Telephone Call/Contact Centers

Service Engineering

Queueing Science

SNC/CRM

July 23, 2004

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

Website: http://ie.technion.ac.il/serveng2004
Contents

1. **Motivation**: "The Right Answer for the Wrong Reason"

2. **Operational Regime (M/M/N):**
   - Quality-Driven
   - Efficiency-Driven
   - The **QED** (Halflin-Whitt) Regime

3. Some Intuition

   Example from a call center, leading to models with

4. **Impatient** (Abandoning) Customers (M/M/N+G)

5. **Time-Varying** Queues with **Time-Stable** Performance

6. **General Service** Times (G/M/N, G/D/N; G/LN/N)

7. **Heterogeneous** Customers and **Multi-skilled** Agents (SBR)

8. **Forecasting** Parameters
Background


M/M/N (Erlang-C); M/G/N

Halfin and Whitt: "Heavy Traffic Limits for Queues with Many Exponential Servers." OR, 1981.


M/M/N+M (ErlangA); M/M/N+G; G/G/N+G


Zeltyn: Ph.D. Thesis on the M/M/N+G Queue (Q/E/QED Asymptotics, Dimensioning, Inference).


In this conference:
- Whitt: Fluid models; Approximations (of G/G/N+G); Uncertainty in arrival rates and staffing levels.

Time-Dependence (Predictable Variability); Service Networks


**In this conference:**
- Helber: Profit Maximization.
- Henken; Retrials (*M-Design*).
- Massey: QED Service Networks.

**SBR**


Gurvich: "Staffing and Control of the M/M/N Queue with Multi-Class Customers and Many Servers." M.Sc. Thesis (*V-Design*)


**In this conference:**
- Armony: Control and Staffing (*Reversed-V*).
- Atar, Shaikhet: QED control, with non-basic activities.
- Harrison and Zeevi: QED control under CRP with linear costs; "ED" staffing under uncertainty.
- Koole: Approximations, with overflow routing.
**Erlang-C** = $M/M/N$

**Erlang-A** <4CallCenters>
Approximating Queueing and Waiting

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

- \( \hat{Q}_N = \{ \hat{Q}_N(t), t \geq 0 \} : \text{stochastic process} \) obtained by centering and rescaling:
  \[
  \hat{Q}_N = \frac{Q_N - N}{\sqrt{N}}
  \]

- \( \hat{Q}_N(\infty) : \text{stationary distribution of } \hat{Q}_N \)

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

Approximating (Virtual) Waiting Time

\[
\hat{V}_N = \sqrt{N} \ V_N \Rightarrow \hat{V} = \left[ \frac{1}{\mu} \hat{Q} \right]^+ \quad \text{(Puhalskii, 1994)}
\]
Staffing Time-Varying Queues:

Two Common Approaches:

**SSA** – Simple Stationary Approximation.
*Constant* staffing levels, based on steady-state $M/M/N$, with $\lambda =$ long-run average number of arrivals.

**PSA** – Point-wise Stationary Approximation.
Time-varying staffing levels, based on steady-state $M/M/N$, with $\lambda = \lambda(t)$ at each time $t$.

Could result in time-varying *(highly oscillating)* performance (utilization, service), which is undesirable.
Simple Stationary Approximation (SSA, $\alpha=0.2$)
Point-wise Stationary Approximation (PSA, $\alpha=0.2$)
Example: "Real" Call Center

Two-hump arrival functions are common

(Adapted from Green L., Kolesar P., Soares J. for benchmarking.)

Assume: Service and abandonment rates are both exponential having mean 0.1 (6 min.)
QED Staffing ($\alpha=0.5$)
Congestion (Queue, Wait)

