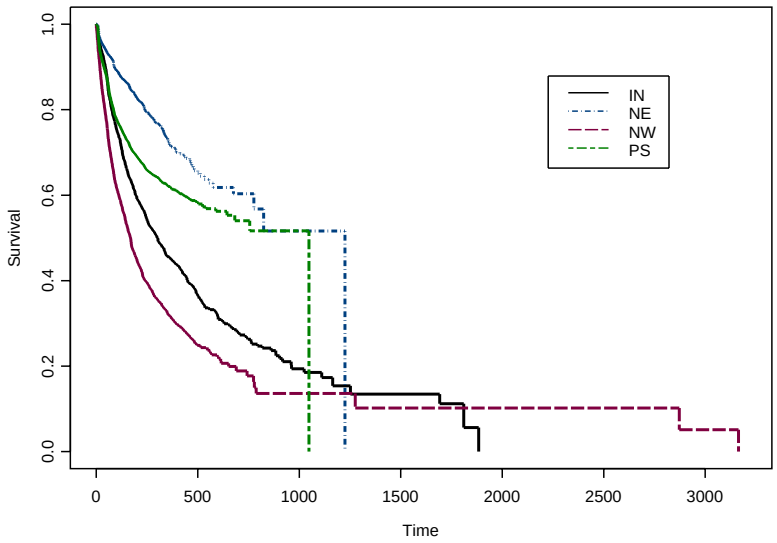


Israeli bank: regular vs. priority customers



Time, sec

Israeli bank: service types



IN – Internet Assistance; NE – Stock Transactions;
NW – New Customers; PS – Regular

Conclusions:

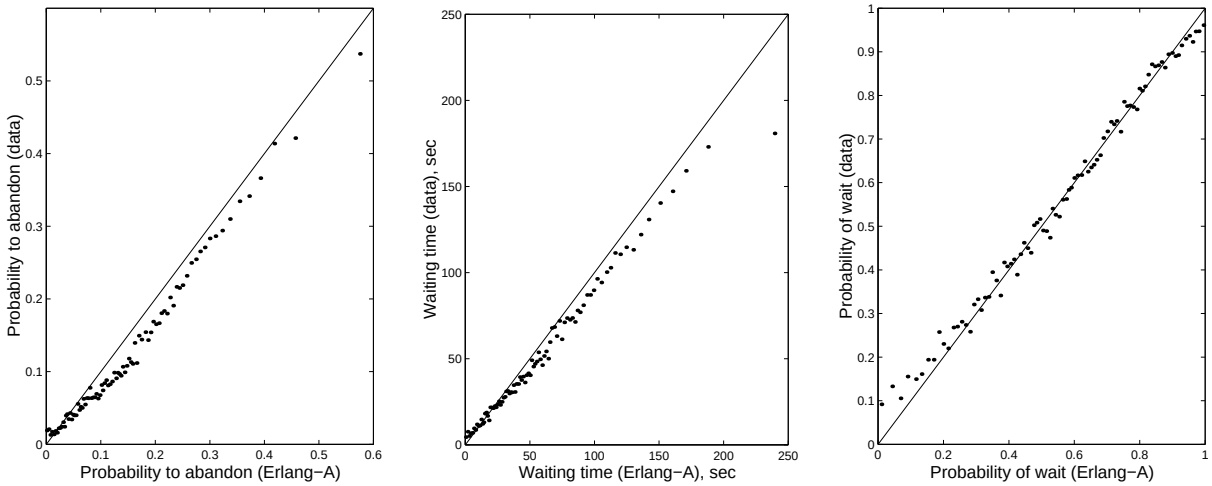
- Patience time are, in general, non-exponential;
- Most tele-customers are **very** patient;
- Kaplan-Meier is very informative concerning patience *qualitative* patterns (abandonment peaks, comparisons, . . .);
- Kaplan-Meier can be problematic concerning estimation of *quantitative* characteristics (mean, variance, median). $\mathrm{E}[\widehat{R}] = \int_0^\infty \widehat{S(x)}dx$, where $S(x)$ - survival function of patience. However, $\widehat{S(x)}$ not reliable for large x.

