Supporting Material (Downloadable)
Call Centers: Research Bibliography with Abstracts.
Version 7, December 2006.

Supporting Material (Downloadable)
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)

By tradition, the method of meeting increased work load in

"First National City Bank Operating Group"
Intuition: at 100% utilization, N servers = 1 fast server

Indeed \( \bar{W} \approx \bar{W}_0 \), \( \bar{W}_0 > 0 = \frac{1}{N} \rho N \cdot E(S) = 8(N - 1) \cdot 7.5 \cdot 60 = \frac{N - 1}{N} 7.5 = 7.30 \)

\[ N = \frac{1}{1 - \rho N} \]

\( P = O C C \)

\( L = \text{Queue} \)

\( W = \text{ASA} \)

\( 730 \)

\( 100\% \)

\( 2,560 \)

\( 400 \)

\( 400 \)

\( 2,560 \)

\( 2,560 \)

\( 400 \)

\( 400 \)

\( 2,560 \)

\( 2,560 \)

\( 400 \)

\( 400 \)

\( 2,560 \)

\( 2,560 \)

\( 400 \)

\( 400 \)

\( 2,560 \)

\( 2,560 \)

\( 400 \)

\( 400 \)

\( 2,560 \)

\( 2,560 \)

\( 400 \)

\( 400 \)

\( 2,560 \)

\( 2,560 \)

\( 400 \)

\( 400 \)

\( 2,560 \)

\( 2,560 \)

\( 400 \)

\( 400 \)

\( 2,560 \)

\( 2,560 \)

\( 400 \)

\( 400 \)

\( 2,560 \)

\( 2,560 \)

\( 400 \)

\( 400 \)

\( 2,560 \)

\( 2,560 \)

\( 400 \)

\( 400 \)

\( 2,560 \)

\( 2,560 \)

\( 400 \)

\( 400 \)

\( 2,560 \)

\( 2,560 \)

\( 400 \)

\( 400 \)

\( 2,560 \)

\( 2,560 \)

\( 400 \)

\( 400 \)
Rough Performance Analysis

2 seconds ASA (Average Speed of Answer)
3:45 minutes average service time
400 calls
Peek 10:00 – 10:30 a.m. with 100 agents

\[ \text{Erlang-C} = \frac{MN}{M+N} \]
Rough Performance Analysis

Peak 10:00 – 10:30 a.m., with 100 agents
400 calls   3:45 minutes average service time
2 seconds ASA

Offered load
\[ R = \frac{E(S)}{g_{117}} \]
\[ E(S) = 400 \text{ calls} \]
\[ 3:45 = 1500 \text{ min./30 min.} \]
\[ = 50 \text{ Erlangs} \]

Occupancy
\[ \rho = \frac{R}{N} \]
\[ \rho = \frac{50}{100} = 50\% \]

Quality-Driven Operation (Light-Traffic)
Classical Queueing Theory (M/G/N approximations)

Above:
\[ R = 50, \quad N = R + 50, \]
all served immediately.

Rule of Thumb:
\[ N = \frac{R}{\rho} \]

<table>
<thead>
<tr>
<th>Offered Load</th>
<th>50/100 = 50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Occupancy</td>
<td>( \rho = \frac{R}{N} )</td>
</tr>
<tr>
<td>Erlang-C N=100</td>
<td>( E(S) = 3:45 \text{ min.} )</td>
</tr>
<tr>
<td>---------------</td>
<td>---------------------------------</td>
</tr>
<tr>
<td>1%</td>
<td>115</td>
</tr>
<tr>
<td>15%</td>
<td>155</td>
</tr>
<tr>
<td>35%</td>
<td>1580</td>
</tr>
<tr>
<td>88%</td>
<td>1585</td>
</tr>
<tr>
<td>100%</td>
<td>1400</td>
</tr>
</tbody>
</table>

Can increase offered load - by how much?

Quality-driven: 100 agents, 50% utilization
## Quality-driven

- 100 agents, 50% utilization

Can increase offered load - by how much?

