Designing a Telephone Call Center with Impatient Customers

with

Ofer Garnett

Marty Reiman

Sergey Zeltyn

Appendix: Strong Approximations of M/M/N + N

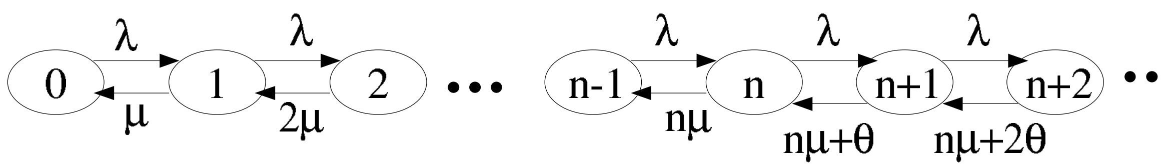
Source: ErlangA_QEDandSTRONG_FINAL.tex

M/M/N + M System

- Poisson arrivals-rate λ
- Service times – $\exp(\mu)$
- N statistically identical agents attending to single queue
- Service policy FCFS
- Customers' patience: $\exp(\theta)$

$Q = \{Q(t), t \geq 0\}$ - number of customers in the system

Birth & Death: transition diagram



"Everything" calculable via stationary distribution

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

where

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

M/M/N + M Characteristics

Notation:

- $P\{Ab\}$ - abandonment probability (fraction)
- $P\{Wait > 0\}$ - waiting probability (fraction)
- $P\{Block\}$ - blocking probability in an $M/M/N/N$ system

1. Stationary distribution always exists

   (Sandwiched between infinite-server models)

2. $P\{Ab\} = \theta \cdot E[Wait]$

Proof: $\lambda P\{Ab\} = \theta \cdot E[\text{Number in queue}]$, now use Little.

3. $P\{Ab\}$ increases monotonically in θ, λ

   $P\{Ab\}$ decreases monotonically in N, μ

   (Bhattacharya and Ephremides (1991))

4. $P\{Ab\} \leq P\{Block\}$

   (Boxma and de Waal (1994))

5. $\lim_{N\uparrow\infty} P_N\{Ab\} = 1 - \left(1 \vee \lim_{N\uparrow\infty} \rho_N\right)^{-1}$

Exact Calculations

- V - virtual waiting time $=$ waiting time of a customer
  with infinite patience (test customer).

- X - customer's patience ($X \sim \exp(\theta)$, independent of V).

- Wait $\equiv V \wedge X$ - actual waiting time.

Performance measures of the form $E[f(V,X)]$:
Calculable, in numerically stable procedures (**4CallCenters**).

$f(v,x)$	$E[f(V,X)]$
$1_{\{v>x\}}$	$P\{V>X\} = P\{Ab\}$
$1_{(t,\infty)}(v \wedge x)$	$P\{Wait > t\}$
$1_{(t,\infty)}(v \wedge x)1_{\{v>x\}}$	$P\{Wait > t;\, Ab\}$
$(v \wedge x)1_{\{v>x\}}$	$E[Wait;\, Ab]$
$(v \wedge x)1_{(t,\infty)}(v \wedge x)1_{\{v>x\}}$	$E[Wait; Wait > t;\, Ab]$
$g(v \wedge x)$	$E[g(Wait)]$

From these obtain more "natural" measures, for example

$$P\{Ab|Wait > t\} = \frac{P\{Wait > t;\, Ab\}}{P\{Wait > t\}}$$

QED M/M/N + M $(\theta_N \equiv \theta)$

Following Halfin-Whitt (1981) and Flemming-Simon-Stolyar (1996)

Theorem: *Let*

$$
\begin{aligned}
\alpha &\equiv \lim_{N\to\infty} P_N\{Wait > 0\} \\
\beta &\equiv \lim_{N\to\infty} \sqrt{N}\ (1-\rho_N) \qquad \Rightarrow \quad N \sim \tfrac{\lambda}{\mu} + \beta\sqrt{\tfrac{\lambda}{\mu}} \\
\gamma &\equiv \lim_{N\to\infty} \sqrt{N} P_N\{Ab\} \qquad \Rightarrow \quad P\{Ab\} \sim \tfrac{\gamma}{\sqrt{N}}
\end{aligned}
$$

