Service Engineering

Class 14

QED (QD, ED) Queues: Extensions
Skills-Based-Routing (SBR)

• Predictable-Variability: Staffing Queues with a Time-Varying Arrival Rate.

• Parameter Uncertainty (Future Topic).

• Heterogeneous Customers and Servers: Skills-Based Routing.

• Additional Important but Uncovered Topics.
Predictable Variability: Motivation

So far: models with constant arrival rate \( \lambda \).
However, prevalent arrival rates are time-varying: \( \lambda(t) \).

**Arrival Rate in an Israeli Bank**

Common: **SIPP** (= Stationary Independent Period by Period):
- Day divided into time-intervals = periods;
- Arrival rates assume constant over each interval;
- System assumed in steady-state during each interval.

Satisfactory:
- Service times are relatively short;
- Variations in the arrival rate are relatively slow;
- Steady-State reached sufficiently fast.
Predictable-Variability: 
Square-Root Staffing

(with Feldman, Massey and Whitt, 2006.)
Consider $M_t/G/n_t + G$.

Objective: $\Pr\{W > 0\} \leq \alpha$, $0 < \alpha < 1$, stably over the day.

Square-Root Staffing: $n(t) = R(t) + \beta(\alpha)\sqrt{R(t)}$;
Here

- $\beta(\alpha)$ is an appropriate Garnett Function (corresponding to a stationary QED Erlang-A).
- $R(t) =$ Offered-Load at time $t$.

We already know that $R(t) = \lambda(t) \cdot \mathbb{E}[S]$ need not hold.

Why? If peak-arrival took place some time ago and those customers arriving then are still in the system (queue or service), one should take them into account for staffing.

The right Definition of the Time-Varying Offered-Load is Amount of work (time-units-of-service) within system at time $r$:

$$R(t) = \mathbb{E}[\lambda(t - S_e)] \cdot \mathbb{E}[S] = \mathbb{E}\left[\int_{t-S}^t \lambda(u)du\right],$$

where $S_e \stackrel{d}{=} \text{Excess Service}$, with mean

$$\mathbb{E}[S_e] = \mathbb{E}[S] \cdot \frac{1 + c_s^2}{2}.$$

How? $R(t) = EL(t)$ in a corresponding $M_t/G/\infty$ model.
Example: "Real" Call Center
(The "Right Answer" for the "Wrong Reasons")

Time-Varying (two-hump) arrival functions common
(Adapted from Green L., Kolesar P., Soares J. for benchmarking.)

Assume: Service and abandonment times are both
Exponential, with mean 0.1 (6 min.)
Time-Varying Arrivals

Model \( M_t / M / N_t + M \)

Parameters \( \lambda(t) \mu \theta \)

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

Offered Load:

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

Average # in \( M_t / M / \infty \)

Gives rise to **TIME-STABLE PERFORMANCE**

(Why? Think \( M_t / M / N_t + M \) with \( \mu = \theta \);
And if \( \mu \neq \theta \), or generally:
use the Iterative Simulation-Based Staffing Algorithm
in Feldman, M., Massey and Whitt, 2005.)
HW/GMR Delay Functions

\[ \alpha \text{ vs. } \beta \]

![Graph showing delay functions for different models](image-url)
QED Staffing \((\beta=0 \text{ iff } \alpha=0.5)\)
Abandon Probability

Abandon Probability

Abandon Probability

Abandon Probability
Real Call Center: Empirical waiting time, given positive wait

(1) $\alpha=0.1 \ (QD)$  
(2) $\alpha=0.5 \ (QED)$  
(3) $\alpha=0.9 \ (ED)$
The "Right Answer" (for the "Wrong Reasons")

Prevalent Practice \[ N_t = \left[ \lambda(t) \cdot E(S) \right] \] (PSA)

"Right Answer"
\[ N_t \approx R_t + \beta \cdot \sqrt{R_t} \] (MOL)

\[ R_t = E\lambda(t - S) \cdot E(S) \]

Practice \approx "Right" \[ \beta \approx 0 \] (QED)

and \[ \lambda(t) \approx \text{stable over service-durations} \]

Practice Improved \[ N_t = \left[ \lambda[t - E(S)] \cdot E(S) \right] \]

When Optimal? for moderately-patient customers:

1. Satisfization \[ \iff \] At least 50% to be serve immediately

2. Optimization \[ \iff \] Customer-Time = 2 x Agent-Salary
Time-Varying Arrivals: √. Safety-Staffing

Model \( M_t / M / N_t + M \)

Parameters \( \lambda(t) \mu \theta \)