### Efficiency-driven Operation (Heavy Traffic)

<table>
<thead>
<tr>
<th>N</th>
<th>% Wait</th>
<th>3:34 min.</th>
<th>99.1%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1585</td>
<td>12%</td>
<td>3:34 min.</td>
<td>99.1%</td>
</tr>
<tr>
<td>1580</td>
<td>15%</td>
<td>2:34 min.</td>
<td>98.8%</td>
</tr>
<tr>
<td>1550</td>
<td>35%</td>
<td>0:48 min.</td>
<td>96.9%</td>
</tr>
<tr>
<td>1540</td>
<td>88%</td>
<td>0:02 min.</td>
<td>87.5%</td>
</tr>
<tr>
<td>1500</td>
<td>100%</td>
<td>0</td>
<td>50%</td>
</tr>
</tbody>
</table>

**Rule of Thumb:**

\[
N = \left\lfloor \frac{\lambda + R}{1 - \rho} \right\rfloor = \frac{N_d - 1}{N_d} \cdot \frac{N}{I} = 0 < \frac{b}{M} | \frac{b}{M} \approx \frac{b}{M}
\]

**Changing N (Staffing) in Erlang-C**

- \(E(W_q) = \frac{\lambda}{\mu \cdot (S) E}\)
- \(R = 99\%\)
- \(N + 1\) delayed

Above:

<table>
<thead>
<tr>
<th>N</th>
<th>% Wait</th>
<th>3:34 min.</th>
<th>99.1%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1585</td>
<td>12%</td>
<td>3:34 min.</td>
<td>99.1%</td>
</tr>
<tr>
<td>1580</td>
<td>15%</td>
<td>2:34 min.</td>
<td>98.8%</td>
</tr>
<tr>
<td>1550</td>
<td>35%</td>
<td>0:48 min.</td>
<td>96.9%</td>
</tr>
<tr>
<td>1540</td>
<td>88%</td>
<td>0:02 min.</td>
<td>87.5%</td>
</tr>
<tr>
<td>1500</td>
<td>100%</td>
<td>0</td>
<td>50%</td>
</tr>
</tbody>
</table>

**Can increase offered load - by how much?**

**Changing N (Staffing) in Erlang-C**

- \(E(W_q) = \frac{\lambda}{\mu \cdot (S) E}\)
- \(R = 99\%\)
- \(N + 1\) delayed

Above:

<table>
<thead>
<tr>
<th>N</th>
<th>% Wait</th>
<th>3:34 min.</th>
<th>99.1%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1585</td>
<td>12%</td>
<td>3:34 min.</td>
<td>99.1%</td>
</tr>
<tr>
<td>1580</td>
<td>15%</td>
<td>2:34 min.</td>
<td>98.8%</td>
</tr>
<tr>
<td>1550</td>
<td>35%</td>
<td>0:48 min.</td>
<td>96.9%</td>
</tr>
<tr>
<td>1540</td>
<td>88%</td>
<td>0:02 min.</td>
<td>87.5%</td>
</tr>
<tr>
<td>1500</td>
<td>100%</td>
<td>0</td>
<td>50%</td>
</tr>
<tr>
<td>Time</td>
<td>% OAS</td>
<td>% Wait</td>
<td>% Avail</td>
</tr>
<tr>
<td>-------</td>
<td>-------</td>
<td>--------</td>
<td>---------</td>
</tr>
<tr>
<td>1599</td>
<td>100%</td>
<td>100%</td>
<td>99.9%</td>
</tr>
<tr>
<td>1599</td>
<td>100</td>
<td>100</td>
<td>99.9%</td>
</tr>
<tr>
<td>3:06</td>
<td>99.9%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>3:06</td>
<td>99.9%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>3:34</td>
<td>99.1%</td>
<td>12%</td>
<td>99.9%</td>
</tr>
<tr>
<td>3:34</td>
<td>99.1%</td>
<td>12%</td>
<td>99.9%</td>
</tr>
</tbody>
</table>

