Service Engineering

Class 12

QED (QD, ED) Queues: Introduction

• Introduction to WFM and Staffing.
• Three Operational Regimes: ED, QD, QED.
• Some History of Square-Root Staffing:
  – Erlang (Erlang-B/C) - 1913/20’s/40’s;
  – Jagerman (Erlang-B) - 1970’s;
  – Halfin-Whitt (Erlang-C) - 1981;
  – Garnett (Erlang-A) - Technion M.Sc. 2001;
  – Gurvich (V-Model; SBR) - Technion M.Sc., 2004;
  – Zeltyn (M/G/n + G) - Technion Ph.D., 2005;
• Some (Asymptotic) Theory.
• Asymptotic Framework/Analysis (Borst et al; Zeltyn 2006-7):
  – Optimization, Constraint Satisfaction;
  – Square-Root Staffing: Economics / Strategy (Pooling);
  – Scenarios.
• Uncertainty: Models (Robustness); Parameters (Forecasting).
Service Engineering: A Subjective View

Goal (Subjective):
Develop scientifically-based design principles (rules-of-thumb) and tools (software) that support the balance of service quality, process efficiency and business profitability, from the (often conflicting) views of customers, servers and managers.

Contrast/Complement the traditional and prevalent

- Service Management (U.S. Business Schools)
- Industrial Engineering (European/Japanese Engineering Schools)

Examples:

- **Staffing** - How many agents required for balancing service-quality with operational efficiency (or, for maximizing profit).

- **Skills-Based Routing (SBR)** - Platinum and Gold and Silver customers, all seeking Information or Purchase or Technical Support, via Telephone or IVR or e.mail of Chat.

- Service Process **Design** + Staffing + SBR.

**Recipe for Progress** in Research, Teaching, Applications: Simple Models at the Service of Complex Realities, with a pinch of a Multidisciplinary View (Operations, HRM, Marketing, MIS) = Service Engineering.
Workforce Management (WFM): Hierarchical Operational View

**Forecasting**  Customers: Statistics, Time-Series
Agents : HRM (Hire, Train; Incentives, Careers)

**Staffing**: Queueing Theory

- Service Level, Costs
  - # FTE’s (Seats) per unit of time

**Shifts**: IP, Combinatorial Optimization; LP

- Union constraints, Costs
  - Shift structure

**Rostering**: Heuristics, AI (Complex)

- Individual constraints
  - Agents Assignments

**Skills-based Routing**: Stochastic Control
The Quality/Efficiency Tradeoff

• Quality and Efficiency are interwind (eg. Healthcare);
• **Personnel Costs:** 65-80% of expenditure (in call centers, and many other services);
• More than 90% of U.S. consumers form a company’s image via their call center experience;

**Objective:** Having, **when** needed, the right **number** of appropriately **skilled** agents/nurses/.../**servers**.

This is a difficult problem, spanning:
**Design, Planning, Forecasting, Staffing, Shifts, Rostering, Control.**

In Lecture: Staffing (later also some Control).
In Recitation: Shifts (Forecasting).
In Homework: almost All.
The Staffing Problem

Central in Services: Call Centers, Healthcare (Nurse, Doctors), ...

Here: Determining Number of Servers (=FTE’s):
Load-Dependent, or (predictable variability) Time-Dependent.

Two Approaches:

1. **Constraint-Satisfaction**: Find the minimal number of agents $n^*$ that satisfies pre-determined performance goal(s) / constraints.

   A specific constraint-satisfaction problem can be solved via 4Call-Centers (goal-seeking). But this solution lacks insight, eg. supporting Rules of thumb:
   “How many servers needed if arrival rate doubles? services pooled?”
   “How sensitive is performance to 25% (50%) error in parameter-estimates?”