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

\( \mu = \theta : \quad L_t \overset{d}{=} \text{Poisson}(R_t) \overset{d}{=} N(R_t, R_t) \quad \# \text{in system} \)

\[ R_t = E\lambda(t - S) \cdot E(S) = E \int_{t-S}^{t} \lambda(u)du \quad \text{offered load} \]

Given \( L_t \approx R_t + Z\sqrt{R_t} \), \( Z \overset{d}{=} N(0,1) \)

choose \( N_t = R_t + \beta \sqrt{R_t} \)

\[ \Rightarrow \alpha = P(W_t > 0) \overset{\text{PASTA}}{=} P(L_t \geq N_t) = P(Z \geq \beta) = 1 - \phi(\beta) \]

\[ \Rightarrow \beta = \phi^{-1} (1 - \alpha) \quad \text{time-stable} \quad \alpha \overset{\text{PASTA}}{=} P(W_t > 0) \]

Indeed, but in fact \text{TIME-STABLE PERFORMANCE} \\
(\mu \neq \theta, \text{or generally : Iterative Simulation-Based Algorithm})
WHY SERVICE STINKS

Companies know just how good a customer you are—and unless you’re a high roller, they would rather lose you than fix your problem.

PAGE 118
Workforce Management: 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
**NationsBank CRM:**

What are the relationship groups?

- The groups
  - RG1: high-value customers
  - RG2: marginally profitable customers (with potential)
  - RG3: unprofitable customer

- What does it mean for a customer in each group to be profitable? Customer Revenue Management

---

**NationsBank’s Design of the Service Encounter**

**Examples of Specifications: Assignable Grade Of Service (AGOS)**

<table>
<thead>
<tr>
<th></th>
<th>RG1</th>
<th>RG2</th>
<th>RG3</th>
</tr>
</thead>
<tbody>
<tr>
<td>VRU Target</td>
<td>70% of calls</td>
<td>85% of calls</td>
<td>90% of calls</td>
</tr>
<tr>
<td>Abandonment rate</td>
<td>&lt; 1%</td>
<td>&lt; 5%</td>
<td>&lt; 9%</td>
</tr>
<tr>
<td>Speed of Answer</td>
<td>100% in 2 rings</td>
<td>80% in 20 seconds</td>
<td>50% in 20 seconds</td>
</tr>
<tr>
<td>Average Talk Time</td>
<td>no limit</td>
<td>4 min. average</td>
<td>2 min. average</td>
</tr>
<tr>
<td>Rep. Training</td>
<td>universal</td>
<td>product experts</td>
<td>basic product</td>
</tr>
<tr>
<td>Rep. Personalization</td>
<td>request rep / callback</td>
<td>FCFS</td>
<td>FCFS</td>
</tr>
<tr>
<td>Trans. Confirmation</td>
<td>call / fax</td>
<td>call / mail</td>
<td>mail</td>
</tr>
<tr>
<td>Problem Resolution</td>
<td>during call</td>
<td>within 2 business days</td>
<td>within 8 business days</td>
</tr>
</tbody>
</table>

---

3. Wharton

---

5. Wharton
Distributed Call Center: Member1

Internal arrivals: 224
- Served at 1: 67 (29.9)
- Served at 2: 41 (18.3)
- Served at 3: 87 (38.8)
- Served at 4:

Internal arrivals: 643
- Served at 1: 157 (24.4)
- Served at 2: 195 (30.3)
- Served at 3: 282 (43.9)
- Served at 4: 4 (0.6)
- Aban at 4:

External arrivals: 1770
- 1755(99.2 Served)+15(0.8 Aban)
  - Not Interqueued: 1503(84.9)
    - Served: 1497 (99.6/84.6)
    - Aban: 6 (0.4/0.4)
  - Interqueued: 258+9 (1.7)
    - Served here: 110 (41.2/6.2)
    - Served at 1: 58 (21.7/3.3)

External arrivals: 1694
- 1687(99.6% Served)+7( 0.4% Aban)
  - Not Interqueued: 1665(98.3)
    - Served: 1659 (99.6/97.9)
    - Aban: 6 (0.4/0.4)
  - Interqueued: 28+1 (1.7)
    - Served here: 17(58.6/1)
    - Served at 1: 3(10.3/0.2)

Internal arrivals: 613
- Served at 1: 41(6.7)
- Served at 2: 513(83.7)
- Served at 3: 55(9.0)
- Aban at 1: 2(0.3)

Internal arrivals: 81
- Served at 1: 17(21)
- Served at 3: 42(51.9)
- Served at 4: 1(12.5)