Then

$$
\begin{aligned}
1 > \alpha > 0 \quad &\textit{iff} \quad \infty > \beta > -\infty \\
&\textit{iff} \quad \infty > \gamma > 0
\end{aligned}
$$

in which case

$$
\begin{aligned}
\alpha &= w(-\beta, \sqrt{\mu/\theta}) \\
\gamma &= \left[\sqrt{\theta/\mu} \cdot h(\beta\sqrt{\mu/\theta}) - \beta\right] \cdot \alpha
\end{aligned}
$$

Here

$$
\begin{aligned}
w(x,y) &= \left[1 + \frac{h(-xy)}{yh(x)}\right]^{-1}, \\
h(x) &= \frac{\phi(x)}{1-\Phi(x)}, \quad \text{hazard rate of std. normal}
\end{aligned}
$$

Designing a QED Call Center (Zeltyn)

(Approximate) Performance Measures

$$P\{Wait > 0\} \approx \left[1 + \frac{h(r\beta)}{rh(-\beta)}\right]^{-1}, \qquad \mathbf{r} = \sqrt{\frac{\mu}{\theta}}$$

$$E\left[\frac{Wait}{\mathsf{E}[S]}\middle| Wait > 0\right] \approx \frac{1}{\sqrt{N}} \cdot r \cdot [h(r\beta) - r\beta]$$

$$P\{Ab\} \approx \frac{1}{\sqrt{N}} \cdot \frac{h(r\beta) - r\beta}{r} \cdot \left[1 + \frac{h(r\beta)}{rh(-\beta)}\right]^{-1}$$

$$P\{Ab|Wait > 0\} \approx \frac{1}{\sqrt{N}} \cdot \frac{h(r\beta) - r\beta}{r}$$

$$P\left\{\frac{Wait}{\mathsf{E}[S]} > \frac{t}{\sqrt{N}}\middle| Wait > 0\right\} \approx \frac{\bar{\Phi}(r\beta + t/r)}{\bar{\Phi}(r\beta)}$$

$$P\left\{\mathsf{Ab}\middle|\frac{Wait}{\mathsf{E}[S]} > \frac{t}{\sqrt{N}}\right\} \approx \frac{1}{\sqrt{N}} \cdot \frac{h(r\beta + t/r) - r\beta}{r}$$

$$E\left[\frac{Wait}{\mathsf{E}[S]}\middle| Ab\right] \approx \frac{1}{\sqrt{N}} \cdot \frac{r}{2} \cdot \left[\frac{1}{h(r\beta) - r\beta} - r\beta\right]$$

Here

$$\bar{\Phi}(x) = 1 - \Phi(x)\,,$$

$$h(x) = \phi(x)/\bar{\Phi}(x)\,, \quad \text{hazard rate of } N(0,1).$$

New Operational Regimes (M/M/N + N)

	N	ρ	$P\{\text{Wait} > 0\}$	$P\{Ab\}$
Efficiency driven	$R - \epsilon R$	$1 + \epsilon$	1	ϵ
Quality driven	$R + \epsilon R$	$1 - \epsilon$	~ 0	~ 0
QED	$R + \beta\sqrt{R}$ $-\infty < \beta < \infty$	$1 - \frac{\beta}{\sqrt{N}}$	$\alpha(\beta)$ $0 < \alpha < 1$	$\frac{\Delta(\beta)}{\sqrt{N}}$

Compare with Previous (M/M/N, Halfin-Whitt)

E	$R + \epsilon$	$1 - \frac{\epsilon}{N}$	1	$-$
Q	$R + \epsilon R$	$\frac{1}{1 + \epsilon}$	~ 0	$-$
QED	$R + \beta\sqrt{R}$ $0 < \beta < \infty$	$1 - \frac{\beta}{\sqrt{N}}$	$\alpha(\beta)$ $0 < \alpha < 1$	$-$

Sequence of M/M/N + M , N = 1, 2, ...

Motivation: Insight, Numerical stability

Focus: Large Call Centers — λ, N large.