2. **Performance-Optimization**: For example,

   **Cost-Minimization**: Find $n^*$ that minimizes
   
   $$C_s \cdot n + (C_a \cdot P_n\{\text{Ab}\} + C_w \cdot E_n[W_q]) \cdot \lambda,$$
   
   where $C_s$, $C_a$ and $C_w$ are the costs of staffing, abandonment and waiting.

   Similarly, which is becoming more and more prevalent,

   **Profit-Maximization**: Find $n^*$ that maximizes
   
   $$r \cdot \lambda \cdot [1 - P_n\{\text{Ab}\}] - [C_s \cdot n + C_w \cdot E_n[W_q]) \cdot \lambda],$$
   
   where $r$ is the revenue from a service.
Our “Solution” to the Staffing Problem

- “Simple Models at the Service of Complex Realities”: Erlang-B, Erlang-C, Erlang-A; then Predictable Variability; SBR; Closed- and Semi-Open Models;

- Many-Servers Approximations (Conceptual Solution): The ED, QD, QED Operational Regimes;

- Determining the Regimes: via Strategy or Operational Constraints;

- Determining Staffing-Levels: via Constraint-Satisfaction or Performance-Optimization;

- Rules-of-Thumb: The same for Constraint-Satisfaction and Performance-Optimization;

- Robustness (mostly) of the QED-Regime: The Square-Root Staffing Rule;

For example, consider the “Basic Service Station $M_t/G/n_t + G$”:
Operational Regimes: Rules-of-Thumb
(The Basic Service Station $M_t/G/n_t +G$)

\[ R_t = \mathbb{E} \int_{t-S_e}^{t} \lambda(u) du = \mathbb{E} \lambda(t - S_e) \cdot ES = \text{Offered-Load} \]

at time $t$, namely “minutes” of work (= service) within the system at time $t$. (Steady-State: $R = \lambda \times \mathbb{E}[S]$ Erlangs, namely “minutes” of work that arrive per “minute”.)

- **Efficiency-Driven (ED) Regime:**
  \[ n_t \approx R_t - \gamma R_t, \quad 0 < \gamma < 1. \]
  Under-staffing with respect to the offered-load.

- **Quality-Driven (QD) Regime:**
  \[ n_t \approx R_t + \delta R_t, \quad \delta > 0. \]
  Over-staffing with respect to the offered-load.

- **Quality- and Efficiency-Driven (QED) Regime:**
  \[ n_t \approx R_t + \beta \sqrt{R_t}, \quad -\infty < \beta < \infty. \]
  Rationalized staffing, or the **Square-Root** Rule:
  - Often all that is needed.
  - Introduced by **Erlang**, already in 1913!
  - Characterized by **Halfin-Whitt**, only in 1981 (Erlang-C);
  - Leads to **Stable Performance**!
Operational Regimes:
Rules-of-Thumb for Performance

If the **Offered-Load** \( R \) is not small (several 10’s or more for QED, more than 100 for ED and QD), then a **relatively time-stable** performance can be expected as follows:

**ED regime:**

\[
 n \approx R_t - \gamma R_t, \quad 0.1 \leq \gamma \leq 0.25.
\]

- Essentially **all** customers delayed prior to service;
- \%Abandoned \( \approx \gamma \) (10-25%);
- Average Wait \( \approx 30 \text{ seconds - 2 minutes} \).

**QD regime:**

\[
 n \approx R_t + \delta R_t, \quad 0.1 \leq \delta \leq 0.25.
\]

Essentially **no** delays.

**QED regime:**

\[
 n \approx R_t + \beta \sqrt{R_t}, \quad -1 \leq \beta \leq 1.
\]

- \%Delayed **constant** over time, with values **25% - 75%**;
- \%Abandoned is 1-5%;
- Average wait is one-order less than average service-time (eg. seconds vs. minutes).
Motivation: QED Erlang-A, or “The Right Answer for the Wrong Reason”

Recall: $R = \frac{\lambda}{\mu}$ is the **offered-load** (measured in Erlangs): “minutes” of work that arrive per “minute”.

