Service Engineering (Science, Management)

Avi Mandelbaum
Technion IE&M

Course Contents

- Introduction to "Services" and "Service-Engineering"
- The Two Prerequisites: Measurements, Models (Operational)
- Empirical (Data-Based) Models
- Fluid (Deterministic) Models
- Stochastic Framework: Dynamic-Stochastic PERT/CPM
- The Building Blocks of a Basic Service Station:
  - Arrivals; Forecasting
  - Service Durations; Workload
  - (Im)Patience; Abandonment
  - Returns (During, After; Positive, Negative)
- Stochatic Models of a Service Station
  - Markovian Queues: Erlang B/C/A,. . .,/R, Jackson
  - Non-Parametric Queues: $G/G/n$, ...
- Operational Regimes and Staffing: ED, QD, QED
- Heterogeneous Customers and Servers (CRM, SBR)

Background Material

Downloadable from the **References** menu in
http://ie.technion.ac.il/serveng/References

Gans (U.S.A.), Koole (Europe), and M. (Israel):
"Telephone Call Centers: Tutorial, Review and Research Prospects."
MSOM, 2003.

Brown, Gans, M., Sakov, Shen, Zeltyn, Zhao:
"**Statistical** Analysis of a Telephone Call Center: A Queueing-Science Perspective." JASA, 2005.

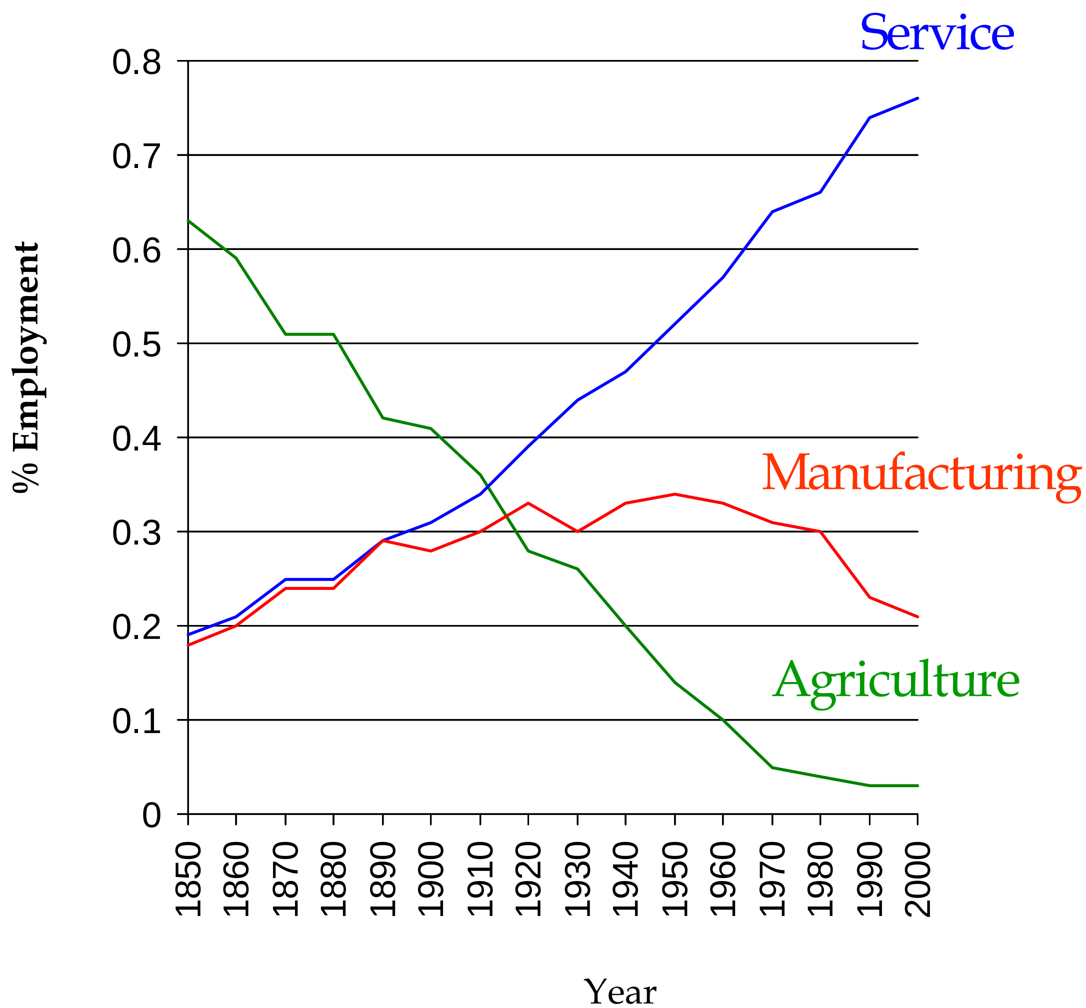
Trofimov, Feigin, M., Ishay, Nadjharov:
"**DataMOCCA**: Models for Call/Contact Center Analysis. (Model Description and Introduction to User Interface.)" Technion Report, 2004-2006.

Technion's **"Service-Engineering" course** lectures: Measurements, Arrivals, Service Times, (Im)Patience, Fluid Models, QED Q's.