QD
$\alpha=0.1$

Negligible

QED
$\alpha=0.5$

Seconds

ED
$\alpha=0.9$

Minutes
When QED? Dimensioning

For example, via $r = \text{value of customer-time} / \text{agent-salary}$

Moderately impatient: QED if $r=2$

Highly impatient: $r=10$
<table>
<thead>
<tr>
<th>Time</th>
<th>Avg Speed</th>
<th>Avg Aban</th>
<th>ACD Calls</th>
<th>Avg ACD</th>
<th>Aban Calls</th>
<th>% ACD</th>
<th>% Aban</th>
<th>% ACD</th>
<th>% ACD</th>
<th>% ACD</th>
<th>% ACD</th>
<th>% ACD</th>
<th>% ACD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>W</td>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Totals</td>
<td>10545</td>
<td>54</td>
<td></td>
<td>46</td>
<td>63</td>
<td>98</td>
<td>70</td>
<td>149</td>
<td>8</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12:00 AM</td>
<td>00:00</td>
<td>00:00</td>
<td>28</td>
<td>00:25</td>
<td>1</td>
<td>76</td>
<td>61</td>
<td>7</td>
<td>4</td>
<td>81</td>
<td>2</td>
<td>18</td>
<td>81</td>
</tr>
<tr>
<td>12:30 AM</td>
<td>00:03</td>
<td>04:10</td>
<td>14</td>
<td>00:33</td>
<td>1</td>
<td>88</td>
<td>52</td>
<td>5</td>
<td>3</td>
<td>48</td>
<td>1</td>
<td>26</td>
<td>83</td>
</tr>
<tr>
<td>1:00 AM</td>
<td>00:00</td>
<td>00:00</td>
<td>9</td>
<td>00:11</td>
<td>1</td>
<td>91</td>
<td>90</td>
<td>1</td>
<td>7</td>
<td>90</td>
<td>0</td>
<td>26</td>
<td>85</td>
</tr>
<tr>
<td>4:30 AM</td>
<td>00:00</td>
<td>00:00</td>
<td>0</td>
<td>00:00</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6:00 AM</td>
<td>00:00</td>
<td>00:00</td>
<td>12</td>
<td>00:18</td>
<td>1</td>
<td>21</td>
<td>100</td>
<td>7</td>
<td>2</td>
<td>100</td>
<td>2</td>
<td>18</td>
<td>6</td>
</tr>
<tr>
<td>8:30 AM</td>
<td>00:00</td>
<td>00:00</td>
<td>27</td>
<td>00:20</td>
<td>2</td>
<td>32</td>
<td>100</td>
<td>14</td>
<td>2</td>
<td>100</td>
<td>5</td>
<td>3</td>
<td>28</td>
</tr>
<tr>
<td>10:00 AM</td>
<td>00:00</td>
<td>00:00</td>
<td>62</td>
<td>00:15</td>
<td>3</td>
<td>38</td>
<td>100</td>
<td>21</td>
<td>3</td>
<td>100</td>
<td>13</td>
<td>4</td>
<td>34</td>
</tr>
<tr>
<td>11:30 AM</td>
<td>00:00</td>
<td>00:00</td>
<td>93</td>
<td>00:14</td>
<td>4</td>
<td>36</td>
<td>100</td>
<td>30</td>
<td>3</td>
<td>100</td>
<td>7</td>
<td>4</td>
<td>32</td>
</tr>
<tr>
<td>1:00 PM</td>
<td>00:00</td>
<td>00:00</td>
<td>120</td>
<td>00:10</td>
<td>3</td>
<td>38</td>
<td>100</td>
<td>47</td>
<td>3</td>
<td>100</td>
<td>8</td>
<td>6</td>
<td>33</td>
</tr>
<tr>
<td>2:30 PM</td>
<td>00:00</td>
<td>00:00</td>
<td>193</td>
<td>00:14</td>
<td>4</td>
<td>44</td>
<td>100</td>
<td>81</td>
<td>3</td>
<td>100</td>
<td>10</td>
<td>7</td>
<td>37</td>
</tr>
<tr>
<td>4:30 PM</td>
<td>00:00</td>
<td>00:00</td>
<td>293</td>
<td>00:25</td>
<td>5</td>
<td>64</td>
<td>99</td>
<td>75</td>
<td>4</td>
<td>97</td>
<td>9</td>
<td>7</td>
<td>47</td>
</tr>
<tr>
<td>7:00 PM</td>
<td>00:00</td>
<td>00:00</td>
<td>381</td>
<td>00:45</td>
<td>6</td>
<td>60</td>
<td>87</td>
<td>21</td>
<td>6</td>
<td>93</td>
<td>8</td>
<td>8</td>
<td>59</td>
</tr>
<tr>
<td>9:30 AM</td>
<td>00:00</td>
<td>00:05</td>
<td>416</td>
<td>00:26</td>
<td>1</td>
<td>63</td>
<td>87</td>
<td>84</td>
<td>4</td>
<td>98</td>
<td>5</td>
<td>6</td>
<td>55</td>
</tr>
<tr>
<td>7:00 AM</td>
<td>00:35</td>
<td>03:55</td>
<td>345</td>
<td>00:33</td>
<td>0</td>
<td>92</td>
<td>99</td>
<td>95</td>
<td>4</td>
<td>99</td>
<td>6</td>
<td>8</td>
<td>44</td>
</tr>
<tr>
<td>10:00 AM</td>
<td>00:00</td>
<td>00:00</td>
<td>352</td>
<td>00:27</td>
<td>1</td>
<td>51</td>
<td>100</td>
<td>102</td>
<td>3</td>
<td>100</td>
<td>7</td>
<td>8</td>
<td>46</td>
</tr>
<tr>
<td>11:30 AM</td>
<td>00:00</td>
<td>00:00</td>
<td>348</td>
<td>00:18</td>
<td>0</td>
<td>48</td>
<td>100</td>
<td>97</td>
<td>4</td>
<td>100</td>
<td>8</td>
<td>6</td>
<td>46</td>
</tr>
<tr>
<td>12:00 PM</td>
<td>00:00</td>
<td>00:00</td>
<td>354</td>
<td>00:18</td>
<td>0</td>
<td>52</td>
<td>99</td>
<td>95</td>
<td>4</td>
<td>99</td>
<td>8</td>
<td>5</td>
<td>47</td>
</tr>
<tr>
<td>1:00 PM</td>
<td>00:00</td>
<td>00:00</td>
<td>366</td>
<td>00:21</td>
<td>0</td>
<td>62</td>
<td>99</td>
<td>97</td>
<td>3</td>
<td>99</td>
<td>9</td>
<td>8</td>
<td>46</td>
</tr>
<tr>
<td>1:30 PM</td>
<td>00:00</td>
<td>00:00</td>
<td>147</td>
<td>00:14</td>
<td>0</td>
<td>51</td>
<td>99</td>
<td>98</td>
<td>4</td>
<td>99</td>
<td>6</td>
<td>6</td>
<td>44</td>
</tr>
<tr>
<td>2:00 PM</td>
<td>00:00</td>
<td>00:00</td>
<td>138</td>
<td>00:17</td>
<td>0</td>
<td>51</td>
<td>100</td>
<td>108</td>
<td>4</td>
<td>100</td>
<td>10</td>
<td>5</td>
<td>46</td>
</tr>
<tr>
<td>2:30 PM</td>
<td>00:00</td>
<td>00:13</td>
<td>403</td>
<td>00:13</td>
<td>0</td>
<td>54</td>
<td>100</td>
<td>113</td>
<td>4</td>
<td>100</td>
<td>10</td>
<td>4</td>
<td>50</td>
</tr>
<tr>
<td>3:00 PM</td>
<td>00:00</td>
<td>00:04</td>
<td>410</td>
<td>00:16</td>
<td>1</td>
<td>57</td>
<td>88</td>
<td>110</td>
<td>4</td>
<td>98</td>
<td>5</td>
<td>5</td>
<td>51</td>
</tr>
<tr>
<td>3:30 PM</td>
<td>00:00</td>
<td>00:14</td>
<td>347</td>
<td>00:14</td>
<td>0</td>
<td>50</td>
<td>100</td>
<td>100</td>
<td>3</td>
<td>100</td>
<td>7</td>
<td>5</td>
<td>45</td>
</tr>
<tr>
<td>4:00 PM</td>
<td>00:00</td>
<td>00:12</td>
<td>382</td>
<td>00:13</td>
<td>0</td>
<td>54</td>
<td>100</td>
<td>98</td>
<td>4</td>
<td>100</td>
<td>6</td>
<td>7</td>
<td>47</td>
</tr>
<tr>
<td>4:30 PM</td>
<td>00:00</td>
<td>00:19</td>
<td>379</td>
<td>00:19</td>
<td>0</td>
<td>65</td>
<td>99</td>
<td>97</td>
<td>4</td>
<td>99</td>
<td>8</td>
<td>5</td>
<td>50</td>
</tr>
<tr>
<td>5:00 PM</td>
<td>00:00</td>
<td>00:19</td>
<td>411</td>
<td>00:19</td>
<td>0</td>
<td>53</td>
<td>100</td>
<td>109</td>
<td>4</td>
<td>100</td>
<td>8</td>
<td>5</td>
<td>46</td>
</tr>
<tr>
<td>5:30 PM</td>
<td>00:01</td>
<td>00:19</td>
<td>387</td>
<td>00:19</td>
<td>0</td>
<td>58</td>
<td>99</td>
<td>98</td>
<td>4</td>
<td>99</td>
<td>10</td>
<td>6</td>
<td>51</td>
</tr>
<tr>
<td>6:00 PM</td>
<td>00:01</td>
<td>00:21</td>
<td>371</td>
<td>00:25</td>
<td>1</td>
<td>53</td>
<td>98</td>
<td>81</td>
<td>4</td>
<td>98</td>
<td>9</td>
<td>8</td>
<td>47</td>
</tr>
<tr>
<td>6:30 PM</td>
<td>00:00</td>
<td>00:13</td>
<td>280</td>
<td>00:13</td>
<td>0</td>
<td>41</td>
<td>100</td>
<td>90</td>
<td>3</td>
<td>100</td>
<td>8</td>
<td>4</td>
<td>37</td>
</tr>
<tr>
<td>7:00 PM</td>
<td>00:00</td>
<td>00:17</td>
<td>269</td>
<td>00:17</td>
<td>0</td>
<td>42</td>
<td>100</td>
<td>78</td>
<td>3</td>
<td>100</td>
<td>9</td>
<td>5</td>
<td>38</td>
</tr>
</tbody>
</table>
Erlang-C = M/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 (Average Speed of Answer)
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 \( R = \lambda \times E(S) \)