**“Naive”** (Deterministic, Stochastic-ignorant) approach:
Staffing at the working-load level: $n = R$.

**Erlang-C**: tele-queue “explodes” ($n > R$ necessary for stability).

But customers do not “think” Erlang-C:
if waiting is excessive they simply **abandon**:

**Erlang-A**: $E[S] = 3$ min, $E[\tau] = 3$ min

<table>
<thead>
<tr>
<th>$\lambda$/hr</th>
<th>$n$</th>
<th>Occupancy</th>
<th>$P{W_q &gt; 0}$</th>
<th>$E[W_q]$</th>
<th>$P{Ab}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>1</td>
<td>63.2%</td>
<td>63.2%</td>
<td>1:06.2</td>
<td>36.8%</td>
</tr>
<tr>
<td>100</td>
<td>5</td>
<td>82.5%</td>
<td>56.0%</td>
<td>0:31.6</td>
<td>17.5%</td>
</tr>
<tr>
<td>500</td>
<td>25</td>
<td>92.0%</td>
<td>52.7%</td>
<td>0:14.3</td>
<td>8.0%</td>
</tr>
<tr>
<td>2,500</td>
<td>125</td>
<td>96.4%</td>
<td>51.2%</td>
<td>0:06.4</td>
<td>3.6%</td>
</tr>
<tr>
<td>9,000</td>
<td>450</td>
<td>98.1%</td>
<td>50.6%</td>
<td>0:03.4</td>
<td>1.9%</td>
</tr>
<tr>
<td>↓</td>
<td>↓</td>
<td>↓</td>
<td>↓</td>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>$\infty$</td>
<td>$\infty$</td>
<td>1 ?</td>
<td>50% ?</td>
<td>0 ?</td>
<td>0 ?</td>
</tr>
</tbody>
</table>
**Motivation: QD Operation, or “What can be Achieved? At what Cost?”**

**U.S. Tele-Retail Company. ACD Report.**

<table>
<thead>
<tr>
<th></th>
<th>Avg Speed</th>
<th>Avg ACD</th>
<th>ACD Calls</th>
<th>Avg ACW</th>
<th>Aban %</th>
<th>ACD</th>
<th>% ACD</th>
<th>Avg Calls</th>
<th>% Serv</th>
<th>% Aux</th>
<th>% ACDW</th>
<th>% ACD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>W Time</td>
<td>A Time</td>
<td>Total</td>
<td>W Time</td>
<td>A Time</td>
<td>Total</td>
<td>W Time</td>
<td>A Time</td>
<td>Total</td>
<td>W Time</td>
<td>A Time</td>
<td>Total</td>
</tr>
<tr>
<td>12:00 AM*</td>
<td>00:00</td>
<td>00:00</td>
<td>00:00</td>
<td>00:00</td>
<td>00:00</td>
<td>00:00</td>
<td>00:00</td>
<td>00:00</td>
<td>00:00</td>
<td>00:00</td>
<td>00:00</td>
<td>00:00</td>
</tr>
<tr>
<td>12:30 AM*</td>
<td>00:00</td>
<td>00:00</td>
<td>00:00</td>
<td>00:00</td>
<td>00:00</td>
<td>00:00</td>
<td>00:00</td>
<td>00:00</td>
<td>00:00</td>
<td>00:00</td>
<td>00:00</td>
<td>00:00</td>
</tr>
<tr>
<td>1:00 AM*</td>
<td>00:00</td>
<td>00:00</td>
<td>00:00</td>
<td>00:00</td>
<td>00:00</td>
<td>00:00</td>
<td>00:00</td>
<td>00:00</td>
<td>00:00</td>
<td>00:00</td>
<td>00:00</td>
<td>00:00</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

*Note: *W = Waiting, A = Answering, Aban = Abandoned, ACW = Average Call Waiting, % = Percentage.
Motivation: QD Performance Analysis

Observed:
10:00-10:30 am, with 94 agents;
416 calls; 2 seconds ASA.