M. "Call Centers: Research **Bibliography** with Abstracts."
Version 7, December 2006.

Introduction to “Services”

U.S. Employment by Sector, 1850 - 2000+



We focus on:

- Function: **Operations** (vs./plus IT, HRM, Marketing)
- Dimension: Accessibility, **Capacity** (vs. RM, SCM,...)
- Modelling Framework: **Queueing** Theory (plus Science)
- Applications: **Call/Contact Centers** (Healthcare,...)

Scope of the Service Industry

- Wholesale and retail trade;
- Government services;
- Healthcare;
- Restaurants and food;
- Financial services;
- Transportation;
- Communication;
- Education;
- Hospitality business:
- Leisure services.

Our Application Focus: **telephone call centers**,
which play an important role in most of these sectors.

Services: Subjective Trends

"Everything is Service"

Rather than buying a **product**, why not **buy only the service it provides?** For example, **car leasing**; or, why setup and run a **help-desk** for technical support, with its costly fast-to-obsolete hardware, growing-sophisticated software, high-skilled peopleware and ever-expanding infoware, rather than let **outsourcing** do it all for you?

"Data; Technology and Human Interaction

Far too little reliance on **data, the language of nature**, in formulating models for the **systems and processes of the deepest importance to human beings**, namely those in which **we are actors**. Systems with fixed rules, such as physical systems, are relatively simple, whereas systems involving human beings expressing their microgoals ... can exhibit incredible complexity; there is yet the hope to devise tractable models through **remarkable collective effects** ...
(Robert Herman: "Reflection on Vehicular **Traffic Science**".)

Fusion of Disciplines: POM/IE, Marketing, IT, HRM

The highest challenge facing banks with respect to efficient and effective innovation lies in the **"New Age Industrial Engineer"** that must combine technological knowledge with process design in order to create the delivery system of the future.
(Frei, Harker and Hunter: "**Innovation in Retail Banking**").

Service-Engineering

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 with the traditional and prevalent

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

Additional Sources (all with websites):

- Fraunhofer **IAO** (Service Engineering, 1995): ... application of engineering science know-how to the service sector ... models, methods and tools for systematic development and design of service products and service systems ...
- **NSF SEE** (Service Enterprise Engineering, 2002): ... Customer Call/Contact Centers ... staff scheduling, dynamic pricing, facilities design, and quality assurance ...
- **IBM SSME** (Services Science, Management and Engineering, 2005): ... new discipline brings together computer science, operations research, industrial engineering, business strategy, management sciences, social and cognitive sciences, and legal sciences ...

Staffing: How Many Servers?

Fundamental problem in service operations: Healthcare, . . . , or **Call Centers**, as a representative example:

- People: $\approx 70\%$ operating costs; $\geq 3\%$ U.S. workforce.
- Business-Frontiers but also Sweat-Shops of the 21^{st} Century.

Reality

- **Complex** and becoming more so
- Staffing is Erlang-based (1913!)

$\Longrightarrow$ Solutions urgently needed

- Technology can accommodate smart protocols
- Theory lags significantly behind needs

$\Longrightarrow$ Ad-hoc methods prevalent: heuristics- or simulation-based.

Research Progress based on

- **Simple Robust Models**, for theoretical insight into complex realities. Their analysis requires and generates:
- Data-Based **Science**: Model, Experiment, Validate, Refine.
- **Management** Principles, Tools: **Service Engineering**.

The First Prerequisite: Data & Measurements

Robert Herman ("Father" of Transportation Science): Far too little reliance on **Data, the language of nature**, in formulating models for the systems of the deepest importance to human beings, namely those in which we are actors.

Empirical "Axiom": The Data One Needs is **Never** There For One To Use (Always Problems with Historical Data).

Averages do NOT tell the whole story
Individual-Transaction Level Data: Time-Stamps of Events

