Class 10

A Stochastic Markovian Service Station in Steady State - Part II; Palm/Erlang-A.

Modelling and Analyzing a Markovian Service Station - Continued

- A Birth & Death Model: stability, MOP's;
  Some concrete models: Erlang B (Loss), Erlang C (Delay), self-service and:
- **Erlang A** = M/M/m + M: The Fundamental Markovian Model of a Service Station (Call Center), namely Poisson arrivals, Exponential services and Exponential (Im)Patience.

Recitation 11: 4CallCenters software.

HW 9: "GazolCo's Call Center".
Carry out the analysis in accordance with the instructions.
Use *4CallCenters*, as described in the assignment. (This software is downloadable from our website: http://iew3.technion.ac.il/serveng/4CallCenters/Downloads.htm.)
Each student should first "experiment" *individually* with the software. Then, continue with the assignment, starting together and, perhaps, dividing the workload as you see fit.

Service Engineering

Class 10

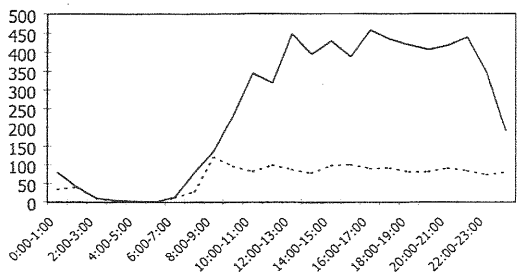
Stochastic Markovian Service Station in Steady State - Part II: The Palm/Erlang-A Queue

- Reviewing Abandonment and (Im)Patience.
- Definition of the Erlang-A Queue.
- Comparison with the Erlang-C Queue.
- Steady-State Distribution and Performance Measures.
- Probability to Abandon vs. Average Wait: $\mathrm{P}\{\mathrm{Ab}\} = \theta \cdot \mathrm{E}[W_q]$.
- Estimating the (Im)Patience Parameter.
- General (Im)Patience Distribution: M/M/n+G Queue.
- Erlang-A: Fitting a Simple Model to a Complex Reality.

Customers' (Im)Patience

Marketing Campaign at a Call Center

Average wait 376 sec, 24% calls **answered**



Abandonment **Important** and **Interesting**

- One of two customer-subjective operational performance measures (Second one is Redials)
- Poor service level (future losses)
- Lost business (present losses)
- 1-800 costs (present gains; out-of-pocket vs. alternative)
- Self-selection: the "fittest survive" and wait less (much less)
- Accurate Robust models (vs. distorted, unstable, sensitive)
- Beyond Operations/OR: Psychology, Marketing, Statistics
- Beyond Telephony: VRU/IVR (Opt-Out-Rates), Internet (over 60%), Hospitals ED (LWBS).

Understanding (Im)Patience

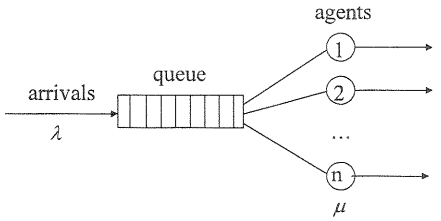
- **Observing** (Im)Patiecne – Heterogeneity:
  Under a single roof, the fraction abandoning varies from 6% to 40%, depending on the type of service/customer.
- **Describing** (Im)Patience Dynamically:
  Irritation proportional to Hazard Rate (Palm's Law).
- **Managing** (Im)Patience:
  - VIP vs. Regulars: who is more "Patient"?
  - What are we actually measuring?
  - (Im)Patience Index:
    "How long Expect to wait" relative to
    "How long Willing to wait".
- **Estimating** (Im)Patience: Censored Sampling.
- **Modeling** (Im)Patience:
  - The "Wait" Cycle:
    Expecting, Willing, Required, Actual, Perceived, etc.
    The case of the Experienced & Rational customer.
  - (Nash) Equilibrium Models.

Basic (Markovian) Queueing Models of a Basic Service Station

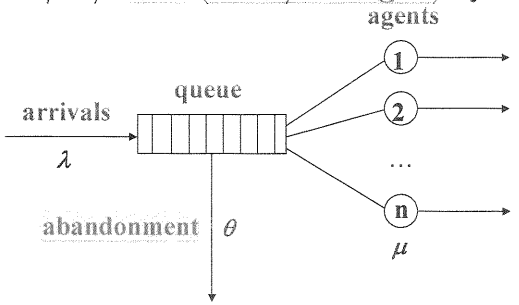
Poisson arrivals, **Exponential** service times, **Exponential** (im)patience.

Mathematical Framework: Markov Jump-Processes (Birth&Death).

M/M/n (Erlang-C) Queue



M/M/n+M (Palm/Erlang-A) Queue



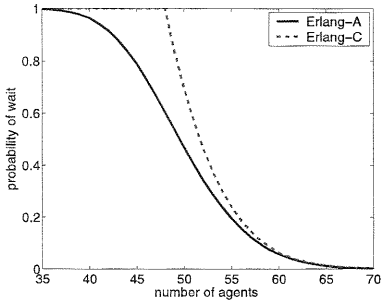
Additional Markovian Models: Balking, Trunks; Retrials.

Applications: Performance Analysis, Design (EOS), Staffing.

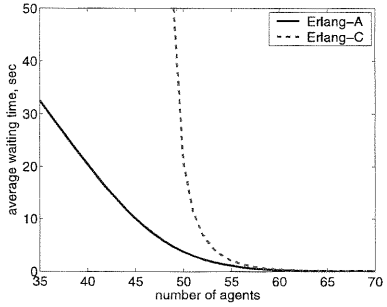
Erlang-A vs. Erlang-C

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.

Predicting (Operational) Performance

Model **Primitives** (Building Blocks):

- Arrivals to service (eg. Poisson);
- (Im)Patience while waiting (eg. Exponential);
- Service times (eg. Exponential);
- Servers (eg. i.i.d.).

Model **Output**: **Offered-Wait** V

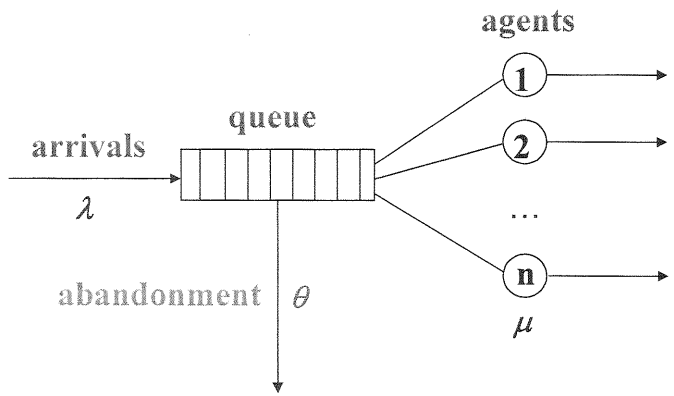
Operational Performance Measure calculable in terms of (τ, V).