Changing N (Staffing) in Erlang-C

\( E(S) = 3:45 \)
No one waits: $\infty = g \iff 0 = \alpha$
Efficiency-driven: $0 = g \iff 1 = \alpha$
Everyone waits: $\alpha \in (0, \infty)$ is the standard normal density/distribution.

$$\left[ \frac{g(\alpha)}{\phi(\alpha)g} + 1 \right] = \alpha$$

Here:
- Manager: $\infty > g > 0 \iff g = (N\gamma - 1)N^{\gamma}$
- Server: $N > \alpha > 0 \iff \alpha = \{0 < \mathcal{W} < 1\}^{\lambda}N^{\gamma}$
- Customer: $0 = \alpha \iff g = (N\gamma - 1)N^{\gamma}$

Then the following 3 points of view are equivalent:

Consider a sequence of M/M/N models, $N = 1, 2, 3, \ldots$

QED Theorem (Halfin-Whitt, 1981)
The Halfin-Whitt Delay Function

\[ P(E) \]

\[ \text{Occupancy} = \frac{N}{R} - 1 \approx \frac{N}{R} \]

\[ \text{Safe, Utilization} = \left\{ \begin{array}{l}
0 < \text{Safety} \leq 1
\end{array} \right\} \%
\]

\[ \text{ASA} = \left[ 0 < \text{ ASA } \right] \%
\]

\[ \text{Congestion index } = E(S) \]

\[ 0 < \gamma \left[ \left( \frac{gS}{R} \right)^{\phi} + 1 \right] = (\gamma)p \approx \text{ Delayed } \%
\]

\[ \text{Expected Performance:}
\]

\[ \text{Safety-Stalling: } \gamma \]

\[ \gamma \left[ R + R \right] = \]

\[ 0 < g \text{ ,, Service-Gra} \leq g \left[ \gamma \right] R + R = N \]

\[ \text{Offered Load (Erlangs) } = R \times E(S) \]

\[ \text{The Halfin-Whitt Delay Function.} \]
QED : Some Intuition  (Assume $\mu = 1$)

M/M/N: $W_N > 0$

...
**Rules of Thumb: Operational Regimes**

**Efficiency-driven**

\[
E = \frac{\text{E(S) units of work per unit of time (load)}}{g_{177}}
\]

**Quality-driven**

\[
N = \frac{R}{g_{170}/g_{186}/g_{77}/g_{14}}, 0 > 0
\]

**Safety-Staffing**

Determine Regimes (Strategy), Parameters (Economics)

**Strategy:**
- Managers, Agents (Unions), Customers

**Economics:**
- Minimize agent salaries + waiting cost

**Strategy:**
- Sustain Regime under Pooling

**Base:**
- \(g_{540} = 300/hr, AHT = 5\) min
- \(R = 40, N = 30\) agents

**Efficiency-driven:**
- Maintain \(N = 100, I\) \(\text{OCC} = 91\%, VSA = 7\) sec
- \(N = 100 + I = 100, \text{OCC} = 91\%, VSA = 7\) sec

**Quality-driven:**
- Maintain \(N = 107, \text{OCC} = 95\%, VSA = 6\) sec
- \(N = 120, \text{OCC} = 83.3\%, VSA = 5\) sec

**Safety-Staffing:**
- \(N = 120, \text{OCC} = 95\%, VSA = 5\) sec
- \(N = 100, \text{ASA} = .5\) sec, \(y = (120 - 100)/10 = 2\)