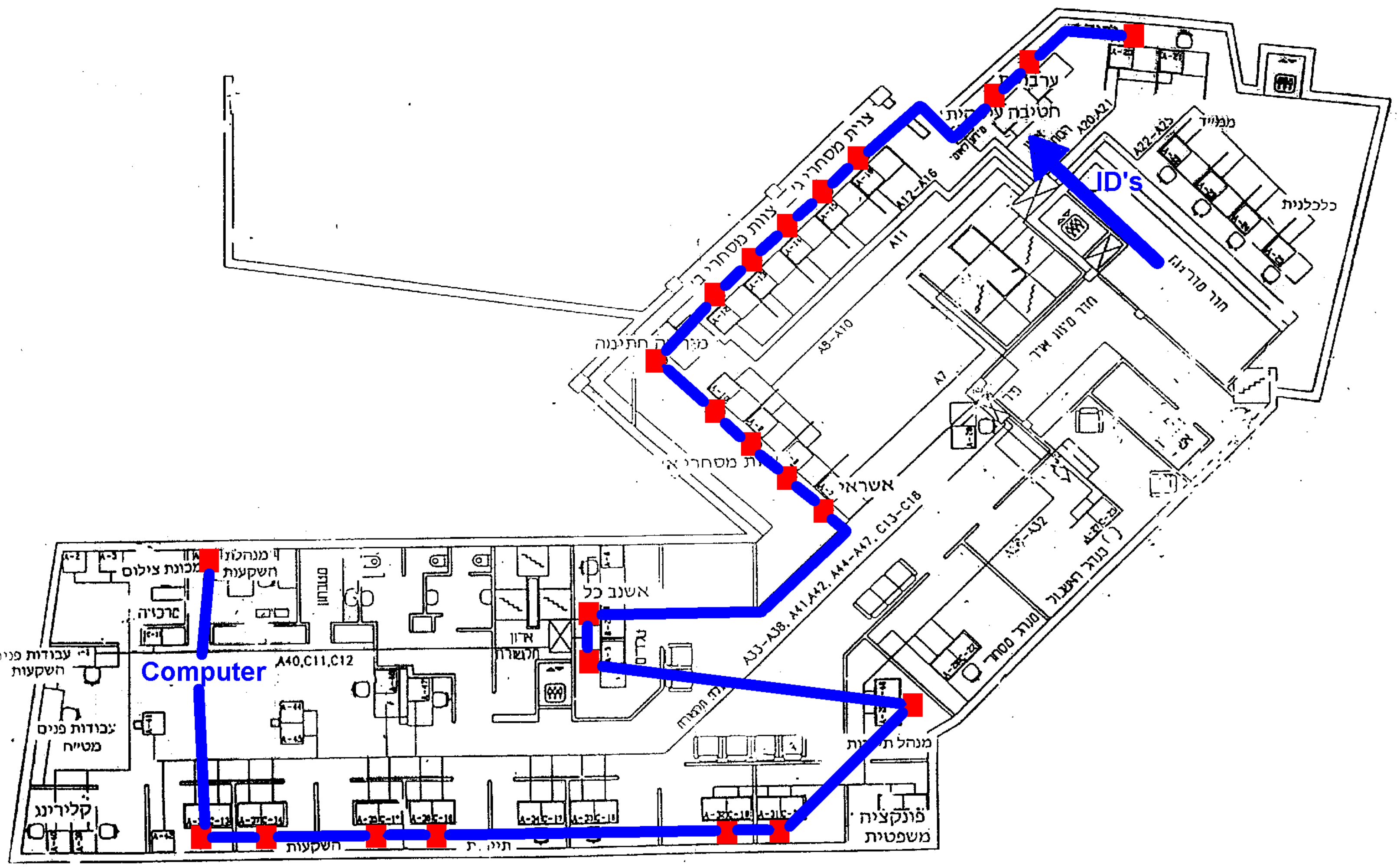
- **Face-to-Face:** T, C, S, I, O, F (QIE, RFID)
- **Telephone:** ACD, CTI/CRM, Surveys
- **Internet:** Log-files
- **Transportation:** measuring devices on highways/intersections

Our Databases: Operations (vs. Marketing, Surveys, ...)

- Face-to-Face data (branch banking) – recitations; QUESTA
- Telephone data (small banking call center) – homework; JASA
- **DataMOCCA** (large cc's: repository, interface) – class/research; Website