- eg. % Abandonment $= \mathrm{P}\{\tau < V\}$ (or $\mathrm{P}\{5 \text{ sec} < \tau < V\}$)
- eg. Average Wait $= \mathrm{E}[\min\{\tau, V\}]$ (or $\mathrm{E}[\tau | \tau < V]$)

Applications:

- **Performance Analysis**
- **Design, Phenomena** (Pooling, Economies of Scale)
- **Staffing – How Many Agents** (FTE's = Full-Time-Equivalent's)
  Note: Within the Basic Model of heterogeneous customers and servers (vs. priorities, SBR - later).

Erlang-A (Palm, M/M/n+M; M-M/M/n)



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

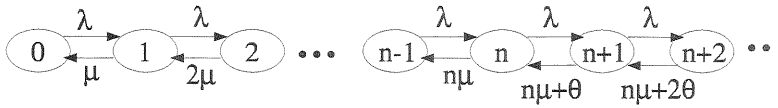
Parameters:

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

Erlang-A = Birth-and-Death Process

$L(t)$ – number-in-system at time t (served plus queued);
$L = \{L(t), t \geq 0\}$ – Markov Birth-and-Death process.

Transition-rate diagram



Steady-state equations:

$$\begin{cases} \lambda\pi_j &= (j+1)\cdot\mu\pi_{j+1}, \quad 0 \leq j \leq n-1 \\ \lambda\pi_j &= (n\mu + (j+1-n)\theta)\cdot\pi_{j+1}, \quad j \geq n. \end{cases}$$

Steady-state distribution:

$$\pi_j = \begin{cases} \dfrac{(\lambda/\mu)^j}{j!}\pi_0\,, & 0 \leq j \leq n \\ \displaystyle\prod_{k=n+1}^{j}\left(\frac{\lambda}{n\mu+(k-n)\theta}\right)\frac{(\lambda/\mu)^n}{n!}\pi_0\,, & j \geq n+1, \end{cases}$$

where

$$\pi_0 = \left[\sum_{j=0}^{n}\frac{(\lambda/\mu)^j}{j!} + \sum_{j=n+1}^{\infty}\prod_{k=n+1}^{j}\left(\frac{\lambda}{n\mu+(k-n)\theta}\right)\frac{(\lambda/\mu)^n}{n!}\right]^{-1}.$$

Numerical drawback: infinite sums.

Erlang-A: Stability

Claim: **Erlang-A is always stable.**

Proof:

$$\pi_0^{-1} = \sum_{j=0}^{n} \frac{(\lambda/\mu)^j}{j!} + \sum_{j=n+1}^{\infty} \prod_{k=n+1}^{j} \left(\frac{\lambda}{n\mu + (k-n)\theta} \right) \frac{(\lambda/\mu)^n}{n!}$$

$$\leq \sum_{j=0}^{\infty} \frac{(\lambda/\min(\mu,\theta))^j}{j!} = e^{-\lambda/\min(\mu,\theta)} .$$

(Used the inequality $n\mu + (k-n)\theta \geq k \min(\mu, \theta)$, for all $k \geq n$.)

Remark: Let d_j = death-rate in state j, $0 < j < \infty$.
Then, in fact,

$$j \cdot \min(\mu, \theta) \leq d_j \leq j \cdot \max(\mu, \theta) .$$

Now observe that the bounds are death-rates of M/M/∞ queues, with service rates $\min(\mu, \theta)$ and $\max(\mu, \theta)$.

This implies that Erlang-A is sandwiched (stochastically) between two M/M/∞ queues.
$\Rightarrow$ The stationary (limiting) distribution is sandwiched (stochastically) between Poisson distributions.

Special case: $\mu = \theta$ $\Rightarrow$ Erlang-A $\stackrel{d}{=}$ M/M/∞.
$\Rightarrow$ Square-Root Staffing
(via Poisson $\approx$ Normal; more on that later).

Properties of P{Ab}

- P{Ab} increases monotonically in θ, λ;
  P{Ab} decreases monotonically in n, μ
  (Bhattacharya and Ephremides, 1991);
- P{Ab} $\leq$ P{Block} in Erlang-B (Boxma and de Waal, 1994)
  (think zero-patience).
- Note: In M/M/n+G. with E[τ] fixed, deterministic patience minimizes P{Ab} but maximizes E[W_q] (Zeltyn's PhD, 2004).

Steady-State Distribution via Special Functions (Palm)

Gamma function:

$$\Gamma(x) \triangleq \int_0^\infty t^{x-1}e^{-t}dt\,, \qquad x>0.$$

Incomplete Gamma function:

$$\gamma(x,y) \triangleq \int_0^y t^{x-1}e^{-t}dt\,, \qquad x>0,\ y\geq 0.$$

$$A(x,y) \triangleq \frac{xe^y}{y^x}\cdot\gamma(x,y) = 1+\sum_{j=1}^{\infty}\frac{y^j}{\Pi_{k=1}^{j}(x+k)}, \qquad x>0,\ y\geq 0.$$

Recall $E_{1,n}$ = *blocking probability* in Erlang-B (M/M/n/n):

$$E_{1,n} = \frac{\frac{(\lambda/\mu)^n}{n!}}{\Sigma_{j=0}^{n}\frac{(\lambda/\mu)^j}{j!}} = \frac{(\lambda/\mu)^n}{e^{\lambda/\mu}}\cdot\frac{1}{\Gamma(n+1)-\gamma(n+1,\lambda/\mu)}.$$

(Can be efficiently calculated via recursion.)
Then

$$\pi_j = \begin{cases} \pi_n\cdot\dfrac{n!}{j!\cdot\left(\frac{\lambda}{\mu}\right)^{n-j}}, & 0\leq j\leq n\,,\\[2ex] \pi_n\cdot\dfrac{\left(\frac{\lambda}{\theta}\right)^{j-n}}{\Pi_{k=1}^{j-n}\left(\frac{n\mu}{\theta}+k\right)}, & j\geq n+1\,, \end{cases}$$

where

$$\pi_n = \frac{E_{1,n}}{1+\left[A\left(\frac{n\mu}{\theta},\frac{\lambda}{\theta}\right)-1\right]\cdot E_{1,n}}.$$

Operational Performance Measures

The most prevalent performance measure is
$P\{W_q\leq T;\mathrm{Sr}\}$ (or "worse" $P\{W_q\leq T\,|\,\mathrm{Sr}\}$).

We recommend:

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

with T determined via *"Waiting less than T is Well-Served"*.

Or even a four-dimensional refinement:

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

with ϵ: *"Abandoning before ϵ is Harmless"*.

Additional Useful Performance Measures

- **ASA** (Average Speed of Answer) – used extensively in call centers; usually taken to be $\mathrm{E}[W_q|\mathrm{Sr}]$ (could be misleading);
- Average Wait $\mathrm{E}[W_q]$;
- Delay Probability $\mathrm{P}\{W_q > 0\}$ - important (later), yet unused;
- Agents' Occupancy $\rho = \dfrac{\lambda \cdot (1 - \mathrm{P}\{\mathrm{Ab}\})}{n\mu}$;
- Average Queue-Length $\mathrm{E}[L_q]$.