**Efficiency-driven:**
- Maintain \(N = 107, \text{OCC} = 95\%, VSA = 6\) sec
- \(N = 100, \text{ASA} = 15\) sec

**Quality-driven:**
- Maintain \(N = 107, \text{OCC} = 95\%, VSA = 6\) sec
- \(N = 100, \text{ASA} = .5\) sec

**Safety-Staffing:**
- \(N = 120, \text{OCC} = 95\%, VSA = 5\) sec
- \(N = 100, \text{ASA} = 15\) sec

**QED:**
- Maintain \(N = 110, \text{OCC} = 91\%, VSA = 6\) sec
- \(N = 100, \text{ASA} = .5\) sec, \(y = 0.8\)

**Economics:**
- Minimize \(R\) + \(N\) \(g_{152}\)

**QED**:
- \(N = 110, \text{OCC} = 91\%, VSA = 6\) sec

**Strategy:**
- Sustain Regime under Pooling

**Base:**
- \(g_{540} = 300/hr, AHT = 5\) min
- \(R = 40, N = 30\) agents

**Efficiency-driven:**
- Maintain \(N = 107, \text{OCC} = 95\%, VSA = 6\) sec
- \(N = 100, \text{ASA} = 15\) sec

**Quality-driven:**
- Maintain \(N = 107, \text{OCC} = 95\%, VSA = 6\) sec
- \(N = 100, \text{ASA} = .5\) sec

**Safety-Staffing:**
- \(N = 120, \text{OCC} = 95\%, VSA = 5\) sec
- \(N = 100, \text{ASA} = 15\) sec

**QED**:
- \(N = 107, \text{OCC} = 95\%, VSA = 6\) sec
- \(N = 100, \text{ASA} = .5\) sec

**Strategy:**
- Sustain Regime under Pooling

**Base:**
- \(g_{540} = 300/hr, AHT = 5\) min
- \(R = 40, N = 30\) agents
### Economics of Scale

**Base case:** M/M/N with parameters λ, μ, N

**Scenario:** λ → mλ (R → mR)

<table>
<thead>
<tr>
<th></th>
<th>Base Case</th>
<th>Efficiency-driven</th>
<th>Quality-driven</th>
<th>Rationalized</th>
</tr>
</thead>
<tbody>
<tr>
<td>Offered load</td>
<td>R = (\frac{\lambda}{\mu})</td>
<td>mR</td>
<td>mR</td>
<td>mR</td>
</tr>
<tr>
<td>Safety staffing</td>
<td>Δ</td>
<td>Δ</td>
<td>mΔ</td>
<td>√mΔ</td>
</tr>
<tr>
<td>Number of agents</td>
<td>N = R + Δ</td>
<td>mR + Δ</td>
<td>mR + mΔ</td>
<td>mR + √mΔ</td>
</tr>
<tr>
<td>Service grade</td>
<td>(\beta = \frac{\Delta}{\sqrt{m}})</td>
<td>(\frac{\beta}{\sqrt{m}})</td>
<td>(\beta)</td>
<td>(\beta)</td>
</tr>
<tr>
<td>Erlang-C = P(Wait &gt; 0)</td>
<td>P((\beta))</td>
<td>P((\frac{\beta}{\sqrt{m}})) (\rightarrow) 1</td>
<td>P((\beta\sqrt{m})) (\rightarrow) 0</td>
<td>[P((\beta))]</td>
</tr>
<tr>
<td>Occupancy</td>
<td>(\rho = \frac{R}{R + \Delta})</td>
<td>(\frac{R}{R + \Delta}) (\rightarrow) 1</td>
<td>(\frac{R}{R + \Delta})</td>
<td>(\frac{R}{R + \Delta}) (\rightarrow) 1</td>
</tr>
<tr>
<td>ASA = E(\left(\text{Wait &gt; 0}\right))</td>
<td>(\frac{1}{\Delta})</td>
<td>(\frac{1}{\Delta} = \text{ASA})</td>
<td>(\frac{1}{m\Delta} = \text{ASA})</td>
<td>(\frac{1}{\sqrt{m}\Delta} = \text{ASA})</td>
</tr>
<tr>
<td>TSP = P(\left(\text{Wait &gt; T}\right)) (\rightarrow) 0</td>
<td>e(-\tau\Delta)</td>
<td>e(-\tau\Delta = \text{TSP})</td>
<td>e(-m\tau\Delta = \text{TSP}^m)</td>
<td>e(-\sqrt{m}\tau\Delta = \text{TSP}^\sqrt{m})</td>
</tr>
</tbody>
</table>