**Service time:** \[ E[S] = \text{ACD Time} + \text{ACW Time}, \]
\[ = 3:49 + 0:26 = 4:15. \]

**Offered load:** \[ R = \lambda \times E[S], \]
\[ = 416 \times (4:15 / 30 \text{ min}), \]
\[ = 1768 \text{ min} / 30 \text{ min} = 59 \text{ Erlangs}. \]

**Occupancy:** \[ \rho = R/n, \]
\[ = 59/94 = 63\%. \]

Compare with the column “% ACD Time” of the ACD report.

**QD Rule-of-Thumb:** \[ n \approx R + \delta \cdot R, \delta > 0, \text{ where} \]
\[ \gamma = \text{Service-Grade} \text{ parameter (or Quality-of-Service (QOS))}. \]

- In the **QD regime** abandonments are rare, in which case there is **hardly any distinction between Erlang-C and Erlang-A**. But this is definitely **not** the case in the QED- and ED-regime, hence our subsequent discussions will be Erlang-specific.
Motivation: ED Erlang-C, or “One-to-One Staffing in City-Bank”

“First National City Bank Operating Group”

“By tradition, the method of meeting increased work load in banking is to increase staff. If an operation could be done at a rate of 80 transactions per day, and daily load increased by 80, then the manager in charge of that operation would hire another person; it was taken for granted…” (Harvard Case)

1:1 Staffing - Classical IE (Erlang-C)

8 transactions per hour \[\Rightarrow\] \(E(S) = 7:30\) minutes (=M)

<table>
<thead>
<tr>
<th>(\lambda/\text{hr})</th>
<th>N Agents</th>
<th>(\rho = \text{OCC})</th>
<th>(L_q = \text{Que})</th>
<th>(W_q = \text{ASA})</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>2</td>
<td>50%</td>
<td>0.3</td>
<td>2:30</td>
</tr>
<tr>
<td>16</td>
<td>3</td>
<td>67%</td>
<td>0.9</td>
<td>3:20</td>
</tr>
<tr>
<td>24</td>
<td>4</td>
<td>75%</td>
<td>1.5</td>
<td>3:49</td>
</tr>
<tr>
<td>32</td>
<td>5</td>
<td>80%</td>
<td>2.2</td>
<td>4:09</td>
</tr>
<tr>
<td>$\lambda$/hr</td>
<td>N</td>
<td>$\rho$ = OCC</td>
<td>$L_q$ = Que</td>
<td>$W_q$ = ASA</td>
</tr>
<tr>
<td>-----------</td>
<td>-----</td>
<td>-------------</td>
<td>-----------</td>
<td>----------</td>
</tr>
<tr>
<td>72</td>
<td>10</td>
<td>90%</td>
<td>60</td>
<td>5:01</td>
</tr>
<tr>
<td>120</td>
<td>16</td>
<td>93.8%</td>
<td>11</td>
<td>5:29</td>
</tr>
<tr>
<td>400</td>
<td>51</td>
<td>98%</td>
<td>42</td>
<td>6:18</td>
</tr>
<tr>
<td>640</td>
<td>81</td>
<td>98.8%</td>
<td>70</td>
<td>6:32</td>
</tr>
<tr>
<td>1,280</td>
<td>161</td>
<td>99.4%</td>
<td>145</td>
<td>6:48</td>
</tr>
<tr>
<td>2,560</td>
<td>321</td>
<td>99.7%</td>
<td>299</td>
<td>7:00</td>
</tr>
<tr>
<td>3,600</td>
<td>451</td>
<td>99.8%</td>
<td>423</td>
<td>7:04</td>
</tr>
</tbody>
</table>

$\Rightarrow$ **Efficiency-Driven Operation** (Heavy-Traffic)