Operational Performance Measures: Calculation via 4CallCenters

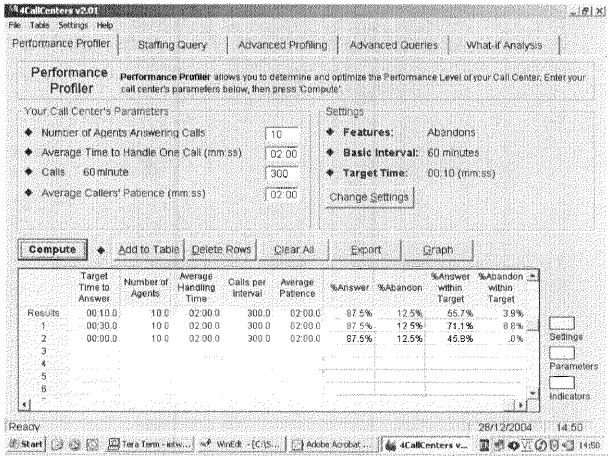
- Performance measures of the form $\mathrm{E}[f(V, \tau)]$.
- Calculable, by numerically stable algorithms.

For example,

$f(v,\tau)$	$\mathrm{E}[f(V,\tau)]$
$1_{\{v>\tau\}}$	$\mathbf{P\{}V > \tau\mathbf{\}} = \mathrm{P}\{\mathrm{Ab}\}$
$1_{(t,\infty)}(v \wedge \tau)$	$\mathbf{P\{}W_q > t\mathbf{\}}$
$1_{(t,\infty)}(v \wedge \tau)1_{\{v>\tau\}}$	$\mathbf{P\{}W_q > t; \mathrm{Ab}\mathbf{\}}$
$(v \wedge \tau)1_{\{v>\tau\}}$	$\mathrm{E}\{W_q; \mathrm{Ab}\}$
$g(v \wedge \tau)$	$\mathrm{E}[g(W_q)]$

From these, one derives additional measures, eg. $\mathrm{E}[W_q|Ab]$.

Operational Performance Measures: Calculation via 4CallCenters



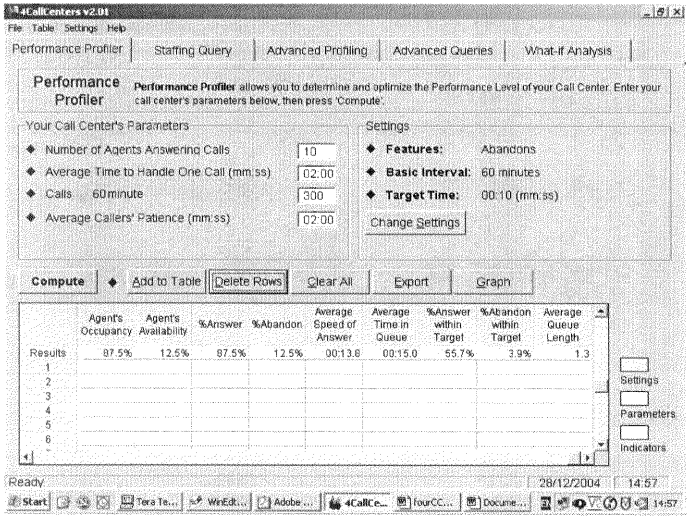
Erlang-A parameters:

$\lambda = 300$ calls/hour, $1/\mu = 2$ min, $n = 10$, $1/\theta = 2$ min.

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

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

Additional Performance Measures: Calculation via 4CallCenters



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

Operational Performance Measures: Calculation via Special Functions

For example,

$$\mathrm{P}\{\boldsymbol{W_q} > 0\} = \sum_{j=n}^{\infty} \pi_j = \frac{\boldsymbol{A}\left(\frac{n\mu}{\theta}, \frac{\lambda}{\theta}\right) \cdot \boldsymbol{E_{1,n}}}{1 + \left(\boldsymbol{A}\left(\frac{n\mu}{\theta}, \frac{\lambda}{\theta}\right) - 1\right) \cdot \boldsymbol{E_{1,n}}},$$

$$\mathrm{P}[\mathrm{Ab}|\boldsymbol{W_q} > 0] = \frac{1}{\rho \boldsymbol{A}\left(\frac{n\mu}{\theta}, \frac{\lambda}{\theta}\right)} + 1 - \frac{1}{\rho},$$

$$\mathrm{E}[\boldsymbol{W_q}|\boldsymbol{W_q} > 0] = \frac{1}{\theta} \cdot \left[\frac{1}{\rho \boldsymbol{A}\left(\frac{n\mu}{\theta}, \frac{\lambda}{\theta}\right)} + 1 - \frac{1}{\rho}\right].$$

Parameter Estimation and Prediction I; 4CallCenters, Erlang-A, and beyond

Estimation: Inference from historical data (e.g. Exp, LogNormal), with parameters assumed fixed over time-periods (overall).
Prediction: Forecast behavior beyond the available data.

Arrivals (λ)

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

Services (μ, or E(S))

- Typically stable from day to day $\Rightarrow$ estimation;
- Can vary, depending on time-of-day;
- Typically, service time $\neq$ talk time, and the former is needed.

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

Second Estimation Approach:

$$\widehat{\mathrm{E}(S)} = \frac{\text{Total Working Time} - \text{Total Accessible (Idle) Time}}{\#\text{ Served Customers}}.$$

Parameter Estimation and Prediction II

Number of Agents (n)

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

(Im)Patience (θ)

- Observations are **censored**! (typically heavy censoring):
  - Customer abandons $\Rightarrow$ patience τ known;
  - Customer served $\Rightarrow$ offered-wait V known ($\Rightarrow \tau > V$).
- Estimate via

$$\widehat{\theta} = \frac{\#\text{ Abandoning}}{\text{Total Waiting Time (Abandoning + Served)}};$$

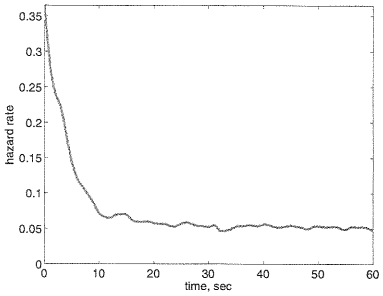
or via slope of the Regression of P{Ab} over E[W_q], as before; or both.

Estimating (Im)Patience Distribution I

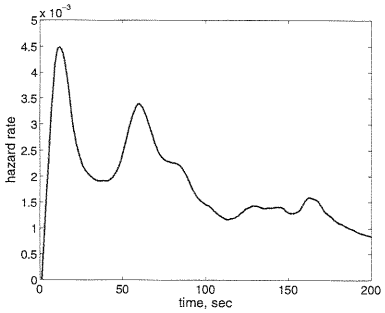
Are patience times really exponential?
To "uncensor data", use the Kaplan-Meier estimator (standard).
Output: Estimates of survival function and hazard-rate function.