\[ = 400 \times 3:45 = 1500 \text{ min./30 min.} \]

\[ = 50 \text{ Erlangs} \]

Occupancy \( \rho = R/N \)

\[ = 50/100 = 50\% \]
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 \( R = \lambda \times E(S) \)
\[ = 400 \times 3:45 = 1500 \text{ min.}/30 \text{ min.} \]
\[ = 50 \text{ Erlangs} \]

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

⇒ **Quality-Driven Operation**  (Light-Traffic)
⇒ Classical Queueing Theory

Above: \( R = 50, \quad N = R + 50, \quad \approx \text{all served immediately.} \)

Rule of Thumb: \( N = \lceil R + \delta R \rceil, \quad \delta > 0 \) service-grade.
**Quality-driven**: 100 agents, 50% utilization

⇒ Can increase offered load - by how much?

**Erlang-C**  
\[ N=100 \quad E(S) = 3:45 \text{ min.} \]

<table>
<thead>
<tr>
<th>λ/hr</th>
<th>ρ</th>
<th>E(W_q) = ASA</th>
<th>% Wait = 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>800</td>
<td>50%</td>
<td>0</td>
<td>100%</td>
</tr>
</tbody>
</table>
**Quality-driven**: 100 agents, 50% utilization

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

**Erlang-C**  \( N=100 \)  \( E(S) = 3:45 \text{ min.} \)

<table>
<thead>
<tr>
<th>( \lambda/\text{hr} )</th>
<th>( \rho )</th>
<th>( E(W_q) = \text{ASA} )</th>
<th>% Wait = 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>800</td>
<td>50%</td>
<td>0</td>
<td>100%</td>
</tr>
<tr>
<td>1400</td>
<td>87.5%</td>
<td>0:02 min.</td>
<td>88%</td>
</tr>
<tr>
<td>1550</td>
<td>96.9%</td>
<td>0:48 min.</td>
<td>35%</td>
</tr>
<tr>
<td>1580</td>
<td>98.8%</td>
<td>2:34 min.</td>
<td>15%</td>
</tr>
<tr>
<td>1585</td>
<td><strong>99.1%</strong></td>
<td><strong>3:34 min.</strong></td>
<td>12%</td>
</tr>
</tbody>
</table>
**Quality-driven:** 100 agents, 50% utilization

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

**Erlang-C**

\[
\text{N}=100 \quad \text{E}(S) = 3:45 \text{ min.}
\]

<table>
<thead>
<tr>
<th>(\lambda/\text{hr})</th>
<th>(\rho)</th>
<th>(\text{E}(W_q) = \text{ASA})</th>
<th>% Wait = 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>800</td>
<td>50%</td>
<td>0</td>
<td>100%</td>
</tr>
<tr>
<td>1400</td>
<td>87.5%</td>
<td>0:02 min.</td>
<td>88%</td>
</tr>
<tr>
<td>1550</td>
<td>96.9%</td>
<td>0:48 min.</td>
<td>35%</td>
</tr>
<tr>
<td>1580</td>
<td>98.8%</td>
<td>2:34 min.</td>
<td>15%</td>
</tr>
<tr>
<td>1585</td>
<td><strong>99.1%</strong></td>
<td><strong>3:34 min.</strong></td>
<td><strong>12%</strong></td>
</tr>
</tbody>
</table>

⇒ **Efficiency-driven Operation** (Heavy Traffic)

\[
\bar{W}_q \approx \bar{W}_q \mid W_q > 0 = \frac{1}{N} \cdot \frac{\rho_N}{1 - \rho_N} \cdot E(S) \rightarrow E(S) = 3:45 !
\]

\[
N(1 - \rho_N) = 1, \quad \rho_N \rightarrow 1
\]

Above: \( R = 99, \quad N = R + 1, \quad \approx \text{all delayed.} \)

Rule of Thumb: \( N = \lceil R + \gamma \rceil, \quad \gamma > 0 \) service grade.
Changing N (Staffing) in Erlang-C

\[ E(S) = 3:45 \]