Framework: Sequence of $M/M/N + M$ systems, indexed by N

- $Q_N = \{Q_N(t),\ t \geq 0\}$: total number in system;
- $V_N = \{V_N(t),\ t \geq 0\}$: virtual wait of an infinite-patience customer
- Parameters $\lambda_N,\ \mu,\quad \theta_N$

$\theta_N \uparrow \infty$	impatient
$\theta_N \downarrow 0$	patient
$\theta_N \equiv \theta$	"rational"

Offered load $R_N = \dfrac{\lambda_N}{\mu} \quad (\rho_N = R_N/N)$

Approximations of Process and Stationary Distribution

$$\hat{Q}_N(t) = \frac{1}{\sqrt{N}} \left[Q_N(t) - N\right], \quad 0 \leq t \leq \infty$$

$$\hat{V}_N(t) = \sqrt{N}\ V_N(t)$$

Approximating Queueing and Waiting

- $Q_N = \{Q_N(t),\ t \geq 0\}$: $Q_N(t) =$ number in system at $t \geq 0$.

- $\widehat{Q}_N = \{\widehat{Q}_N(t), t \geq 0\}$: stochastic process obtained by centering and rescaling:

$$\widehat{Q}_N = \frac{Q_N - N}{\sqrt{N}}$$

- $\widehat{Q}_N(\infty)$: stationary distribution of $\widehat{Q}_N$

- $\widehat{Q} = \{\widehat{Q}(t), t \geq 0\}$: process defined by: $\widehat{Q}_N(t) \xrightarrow{d} \widehat{Q}(t)$.

$$\begin{tikzcd}
\widehat{Q}_N(t) \arrow[r, "t \to \infty"] \arrow[d, "N \to \infty"'] & \widehat{Q}_N(\infty) \arrow[d, "N \to \infty"] \\
\widehat{Q}(t) \arrow[r, "t \to \infty"'] & Q(\infty)
\end{tikzcd}$$

Approximating (Virtual) Waiting Time

$$\widehat{V}_N = \sqrt{N}\ V_N \Rightarrow \widehat{V} = \left[\frac{1}{\mu}\,\widehat{Q}\right]^+ \qquad \text{(Puhalskii, 1994)}$$

Weak Convergence of Stationary Distribution

Assume $\lim_{N\to\infty} \sqrt{N}\,(1-\rho_N) = \beta$, $-\infty < \beta < \infty$.

Allow θ_N to vary with $N \uparrow \infty$.

$\widehat{Q}_N(\infty)$ converges iff at least one of the following prevails:

1. $\theta_N \equiv \theta$ or $\theta_N \uparrow \infty$

2. $\beta > 0$

in which case $\widehat{Q}_N(\infty) \xrightarrow{d} \widehat{Q}(\infty)$, where

Density function $f(x)$ of $\widehat{Q}(\infty)$:

$$\theta_N \downarrow 0,\ \beta > 0 \qquad f(x) = \begin{cases} A_1\phi(\beta + x) & x \le 0 \\ A_2 \exp(-\beta x) & x > 0 \end{cases}$$

$$\theta_N \equiv \theta \qquad f(x) = \begin{cases} B_1\phi(\beta + x) & x \le 0 \\ B_2\phi\left(\beta\sqrt{\mu/\theta} + x\sqrt{\theta/\mu}\right) & x > 0 \end{cases}$$

$$\theta_N \uparrow \infty \qquad f(x) = \begin{cases} C\phi(\beta + x) & x \le 0 \\ 0 & x > 0 \end{cases}$$

Weak Convergence of Processes

$$\left(\boldsymbol{\rho_N \sim 1 - \frac{\beta}{\sqrt{N}}\ ;\ N \sim \frac{\lambda}{\mu} + \beta\sqrt{\frac{\lambda}{\mu}}}\right)$$

Theorem: $\hat{Q}_N \xrightarrow{d} \hat{Q}$ *if* $\hat{Q}_N(0) \xrightarrow{d} \hat{Q}(0)$.

Theorem: $\hat{V}_N \xrightarrow{d} \hat{V} \stackrel{d}{=} \frac{1}{\mu}\,\hat{Q}^+$.