Intuition: at 100% utilization, N servers = 1 fast server

Indeed  \[ \overline{W}_q \approx \overline{W}_q \mid W_q > 0 = \frac{1}{N} \cdot \frac{\rho_N}{1 - \rho_N} \cdot E(S) \rightarrow E(S) = 7:30 ! \]

since  \[ \rho_N = \frac{\lambda_N \times E(S)}{N} = \frac{8(N-1) \times 7.5 / 60}{N} = \frac{N-1}{N} = 1 - \frac{1}{N} \]

\[ N(1 - \rho_N) = 1 \quad , \quad \rho_N \rightarrow 1 . \]
Motivation: Operational Regimes

Health insurance company. ACD Report.

<table>
<thead>
<tr>
<th>Time</th>
<th>Calls</th>
<th>Answered</th>
<th>Abandoned%</th>
<th>ASA</th>
<th>AHT</th>
<th>Occ%</th>
<th># of agents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>20,577</td>
<td>19,860</td>
<td>3.5%</td>
<td>30</td>
<td>307</td>
<td>95.1%</td>
<td></td>
</tr>
<tr>
<td>8:00</td>
<td>332</td>
<td>308</td>
<td>7.2%</td>
<td>27</td>
<td>302</td>
<td>87.1%</td>
<td>59.3</td>
</tr>
<tr>
<td>8:30</td>
<td>653</td>
<td>615</td>
<td>5.8%</td>
<td>58</td>
<td>293</td>
<td>96.1%</td>
<td>104.1</td>
</tr>
<tr>
<td>9:00</td>
<td>866</td>
<td>796</td>
<td>8.1%</td>
<td>63</td>
<td>308</td>
<td>97.1%</td>
<td>140.4</td>
</tr>
<tr>
<td>9:30</td>
<td>1,152</td>
<td>1,138</td>
<td>1.2%</td>
<td>28</td>
<td>303</td>
<td>90.8%</td>
<td>211.1</td>
</tr>
<tr>
<td>10:00</td>
<td>1,330</td>
<td>1,286</td>
<td>3.3%</td>
<td>22</td>
<td>307</td>
<td>98.4%</td>
<td>223.1</td>
</tr>
<tr>
<td>10:30</td>
<td>1,364</td>
<td>1,338</td>
<td>1.9%</td>
<td>33</td>
<td>296</td>
<td>99.0%</td>
<td>222.5</td>
</tr>
<tr>
<td>11:00</td>
<td>1,380</td>
<td>1,280</td>
<td>7.2%</td>
<td>34</td>
<td>306</td>
<td>98.2%</td>
<td>222.0</td>
</tr>
<tr>
<td>11:30</td>
<td>1,272</td>
<td>1,247</td>
<td>2.0%</td>
<td>44</td>
<td>298</td>
<td>94.6%</td>
<td>218.0</td>
</tr>
<tr>
<td>12:00</td>
<td>1,179</td>
<td>1,177</td>
<td>0.2%</td>
<td>1</td>
<td>306</td>
<td>91.6%</td>
<td>218.3</td>
</tr>
<tr>
<td>12:30</td>
<td>1,174</td>
<td>1,160</td>
<td>1.2%</td>
<td>10</td>
<td>302</td>
<td>95.5%</td>
<td>203.8</td>
</tr>
<tr>
<td>13:00</td>
<td>1,018</td>
<td>999</td>
<td>1.9%</td>
<td>9</td>
<td>314</td>
<td>95.4%</td>
<td>182.9</td>
</tr>
<tr>
<td><strong>13:30</strong></td>
<td><strong>1,061</strong></td>
<td><strong>961</strong></td>
<td><strong>9.4%</strong></td>
<td><strong>67</strong></td>
<td><strong>306</strong></td>
<td><strong>100.0%</strong></td>
<td><strong>163.4</strong></td>
</tr>
<tr>
<td>14:00</td>
<td>1,173</td>
<td>1,082</td>
<td>7.8%</td>
<td>78</td>
<td>313</td>
<td>99.5%</td>
<td>188.9</td>
</tr>
<tr>
<td><strong>14:30</strong></td>
<td><strong>1,212</strong></td>
<td><strong>1,179</strong></td>
<td><strong>2.7%</strong></td>
<td><strong>23</strong></td>
<td><strong>304</strong></td>
<td><strong>96.6%</strong></td>
<td><strong>206.1</strong></td>
</tr>
<tr>
<td>15:00</td>
<td>1,137</td>
<td>1,122</td>
<td>1.3%</td>
<td>15</td>
<td>320</td>
<td>96.9%</td>
<td>205.8</td>
</tr>
<tr>
<td>15:30</td>
<td>1,169</td>
<td>1,137</td>
<td>2.7%</td>
<td>17</td>
<td>311</td>
<td>97.1%</td>
<td>202.2</td>
</tr>
<tr>
<td>16:00</td>
<td>1,107</td>
<td>1,059</td>
<td>4.3%</td>
<td>46</td>
<td>315</td>
<td>99.2%</td>
<td>187.1</td>
</tr>
<tr>
<td>16:30</td>
<td>914</td>
<td>892</td>
<td>2.4%</td>
<td>22</td>
<td>307</td>
<td>95.2%</td>
<td>160.0</td>
</tr>
<tr>
<td><strong>17:00</strong></td>
<td><strong>615</strong></td>
<td><strong>615</strong></td>
<td><strong>0.0%</strong></td>
<td><strong>2</strong></td>
<td><strong>328</strong></td>
<td><strong>83.0%</strong></td>
<td><strong>135.0</strong></td>
</tr>
<tr>
<td>17:30</td>
<td>420</td>
<td>420</td>
<td>0.0%</td>
<td>0</td>
<td>328</td>
<td>73.8%</td>
<td>103.5</td>
</tr>
<tr>
<td>18:00</td>
<td>49</td>
<td>49</td>
<td>0.0%</td>
<td>14</td>
<td>180</td>
<td>84.2%</td>
<td>5.8</td>
</tr>
</tbody>
</table>
Quality-Driven (QD) Erlang-A