Empirical Hazard Rates of (Im)Patience

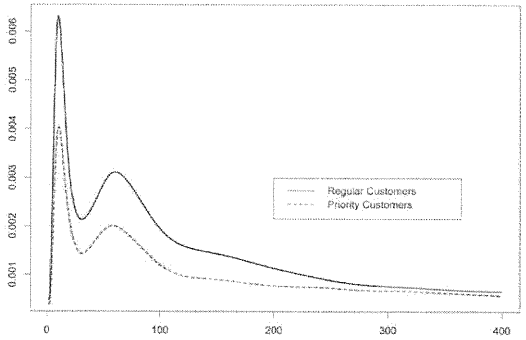
U.S. Bank



Israeli Bank



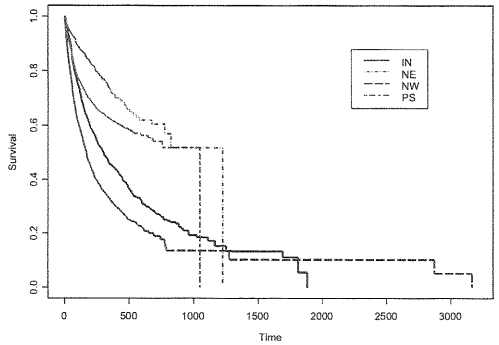
Israeli Bank: Regulars vs. VIP's



Time, sec

Estimating (Im)Patience Distribution II

Israeli Bank: Service Types



IN – Internet; NE – Stocks; NW – New; PS – Regulars

Conclusions:

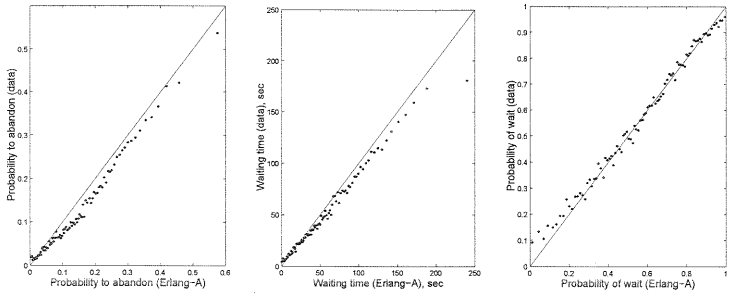
- Patience time are, in general, non-exponential;
- Tele-customers are (perhaps surprisingly) **very** patient;
- Hazard-Rates very informative concerning dynamic *qualitative* evolution of (im)patience (peaks, IFR, DFR). (Palm: proportional to irritation);
- Survival functions useful for (stochastic) comparisons;
- Kaplan-Meier often problematic for estimating *quantitative* characteristics (mean, variance, median). (Eg. $\widehat{\mathrm{E}[\tau]} = \int_0^\infty \widehat{S(x)}dx$.)

Question: Can Erlang-A be applied with non-exponential (im)patience?

Erlang-A: Simple Model at the Service of Complex Realities

- Small Israeli bank (10 agents);
- Data-Based Estimation of $\widehat{\theta} = \dfrac{\#\text{ Abandoning}}{\text{Total Waiting Time}}$;
- Graph: Actual Performance vs. Erlang-A Predictions (aggregation of 40 similar hours): Model provides tight upper bounds.

$\mathbf{P\{Ab\}}$ $\qquad$ $\mathbf{E[}W_q\mathbf{]}$ $\qquad$ $\mathbf{P\{}W_q > 0\mathbf{\}}$

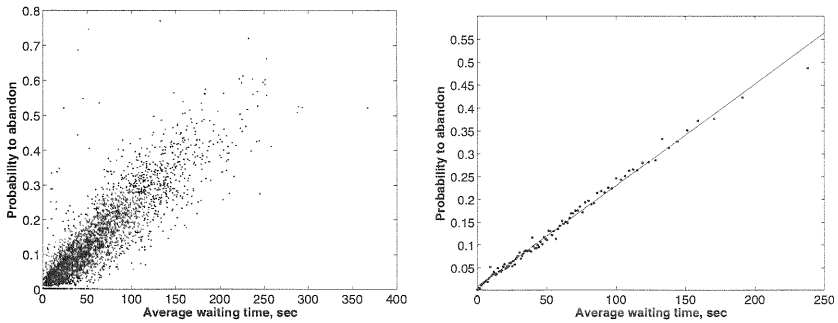


- **Question:** Why does Erlang-A works? indeed, **all** its underlying assumptions fail (Arrivals, Services, Impatience).
- **Towards a Theoretical Answer:** Robustness and Limitations, via Asymptotic (QED/QD) Analysis - later.
- **Practical Significance:** Asymptotic results applicable in small systems (eg. healthcare).

Queueing Science: In Support of Erlang-A

Israeli Bank: Yearly Data

Hourly Data $\qquad$ Aggregated



Data: $\mathbf{P\{Ab\} \propto E[}W_q\mathbf{]}$.

Theory: $\mathbf{P\{Ab\} = \theta \cdot E[}W_q\mathbf{]}$, **if** (Im)Patience = Exp(θ).

Proof: Let λ = Arrival Rate. Then, by Conservation & Little:

$$\lambda \cdot \mathrm{P\{Ab\}} \;=\; \theta \cdot \mathrm{E}[L_q] \;=\; \theta \cdot \lambda \cdot \mathrm{E}[W_q], \quad \text{q.e.d.}$$

Recipe: Use Erlang-A, with $\widehat{\theta} = \mathrm{P}\{\widehat{\mathrm{Ab}}\}/\mathrm{E}[\widehat{W}_q]$ (slope above).

But (Im)Patience is **not** Exponentially distributed !?

Queueing Science: via Data & Theory, Linearity Robust.
Service Engineering: via Theory & Simulations, often-enough,

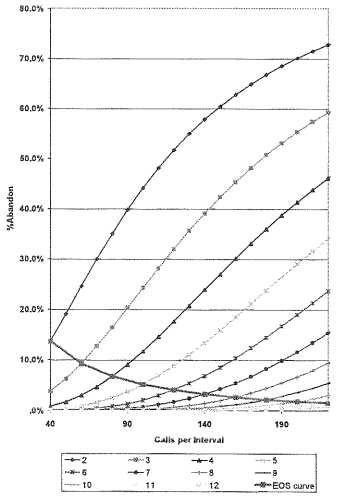
- Reality $\approx$ M/G/n + G $\approx$ Erlang-A, in which $\theta = g(0)$;
- $\mathrm{P\{Ab\}} \approx g(0) \cdot \mathrm{E}[W_q]$, hence recipe prevails, often enough.

4CallCenters: Congestion Curves

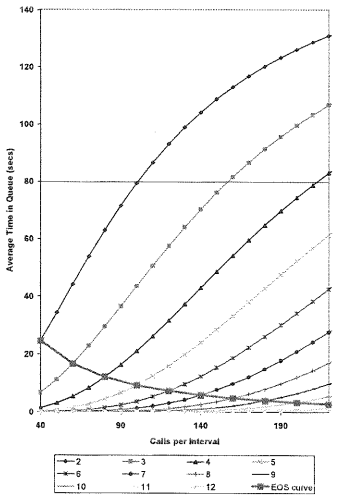
Vary input parameters of Erlang-A and display output (performance measures) in a table or graphically.

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: offered load per server fixed.
EOS (Economies-Of-Scale) observed.
Why are the two graphs similar?

