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; Columbia Ph.D., 2007.
  - Zeltyn (M/G/n + G) - Technion Ph.D., 2005;
  - Feldman (Predictable Queues) - Technion M.Sc., 2006-7.
- 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).

Queueing Science: Data-Based QED's Q's

Traditional Queueing Theory predicts that **Service-Quality** and **Servers' Efficiency** **must** be traded off against each other.

For example, **M/M/1 in heavy-traffic**: **91%** server's utilization goes with

$$\textbf{Congestion Index} = \frac{E[Wait]}{E[Service]} = \mathbf{10},$$

and only 9% of the customers are served immediately upon arrival.

Yet, heavily-loaded queueing systems with **Congestion Index = 0.1** (Waiting one order of magnitude less than Service) are prevalent:

- **Call Centers**: Wait **"seconds"** for **minutes** service;
- **Transportation**: Search **"minutes"** for **hours** parking;
- **Hospitals**: Wait **"hours"** in ED for **days** hospitalization in IW's;

and, moreover, a significant fraction are not delayed in queue. (For example, in well-run call-centers, **50%** served "immediately", along with over **90%** agents' utilization, is not uncommon) **?**

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.

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 $\mathbf{M}_t/\mathbf{G}/\mathbf{n}_t + \mathbf{G}$":

Operational Regimes: Rules-of-Thumb

Constraint	P{Ab}		E[W]		P{$W > T$}	
Offered Load	Tight 1-10%	Loose $\geq$ 10%	Tight $\leq 10\%\mathrm{E}[\tau]$	Loose $\geq 10\%\mathrm{E}[\tau]$	Tight $0 \leq T \leq 10\%\mathrm{E}[\tau]$ $5\% \leq \alpha \leq 50\%$	Loose $T \geq 10\%\mathrm{E}[\tau]$ $5\% \leq \alpha \leq 50\%$
Small (10's)	QED	QED	QED	QED	QED	QED
Moderate-to-Large (100's-1000's)	QED	ED, QED	QED	ED, QED if $\tau \overset{d}{=} \exp$	QED	ED+QED

ED: $\boldsymbol{N} \approx \boldsymbol{R} - \gamma \boldsymbol{R}$ $\quad (0.1 \leq \gamma \leq 0.25)$.

QD: $\boldsymbol{N} \approx \boldsymbol{R} + \delta \boldsymbol{R}$ $\quad (0.1 \leq \delta \leq 0.25)$.

QED: $\boldsymbol{N} \approx \boldsymbol{R} + \beta\sqrt{\boldsymbol{R}}$ $\quad (-1 \leq \beta \leq 1)$.

ED+QED: $\boldsymbol{N} \approx (1-\gamma)\boldsymbol{R} + \beta\sqrt{\boldsymbol{R}}$ $\quad (\gamma, \beta$ as above).

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

$$C_s \cdot n + (C_a \cdot \mathrm{P}_n\{\mathrm{Ab}\} + C_w \cdot \mathrm{E}_n[W_q]) \cdot \lambda \,,$$

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

$$r \cdot \lambda \cdot [1 - \mathrm{P}_n\{\mathrm{Ab}\}] - [C_s \cdot n + C_w \cdot \mathrm{E}_n[W_q]) \cdot \lambda] \,,$$

where r is the **revenue** from a service.

Operational Regimes: Rules-of-Thumb
(The Basic Service Station $M_t/G/n_t + G$)

$\boldsymbol{R_t} = \mathrm{E}\int_{t-S}^{t} \boldsymbol{\lambda(u)du} = \mathrm{E}\boldsymbol{\lambda(t - S_e)} \cdot \mathrm{E}\boldsymbol{S}$ = **Offered-Load** at time t, namely **"minutes" of work (=service) within the system at time t**. (Steady-State: $\boldsymbol{R} = \boldsymbol{\lambda} \times \mathrm{E}[\boldsymbol{S}]$ Erlangs, namely "minutes" of work that arrive per "minute".)