Future Research:
Healthcare, Multimedia, Field-Support; Operation+Marketing,

Measurements: Face-to-Face Services
23 Bar-Code Readers at an Israeli Bank



Measurements: Telephone Services
Log-File of Call-by-Call Data

vru+line	call_id	customer_id	priority	type	date	vru_entry	vru_exit	vru_time	q_start	q_exit	q_time	outcome	ser_start	ser_exit	ser_time	server
AA0101	44749	27644400	2	PS	990901	11:45:33	11:45:39	6	11:45:39	11:46:58	79	AGENT	11:46:57	11:51:00	243	DORIT
AA0101	44750	12887816	1	PS	990905	14:49:00	14:49:06	6	14:49:06	14:53:00	234	AGENT	14:52:59	14:54:29	90	ROTH
AA0101	44967	58660291	2	PS	990905	14:58:42	14:58:48	6	14:58:48	15:02:31	223	AGENT	15:02:31	15:04:10	99	ROTH
AA0101	44968	0	0	NW	990905	15:10:17	15:10:26	9	15:10:26	15:13:19	173	HANG	00:00:00	00:00:00	0	NO_SERVER
AA0101	44969	63193346	2	PS	990905	15:22:07	15:22:13	6	15:22:13	15:23:21	68	AGENT	15:23:20	15:25:25	125	STEREN
AA0101	44970	0	0	NW	990905	15:31:33	15:31:47	14	00:00:00	00:00:00	0	AGENT	15:31:45	15:34:16	151	STEREN
AA0101	44971	41630443	2	PS	990905	15:37:29	15:37:34	5	15:37:34	15:38:20	46	AGENT	15:38:18	15:40:56	158	TOVA
AA0101	44972	64185333	2	PS	990905	15:44:32	15:44:37	5	15:44:37	15:47:57	200	AGENT	15:47:56	15:49:02	66	TOVA
AA0101	44973	3.06E+08	1	PS	990905	15:53:05	15:53:11	6	15:53:11	15:56:39	208	AGENT	15:56:38	15:56:47	9	MORIAH
AA0101	44974	74780917	2	NE	990905	15:59:34	15:59:40	6	15:59:40	16:02:33	173	AGENT	16:02:33	16:26:04	1411	ELI
AA0101	44975	55920755	2	PS	990905	16:07:46	16:07:51	5	16:07:51	16:08:01	10	HANG	00:00:00	00:00:00	0	NO_SERVER
AA0101	44976	0	0	NW	990905	16:11:38	16:11:48	10	16:11:48	16:11:50	2	HANG	00:00:00	00:00:00	0	NO_SERVER
AA0101	44977	33689787	2	PS	990905	16:14:27	16:14:33	6	16:14:33	16:14:54	21	HANG	00:00:00	00:00:00	0	NO_SERVER
AA0101	44978	23817067	2	PS	990905	16:19:11	16:19:17	6	16:19:17	16:19:39	22	AGENT	16:19:38	16:21:57	139	TOVA
AA0101	44764	0	0	PS	990901	15:03:26	15:03:36	10	00:00:00	00:00:00	0	AGENT	15:03:35	15:06:36	181	ZOHARI
AA0101	44765	25219700	2	PS	990901	15:14:46	15:14:51	5	15:14:51	15:15:10	19	AGENT	15:15:09	15:17:00	111	SHARON
AA0101	44766	0	0	PS	990901	15:25:48	15:26:00	12	00:00:00	00:00:00	0	AGENT	15:25:59	15:28:15	136	ANAT
AA0101	44767	58859752	2	PS	990901	15:34:57	15:35:03	6	15:35:03	15:35:14	11	AGENT	15:35:13	15:35:15	2	MORIAH
AA0101	44768	0	0	PS	990901	15:46:30	15:46:39	9	00:00:00	00:00:00	0	AGENT	15:46:38	15:51:51	313	ANAT
AA0101	44769	78191137	2	PS	990901	15:56:03	15:56:09	6	15:56:09	15:56:28	19	AGENT	15:56:28	15:59:02	154	MORIAH
AA0101	44770	0	0	PS	990901	16:14:31	16:14:46	15	00:00:00	00:00:00	0	AGENT	16:14:44	16:16:02	78	BENSION
AA0101	44771	0	0	PS	990901	16:38:59	16:39:12	13	00:00:00	00:00:00	0	AGENT	16:39:11	16:43:35	264	VICKY
AA0101	44772	0	0	PS	990901	16:51:40	16:51:50	10	00:00:00	00:00:00	0	AGENT	16:51:49	16:53:52	123	ANAT
AA0101	44773	0	0	PS	990901	17:02:19	17:02:28	9	00:00:00	00:00:00	0	AGENT	17:02:28	17:07:42	314	VICKY
AA0101	44774	32387482	1	PS	990901	17:18:18	17:18:24	6	17:18:24	17:19:01	37	AGENT	17:19:00	17:19:35	35	VICKY
AA0101	44775	0	0	PS	990901	17:38:53	17:39:05	12	00:00:00	00:00:00	0	AGENT	17:39:04	17:40:43	99	TOVA
AA0101	44776	0	0	PS	990901	17:52:59	17:53:09	10	00:00:00	00:00:00	0	AGENT	17:53:08	17:53:09	1	NO_SERVER
AA0101	44777	37635950	2	PS	990901	18:15:47	18:15:52	5	18:15:52	18:16:57	65	AGENT	18:16:56	18:18:48	112	ANAT
AA0101	44778	0	0	NE	990901	18:30:43	18:30:52	9	00:00:00	00:00:00	0	AGENT	18:30:51	18:30:54	3	MORIAH
AA0101	44779	0	0	PS	990901	18:51:47	18:52:02	15	00:00:00	00:00:00	0	AGENT	18:52:02	18:55:30	208	TOVA
AA0101	44780	0	0	PS	990901	19:19:04	19:19:17	13	00:00:00	00:00:00	0	AGENT	19:19:15	19:20:20	65	MEIR
AA0101	44781	0	0	PS	990901	19:39:19	19:39:30	11	00:00:00	00:00:00	0	AGENT	19:39:29	19:41:42	133	BENSION
AA0101	44782	0	0	NW	990901	20:08:13	20:08:25	12	00:00:00	00:00:00	0	AGENT	20:08:28	20:08:41	13	NO_SERVER
AA0101	44783	0	0	PS	990901	20:23:51	20:24:05	14	00:00:00	00:00:00	0	AGENT	20:24:04	20:24:33	29	BENSION
AA0101	44784	0	0	NW	990901	20:36:54	20:37:14	20	00:00:00	00:00:00	0	AGENT	20:37:13	20:38:07	54	BENSION
AA0101	44785	0	0	PS	990901	20:50:07	20:50:16	9	00:00:00	00:00:00	0	AGENT	20:50:15	20:51:32	77	BENSION
AA0101	44786	0	0	PS	990901	21:04:41	21:04:51	10	00:00:00	00:00:00	0	AGENT	21:04:50	21:05:59	69	TOVA
AA0101	44787	0	0	PS	990901	21:25:00	21:25:13	13	00:00:00	00:00:00	0	AGENT	21:25:13	21:28:03	170	AVI
AA0101	44788	0	0	PS	990901	21:50:40	21:50:54	14	00:00:00	00:00:00	0	AGENT	21:50:54	21:51:55	61	AVI
AA0101	44789	9103060	2	NE	990901	22:05:40	22:05:46	6	22:05:46	22:09:52	246	AGENT	22:09:51	22:13:41	230	AVI
AA0101	44790	14558621	2	PS	990901	22:24:11	22:24:17	6	22:24:17	22:26:16	119	AGENT	22:26:15	22:27:28	73	VICKY
AA0101	44791	0	0	PS	990901	22:46:27	22:46:37	10	00:00:00	00:00:00	0	AGENT	22:46:36	22:47:03	27	AVI
AA0101	44792	67158097	2	PS	990901	23:05:07	23:05:13	6	23:05:13	23:05:30	17	AGENT	23:05:29	23:06:49	80	VICKY
AA0101	44793	15317126	2	PS	990901	23:28:52	23:28:58	6	23:28:58	23:30:08	70	AGENT	23:30:07	23:35:03	296	DARMON
AA0101	44794	0	0	PS	990902	00:10:47	00:12:05	78	00:00:00	00:00:00	0	HANG	00:00:00	00:00:00	0	NO_SERVER
AA0101	44795	0	0	PS	990902	07:16:52	07:17:01	9	00:00:00	00:00:00	0	AGENT	07:17:01	07:17:44	43	ANAT
AA0101	44796	0	0	PS	990902	07:50:05	07:50:16	11	00:00:00	00:00:00	0	AGENT	07:50:16	07:53:03	167	STEREN