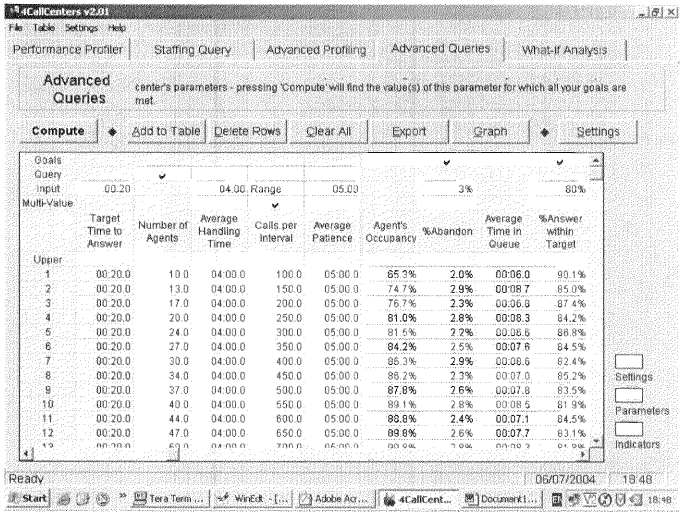
4CallCenters: Advanced Staffing Queries I

Set multiple performance goals.

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

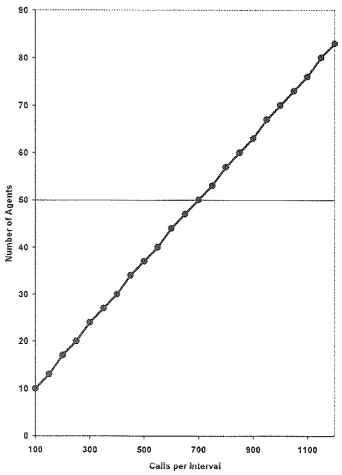
Performance targets:
$P\{\text{Ab}\} \leq 3\%$; $\quad \mathbf{P\{}W_q < \mathbf{20}$ sec; $\text{Sr}\mathbf{\} \geq 0.8}$.

4CallCenters output

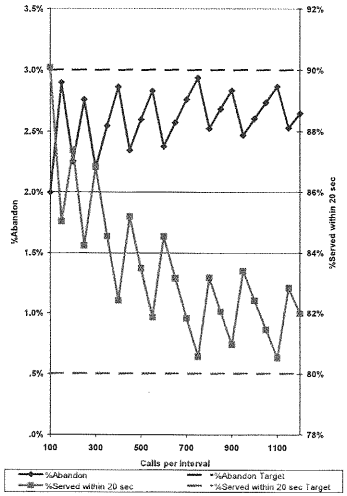


4CallCenters: Advanced Staffing Queries II

Recommended staffing level



Target performance measures



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

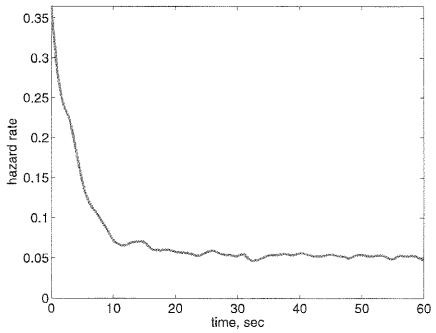
Back to General (Im)Patience: Empirical Patience Distributions

Are patience times Exponential?

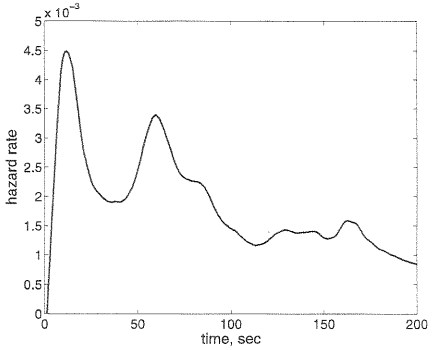
In the call centers that we studied, they are not!

Empirical hazard rates of patience times

U.S. bank

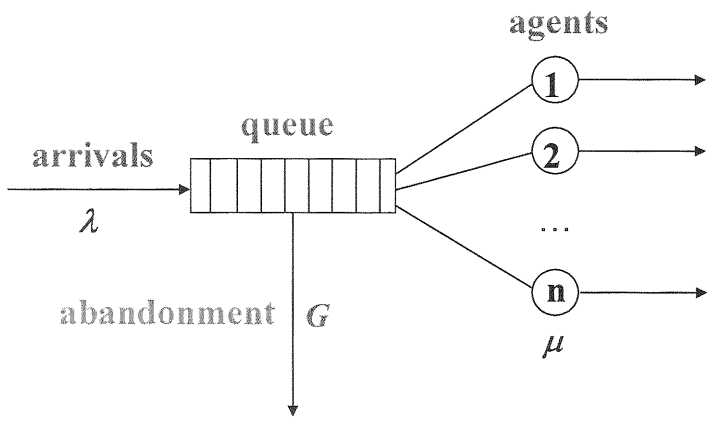


Israeli bank



To "uncensor data" use Kaplan-Meier (product-limit) estimator.
Output: estimates of **survival function** and **hazard rate**.

The M/M/n+G Queue



Patience times $\stackrel{d}{=}$ **G(eneral)**, i.i.d., independent of all else.

Performance measures can be computed, but calculations are cumbersome.

M/M/n+G: Building Blocks, for calculating Performance Measures

Reference (Support Material in website): with Zeltyn, prepared for Bank of America.

$$H(x) \triangleq \int_0^x \bar{G}(u)du\,,$$

where $\bar{G}(\cdot) = 1 - G(\cdot)$ is the survival function of (im)patience.

$$\begin{aligned}
J &\triangleq \int_0^\infty \exp\{\lambda H(x) - n\mu x\}\,dx\,,\\
J_1 &\triangleq \int_0^\infty x\cdot\exp\{\lambda H(x) - n\mu x\}\,dx\,,\\
J_H &\triangleq \int_0^\infty H(x)\cdot\exp\{\lambda H(x) - n\mu x\}\,dx\,,\\
J(t) &\triangleq \int_t^\infty \exp\{\lambda H(x) - n\mu x\}\,dx\,.\\
J_1(t) &\triangleq \int_t^\infty x\cdot\exp\{\lambda H(x) - n\mu x\}\,dx\,,\\
J_H(t) &\triangleq \int_t^\infty H(x)\cdot\exp\{\lambda H(x) - n\mu x\}\,dx\,.
\end{aligned}$$

Finally,

$$\mathcal{E} \triangleq \frac{\sum_{j=0}^{n-1}\frac{1}{j!}\left(\frac{\lambda}{\mu}\right)^j}{\frac{1}{(n-1)!}\left(\frac{\lambda}{\mu}\right)^{n-1}}\,.$$

M/M/n+G: Performance Measures

{Ab} = {Abandonment}, {Sr} = {Served},
W – waiting time, V – offered wait,
Q – queue length.