<table>
<thead>
<tr>
<th>( \lambda/\text{hr} )</th>
<th>( N )</th>
<th>OCC</th>
<th>ASA</th>
<th>% Wait = 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1585</td>
<td>100</td>
<td>99.1%</td>
<td>3:34</td>
<td>12%</td>
</tr>
</tbody>
</table>
Changing N (Staffing) in Erlang-C

$$E(S) = 3:45$$

<table>
<thead>
<tr>
<th>$\lambda$/hr</th>
<th>N</th>
<th>OCC</th>
<th>ASA</th>
<th>% Wait = 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1585</td>
<td>100</td>
<td>99.1%</td>
<td>3:34</td>
<td>12%</td>
</tr>
<tr>
<td>1599</td>
<td><strong>100</strong></td>
<td>99.9%</td>
<td><strong>59:33</strong></td>
<td>0%</td>
</tr>
</tbody>
</table>
## Changing N (Staffing) in Erlang-C

E(S) = 3:45

<table>
<thead>
<tr>
<th>λ/hr</th>
<th>N</th>
<th>OCC</th>
<th>ASA</th>
<th>% Wait = 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1585</td>
<td>100</td>
<td>99.1%</td>
<td>3:34</td>
<td>12%</td>
</tr>
<tr>
<td>1599</td>
<td>100</td>
<td>99.9%</td>
<td>59:33</td>
<td>0%</td>
</tr>
<tr>
<td>1599</td>
<td>100</td>
<td>98.9%</td>
<td>3:06</td>
<td>13%</td>
</tr>
<tr>
<td>1599</td>
<td>102</td>
<td>98.0%</td>
<td>1:24</td>
<td>24%</td>
</tr>
<tr>
<td>1599</td>
<td>105</td>
<td>95.2%</td>
<td>0:23</td>
<td>50%</td>
</tr>
</tbody>
</table>
Changing N (Staffing) in Erlang-C

\[ E(S) = 3:45 \]

<table>
<thead>
<tr>
<th>( \lambda/\text{hr} )</th>
<th>( N )</th>
<th>OCC</th>
<th>ASA</th>
<th>% Wait = 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1585</td>
<td>100</td>
<td>99.1%</td>
<td>3:34</td>
<td>12%</td>
</tr>
<tr>
<td>1599</td>
<td>100</td>
<td>99.9%</td>
<td>59:33</td>
<td>0%</td>
</tr>
<tr>
<td>1599</td>
<td>100+1</td>
<td>98.9%</td>
<td>3:06</td>
<td>13%</td>
</tr>
<tr>
<td>1599</td>
<td>102</td>
<td>98.0%</td>
<td>1:24</td>
<td>24%</td>
</tr>
<tr>
<td>1599</td>
<td>105</td>
<td>95.2%</td>
<td>0:23</td>
<td>50%</td>
</tr>
</tbody>
</table>

⇒ New Rationalized Operation

Efficiently driven, in the sense that \( \text{OCC} > 95\% \);

Quality-Driven, \( 50\% \) answered immediately

QED Regime = Quality- and Efficiency-Driven Regime

Above: \( R = 100, \quad N = R + 5, \quad 50\% \) delayed.

\[ \sqrt{\text{Safety-Staffing}} \quad N = \left\lfloor R + \beta \sqrt{R} \right\rfloor, \quad \beta > 0. \]
**QED Theorem (Halfin-Whitt, 1981)**

Consider a sequence of M/M/N models, N=1,2,3,...

Then the following 3 points of view are equivalent:

- **Customer** \( \lim_{N \to \infty} P_N \{ \text{Wait} > 0 \} = \alpha, \quad 0 < \alpha < 1; \)

- **Server** \( \lim_{N \to \infty} \sqrt{N}(1 - \rho_N) = \beta, \quad 0 < \beta < \infty; \)

- **Manager** \( N \approx R + \beta \sqrt{R}, \quad R = \lambda \times \text{E}(S) \text{ large}; \)

Here \( \alpha = \left[ 1 + \frac{\beta \phi(\beta)}{\phi(\beta)} \right]^{-1} \),

where \( \phi(\cdot) / \phi(\cdot) \) is the standard normal density/distribution.

Extremes:

- **Everyone waits**: \( \alpha = 1 \iff \beta = 0 \) Efficiency-driven
- **Quality-driven** \( \alpha = 0 \iff \beta = \infty \) No one waits:
The Halfin-Whitt Delay Function
Economics: Quality vs. Efficiency

(Dimensioning: with S. Borst and M. Reiman)

Quality \( D(t) \) delay cost (\( t = \) delay time)
Efficiency \( C(N) \) staffing cost (\( N = \# \) agents)

Optimization: \( N^* \) minimizes Total Costs

- \( C \gg D \): Efficiency-driven
- \( C \ll D \): Quality-driven
- \( C \approx D \): Rationalized - QED

Satisfization: \( N^* \) minimal s.t. Service Constraint

Eg. \( \%\text{Delayed} < \alpha \).

- \( \alpha \approx 1 \): Efficiency-driven
- \( \alpha \approx 0 \): Quality-driven
- \( 0 < \alpha < 1 \): Rationalized - QED

Framework: Asymptotic theory of M/M/N, \( N \uparrow \infty \)
QED : Some Intuition  

( Assume $\mu = 1$ )

$M/M/N$:

$W_N | W_N > 0 \overset{d}{=} \exp \left( \text{mean} = \frac{1}{N} \frac{1}{1 - \rho_N} \right)$

$\sqrt{N} W_N | W_N > 0 \overset{d}{=} \exp \left( \sqrt{N} \left( 1 - \rho_N \right) \right) \Rightarrow \exp(\beta)$

But why $P(W_N > 0) \rightarrow \alpha, \quad 0 < \alpha < 1$?
M/M/N (Erlang-C) with Many Servers: $N \uparrow \infty$

$Q(0) = N$: all servers busy, no queue.

Recall $E_{2,N} = \left[1 + \frac{T_{N-1,N}}{T_{N,N-1}}\right]^{-1} = \left[1 + \frac{1 - \rho_N}{\rho_N E_{1,N-1}}\right]^{-1}$.