Measurements:
Prevalent Averages (ACD Data)

Command Center Intraday Report

Date
06/13 - Tue

Updated Through: All Day

		Recvd	Answ	Abn %	ASA	AHT	Occ %	On Prod%	On Prod FTE	Sch Open FTE	Sch Avail %
	Total:	**129,960**	**126,321**	**2.8%**	**31**	**318**	**90.9%**	**88.4%**	**1531.7**	**1585.0**	**96.6%**
INQ	Charlotte	20,577	19,860	3.5%	30	307	95.1%	85.4%	222.7	234.6	95.0%
INQ	Columbus MCSC	7,973	7,773	2.5%	36	314	94.9%	89.8%	89.2	94.5	94.4%
INQ	Phoenix	17,102	16,757	2.0%	31	298	92.7%	91.8%	187.3	194.8	96.2%
INQ	Scranton	1,257	1,254	0.2%	6	515	78.6%	28.9%	28.5	35.1	81.2%
INQ	Tampa	9,174	8,859	3.4%	42	366	91.5%	93.6%	123.1	125.9	97.8%
CEN	Bourbonnais	6,070	5,937	2.2%	33	362	86.7%	90.2%	86.0	88.4	97.3%
CEN	Bristol	10,667	10,505	1.5%	25	355	95.1%	93.1%	136.3	139.6	97.6%
CEN	Columbus Claims	5,258	5,153	2.0%	27	293	86.7%	89.8%	60.5	62.2	97.3%
STH	Atlanta	7,514	7,338	2.3%	40	318	82.1%	89.5%	98.6	99.8	98.8%
STH	Sherman	19,669	18,833	4.3%	46	252	93.8%	90.6%	175.5	174.9	100.4%
STH	Wilmington	10,422	9,888	5.1%	21	285	89.9%	92.1%	108.7	114.6	94.8%
WST	Visalia	14,277	14,164	0.8%	10	382	87.2%	85.0%	215.2	220.6	97.6%