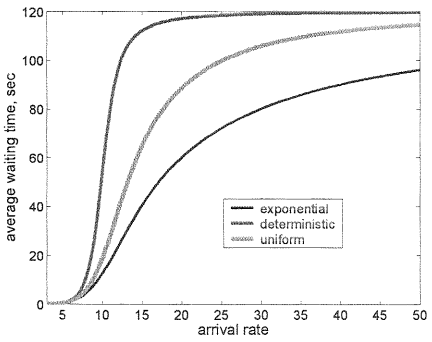
$$
\begin{aligned}
\mathrm{P}\{V>0\} &= \frac{\lambda J}{\mathcal{E}+\lambda J},\\
\mathrm{P}\{W>0\} &= \frac{\lambda J}{\mathcal{E}+\lambda J}\cdot \bar{G}(0),\\
\mathrm{P}\{\mathrm{Ab}\} &= \frac{1+(\lambda-n\mu)J}{\mathcal{E}+\lambda J},\\
\mathrm{P}\{\mathrm{Sr}\} &= \frac{\mathcal{E}+n\mu J-1}{\mathcal{E}+\lambda J},\\
\mathrm{E}[V] &= \frac{\lambda J_1}{\mathcal{E}+\lambda J},\\
\mathrm{E}[W] &= \frac{\lambda J_H}{\mathcal{E}+\lambda J},\\
\mathrm{E}[Q] &= \frac{\lambda^2 J_H}{\mathcal{E}+\lambda J},\\
\mathrm{E}[W\mid \mathrm{Ab}] &= \frac{J+\lambda J_H-n\mu J_1}{(\lambda-n\mu)J+1},\\
\mathrm{E}[W\mid \mathrm{Sr}] &= \frac{n\mu J_1-J}{\mathcal{E}+n\mu J-1},\\
\mathrm{P}\{W>t\} &= \frac{\lambda \bar{G}(t)J(t)}{\mathcal{E}+\lambda J},\\
\mathrm{E}[W\mid W>t] &= \frac{J_H(t)-(H(t)-t\bar{G}(t))\cdot J(t)}{\bar{G}(t)J(t)},\\
\mathrm{P}\{\mathrm{Ab}\mid W>t\} &= \frac{\lambda-n\mu-G(t)}{\lambda\bar{G}(t)}+\frac{\exp\{\lambda H(t)-n\mu t\}}{\lambda\bar{G}(t)J(t)}.
\end{aligned}
$$

M/M/n+G: Impact on Performance of Patience-Distribution

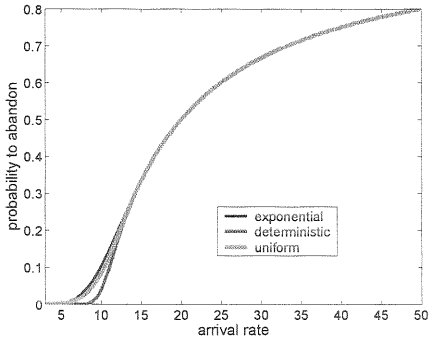
Parameters: 1 min average service time, 2 min average patience, 10 agents, arrival rate varies from 3 to 50 per minute.

G = Exponential, Deterministic, Uniform (mean = 2 min)

Average Wait

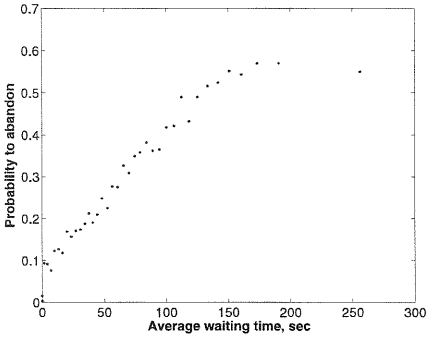


Probability to Abandon

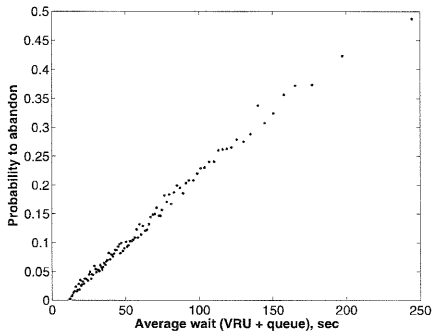


Applications of M/M/n+G Model: Linear Patterns of P{Ab}/E[W_q] with Non-Zero Intercepts

Israeli data: new customers



VRU-time included in wait



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

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

Customer-Focused Queueing Theory

Waiting experience of experienced customer often cycles through:

1. Time that a customer *expects* to wait;
2. Time that a customer is *willing* to wait ($\boldsymbol{\tau}$, patience or need);
3. Time that a customer *required* wait ($\boldsymbol{V}$, offered wait);
4. Time that a customer *actually* waits ($\boldsymbol{W_q = \min(\tau, V)}$);
5. Time that a customer *perceives* waiting.

Experienced customers $\Rightarrow$ 1=3.
Rational customers $\Rightarrow$ 4=5.
Thus left with $\boldsymbol{(\tau, V)}$, as in Erlang-A.

Eg. 200 abandonment in Direct-Banking: Perceived vs. Actual.

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

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 Palm/Erlang-A Queue, with Applications to Call Centers*

Avishai Mandelbaum and Sergey Zeltyn

Faculty of Industrial Engineering & Management
Technion.
Haifa 32000, ISRAEL

emails: avim@tx.technion.ac.il, zeltyn@ie.technion.ac.il

December 28, 2004

Contents

*Parts of the text are adapted from [11], [19], [22], [28] and [40]

4 Operational measures of performance

In order to understand and apply the Erlang-A model, one must first define its measures of performance, and then be able to calculate them. Moreover, since call centers can get very large (thousands of agents). the implementation of these calculations must be both fast and numerically stable.

4.1 Practical measures: Waiting Time

The most popular measure of operational (positive) performance is the fraction of served customers that have been waiting less than some given time, or formally $\mathrm{P}\{W \leq T, \mathrm{Sr}\}$, where W is the (random) waiting time in steady-state, $\{\mathrm{Sr}\}$ is the event "customer gets service" and T is a target time that is determined by Management/Marketing. For example, in a call center that caters to emergency calls, $T = 0$ (or T very small) would be appropriate. A common rule of thumb (without any theoretical backing, as far as we know) is the goal that at least 80% of the customers be served within 20 seconds; formally, $P\{W \leq 20, \mathrm{Sr}\} \geq 0.8$. To this, one sometimes adds E[W], or $\mathrm{E}[W|W > 0]$, as some measure of an average (negative) experience for those who waited.

An important measure that is rarely used in practice is $P\{W > 0\}$, the fraction of customers who encounter a delay. This is a useful stable measure of congestion. Its importance stems from the fact that it identifies an organization's operational focus, in the following sense:

- $P\{W > 0\}$ close to 0 indicates a Quality-Driven operation, where the focus is on *service quality*;
- $P\{W > 0\}$ close to 1 indicates an Efficiency-Driven operation, where the focus is on *servers' efficiency* (in the sense of high servers' utilization);
- $P\{W > 0\}$ strictly between 0 and 1 (for example 0.5) indicates a careful *balance* between Quality and Efficiency, which we abbreviate to **QED = Quality & Efficiency Driven** operational regime.

The above three-regime dichotomy is rather delicate. For example, consider a system in which customers' average patience is close to the average service duration (for example, let both be equal to one minute), and assume that its offered load λ/μ is 100 Erlangs. Then, staffing of 100 servers would lead to the QED regime, with high levels of both service and efficiency that are balanced as follows: about 50% of the customers are served immediately upon arrival, the average wait is 2.3 seconds, 4% of the customers abandon due to their impatience, and servers' utilization levels are 96%. The QED regime still prevails at staffing levels between 95 and 105. With 90 servers, the system is efficiency-driven: 11% of the customers abandon, only 15% are served immediately, and utilization is over 99%. With 110 agents, it is quality-driven: abandonment is less than 1%, and 83% are served immediately.