Here $T_{N-1,N} = \frac{1}{\lambda_N E_{1,N-1}} \sim \frac{1}{N\mu \times h(-\beta)/\sqrt{N}} \sim \frac{1}{\mu} \frac{1}{h(-\beta)\sqrt{N}}$

which applies as $\sqrt{N} (1 - \rho_N) \to \beta$, $-\infty < \beta < \infty$.

Also $T_{N,N-1} = \frac{1}{N\mu (1 - \rho_N)} \sim \frac{1}{\mu} \frac{1}{\beta\sqrt{N}}$

which applies as above, but for $0 < \beta < \infty$.

Hence, $E_{2,N} \sim \left[1 + \frac{\beta}{h(-\beta)}\right]^{-1}$, assuming $\beta > 0$. 

QED: $N \sim R + \beta \sqrt{R}$ for some $\beta$, $0 < \beta < \infty$.
Figure 12: Mean Service Time (Regular) vs. Time-of-day (95% CI) \((n = 42613)\)
Predictable Variability

Arrivals

Queues

Waiting
Service Time

Survival curve, by Types

Means (In Seconds)
NW (New) = 111
PS (Regular) = 181
NE (Stocks) = 269
IN (Internet) = 381
Hazard Rate: Empirical (Im)Patience

- **Regular Customers**
- **Priority Customers**

![Graph showing hazard rates for regular and priority customers](image)
Contents

1. Motivation: "The Right Answer for the Wrong Reason"

2. Operational Regime (M/M/N):
   - Quality-Driven
   - Efficiency-Driven
   - The QED (Halfin-Whitt) Regime

3. Some Intuition
   Example from a call center, leading to models with

4. Impatient (Abandoning) Customers (M/M/N+G)

5. Time-Varying Queues with Time-Stable Performance

6. General Service Times (G/M/N, G/D/N; G/LN/N)

7. Heterogeneous Customers and Multi-skilled Agents (SBR)

8. Forecasting Parameters
Operational Aspects of Impatience

Recall earlier Q, E and QED Scenarios \((E(S) = 3:45)\):

<table>
<thead>
<tr>
<th>(\lambda/\text{hr} )</th>
<th>(N)</th>
<th>OCC</th>
<th>ASA</th>
<th>% Wait = 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1599</td>
<td>100</td>
<td>99.9%</td>
<td>59:33</td>
<td>1%</td>
</tr>
<tr>
<td>1599</td>
<td>105</td>
<td>95.2%</td>
<td>0:23</td>
<td>51%</td>
</tr>
<tr>
<td>1600</td>
<td>100</td>
<td>100%</td>
<td>infinity</td>
<td>0%</td>
</tr>
</tbody>
</table>

BUT with Patience \(= E(S)\):

<table>
<thead>
<tr>
<th>(\lambda/\text{hr} )</th>
<th>(N)</th>
<th>OCC</th>
<th>ASA</th>
<th>% Wait = 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1600</td>
<td>100</td>
<td>96%</td>
<td>0:09</td>
<td>50%</td>
</tr>
<tr>
<td>1600</td>
<td>95</td>
<td>98.4%</td>
<td>0:23</td>
<td>6.5%</td>
</tr>
<tr>
<td>1800</td>
<td>105</td>
<td>97.7%</td>
<td>0:23</td>
<td>3.4%</td>
</tr>
</tbody>
</table>

QED with \(\text{Im} \)patient Customers:
The “fittest” survive and wait much less!

Erlang-A: Erlang-C with Exponential Patience / Abandonment

Downloadable implementation: 4CallCenters(.com)
### Operational Aspects of Impatience

Recall earlier Q, E and QED Scenarios \( \text{E(S) = 3:45} \):

<table>
<thead>
<tr>
<th>( \lambda )/hr</th>
<th>N</th>
<th>OCC</th>
<th>ASA</th>
<th>% Wait = 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1599</td>
<td>100</td>
<td>99.9%</td>
<td>59:33</td>
<td>1%</td>
</tr>
<tr>
<td>1599</td>
<td>105</td>
<td>95.2%</td>
<td>0:23</td>
<td>51%</td>
</tr>
<tr>
<td>1600</td>
<td>100</td>
<td>100%</td>
<td>infinity</td>
<td>0%</td>
</tr>
<tr>
<td>BUT with Patience=E(S)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1600</td>
<td>100</td>
<td>96%</td>
<td>0:09</td>
<td>50%</td>
</tr>
</tbody>
</table>

Erlang-A: Erlang-C with Exponential Patience / Abandonment

Downloadable implementation: [4CallCenters.com](http://4CallCenters.com)
Operational Aspects of *Impatience*

Recall earlier Q, E and QED Scenarios (E(S) = 3:45):

<table>
<thead>
<tr>
<th>λ/hr</th>
<th>N</th>
<th>OCC</th>
<th>ASA</th>
<th>% Wait = 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1599</td>
<td>100</td>
<td>99.9%</td>
<td>59:33</td>
<td>1%</td>
</tr>
<tr>
<td>1599</td>
<td>105</td>
<td>95.2%</td>
<td>0:23</td>
<td>51%</td>
</tr>
<tr>
<td>1600</td>
<td>100</td>
<td>100%</td>
<td>infinity</td>
<td>0%</td>
</tr>
</tbody>
</table>

**BUT** with Patience=E(S)

<table>
<thead>
<tr>
<th>λ/hr</th>
<th>N</th>
<th>OCC</th>
<th>ASA</th>
<th>% Abandon</th>
</tr>
</thead>
<tbody>
<tr>
<td>1600</td>
<td>100</td>
<td>96%</td>
<td>0:09</td>
<td>50%</td>
</tr>
<tr>
<td>1600</td>
<td>100</td>
<td>97.3%</td>
<td>0:23</td>
<td>2.7 %</td>
</tr>
<tr>
<td>1600</td>
<td>95</td>
<td>98.4%</td>
<td>0:23</td>
<td>6.5%</td>
</tr>
<tr>
<td>1800</td>
<td>105</td>
<td>97.7%</td>
<td>0:23</td>
<td>3.4%</td>
</tr>
</tbody>
</table>

The "fittest" survive and wait less – much less!