<table>
<thead>
<tr>
<th>Time</th>
<th>Calls</th>
<th>Answered</th>
<th>Abandoned%</th>
<th>ASA</th>
<th>AHT</th>
<th>Occ%</th>
<th># of agents</th>
</tr>
</thead>
<tbody>
<tr>
<td>17:00</td>
<td>615</td>
<td>615</td>
<td>0.0%</td>
<td>2</td>
<td>328</td>
<td>83.0%</td>
<td>135.0</td>
</tr>
</tbody>
</table>

- Occupancy far below 100% (for a many-server system);
- Negligible \( P\{\text{Ab}\}\);
- Very short ASA;
- \( P\{W_q > 0\} \approx 0 \).

**Offered Load:**

\[
R = \frac{\lambda}{\mu} = \frac{615}{1,800} \times 328 = 112.07 \text{ Erlangs.}
\]

**Characterization:**

\[
n = R \cdot (1 + \delta), \quad \delta > 0.
\]

**QOS parameter:**

\[
\delta = \frac{n}{R} - 1 = \frac{135}{112.07} - 1 = 0.205.
\]

**Note:** With offered-load \( R \) higher than 100 Erlangs, staffing of 20% over \( R \) (\( \delta = 0.2 \)) already suffices for QD service.
Efficiency-Driven (ED) Erlang-A

<table>
<thead>
<tr>
<th>Time</th>
<th>Calls</th>
<th>Answered</th>
<th>Abandoned%</th>
<th>ASA</th>
<th>AHT</th>
<th>Occ%</th>
<th># of agents</th>
</tr>
</thead>
<tbody>
<tr>
<td>13:30</td>
<td>1,061</td>
<td>961</td>
<td>9.4%</td>
<td>67</td>
<td>306</td>
<td>100.0%</td>
<td>163.4</td>
</tr>
</tbody>
</table>

- 100% occupancy;
- High $P\{\text{Ab}\}$;
- Considerable ASA;
- $P\{W_q > 0\} \approx 1$.