- Efficiency-Driven (ED) Regime:

$$n_t \approx R_t - \gamma R_t \,, \qquad 0 < \gamma < 1 \,.$$

Under-staffing with respect to the offered-load.

- Quality-Driven (QD) Regime:

$$n_t \approx R_t + \delta R_t \,, \qquad \delta > 0 \,.$$

Over-staffing with respect to the offered-load.

- Quality- and Efficiency-Driven (QED) Regime:

$$n_t \approx R_t + \beta\sqrt{R_t}, \qquad -\infty < \beta < \infty \,.$$

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);
- Above version: Garnett, Zeltyn, Feldman (Technion theses).
- 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:

$$n \;\approx\; R_t - \gamma R_t\,, \qquad 0.1 \le \gamma \le 0.25\,.$$

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

QD regime:

$$n \;\approx\; R_t + \delta R_t\,, \qquad 0.1 \le \delta \le 0.25\,.$$

Essentially **no** delays.

QED regime:

$$n \;\approx\; R_t + \beta\sqrt{R_t}\,, \qquad -1 \le \beta \le 1\,.$$

- %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: $\boldsymbol{R = \lambda/\mu}$ 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: $\boldsymbol{n = R}$.

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[τ]=3 min

λ/hr	n	Occupancy	P$\{W_q > 0\}$	E[W_q]	P{Ab}
20	1	63.2%	63.2%	1:06.2	36.8%
100	5	82.5%	56.0%	0:31.6	17.5%
500	25	92.0%	52.7%	0:14.3	8.0%
2,500	125	96.4%	51.2%	0:06.4	3.6%
9,000	450	98.1%	50.6%	0:03.4	1.9%
$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$
∞	∞	1 ?	50% ?	0 ?	0 ?

~~Motivation: QD Operation, or “What can be Achieved? At what Cost?”~~

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

ıe	Avg Speed Ans W	Avg Aban Time	ACD Calls λ	Avg ACD Time 1/μ	Avg ACW Time	Aban Calls #Aban	% ACD Time	% Ans Calls	Avg Pos Staff N	Calls Per Pos	%Serv Lev	%Aux Time	%ACW Time	%ACD Time P
Totals	:00:02	:00:28	10456	:03:47	:00:25	46	53	98	70	149		8		
12:00 AM*	:00:00	:00:00	26	:04:31	:00:02	1	76	61	7	4	61	2	16	61
12:30 AM*	:00:03	:04:10	14	:07:27	:00:33	1	89	52	5	3	48	1	26	63
1:00 AM*	:00:00		9	:04:54	:11:29	0	91	90	1	7	90	0	26	65
5:30 AM*			0			0	0		0	0		33	0	0
6:00 AM*	:00:00		12	:03:21	:00:19	0	21	100	7	2	100	9	2	19
6:30 AM*	:00:00		27	:02:51	:00:20	0	32	100	14	2	100	5	3	29
7:00 AM*	:00:00		62	:03:34	:00:15	0	38	100	21	3	100	13	4	34
7:30 AM*	:00:00		93	:03:11	:00:34	0	36	100	30	3	100	7	4	32
8:00 AM*	:00:00		120	:03:37	:00:40	0	39	100	47	3	100	8	6	33
8:30 AM*	:00:00		193	:03:04	:00:14	0	44	100	61	3	100	10	7	37
9:00 AM*	:00:01		293	:03:25	:00:25	0	54	99	75	4	97	9	7	47
9:30 AM*	:00:02	:00:06	381	:03:45	:00:22	2	60	97	91	4	93	8	8	52
10:00 AM*	:00:02	:00:01	416	:03:49	:00:26	1	63	97	94	4	96	5	8	55
10:30 AM*	:00:00		349	:03:35	:00:33	0	52	99	96	4	99	6	8	44
11:00 AM*	:00:00		352	:03:50	:00:27	0	51	100	102	3	100	7	8	45
11:30 AM*	:00:00		349	:03:44	:00:18	0	49	100	97	4	100	8	5	45
12:00 PM*	:00:01		354	:03:59	:00:18	0	52	95	95	4	95	8	5	47
12:30 PM*	:00:00		336	:03:38	:00:21	0	52	99	97	3	99	9	8	46
1:00 PM*	:00:00		347	:03:53	:00:32	0	51	99	98	4	99	11	8	44
1:30 PM*	:00:00		366	:03:52	:00:14	0	56	99	99	4	99	11	7	50
2:00 PM*	:00:01		393	:03:55	:00:17	0	51	100	106	4	100	10	5	46
2:30 PM*	:00:00		403	:03:58	:00:13	0	54	100	112	4	100	10	4	50
3:00 PM*	:00:00	:00:04	410	:04:02	:00:16	1	57	98	110	4	98	8	5	51
3:30 PM*	:00:00		347	:03:59	:00:14	0	60	100	100	3	100	7	5	45
4:00 PM*	:00:00		382	:03:46	:01:37	0	64	100	98	4	100	6	7	47
4:30 PM*	:00:00		379	:03:41	:00:19	0	55	99	97	4	99	8	5	50
5:00 PM*	:00:00		411	:03:53	:00:19	0	53	100	109	4	100	9	5	46
5:30 PM*	:00:01		387	:03:58	:00:19	0	58	99	98	4	99	10	6	51
6:00 PM*	:00:01	:00:21	371	:03:28	:00:25	1	53	98	91	4	98	9	6	47
6:30 PM*	:00:00		260	:03:26	:00:13	0	41	100	90	3	100	8	4	37
7:00 PM*	:00:00		269	:03:24	:00:17	0	42	100	78	3	100	9	5	38