Erlang-A: Erlang-C with Exponential Patience / Abandonment

Downloadable implementation: 4CallCenters.com
Operational Aspects of *Impatience*

Recall earlier Q, E and QED Scenarios \((E(S) = 3:45)\):

<table>
<thead>
<tr>
<th>(\lambda/hr)</th>
<th>N</th>
<th>OCC</th>
<th>ASA</th>
<th>% Wait = 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1599</td>
<td>100</td>
<td>99.9%</td>
<td>59:33</td>
<td>1%</td>
</tr>
<tr>
<td>1599</td>
<td>105</td>
<td>95.2%</td>
<td>0:23</td>
<td>51%</td>
</tr>
<tr>
<td>1600</td>
<td>100</td>
<td>100%</td>
<td>infinity</td>
<td>0%</td>
</tr>
</tbody>
</table>

**BUT** with Patience=\(E(S)\)

| 1600          | 100| 96%  | 0:09  | 50%        |

**AND** could have \(%\text{Abandon}\)

| 1600          | 100| 97.3%| 0:23  | 2.7 \%     |
| 1600          | 95 | 98.4%| 0:23  | 6.5%       |
| 1800          | 105| 97.7%| 0:23  | 3.4%       |

**QED** with *(Im)*patient Customers:

The "fittest" survive and wait less – much less!

**Erlang-A**: Erlang-C with Exponential Patience / Abandonment

Downloadable implementation: 4CallCenters(.com)
Erlang-A (with G-Patience): M/M/N+G
**QED Theorem** (Garnett, M. and Reiman '02; Zeltyn '03)

Consider a sequence of $M/M/N+G$ models, $N=1,2,3,...$

Then the following **points of view** are equivalent:

- **QED**  
  \[
  \%\{\text{Wait} > 0\} \approx \alpha, \quad 0 < \alpha < 1; 
  \]

- **Customers**  
  \[
  \%\{\text{Abandon}\} \approx \frac{\gamma}{\sqrt{N}}, \quad 0 < \gamma; 
  \]

- **Agents**  
  \[\text{OCC} \approx 1 - \frac{\beta + \gamma}{\sqrt{N}}, \quad -\infty < \beta < \infty;\]

- **Managers**  
  \[N \approx R + \beta \sqrt{R}, \quad R = \lambda \times E(S) \text{ not small};\]

QED performance (ASA, ...) is easily computable, all in terms of $\beta$ (the square-root safety staffing level) – see later.

Covers also the Extremes:

- $\alpha = 1 : \quad N = R - \gamma R$ **Efficiency-driven**
- $\alpha = 0 : \quad N = R + \gamma R$ **Quality-driven**
QED Approximations (Zeltyn)

λ – arrival rate,

µ – service rate,

N – number of servers,

G – patience distribution,

g₀ – patience density at origin \( (g_0 = \theta, \text{ if exp}(\theta)) \).

\[ N = \frac{\lambda}{\mu} + \beta \sqrt{\frac{\lambda}{\mu}} + o(\sqrt{\lambda}), \quad -\infty < \beta < \infty. \]

\[
\begin{align*}
P\{\text{Ab}\} & \approx \frac{1}{\sqrt{N}} \cdot [h(\beta) - \bar{\beta}] \cdot \left[ \sqrt{\frac{\mu}{g_0}} + \frac{h(\beta)}{h(-\beta)} \right]^{-1}, \\
P\left\{ W > \frac{T}{\sqrt{N}} \right\} & \approx \left[ 1 + \sqrt{\frac{g_0}{\mu}} \cdot \frac{h(\beta)}{h(-\beta)} \right]^{-1} \cdot \frac{\Phi(\beta + \sqrt{g_0 \mu} \cdot T)}{\bar{\Phi}(\beta)}, \\
P\left\{ \text{Ab} \mid W > \frac{T}{\sqrt{N}} \right\} & \approx \frac{1}{\sqrt{N}} \cdot \sqrt{\frac{g_0}{\mu}} \cdot [h(\beta + \sqrt{g_0 \mu} \cdot T) - \bar{\beta}] .
\end{align*}
\]

\( \bar{\beta} = \beta \sqrt{\frac{\mu}{g_0}} \)

\( \Phi(x) = 1 - \Phi(x), \)

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

- No Process Limits
Efficiency-Driven Approximations
(Zeltyn; Whitt)

\( G = \text{(Im)Patience distribution} \)

\[ N = \frac{\lambda}{\mu} \cdot (1 - \gamma) + o(\lambda), \quad \gamma > 0. \]

Assume the equation
\[ G(x) = \gamma \]
has a unique solution \( x^\ast \).

Then
\[
\begin{align*}
\mathbb{P}\{\text{Ab}\} & \approx \gamma \quad \text{(insensitive to } G) \\
\mathbb{P}\{W > T\} & \approx \begin{cases} 
1 - G(T), & T < x^\ast \\
0, & T > x^\ast
\end{cases} \\
\mathbb{P}\{\text{Ab} \mid W > T\} & \approx \gamma - G(T), \quad 0 \leq T < x^\ast.
\end{align*}
\]

- **Derivation**: Laplace Method, based on Baccelli & Hebuterne (1981)
- Towards Dimensioning (with Borst, Reiman)
**Erlang-A: Moderate (Im)patience**

- M/M/N + M queue, with
  service rate $\mu$ equals $\theta$ abandonment rate

- $L_t$: number-in-system at time $t$ (Birth & Death)

- For any $N$, transition-rates for $\{L_t, t \geq 0\}$:

  - $0 \rightarrow 1$: $\lambda$
  - $1 \rightarrow 2$: $\lambda$
  - $2 \rightarrow \ldots$: $\mu$
  - $\ldots \rightarrow N-1$: $\mu$
  - $N-1 \rightarrow N$: $\lambda$
  - $N \rightarrow N+1$: $\lambda$
  - $N+1 \rightarrow \ldots$: $\mu$
  - $N \rightarrow N$: $\mu$
  - $N+1 \rightarrow N+1$: $\mu + \theta$