**Offered Load:**

$$R \triangleq \frac{\lambda}{\mu} = \frac{1,061}{1,800} \times 306 = 180.37 \text{ Erlangs. (Rates: per 30 min.)}$$

**Characterization:**

$$n = R \cdot (1 - \gamma), \quad \gamma > 0.$$  

**Service-Grade (or Quality-of-Service (QOS)) parameter:**

$$\gamma = 1 - \frac{n}{R} = 1 - \frac{163.4}{180.37} = 0.094 \approx P\{\text{Ab}\}.$$  

**Proof** via flow conservation (fluid-view):

$$\lambda \cdot (1 - P\{\text{Ab}\}) = n \cdot \mu, \quad \text{hence } P\{\text{Ab}\} = 1 - \frac{n}{R} = \gamma.$$
<table>
<thead>
<tr>
<th>Time</th>
<th>Calls</th>
<th>Answered</th>
<th>Abandoned%</th>
<th>ASA</th>
<th>AHT</th>
<th>Occ%</th>
<th># of agents</th>
</tr>
</thead>
<tbody>
<tr>
<td>14:30</td>
<td>1,212</td>
<td>1,179</td>
<td>2.7%</td>
<td>23</td>
<td>304</td>
<td>96.6%</td>
<td>206.1</td>
</tr>
</tbody>
</table>

- High occupancy, yet not 100%;
- Small $P\{\text{Ab}\}$ and ASA, yet not negligible;
- $P\{W_q > 0\} \approx \alpha$, $0 < \alpha < 1$.

**Offered Load:**

$$R = \frac{\lambda}{\mu} = \frac{1212}{1800} \times 304 = 204.69 \text{ Erlangs};$$

(very close to $n = 206.1$; recall stochastic-ignorant staffing).

**Characterization:**

$$n = R + \beta \sqrt{R}, \quad -\infty < \beta < \infty.$$  

**QOS parameter:**

$$\beta = \frac{n - R}{\sqrt{R}} = \frac{206.1 - 204.69}{\sqrt{204.69}} = 0.10.$$  

**Square-Root Staffing Rule:**

- Described by Erlang already in 1924 (used in 1913);
- Folklore till Halfin & Whitt, 1981 (Erlang-C);
The QED Regime in Practice

Two call centers: U.S. (Health-Insurance) and Italian (Tele-Banking). Calculate hourly $\beta = \frac{n-R}{\sqrt{R}}$, then compare against performance.

**QOS $\beta$ vs. Abandonment**

**U.S. data**

**Italian data**

**QOS $\beta$ vs. Average Wait**

**U.S. data**

**Italian data**
Yet to Come:

- Jagerman (Erlang-B) - 1970’s;
- The Halfin-Whitt (Erlang-C) Theorem - 1981;
- Intuition via Excursions (Busy- and Idle-Periods);
- QD Erlang-C;
- Pooling Scenarios;
- Motivating Erlang-A via \( M/M/\infty \);
- Garnett’s Theorem (Erlang-A) - Technion M.Sc. 2001;
- Zeltyn’s Theorem \( (M/M/n + G) \) - Technion Ph.D., 2005;
- Cost Minimization (Erlang-C, Erlang-A);
- Constraint Satisfaction (Erlang-A): the 80-20 rule;
- Feldman’s Algorithm (Predictable Queues) - Technion M.Sc., 2006-7.
- Gurvich (V-Model; SBR) - Technion M.Sc., 2004; Columbia Ph.D., 2007.