$\theta \downarrow 0$ Patient

$$\begin{cases} d\hat{Q}(t) = f(\hat{Q})dt + \sqrt{2\mu}\,dB(t) \\ f(x) = \begin{cases} -\mu(\beta + x) & x \le 0 \\ -\mu\beta & x > 0 \end{cases} \end{cases} \qquad \begin{pmatrix} OU & \hat{Q} \le 0 \\ BM & \hat{Q} > 0 \end{pmatrix}$$

M/M/N
Erlang C

$\theta \uparrow \infty$ Impatient

$$\begin{cases} d\hat{Q}(t) = -\mu(\beta + \hat{Q}(t))dt \\ \qquad\qquad + \sqrt{2\mu}\,dB(t) - dY(t) \qquad (ROU\ \hat{Q} \le 0) \\ Y(0) = 0\ ,\ Y \uparrow 0\ ,\ \hat{Q}\,dY = 0 \end{cases}$$

M/M/N/N
Erlang B

θ fixed Rational

$$\begin{cases} d\hat{Q}(t) = f(\hat{Q})dt + \sqrt{2\mu}\,dB(t) \\ f(x) = \begin{cases} -\mu(\beta + x) & x \le 0 \\ -(\mu\beta + \theta x) & x > 0 \end{cases} \end{cases} \qquad \begin{pmatrix} OU & \hat{Q} \le 0 \\ OU & \hat{Q} > 0 \end{pmatrix}$$

(B - standard Brownian Motion)

(β - as before)

Designing a Call Center - Selecting a Model

- Very impatient customers - $M/M/N/N$ model
- Very patient customers - $M/M/N$ model
- “Balanced” abandoning - $M/M/N + M$, θ fixed

What if General Patience with distribution function G?

Steady-state formulae prevail with

$$\theta \leftrightarrow \mathbf{G'(0)}.$$

M/M/N+G System

Customers' patience: general distribution G, with $g_0 = G'(0)$.

Steady-State Results

- Baccelli & Hebuterne (1981) - virtual wait, abandon probability.

- Brandt & Brandt (1999, 2002) - stationary-queue distribution.

QED Regime

Main Case: $G(0) = 0,\ g_0 > 0$ (no balking).

Use $M/M/N + M$ formulae: $\theta \to g_0,\ r = \sqrt{\frac{\mu}{g_0}}$.

Hence, $P\{Ab\} \approx g_0 \cdot E[W]$.

Special Cases:

- Balking $(G(0) > 0,\ g_0 > 0)$;

- $g_0 = 0$: k-th derivative of G at zero positive, $k \geq 2$ (Erlang, Phase-type);

- $g_0 = 0$: Delayed patience distributions (const, c+exp(θ)).

Asymptotic analysis possible for all special cases, which yields varying convergence rates.

Appendix

M/M/N + M : Strong Approximations

Parameters: λ, μ, θ, N

Building Blocks: A_i, independent Poisson (1) processes

Model: $Q = \{Q(t),\ t \geq 0\}$ total number in system

$$\begin{aligned} Q(t) = Q(0) \quad & + A_1(\lambda t) && \text{arrivals} \\ & - A_2\left(\int_0^t \mu \cdot [Q(u) \wedge N] du\right) && \text{services} \\ & - A_3\left(\int_0^t \theta \cdot [Q(u) - N]^+ du\right) && \text{abandons} \end{aligned}$$

Strong Approximations: on the same probability space with A_i's, there are SBM B_i's such that

$$A_i(t) = t + B_i(t) + O(\log t), \quad t \uparrow \infty .$$

$\Rightarrow$ Approximate Q by substituting above

$$A_i(t) \leftrightarrow t + B_i(t) .$$

Approximation: $Q(t) = \tilde{Q}(t) + o\left(\sqrt{N}\,\right)$, u.o.c.,

via $\quad A_i(t) \leftrightarrow t + B_i(t)$

$$
\begin{aligned}
\tilde{Q}(t) = Q(0) + \lambda t \quad &+ \quad B_1(\lambda t) \\
- \int_0^t \mu \cdot [\tilde{Q}(u) \wedge N] du \quad &- \quad B_2\left(\int_0^t \mu \cdot [\tilde{Q}(u) \wedge N] du\right) \\
- \int_0^t \theta \cdot [\tilde{Q}(u) - N]^+ du \quad &- \quad B_3\left(\int_0^t \theta \cdot [\tilde{Q}(u) - N]^+ du\right)
\end{aligned}
$$

Insight?