Internal arrivals: 122
- 112(91.8 Served)+10(8.2 Aban)
  - Not Interqueued: 93 (76.2)
    - Served: 85 (91.4/69.7)
    - Aban: 8 (8.6/6.6)
  - Interqueued: 27+2 (23.8)
    - Served here: 14(48.3/11.5)
    - Served at 1: 6

10 AM – 11 AM (03/19/01): Interflow Chart Among the 4 Call Centers of Fleet Bank

External arrivals: 2092
- 2063(98.6% Served)+29(1.4% Aban)
  - Not Interqueued: 1209(57.8%)
    - Served: 1184(97.9/56.6)
    - Aban: 25(2.1/1.2)
  - Interqueued: 883(42.2)
    - Served here: 174(19.7/8.3)
    - Served at 2: 438(49.6/20.9)
Example of a Routing Protocol

U.S. Bank: Histogram of Waiting Times
Retail Customers

![Histogram of Waiting Times for Retail Customers]

Business Customers

![Histogram of Waiting Times for Business Customers]
An Introduction to Skills-Based Routing
and its Operational Complexities

By Ofer Garnett and Avishai Mandelbaum
Technion, ISRAEL

( Full Version )

Contents:
1. Introduction
2. N-design with single servers
3. X-design with multi-server pools and impatient customers
4. Technical Appendix: Simulations – the computational effort

Acknowledgement: This teaching-note was written with the financial support of the Fraunhofer IAO Institute in Stuttgart, Germany. The authors are grateful to Dr. Thomas Meiren and Prof. Klaus-Peter Fähnrich of the IAO for their assistance and encouragement.
Introduction

Multi-queue parallel-server system = schematic depiction of a telephone call-center:

Here the $\lambda$'s designate arrival rates, the $\mu$'s service rates, the $\theta$'s abandonment rates, and the $S$'s are the number of servers in each server-pool.

Skills-Based Design:
- **Queue**: "customer-type" requiring a specific type of service;
- **Server-Pool**: "skills" defining the service-types it can perform;
- **Arrow**: leading into a server-pool define its skills / constituency.

For example, a server with skill 2 ($S2$) can serve customers of type 3 ($C3$) at rate $\mu_6$ customers/hour.

Customers of type 3 arrive randomly at rate $\lambda_3$ customers/hour, equipped with an impatience rate of $\theta_3$. 
Some Canonical Designs - Animation

**I** – dedicated (specialized) agents

**N**: for example,
- C1 = VIP, then S2 are serving C1 to improve service level.
- C2 = VIP, then S2 serve C1 to improve efficiency.
- S2 = Bilingual.

**X**: for example, S1 has C1 as Primary and C2 as Secondary Types.

**V**: Pure Scheduling; **Upside-down V**: Pure Routing.
Major Design / Engineering Decisions

1. Classifying customers into types (Marketing):
   Tech. support vs. Billing, VIP vs. Members vs. New

2. Determining server skills, incentives, numbers (HRM, OM, OR)
   Universal vs. Specialist, Experienced / Novice, Uni- / Multi-lingual;
   **Staffing**: how many servers?

3. Prerequisite Infrastructure - MIS / IT / Data-Bases (CS, Statistics)
   CTI, ERP, Data-Mining

Major Control Decisions

4. Matching customers and agents (OR)
   - **Customer Routing**: Whenever an agent turns idle and there
     are queued customers, which customer (if any) should be routed
     to this agent.
   - **Agent Scheduling**: Whenever a customer arrives and there
     are idle agents, which agent (if any) should serve this customer.

5. **Load Balancing**
   - Routing of customers to distributed call centers (eg. nation-wide)

**Multidisciplinary Challenge**
Skills-Based Routing: protocol for online matching of S's and C's.

- **Prevalent**: Static Priorities of customer types and agent skills
- **Index-based**: Dynamic Priorities via continuous review
- **Threshold-based**: Dynamic Management by Exception
- **Others**: discrete review, credit schemes (SLA), scripts; call backs

Example: **Scripts** for Staffing, Scheduling, Routing

**Setup A** : (X-design)

"VIP" servers : S₁ = 20
- If "VIP" queue not empty serve the "VIP" queue + all "Members" waiting more than 40 seconds, as a single FIFO queue.
- If "VIP" queue is empty, serve the first in the "Member" queue.

"Member" servers : S₂ = 15
- If "Member" queue not empty serve the "Member" queue + all "VIPs" waiting more than 6 seconds, as a single FIFO queue.
- If "Member" queue is empty, serve the first in the "VIP" queue.