Question: can we apply Erlang-A with non-exponential patience?

Fitting a simple model to a complex reality

Erlang-A Formulae vs. Data Averages

P{Ab} E[W] P{$W > 0$}



- Points: hourly data vs. Erlang-A output;
- Formulae with continuous n used;
- Patience estimated via P{Ab}/E[W] relation;
- Erlang-A estimates – close upper bounds.

Fitting a simple model to a complex reality: Patience index

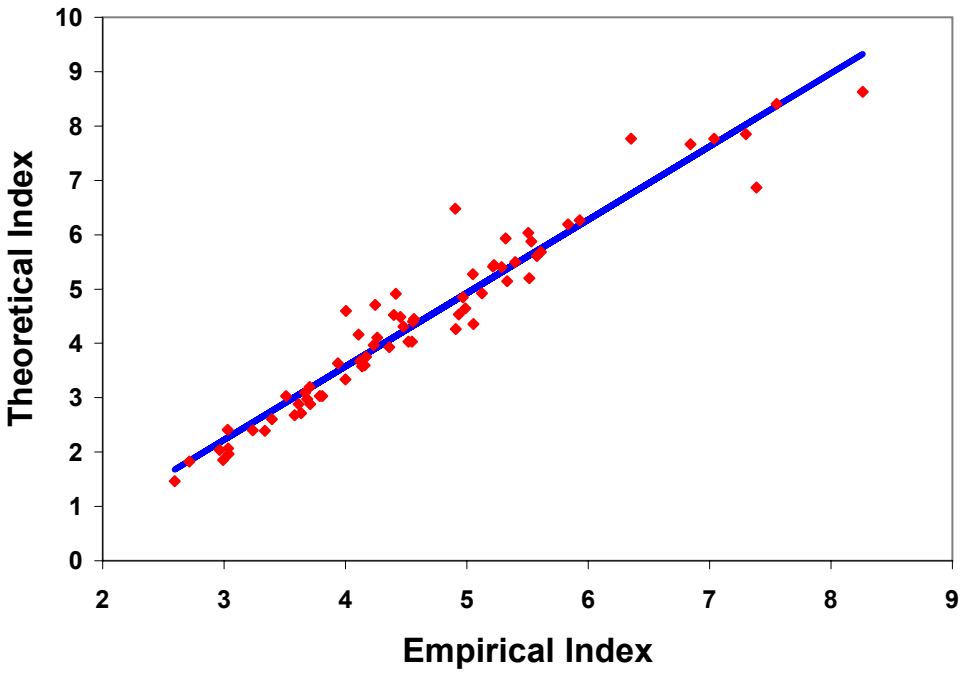
$$\text{Theoretical Patience Index} \triangleq \frac{\text{time willing to wait}}{\text{time required to wait}} = \frac{\text{average patience}}{\text{average offered wait}}. \quad (3)$$

Calculation is difficult.

$$\text{Empirical Patience Index} \triangleq \frac{\%\text{ served}}{\%\text{ abandoned}}. \quad (4)$$

Easily calculable from ACD reports.
If R and V exponentially distributed, (4) is MLE of (3).

Patience index – empirical vs. theoretical



Customer-Focused Queueing Theory

Waiting experience can be summarized by:

1. Time that a customer *expects* to wait;
2. Time that a customer is *willing* to wait (R, patience or need);
3. Time that a customer *must* wait (V, offered wait);
4. Time that a customer *actually* waits ($W = \min(R, V)$);
5. Time that a customer *perceives* waiting.

If 1=3, 4=5: (R, V, W) from the basic model.