Motivation: QD Performance Analysis

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

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

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

Occupancy: $\rho = R/n,$
$= 59/94 = 63\%.$

Compare with the column "% ACD Time" of the ACD report.

QD Rule-of-Thumb: $n \approx R + \delta \cdot R,\ \ \delta > 0,$ where
γ = **Service-Grade** 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**)

λ/hr	N Agents	ρ = OCC	L_q = Que	W_q = ASA
8	2	50%	0.3	2:30
16	3	67%	0.9	3:20
24	4	75%	1.5	3:49
32	5	80%	2.2	4:09

λ/hr	N	ρ = OCC	L_q = Que	W_q = ASA
72	10	90%	60	5:01
120	16	93.8%	11	5:29
400	51	98%	42	6:18
640	81	98.8%	70	6:32
1,280	161	99.4%	145	6:48
2,560	321	99.7%	299	7:00
3,600	451	**99.8%**	423	**7:04**
↓	↓	↓	↓	↓
∞	∞	**1**	∞	**7:30 !**

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

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

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

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

$N(1-\rho_N) = 1 \quad , \quad \rho_N \to 1 .$

Motivation: Operational Regimes

Health insurance company. ACD Report.

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

Quality-Driven (QD) Erlang-A

Time	Calls	Answered	Abandoned%	ASA	AHT	Occ%	# of agents
17:00	**615**	**615**	**0.0%**	**2**	**328**	**83.0%**	**135.0**

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

Offered Load:

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

Characterization:

$$n \;=\; R \cdot (1+\delta)\,, \qquad \delta > 0.$$

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

Time	Calls	Answered	Abandoned%	ASA	AHT	Occ%	# of agents
13:30	**1,061**	**961**	**9.4%**	**67**	**306**	**100.0%**	**163.4**

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

Offered Load:

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

Characterization:

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

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

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

Proof via flow conservation (fluid-view):

$\lambda \cdot (1 - P\{Ab\}) = n \cdot \mu$, hence $P\{Ab\} = 1 - \frac{n}{R} = \gamma$.