**Setup C** : (V-design; feasible since servers are assumed equally skilled.)

Total servers: 35
- Serve as a FIFO queue, but "VIPs" enter the queue with a virtual 15 second wait (i.e. as if they had joined the queue 15 seconds earlier).
Chart 2: 1000 Calls/hour - ASA

Overall
- A: 22.8
- B: 22.7
- C: 22.9
- D: 12.5

Members
- A: 24.6
- B: 24.6
- C: 24.6
- D: 11

VIP
- A: 18.2
- B: 16.6
- C: 16.1
- D: 16.8

Chart 3: 1000 Calls - Abandonment

Overall
- A: 17%
- B: 17%
- C: 17%
- D: 17%

Members
- A: 18%
- B: 20%
- C: 20%
- D: 20%

VIP
- A: 13%
- B: 7%
- C: 7%
- D: 7%

Chart 4: 1000 Calls - Overflows

Overall
- A: 24%
- B: 19%

VIP 2 Members
- A: 27%
- B: 14%

Members 2 VIP
- A: 39%
- B: 13%

VIP Members Overall
- A
- B
- C
- D
WHAT IF: 1500 Calls/hour - ASA

Chart 7: 1500 Calls - Abandonment

Chart 8: 1500 Calls - Overflows
Reality

- Technology enables smart systems
- Reality becomes increasingly complex
- Solutions are urgently needed
- Theory lags significantly behind needs
- **Ad-hoc methods**: heuristics, simulation-based

Research Status

- Efficiency-driven SBR well understood and solved
- QED SBR is challenging and advancing
- **Small yet significant models for theoretical insight**
- Principles/Guidelines for design, staffing, control
- Implementation: fine-tuning of parameters, scale-up
Static Priorities (Cross-Training): Some Subtleties

- C1 are **VIP**, hence S2 helps S1 by giving priority to C1 over C2.
- If both servers are idle - Ci customers are routed to server Si

Queue length: S2 helps with VIP C1, Heavy Loading -

Q2 "explodes, while Q2 is negligibly small – why ?
Servers' utilization profiles

Instability: S2 **overworked** serving C1 and neglecting C2, while S1 is **20% idle**.

To avoid "overzealous help", apply **Threshold Control**: S2 assists S1 only when Q1 is at or above a certain threshold

Both Q1 and Q2 are stable.

Now fine-tuning of the threshold value
Efficiency-Driven SBR - the "EASY" Case (Stolyar)

Examples: Scarce agents, hence must be well utilized. Email-dominance, hence can delay response.

Classical special case: $V$-design
- Agent Scheduling: upon service completion, if
  1. Same mean service times: serve the costliest queue (largest $c$)
  2. Same delay costs: serve the shortest service (smallest $m$)
  3. Generally: serve the largest $c/m$ (= index).

General (N, X, W, M, … ) solution: Index Control is optimal, under sufficient skills-overlap (complete resource pooling; Harrison, Lopez).
- Customer Routing: irrelevant, since essentially all customers wait.
- Agent Scheduling: upon service completion, the server chooses the queue with the largest index and serves its "oldest" customer.
- Index: marginal waiting-cost per unit of average service-time (Example: "waiting-time" of "oldest" customer in queue)

However: well-managed telephone services are not
(or, typically, should not be) E-Driven !?
**V-Design: Pure Scheduling**

N agents, fully flexible

C1 = VIP

Optimal Scheduling: **Agent Reservation** (Yahalom)

- C1 (=VIP) always served, when possible;
- C2 served only if # of idle agents exceeds a threshold.

**QED** regime: $\sqrt{\cdot}$ **Safety-Staffing**, as before (Gurvich)

Threshold Size (relative to N) determines Service Levels:

- Large: C1 is Q-served, C2 is E-served
- Small: C1 and C2 indistinguishable QED
- Moderate: C1 is Q-served, C2 is QED

$\sqrt{\cdot}$ Safety-Staffing is asymptotically optimal.
**Reversed-V Design: Pure Routing**

Homogeneous Customers

Heterogeneous Agents: $S_2 = \text{Faster}$

Optimal Routing: "Slow-Server" phenomenon (Rykov)

- $S_2(=\text{Fast})$ always employed, if possible;

- $S_1(= \text{Slow})$ employed if # in queue exceeds a threshold.

**QED** regime: $\sqrt{\cdot}$ Safety-Staffing – see below (Armony)

- No threshold needed: just have all servers work when possible, ensuring that the "fast" get the priority.

**Asymptotically optimal staffing:**

1. Given a delay probability, determine $S_1 + S_2$ via $\sqrt{\cdot}$ Safety.
2. Given staffing costs, determine $S_1 / S_2$.

**Distributed** call centers: in progress.