Limit theorems (FSLLN, FCLT) as $N \uparrow \infty$,

for $\tilde{Q}_N$, hence for Q_N:

For example, $\qquad \dfrac{1}{N} Q_N(t) = \dfrac{1}{N} \tilde{Q}_N(t) + o(1).$

QED M/M/N + N : Approximations, Limits

For simplicity $Q_N(0) \equiv N$: all servers busy, no queue.

Recall $\lambda_N = \mu N - \mu B\sqrt{N}, \ \mu, \theta$

Theorem: Strong **Approximation**.

$$Q_N(t) = N + \sqrt{N}\,\hat{Q}(t) + o\left(\sqrt{N}\right) \text{ u.o.c., as } N \uparrow \infty$$

or
$$\frac{1}{\sqrt{N}}\left[Q(t) - N\right] \stackrel{d}{\approx} \hat{Q}(t) \qquad 0 \le t \le \infty \ ,$$

where

$$d\hat{Q}(t) = \left[-\mu\beta + \mu\hat{Q}^-(t) - \theta\hat{Q}^+(t)\right] dt + \sqrt{2\mu}\, dB(t) \ ;$$

namely $\hat{Q}$ is a diffusion with

$$\mu(x) = \begin{cases} -\mu\beta - \theta x & x \ge 0 \\ -\mu\beta - \mu x & x \le 0 \end{cases} , \quad \sigma^2(x) \equiv 2\mu \,.$$

$\Rightarrow$ **FSLLN** $\frac{1}{N}\, Q_N(t) \to 1$ u.o.c., a.s.

$\Rightarrow$ **FCLT** $\sqrt{N}\,\left[\frac{1}{N}\, Q_N - 1\right] = \frac{1}{\sqrt{N}}\,[Q_N - N] \xrightarrow{d} \hat{Q}$

QED M/M/N + N : FSLLN

$$
\begin{aligned}
\tilde{Q}_N(t) &= N + \lambda_N t + B_1(\lambda_N t) \\
&\quad - \int_0^t \mu \cdot [\tilde{Q}_N(u) \wedge N] du - B_2\left(\int_0^t \mu \cdot [\tilde{Q}_N(u) \wedge N] du\right) \\
&\quad - \int_0^t \theta \cdot [\tilde{Q}_N(u) - N]^+ du - B_3(\cdots)
\end{aligned}
$$

FSLLN: $\frac{1}{N}\,\tilde{Q}_N(t) = 1 + \frac{1}{N}\,\lambda_N t + \frac{1}{N}\,B_1\left(N \cdot \frac{\lambda_N}{N}\,t\right)$

$$
- \int_0^t \mu \cdot \left[\frac{1}{N}\,\tilde{Q}_N(u) \wedge 1\right] du - \frac{1}{N}\,B_2\left(N \int_0^t \mu \cdot \left[\frac{1}{N}\,\tilde{Q}_N(u) \wedge 1\right] du\right)
$$

$- \ \ldots$

Observe	$\frac{1}{N}\,\lambda_N t \to \mu t$	since $\lambda_N = \mu N - \mu\beta\sqrt{N}$
and	$\frac{1}{N}\,B_i(Nt) \to 0$	FSLLN for SBM.
If indeed	$\frac{1}{N}\,\tilde{Q}_N(t) \to \overline{Q}(t)$,	then

$$
\overline{Q}(t) = 1 + \mu t - \mu \int_0^t [\overline{Q}(u) \wedge 1] du - \theta \int_0^t [\overline{Q}(u) - 1]^+ du
$$

which has a unique solution $\overline{Q}(t) \equiv 1,\ t \geq 0$.

Theorem (FSLN): As $N \uparrow \infty$, $\frac{1}{N}\,Q_N(t) \to 1$ u.o.c. a.s.