QED Erlang-A

Time	Calls	Answered	Abandoned%	ASA	AHT	Occ%	# of agents
14:30	**1,212**	**1,179**	**2.7%**	**23**	**304**	**96.6%**	**206.1**

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

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:

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

QOS parameter:

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

Square-Root Staffing Rule:

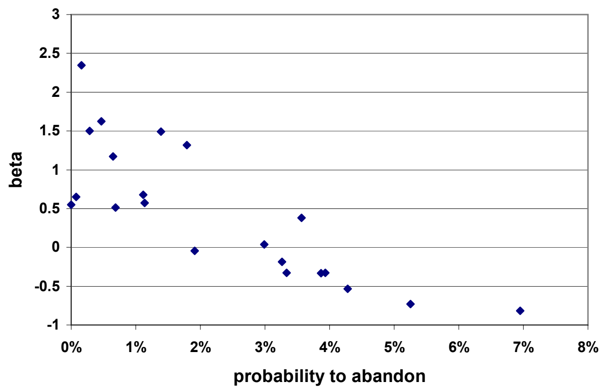
- Described by Erlang already in 1924 (used in 1913);
- Folklore till Halfin & Whitt, 1981 (Erlang-C);
- Above (Erlang-A) from Garnett's Technion M.Sc. thesis, 2001.

The QED Regime in Practice

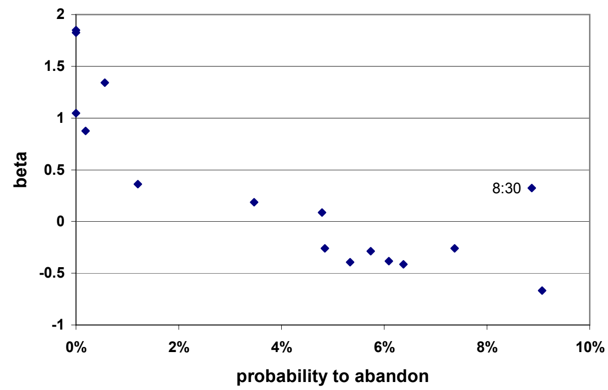
Two call centers: U.S. (Health-Insurance) and Italian (Tele-Banking). Calculate hourly $\beta = \frac{n-R}{\sqrt{R}}$, then compare against performance.

QOS β vs. Abandonment

U.S. data

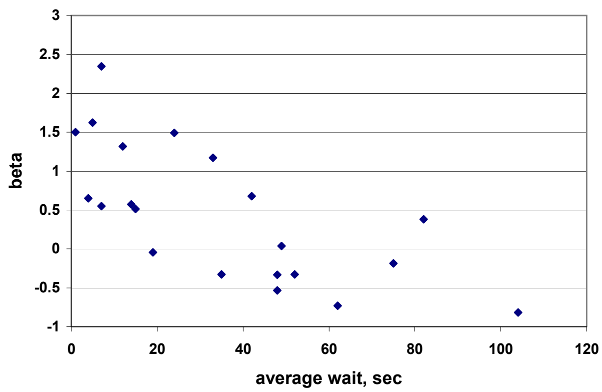


Italian data

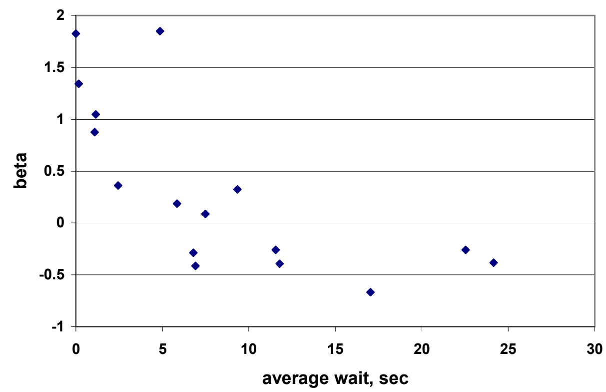


QOS β vs. Average Wait

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.