In Section 6, we shall add details about the three operational regimes. This will be done in the context of describing regime-specific approximations for performance measures. However, there is much more to say about this important subject, and readers are referred to [19, 9] and Section 4 in the review [17] for details.

4.2 Practical measures: accounting for Abandonment

In a quality-driven service, $P\{W > 0\}$ seems the "right" measure of operational performance. We thus turn to alternative modes of operations and consider hereafter services in which $P\{W > 0\}$ is not close to vanishing.

As explained before, performance measures must take into account those customers who abandon. Indeed, if forced into choosing a *single* number as a proxy for operational performance, we recommend the probability to abandon $\mathbf{P\{Ab\}}$, the fraction of customers who explicitly declare that the service offered is not worth its wait. Some managers actually opt for a refinement that excludes those who abandon within a very short time, formally $\mathrm{P}\{W > \epsilon; \mathbf{Ab}\}$, for some small $\epsilon > 0$, for example $\epsilon = 3$ seconds. The justification is that those who abandon within 3 seconds can not be characterized as poorly served. There is also a practical rational that arises from physical limitations, specifically that such "immediate" abandonment could in fact be a malfunction or an inaccuracy of the measurement devices.

The single abandonment measure $\mathrm{P}\{\mathrm{Ab}\}$ can be in fact refined to account explicitly for those customers who were or were not well-served. Thus, we propose:

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

A further refinement, that yields a four-dimensional service measure, could be:

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

Remark 4.1 4CallCenters [16] calculates, for a given target time, both $P\{W \le T; \mathrm{Sr}\}$, the fraction of customers who are served within target, and $P\{W \le \epsilon; \mathrm{Ab}\}$, those who abandon within target. To calculate the other two measures, it suffices to have $\mathrm{P}\{\mathrm{Ab}\}$, also calculated by 4CallCenters. Indeed,

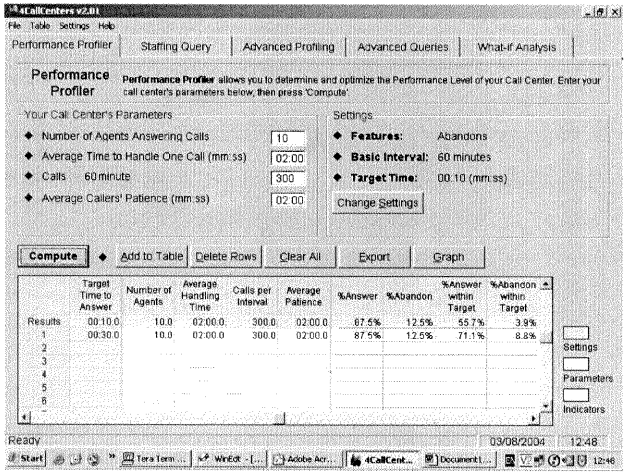
$$\begin{aligned} \mathrm{P}\{W > T; \mathrm{Sr}\} &= 1 - \mathrm{P}\{\mathrm{Ab}\} - \mathrm{P}\{W \le T; \mathrm{Sr}\}, \\ \mathrm{P}\{W > \epsilon; \mathrm{Ab}\} &= \mathrm{P}\{\mathrm{Ab}\} - \mathrm{P}\{W \le \epsilon; \mathrm{Ab}\}. \end{aligned}$$

Since a single target must be used ($T = \epsilon$ above), one must apply the program twice if different targets are required.

4.3 Calculations: the 4CallCenters software

Black-box Erlang-A calculations, as well as many other useful features, are provided by the free-to-use software 4CallCenters [16]. (This software is being regularly debugged and upgraded.) The calculation methods are described in Appendix B of [19]; they were developed in the Technion's M.Sc. thesis of the first author, Ofer Garnett.

Figure 5: **4Callcenters. Example of output.**



These calculations are in fact for measures of the form $\mathrm{E}[f(V,\tau)]$, for various functions f (Table 3 in [19]). For example,

$$\mathrm{E}[W] = \mathrm{E}\left[\min\{V,\tau\}\right] \ , \qquad \mathrm{P}\{\mathrm{Abandon}\} = \mathrm{E}\left[1_{\{\tau<V\}}\right] \ .$$

Figure 5 displays a 4CallCenters output and demonstrates how to calculate the four-dimensional service measure, introduced in Subsection 4.2.

The values of the four Erlang-A parameters are displayed in the middle of the upper half of the screen: $n = 10$, $1/\mu = 2$ minutes, $\lambda = 300$ calls per hour, $1/\theta = 2$ minutes. Let $T = 30$

seconds and $\epsilon = 10$ seconds. Then one should perform computations twice: with *Target Time* 30 and 10 seconds. (Both computations appear in Figure 5.) We get:

- $P\{W \le T; \mathrm{Sr}\}$ - fraction of well-served is equal to 71.1%;
- $P\{W > T; \mathrm{Sr}\}$ - fraction of served, with a potential for improvement, is 16.4% (87.5% − 71.1%);
- $P\{W > \epsilon; \mathrm{Ab}\}$ - fraction of poorly-served is 8.6% (12.5% − 3.9%);
- $P\{W \le \epsilon; \mathrm{Ab}\}$ - fraction of those whose service-level is undetermined is 3.9%.

Note that the 4CallCenters output includes many more performance measures than those displayed in Figure 5: one could scroll the screen to values of agents' occupancy, average waiting time, average queue length, etc.

In Section 9 we describe several examples of the more advanced capabilities of 4CallCenters.

4.4 Delay probability P{W>0}

In this note, we content ourselves with few representative insightful calculations, based on conditioning and the incomplete gamma function introduced above. We start with the *delay probability* $\mathrm{P}\{W > 0\}$, which represents the fraction of customers who are forced to actually wait for service. (The others are served immediately upon calling.) Recall that this measure identifies operational regimes of performance.

Following Palm [31], we show in the Appendix that the representations (3.5) and (3.7) immediately imply

$$\mathrm{P}\{W > 0\} \ = \ \sum_{j=n}^{\infty} \pi_j \ = \ \frac{A\left(\frac{n\mu}{\theta}, \frac{\lambda}{\theta}\right) \cdot E_{1,n}}{1 + \left(A\left(\frac{n\mu}{\theta}, \frac{\lambda}{\theta}\right) - 1\right) \cdot E_{1,n}} \ ; \tag{4.1}$$

here, the first equality in (4.1) follows from PASTA.