12 CC's

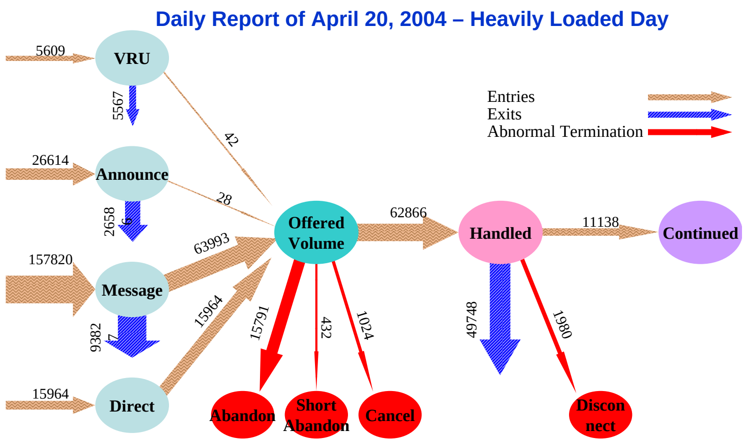
6/13/00 - Tue

[redacted] - Center

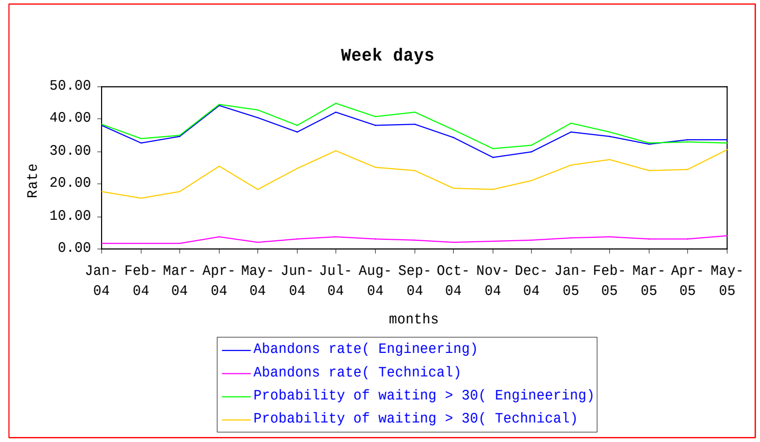
Time	Recvd	Answ	Abn %	ASA	AHT	Occ %	On Prod%	On Prod FTE	Sch Open FTE	Sch Avail %
0	**20,577**	**19,860**	**3.5%**	**30**	**307**	**95.1%**	**85.4%**	**222.7**	**234.6**	**95.0%**
8:00	332	308	7.2%	27	302	87.1%	79.5%	59.3	66.9	88.5%
8:30	653	615	5.8%	58	293	96.1%	81.1%	104.1	111.7	93.2%
9:00	866	796	8.1%	63	308	97.1%	84.7%	140.4	145.3	96.6%
9:30	1,152	1,138	1.2%	28	303	90.8%	81.6%	211.1	221.3	95.4%
10:00	1,330	1,286	3.3%	22	307	98.4%	84.3%	223.1	229.0	97.4%
10:30	1,364	1,338	1.9%	33	296	99.0%	84.1%	222.5	227.9	97.6%
11:00	1,380	1,280	7.2%	34	306	98.2%	84.0%	222.0	223.9	99.2%
11:30	1,272	1,247	2.0%	44	298	94.6%	82.8%	218.0	233.2	93.5%
12:00	1,179	1,177	0.2%	1	306	91.6%	88.6%	218.3	222.5	98.1%
12:30	1,174	1,160	1.2%	10	302	95.5%	93.6%	203.8	209.8	97.1%
13:00	1,018	999	1.9%	9	314	95.4%	91.2%	182.9	187.0	97.8%
13:30	1,061	961	9.4%	67	306	100.0%	88.9%	163.4	182.5	89.5%
14:00	1,173	1,082	7.8%	78	313	99.5%	85.7%	188.9	213.0	88.7%
14:30	1,212	1,179	2.7%	23	304	96.6%	86.0%	206.1	220.9	93.3%
15:00	1,137	1,122	1.3%	15	320	96.9%	83.5%	205.8	222.1	92.7%
15:30	1,169	1,137	2.7%	17	311	97.1%	84.6%	202.2	207.0	97.7%
16:00	1,107	1,059	4.3%	46	315	99.2%	79.4%	187.1	192.9	97.0%
16:30	914	892	2.4%	22	307	95.2%	81.8%	160.0	172.3	92.8%
17:00	615	615	0.0%	2	328	83.0%	93.6%	135.0	146.2	92.3%
17:30	420	420	0.0%	0	328	73.8%	95.4%	103.5	116.1	89.2%
18:00	49	49	0.0%	14	180	84.2%	89.1%	5.8	1.4	416.2%