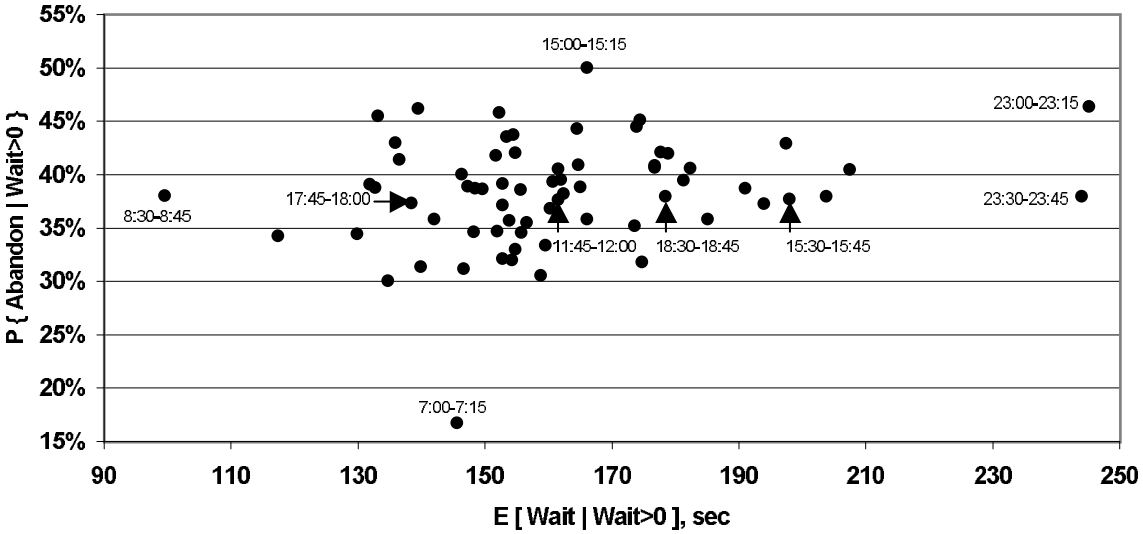
200 abandonment in Direct-Banking: perceived vs. actual waiting.

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

Adaptive behavior of impatient customers

Question: Do customers adapt their patience to system performance (offered wait)?

Israeli bank: Internet-support customers



Rational abandonment from invisible queues: Mandelbaum, Shimkin, Zohar.