4.5 Fraction abandoning P{Ab}

We proceed with calculating the probability to abandon, which represents the fraction abandoning. Define $\mathrm{P}_j\{\mathrm{Sr}\}$ to be the probability of ultimately getting served, for a customer that encounters all servers busy and j customers in queue, upon arrival (equivalently, $n + j$ in the system). "Competition among exponentials" now implies that

$$\mathrm{P}_0\{\mathrm{Sr}\} = \frac{n\mu}{n\mu + \theta} \ .$$

Then,

$$\mathrm{P}_1\{\mathrm{Sr}\} = \frac{n\mu+\theta}{n\mu+2\theta}\cdot \mathrm{P}_0\{\mathrm{Sr}\} \;=\; \frac{n\mu}{n\mu+2\theta}\,,$$

where we conditioned on the first event, after an arrival that encounters all servers busy and a single customer in queue; this event is either a service completion (with probability $\frac{n\mu+\theta}{n\mu+2\theta}$) or an abandonment. More generally, via induction:

$$\mathrm{P}_j\{\mathrm{Sr}\} = \frac{n\mu+j\theta}{n\mu+(j+1)\theta}\cdot \mathrm{P}_{j-1}\{\mathrm{Sr}\} \;=\; \frac{n\mu}{n\mu+(j+1)\theta}\,, \qquad j\geq 1\,.$$

The probability to abandon service, given all servers busy and j customers in the queue upon arrival, finally equals

$$\mathrm{P}_j\{\mathrm{Ab}\} \;=\; 1-\mathrm{P}_j\{\mathrm{Sr}\} \;=\; \frac{(j+1)\theta}{n\mu+(j+1)\theta}\,, \qquad j\geq 0\,. \tag{4.2}$$

It follows that

$$\mathrm{P}[\mathrm{Ab}|W>0] \;=\sum_{j=n}^{\infty}\pi_j \mathrm{P}_{j-n}\{\mathrm{Ab}\}/\,\mathrm{P}\{W>0\} \;=\; \frac{1}{\rho A\left(\frac{n\mu}{\theta},\frac{\lambda}{\theta}\right)}+1-\frac{1}{\rho}\,. \tag{4.3}$$

The first equality in (4.3) is a consequence of PASTA, and the second is derived in the Appendix. The fraction abandoning, $\mathrm{P}\{\mathrm{Ab}\}$, is simply the product $\mathrm{P}[\mathrm{Ab}|W>0]\times\mathrm{P}\{W>0\}$.

4.6 Theoretical relations among P{Ab}, E(W), E(Q)

A remarkable property of Erlang-A, which in fact generalizes to other models with patience that is *exp*(θ), is the following linear relation between the fraction abandoning $\mathrm{P}\{\mathrm{Ab}\}$ and average wait $\mathrm{E}[W]$:

$$\mathrm{P}\{\mathrm{Ab}\} \;=\; \theta\cdot\mathrm{E}[W]\,. \tag{4.4}$$

Proof: The proof is based on the balance equation

$$\theta\cdot\mathrm{E}[Q] \;=\; \lambda\cdot\mathrm{P}\{\mathrm{Ab}\}\,, \tag{4.5}$$

and on Little's formula

$$\mathrm{E}[Q] \;=\; \lambda\cdot\mathrm{E}[W]\,, \tag{4.6}$$

where Q is the steady-state queue length. The balance equation (4.5) is a steady-state equality between the rate that customers abandon the queue (left hand side) and the rate that abandoning customers (i.e. - customers who eventually abandon) enter the system. Substituting Little's formula (4.6) into (4.5) yields formula (4.4). ■

Observe that (4.4) is equivalent to

$$\mathrm{P}[\mathrm{Ab}|W>0] \;=\; \theta\cdot\mathrm{E}[W|W>0]\,. \tag{4.7}$$

Then, the average waiting time of delayed customers is computed via (4.3) and (4.7):

$$\mathrm{E}[W|W>0] \;=\; \frac{1}{\theta}\cdot\left[\frac{1}{\rho A\left(\frac{n\mu}{\theta},\frac{\lambda}{\theta}\right)}+1-\frac{1}{\rho}\right]. \tag{4.8}$$

The unconditional average wait $\mathrm{E}[W]$ equals the product of (4.1) with (4.8).

4.7 A general approach for computing operational performance measures

Expressions for additional performance measures of Erlang-A are derived in Riordan [32]. However, we recommend to use more general M/M/n+G formulae, as the main alternative to the 4CallCenters software. Indeed, M/M/n+G is a generalization of Erlang-A, in which patience times are generally distributed. A comprehensive list of M/M/n+G formulae, as well as guidance for their application, appears in Mandelbaum and Zeltyn [30]. The preparation of [30] was triggered by a request from a large U.S. bank. Consequently, this bank has been routinely applying Erlang-A in the workforce management of its 10,000 telephone agents, who handle close to 150 millions calls yearly. (In fact, Erlang-A replaced a simulation tool that had been used before.)

The handout [30] also explains how to adapt the M/M/n+G formulae to Erlang-A, in which patience is exponentially distributed:

$$G(x) = 1-e^{-\theta x}\,, \qquad \theta>0.$$

Specifically, see Sections 1.2 and 5 of [30].

Finally, we explain how to calculate the four service measures from Section 4.2. The list on page 4 of [30] contains formulae for $\mathrm{P}\{\mathrm{Ab}\}$, $\mathrm{P}\{W>T\}$ and $\mathrm{P}\{\mathrm{Ab}|W>T\}$. The product of the last two provides us with $\mathrm{P}\{W>T;\mathrm{Ab}\}$. The other three service measures are easily derived. For example,

$$\mathrm{P}\{W>T;\mathrm{Sr}\} \;=\; \mathrm{P}\{W>T\}-\mathrm{P}\{W>T;\mathrm{Ab}\}\,.$$

4.8 Empirical relations between E(W) and P{Ab}

Figure 6 illustrates the relation (4.4). It was plotted using yearly data of an Israeli bank call center [12]. (See also Brown et al. [11] for statistical analysis of this call center data.) First,

The M/M/n+G Queue: Summary of Performance Measures

Avishai Mandelbaum and Sergey Zeltyn

Faculty of Industrial Engineering & Management
Technion
Haifa 32000, ISRAEL

emails: avim@tx.technion.ac.il, zeltyn@ie.technion.ac.il

May 10, 2004

Contents

Designing a Call Center with Impatient Customers

O. Garnett* A. Mandelbaum*† M. Reiman ‡

March **26**, 2002

ABSTRACT. The most common model to support workforce management of telephone call centers is the $M/M/N/B$ model, in particular its special cases $M/M/N$ (Erlang C, which models out busy-signals) and $M/M/N/N$ (Erlang B, disallowing waiting). All of these models lack a central prevalent feature, namely that impatient customers might decide to leave (abandon) before their service begins.

In this paper we analyze the simplest abandonment model, in which customers' patience is exponentially distributed and the system's waiting capacity is unlimited ($M/M/N+M$). Such a model is both rich and analyzable enough to provide information that is practically important for call center managers. We first outline a method for exact analysis of the $M/M/N+M$ model, that while numerically tractable is not very insightful. We then proceed with an asymptotic analysis of the $M/M/N+M$ model, in a regime that is appropriate for large call centers (many agents, high efficiency, high service level). Guided by the asymptotic behavior, we derive approximations for performance measures and propose "rules of thumb" for the design of large call centers. We thus add support to the growing acknowledgment that insights from diffusion approximations are directly applicable to management practice.

*Davidson Faculty of Industrial Engineering and Management, Technion, Haifa 32000, ISRAEL.

†Research supported by the fund for the promotion of research at the Technion, by the Technion V.P.R. funds - Smoler Research Fund, and B. and G. Greenberg Research Fund (Ottawa), and by the Israel Science Foundation (grant no. 388/99).

‡Bell Laboratories, Murray Hill, NJ 07974, USA.