DataMOCCA

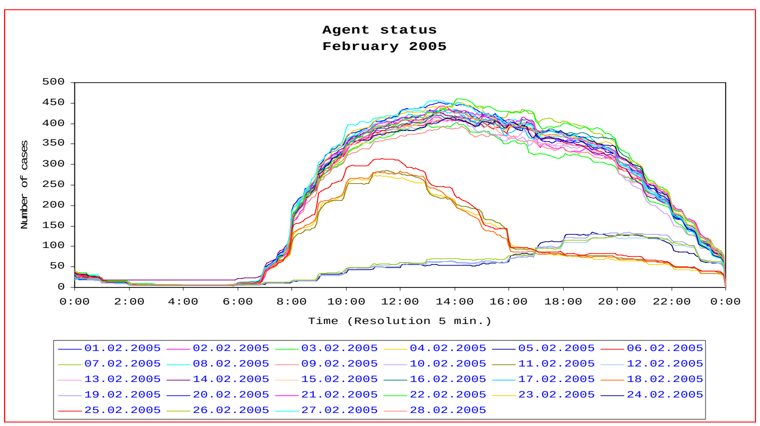
Daily Report



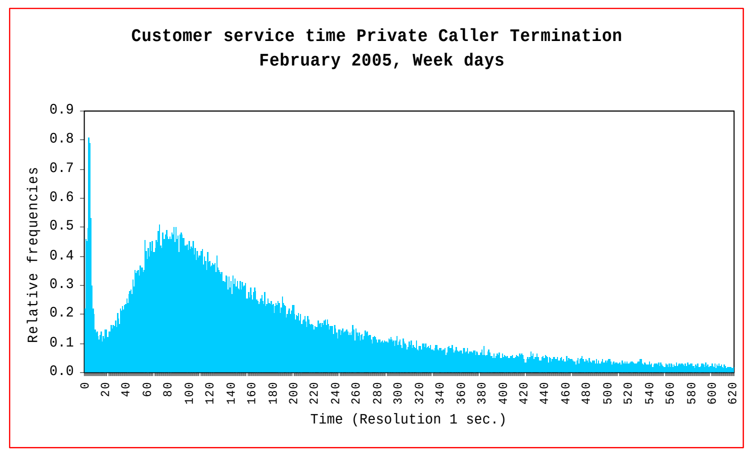
Time Series



Cross Tabulation

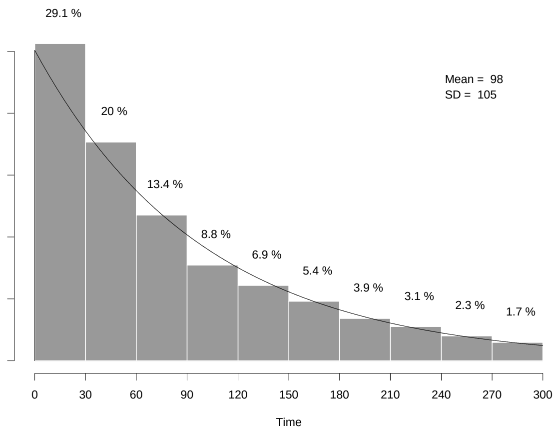


Histogram

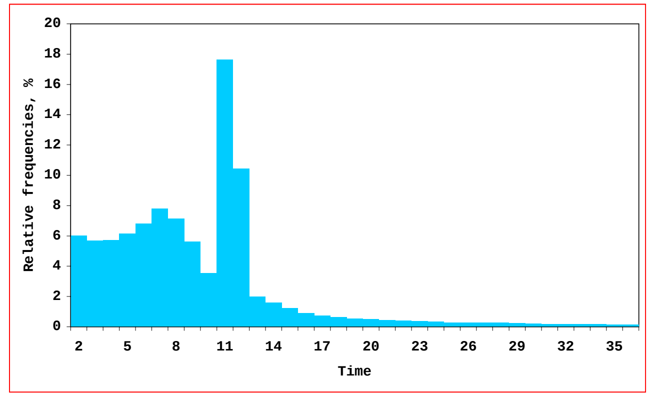


Beyond Averages: Waiting Times in a Call Center

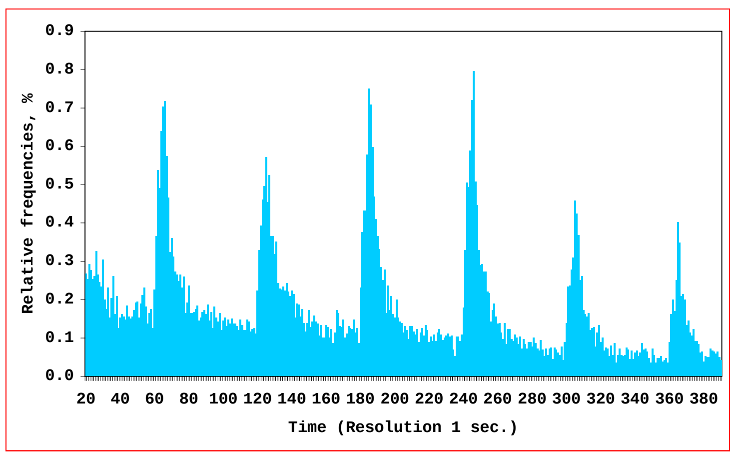
Small Israeli Bank



Large U.S. Bank



Medium Israeli Bank



The Second Prerequisite: (Operational) Models

Empirical Models

- Conceptual
  - Service-Process **Data** = **Flow** Network
  - **Service Networks = Queueing Networks**
- Descriptive
  - QC-Tools: Pareto, Gantt, Fishbone Diagrams,...
  - Histograms, Hazard-Rates, ...
  - Data-MOCCA: Repository + Interface
- Explanatory
  - Nonparametric: Comparative Statistics, Regression,...
  - Parametric: Log-Normal Services, (Doubly) Poisson Arrivals, Exponential (Im)Patience

Analytical Models

- Fluid (Deterministic) Models
- Stochastic Models (Birth & Death, $G/G/n$, Jackson,...)

Conceptual Model: Service Networks = Queueing Networks

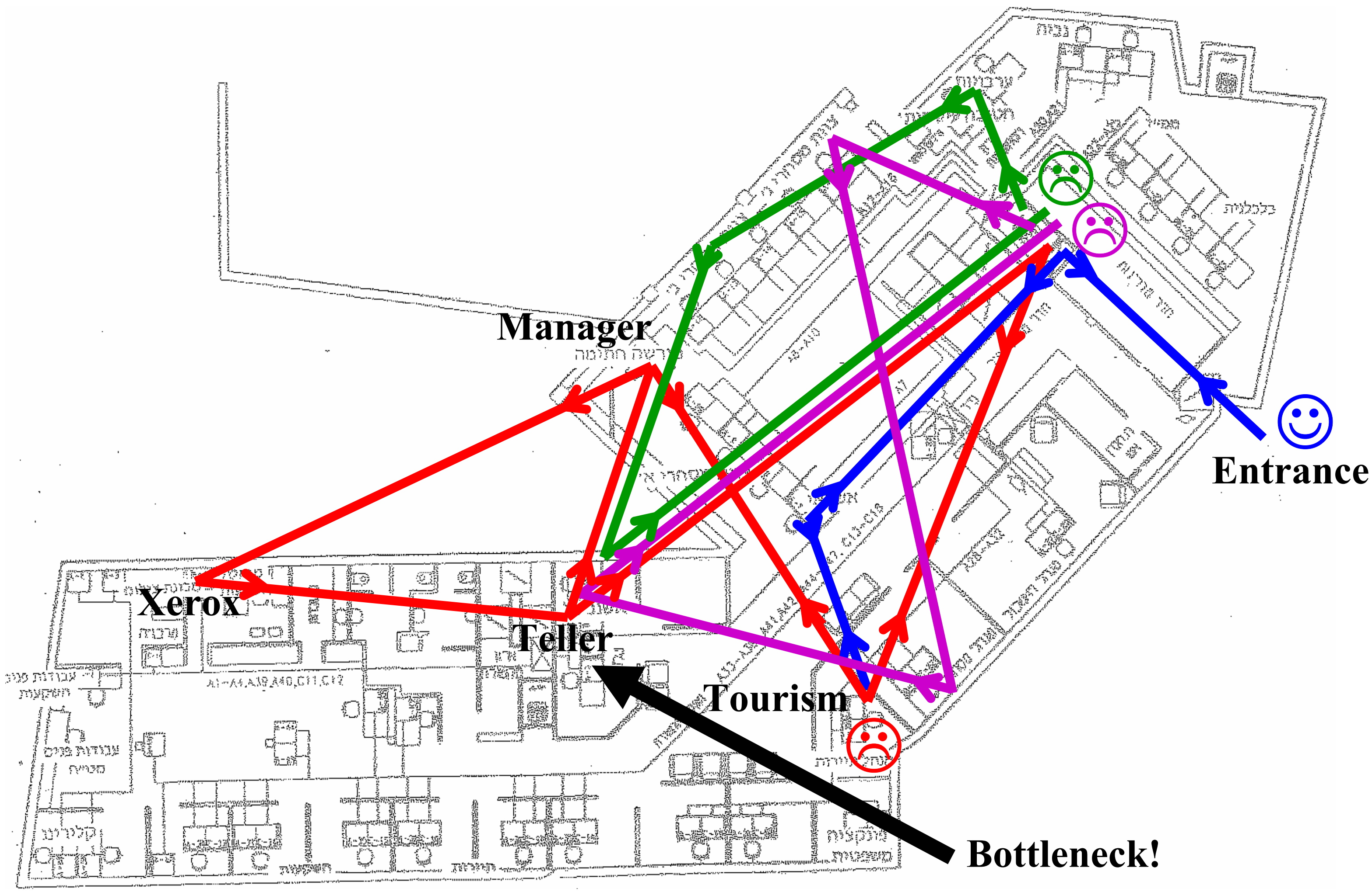
- People, waiting for service: teller, repairman, ATM
- Telephone-calls, to be answered: busy, music, info.
- Forms, to be sent, processed, printed; for a partner
- Projects, to be developed, approved, implemented
- Justice, to be made: pre-trial, hearing, retrial
- Ships, for a pilot, berth, unloading crew
- Patients, for an ambulance, emergency room, operation
- Cars, in rush hour, for parking
- Checks, waiting to be processed, cashed