Note: The same transition rates as $\textbf{M/M/}\infty$
Square-Root Staffing: Motivation

\[ P\{W_q(M / M / N + M) > 0\} = PASTA \]
\[ P\{L(M / M / N + M) \geq N\} = \theta = \mu \]
\[ P\{L(M / M / \infty) \geq N\} \]

Fact: \( L(M / M / \infty) \sim \text{Poisson}(R) \); \( R = \lambda / \mu \) offered load
Square-Root Staffing: Motivation

\[ P\{W_q (M / M / N + M) > 0\} = \text{PASTA} \]
\[ P\{L(M / M / N + M) \geq N\} = \theta = \mu \]
\[ P\{L(M / M / \infty) \geq N\} \]

Fact: \( L(M / M / \infty) \sim \text{Poisson}(R); \quad R = \frac{\lambda}{\mu} \) offered load

For \( R \) not too small:

\[ \frac{d}{L(M/M/\infty)} \approx \text{Normal}(R,R) = R + Z\sqrt{R} \]

\[ \Rightarrow \quad P\{W_q > 0\} \approx P\left\{ Z \geq \frac{N-R}{\sqrt{R}} \right\} = 1 - \phi\left(\frac{N-R}{\sqrt{R}}\right) \]
Square-Root Staffing: Motivation

\[
P\{W_q \ (M / M / N + M) > 0\} = PASTA
\]
\[
P\{L(M / M / N + M) \geq N\} = \theta = \mu
\]
\[
P\{L(M / M / \infty) \geq N\}
\]

Fact: \( L(M / M / \infty) \sim \text{Poisson}(R) \); \( R = \lambda / \mu \) offered load

For \( R \) not too small:

\[
\text{d} \quad \text{d}
L(M/M/\infty) \approx \text{Normal}(R, R) = R + Z\sqrt{R}
\]

\[
\Rightarrow \quad P\{W_q > 0\} \approx P\left\{Z \geq \frac{N-R}{\sqrt{R}}\right\} = 1 - \phi\left(\frac{N-R}{\sqrt{R}}\right)
\]

Given target delay-probability \( \alpha = 1 - \phi\left(\frac{N-R}{\sqrt{R}}\right) \)

\[
\Rightarrow \quad N = R + \beta \cdot \sqrt{R}, \text{ with } \beta = \phi^{-1}(1-\alpha)
\]

\( N \) is the "least integer for which" \( P\{W_q > 0\} \leq \alpha \)
Time-Varying Arrivals

Extension: \( M_t / M / N_t + M \) \( (\mu=\theta) \)

\[ N_t = R_t + \beta \cdot \sqrt{R_t} \]

Fact: \( L_t \sim \text{Poisson}(R_t) \)

\( R_t \) – the offered load at time \( t \), namely:

\[ R_t = E\lambda(t-S_e) \cdot E(S) = E \int_{t-S}^{t} \lambda(u)du \]

\( S_e \) – excess service \( E(S_e) = E(S) \frac{1+c_s^2}{2} \)
Time-Varying Arrivals

Extension: \( M_t / M / N_t + M \) \( (\mu=\theta) \)

\[ N_t = R_t + \beta \cdot \sqrt{R_t} \]

Fact: \( L_t \sim \text{Poisson}(R_t) \)

\( R_t \) – the offered load at time \( t \), namely:

\[ R_t = E \lambda(t - S_e) \cdot E(S) = E \int_{t-s}^{t} \lambda(u) du \]

\( S_e \) – excess service

\[ E(S_e) = E(S) \frac{1+c_s^2}{2} \]

\( L_t \overset{d}{\approx} N(R_t, R_t) \) hence, as before:

\[ N_t = \left\lfloor R_t + \beta \cdot \sqrt{R_t} \right\rfloor, \quad \beta = \phi^{-1}(1-\alpha) \]

hopefully yields time-stable delay probability \( \alpha \):

Indeed, but in fact **TIME-STABLE PERFORMANCE**!

What if \( \mu \neq \theta \)?

Use an *Iterative Algorithm* that is *Simulation-Based*
Performance Measures

• **Delay probability in interval** $t$, calculated by the fraction of customers who are not served immediately upon arrival, out of all arriving customers during the $t$ time-interval

• **Average waiting time in interval** $t$, calculated by the average waiting time of all customers arriving during the $t$ time-interval.

• **Average queue length in interval** $t$, taken constant over the time-interval. The queue length is averaged over all replications.

• **Tail probability in interval** $t$, calculated as the probability that queue size equals or exceeds some threshold (e.g. 3 times average queue)

• **Servers' Utilization in interval** $t$, calculated as the fraction of busy-servers during a time-interval (accounting for servers who are busy only a fraction of the interval)

• **Service grade** $\beta_t$ in interval $t$, which arises from the following "Square-Root Staffing" rule:

$$N_t = R_t + \beta_t \sqrt{R_t}$$
Example: "Real" Call Center

Two-hump arrival functions are typical
(Adapted from Green L., Kolesar P., Soares J. for benchmarking.)

- Service and abandonment rates are both exponential having mean 0.1 (6 min.)
Delay Probability $\alpha$

Service Grade $\beta$

Beta
QED Staffing (α=0.5)
Erlang-A: Theoretical vs. Empirical

\[ P\{Wait>0\} = \alpha \text{ vs. } \beta \quad (N=R+\beta\sqrt{R}) \]

Moderate Patience
Erlang-A: \( P\{\text{Wait}>0\}=\alpha \) vs. \( \beta \) \( (N=R+\beta \sqrt{R}) \)

GMR(\( x \)) describes the asymptotic probability of delay as a function of \( \beta \) when \( \frac{\theta}{\mu} = x \). Here, \( \theta \) and \( \mu \) are the abandonment and service rate, respectively.
Erlang-A: $P\{\text{Abandon}\}^*\sqrt{N}$ vs. $\beta$
Beyond Data Averages
Short Service Times

Jan – Oct:

AVG: 185
STD: 238

Nov – Dec:

AVG: 200
STD: 249

Log-Normal
E(W_q|W_q>0) vs. $\beta$

M/M/100, M/D/100 and M/LN/100 with CV=1

- $E(W_q|W_q>0)$ (M/D/100)
- $E(W_q|W_q>0)$ (M/V/M/100)
- $E(W_q|W_q>0)$ (M/LN/100, CV=1)
QED: Some Intuition (Assume $\mu = 1$)