#### Framework: Asymptotic theory of M/M/N, N \(\rightarrow\) ∞

- **Eg.**: % Delayed < \(\alpha\).
- **Optimization**: N* minimizes Total Costs
  - Quality Efficiency D(\(\alpha\))
  - Efficiency Cost C(N) (delay cost = delay time)
  - Efficiency Staffing Cost (N = # agents)

#### Satisfaktion: N* minimal s.t. Service Constraint

- Efficiency-driven QED
- Quality-driven QED
- Rationalized - QED

#### Strategy: Sustain Regime under Pooling

See: Whitt's "How multi-server queues scale with demand"
Square-Root Safety Staffing: $\frac{R}{P\frac{1}{2}} = p$
Rules-of-Thumb in an "Erlang-C World"

Efficiency-Driven: \( N = R + 2 \) (or 3, or…).
Expect that essentially all customers are delayed in queue, that average delay is about 1/2 (or 1/3, or…) average service-time, and that agents utilization is extremely high (close to 100%).

Quality-Driven: \( N = R + (10\% - 20\%) R \)
Expect essentially no delays of customers.

Economics: Minimize agent salaries + congestion cost, or avoid higher service-time (seconds vs. minutes), and keep agents utilization is extremely high (close to 100%).
Expect that essentially all customers are delayed in queue, that average delay is about 1/2 (or 1/3, or…) average service-time.

Satisficing: Least Number of Agents s.t. Constraints
Can determine regime scientifically.

\[ QED: \quad N = \sqrt{R + 0.5 R} \]

Strategy: Retain performance levels under Pooling (4C demo)
Economics: Minimize agent salaries + congestion cost, or least Number of Agents s.t. Constraints.
Note: Satisfiation easier to model but Costs easier to grasp.

Costly delays (Emergency)
Cheap servers (IVR)

Costly delays (Emergency) •
Cheap servers (IVR) •

Scenario Analysis: 80:20 Rule (Large Call Center)
Prevalent std: at least 80% customers wait less than 20 sec.
Formally: \( \text{%}(\text{Wait} > 20 \text{ sec.}) < 0.2 \)

Base Case: \( 100 \) calls per min (avg)
\( M = 4 \) min. service time (avg)
\( R = 400 \) Erlangs offered load (large)
\( y = \left( 0.10 \right)^{-1} \) \( (0.53) \) = 0.53, \( \text{by } \%\text{(Wait} > 20 \text{ sec.}) = P(y) \)
\( y^* = 0.2 \)
Hence: \( N^* = 400 + 0.53 \) \( 400 = 411 \), by safety-staffing

And \( c_d = (y^*)^{-1} (0.53) = 0.32 \), \( \text{by inverting } y^* \)

Low valuation of customers' time, at 31 of server's time, yet reasonable 80:20 performance? enabled by scale!

What if \( c_d = 5 \) ?
\( N^* = 429 \) agents (vs. 411 before)
Agents' accessibility (idelness) = 7% (vs. 3% before)
Hence, 1 out of 100 waits over 20 sec. (vs. 1 out of 5)

Scenario Analysis: "Reasonable" Service Level?

Theory: The least \( N \), \% delayed that guarantees that 99% answered immediately.
Service level constraint: 1 out of 100 delayed (avg), namely
\( \%\text{(Wait} > 0\text{ sec.}) < 0.2 \)

Example: \( R = 1800 \) Erlangs offered-load
\( M = 4 \) min. service time (avg)
\( \text{(a/vh)} \)
\( N = R + p \cdot \left( y \right)^{-1} \) \( (0.53) \) = 411 agents
99% answered immediately.
Hence: \( N^* = 411 \) agents

Valuation of customers' time as being worth 75-fold of agents' time seems reasonable only in extreme circumstances:

- Very high service index
- Service level constraint: 1 out of 100 delayed (avg), namely
- \( \%\text{(Wait} > 0\text{ sec.}) < 0.2 \)

Formally: \( \%\text{(Wait} > 20 \text{ sec.}) > 0.2 \)
Prevalent std: at least 80% customers wait less than 20 sec.

Scenario Analysis: 80:20 Rule (Large Call Center)