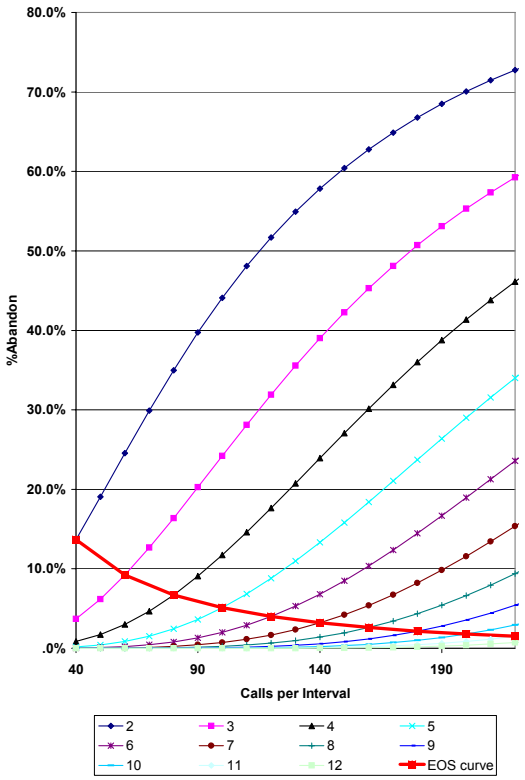
Advanced features of 4CallCenters

Advanced profiling

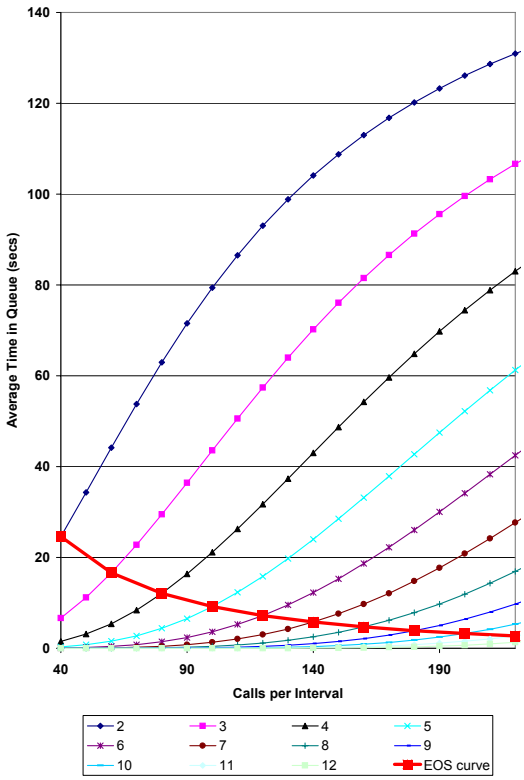
Vary input parameters of Erlang-A and display output (performance measures) in a table or graphically.

Example: $1/\mu = 2$ minutes, $1/\theta = 3$ minutes;
λ varies from 40 to 230 calls per hour, in steps of 10;
n varies from 2 to 12.

Probability to abandon



Average wait



Red curve: EOS (Economies-Of-Scale)

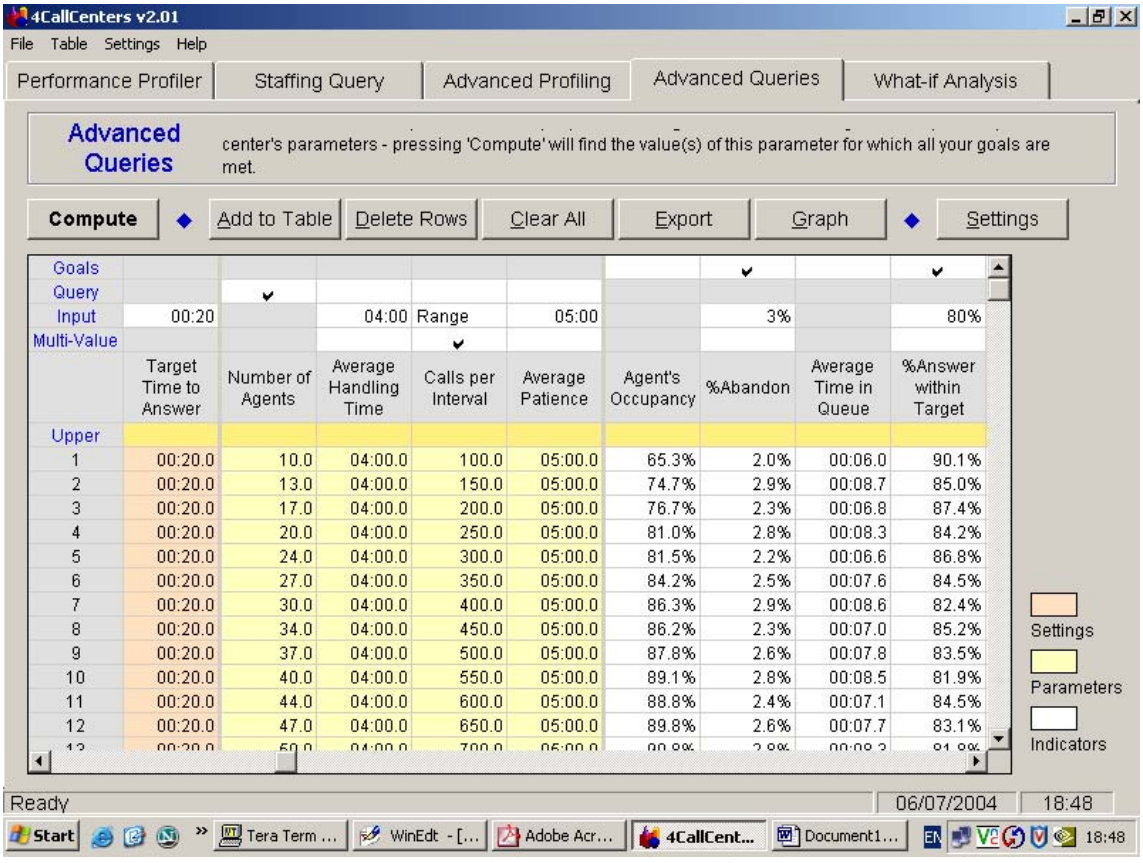
Advanced staffing queries

Multiple performance goals.