Proof: Gronwall's inequality

QED M/M/N: FCLT

$$\tilde{Q}_N(t) = N + \lambda_N t + B_1(\lambda_N t)$$

$$- \int_0^t \mu \cdot [N \wedge \tilde{Q}_N(u)]du - B_2\left(\int_0^t \mu \cdot [N \wedge \tilde{Q}_N(u)]du\right)$$

$$- \int_0^t \theta \cdot [\tilde{Q}_N(u) - N]^+ du - B_3(\cdots)$$

FCLT $\quad \hat{Q}_N(t) = \frac{1}{\sqrt{N}} [\tilde{Q}_N(t) - N] = \sqrt{N} \left[\frac{1}{N} \tilde{Q}_N(t) - 1\right]$

consists of the following ingredients:

1. $\quad -\mu\beta t + \frac{1}{\sqrt{N}} B_1 \left[N\mu \left(1 - \frac{\beta}{\sqrt{N}}\right) t\right]$

2. $\quad +\mu\sqrt{N}\, t - \mu\sqrt{N} \int_0^t \left[1 \wedge \frac{1}{N} \tilde{Q}_N(u)\right] du - \frac{1}{\sqrt{N}} B_2(\cdot)$

$$= \mu \int_0^t \hat{Q}_N^-(u)du - \frac{1}{\sqrt{N}} B_2\left(N\mu \int_0^t \left[1 \wedge \frac{1}{N} \tilde{Q}_N(u)\right] du\right)$$

3. $\quad -\theta \int_0^t \hat{Q}_N^+(u)du - \frac{1}{\sqrt{N}} B_3\left(N\theta \frac{1}{\sqrt{N}} \int_0^t \tilde{Q}_N^+(u)du\right)$

Basic Properties of Brownian Motion

Let $B \stackrel{d}{=}$ Standard Brownian Motion. Then,

Self-similarity $$\frac{1}{\sqrt{N}} B(Nt) \stackrel{d}{=} B(t)$$

Additivity $$\sum_i B_i(C_i t) \stackrel{d}{=} B\left(\sum_i C_i t\right) \stackrel{d}{=} \sqrt{\sum_i C_i} \cdot B(t),$$

if B_i independent SBM

Time-Change $$\frac{1}{\sqrt{N}} B(N \cdot \tau_N(t)) \stackrel{d}{\to} B(\tau(t)),$$

if $\tau_N \stackrel{d}{\to} \tau$ and, say, τ is continuous deterministic.

Recall

1. $$-\mu\beta t + \frac{1}{\sqrt{N}}\, B_1\left[N\mu\left(1-\frac{\beta}{\sqrt{N}}\right)t\right]$$

2. $$+\mu\sqrt{N}\, t - \mu\sqrt{N}\int_0^t \left[1 \wedge \frac{1}{N}\,\tilde{Q}_N(u)\right] du - \frac{1}{\sqrt{N}}\, B_2(\cdot)$$

$$= \mu\int_0^t \hat{Q}_N^-(u)du - \frac{1}{\sqrt{N}}\, B_2\left(N\mu\int_0^t \left[1 \wedge \frac{1}{N}\,\tilde{Q}_N(u)\right] du\right)$$

3. $$-\theta\int_0^t \hat{Q}_N^+(u)du - \frac{1}{\sqrt{N}}\, B_3\left(N\theta\,\frac{1}{\sqrt{N}}\int_0^t \tilde{Q}_N^+(u)du\right)$$

If indeed $\hat{Q}_N \xrightarrow{d} \hat{Q}$, then

1. $$\xrightarrow{d} -\mu\beta t + B_1(\mu t)$$

2. $$\xrightarrow{d} \mu\int_0^t \hat{Q}^-(u)du - B_2(\mu t)$$

3. $$\xrightarrow{d} -\theta\int_0^t \hat{Q}^+(u)du - B_3(0)$$

Theorem (FCLT) As $N \uparrow \infty$, $\hat{Q}_N \xrightarrow{d} \hat{Q}$, where

$$d\hat{Q}(t) = [-\mu\beta + \mu\hat{Q}^-(t) - \theta\hat{Q}^+(t)]dt + \sqrt{2\mu}\, dB(t)$$

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).