**M/M/N:** $W_N \mid W_N > 0 \overset{d}{\sim} \exp\left(\text{mean} = \frac{1}{N} \frac{1}{1 - \rho_N}\right)$

$$\sqrt{N} W_N \mid W_N > 0 \overset{d}{\sim} \exp\left(\sqrt{N} (1 - \rho_N)\right) \Rightarrow \exp(\beta)$$

But why $P(W_N > 0) \rightarrow \alpha, \quad 0 < \alpha < 1$? answer via

**M/D/N:** (with P. Jelenkovic and P. Momcilovic)

Observation: Cyclic assignment does not alter waiting times

$$\Rightarrow \text{Same waiting as in } E_N/D/1!$$

QED $N = R + \beta \sqrt{R}$ and consider one of the $E_N/D/1$:

Interarrivals $A_N \approx 1 + \frac{\beta}{\sqrt{N}} + \frac{Z}{\sqrt{N}}, \quad Z \overset{d}{\sim} \mathcal{N}(0,1)$

Lindley $W_N = (W_N + 1 - A_N)^+ \quad (\sqrt{N} \ W_N \Rightarrow W)$

$$P(W_N \leq 0) = P(W_N + 1 - A_N \leq 0) \approx$$
$$\approx P\left(W/\sqrt{N} + 1 - 1 - \beta/\sqrt{N} - Z/\sqrt{N} \leq 0\right) = P(W - \beta \leq Z)$$

$$P(W_N > 0) \rightarrow P(Z < W - \beta) = E\phi(W - \beta) < 1$$

( Efficiency: $N = R+c$ (HT); Quality: $N = R+bR$ (Stable) )
Forecasting the Arrival Function

Theoretical & Empirical

Prob. of Delay vs. $\beta$
Dimensioning the $V$-Model

- $J$ customer classes: arrivals $Poisson(\lambda_j)$.
- $N$ iid servers: service durations $Exp(\mu)$.

The staffing problem:

Given $0 < \alpha_1 < \alpha_2 < \ldots \alpha_J < 1$,

$$\text{Min } N$$

$$\text{s.t. } P_\pi(W_j(\infty) > 0) \leq \alpha_j, \quad j = 1, ..., J$$

for some scheduling policy $\pi$

(Could also minimize $cN + \sum_j d_j \lambda_j EW_j(\infty)$)
Multi-Class $M/M/N$ (V-Design)
Threshold-Based Priorities (Schaack & Larson)

Static priorities $1 > 2 > \ldots$ with thresholds

$$0 = K_1 \leq K_2 \leq \ldots K_J \leq N$$

i.e. a class-$j$ customer is served when it is of the present highest-priority and the number of idle servers is more than $K_j$.

Note:
Optimal Control with State-Dependent Thresholds (Yahalom)
QED Multi-Class $M/M/N$ (V-Design)

Threshold-Based Priorities (Gurvich)

Thresholds: $0 = K_1^N \leq K_2^N \leq ... \leq K_J^N \leq N$

Service Levels $0 \leq \alpha_1 \leq ... \leq \alpha_J \leq 1$: Delay Probabilities

Consider a sequence indexed by $N = 1, 2, ...$

Assume: $\lambda_j^N / \lambda^N \mu \to \rho_j > 0 \ , \forall j$ (all classes non-negligible)

Assume: $K_J^N = o(\sqrt{N})$

Then the following conditions are equivalent:

- **Customer:** $\lim_{N \to \infty} P\{W_j^N > 0\} = \alpha_j$, $0 < \alpha_j < 1$;
- **Server:** $\lim_{N \to \infty} \sqrt{N} (1 - \rho^N) = \beta$, $0 < \beta < \infty$;
- **Manager:** $N \approx R + \beta \sqrt{R}$, $R = \lambda^N / \mu$ large;

in which case $\alpha_j = \left[1 + \frac{\beta \Phi(\beta)}{\phi(\beta)}\right]^{-1}$ and, moreover,

the following two conditions are equivalent

- **Customer:** $\lim_{N \to \infty} P\{W_j^N > 0\} = \alpha_j$, $0 < \alpha_j < 1$, $\forall j$;
- **Manager:** $K_{j+1}^N - K_j^N \to \frac{\ln \alpha_j / \alpha_j + 1}{\ln \sum_{i=1}^j \rho_i}$, $1 \leq j \leq J - 1$;

Solves the Dimensioning Problem
Iterative Algorithm

Inputs

- System primitives:
  - arrival function, service-time distribution, patience distribution (when relevant);
- Target delay probability $\alpha$;
- Time horizon $[0,T]$.

Outputs

- Staffing function, aiming at
  - a delay probability $\alpha$ is over $[0,T]$.

Starting point: The infinite-server heuristics by Jennings, M., Massey, Whitt (1996)
Algorithm (cont.)

**Notation:** \( \forall t \in [0,T] \) (practically \( t=0, \Delta, 2 \Delta, \ldots \))

- \( N_i(t) \) – staffing level at time \( t \), determined in iteration \( i=1,2,\ldots \)
- \( L_i(t) \) – number in the system at \( t \), under staffing function \( s_i(t) \).

**Algorithm:**

1. \( i=0; \ N_0(t) \equiv 0 \) (delay probability =0)
2. Evaluate the distribution of \( L_i(t) \), using simulation.
3. Determine \( N_{i+1}(t) \) as follows:
   \[
   N_{i+1}(t) = \arg \min \{ c : P\{L_i(t) \geq c\} < \alpha \} , \quad 0 \leq t \leq T .
   \]
4. Check stopping condition:
   - if \( \| N_{i+1}(\cdot) - N_i(\cdot) \|_\infty \leq 1 \), then \( N_{i+1}(\cdot) \) is our staffing level;
   - else \( i := i+1 \), and go back to (2).

\((\infty)\) Last iteration. The algorithm converges to a Staffing Function \( N_\infty(\cdot) \) least for which

\[
P\{L_\infty(t) \geq N_\infty(t)\} \leq \alpha , \quad 0 \leq t \leq T .
\]