Example: $1/\mu = 4$ minutes, $1/\theta = 5$ minutes;
λ varies from 100 to 1200, in steps of 50.

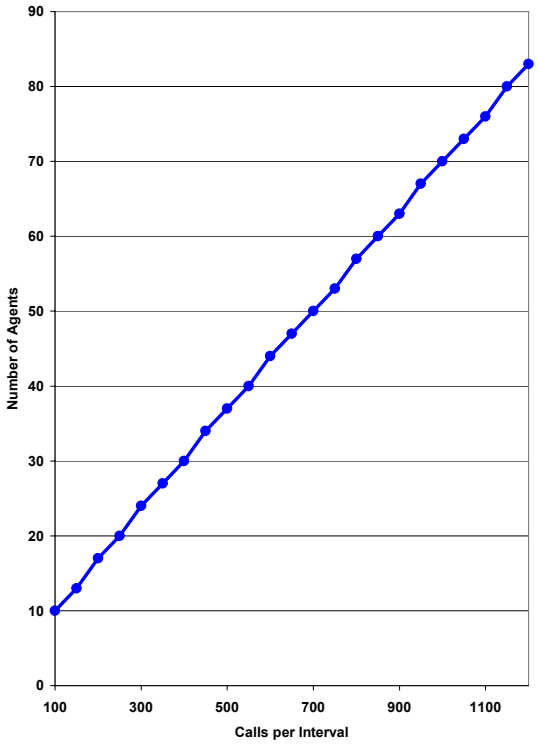
Performance targets:
$\mathrm{P}\{\mathrm{Ab}\} \leq 3\%$; $\mathrm{P}\{W < 20 \text{ sec}; \ \mathrm{Sr}\} \geq 0.8$.

4CallCenters output

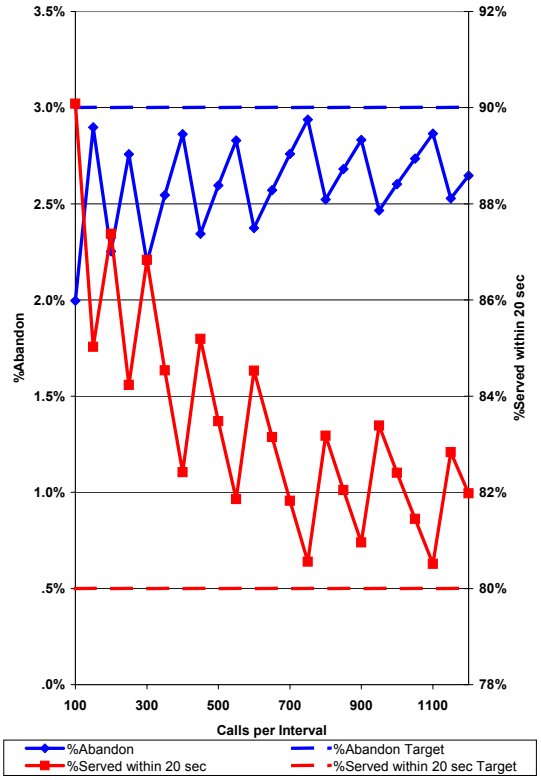


4CallCenters. Advanced staffing queries. Dynamics of staffing level and performance.

Recommended staffing level



Target performance measures



EOS: 10 agents needed for 100 calls per hour but only 83 for 1200 calls per hour.