- Queues Scarce Resources, Synchronization Gaps

  Costly, but here to stay

  - Face-to-face Nets (Chat) (min.)
  - Tele-to-tele Nets (Telephone) (sec.)
  - Administrative Nets (Letter-to-Letter) (days)
  - Fax, e.mail (hours)
  - Face-to-ATM, Tele-to-IVR
  - Mixed Networks (Contact Centers)

Conceptual Model:
Bank Branch = Queueing Network



Bank Branch: A Queuing Network

Transition Frequencies Between Units in The Private and Business Sections:

	To Unit / From Unit	Private Banking				Business				
		Bankers	Authorized Personal	Compens - - ations	Tellers	Tellers	Overdrafts	Authorized Personal	Full Service	Exit
Private Banking	Bankers		1%	1%	4%	4%	0%	0%	0%	90%
	Authorized Personal	12%		5%	4%	6%	0%	0%	0%	73%
	Compensations	7%	4%		18%	6%	0%	0%	1%	64%
	Tellers	6%	0%	1%		1%	0%	0%	0%	90%
Services	Tellers	1%	0%	0%	0%		1%	0%	2%	94%
	Overdrafts	2%	0%	1%	1%	19%		5%	8%	64%
	Authorized Personal	2%	1%	0%	1%	11%	5%		11%	69%
	Full Service	1%	0%	0%	0%	8%	1%	2%		88%
	Entrance	13%	0%	3%	10%	58%	2%	0%	14%	0%

Legend:

0%-5%	5%-10%	10%-15%	>15%

Dominant Paths - Business:

Unit Parameter	Station 1 Tourism	Station 2 Teller	Total Dominant Path
Service Time	12.7	4.8	17.5
Waiting Time	8.2	6.9	15.1
Total Time	20.9	11.7	32.6
Service Index	**0.61**	**0.41**	**0.53**

Dominant Paths - Private:

Unit Parameter	Station 1 Banker	Station 2 Teller	Total Dominant Path
Service Time	12.1	3.9	16.0
Waiting Time	6.5	5.7	12.2
Total Time	18.6	9.6	28.2
Service Index	**0.65**	**0.40**	**0.56**

Service Index = % time being served

Mapping the Offered Load (Bank Branch)

Department / Time	Business Services		Private Banking	Banking Services	
	Tourism	Teller	Teller	Teller	Comprehensive
8:30 – 9:00	Not Busy	Very Busy	Very Busy	Very Busy	Very Busy
9:00 – 9:30	Not Busy	Busy	Not Busy	Not Busy	Not Busy
9:30 – 10:00	Very Busy	Busy	Very Busy	Very Busy	Busy
10:00 – 10:30	Very Busy	Very Busy	Very Busy	Busy	Busy
10:30 – 11:00	Very Busy	Very Busy	Busy	Busy	Not Busy
11:00 – 11:30	Not Busy	Very Busy	Busy	Not Busy	Not Busy
11:30 – 12:00	Not Busy	Busy	Not Busy	Not Busy	Not Busy
12:00 – 12:30	Not Busy	Not Busy	Not Busy	Not Busy	Not Busy
Break					
16:00 – 16:30	Very Busy	Not Busy	Not Busy	Very Busy	Very Busy
16:30 – 17:00	Very Busy	Not Busy	Not Busy	Busy	Not Busy
17:00 – 17:30	Very Busy	Not Busy	Not Busy	Busy	Busy
17:30 – 18:00	Very Busy	Not Busy	Not Busy	Not Busy	Not Busy

Legend:

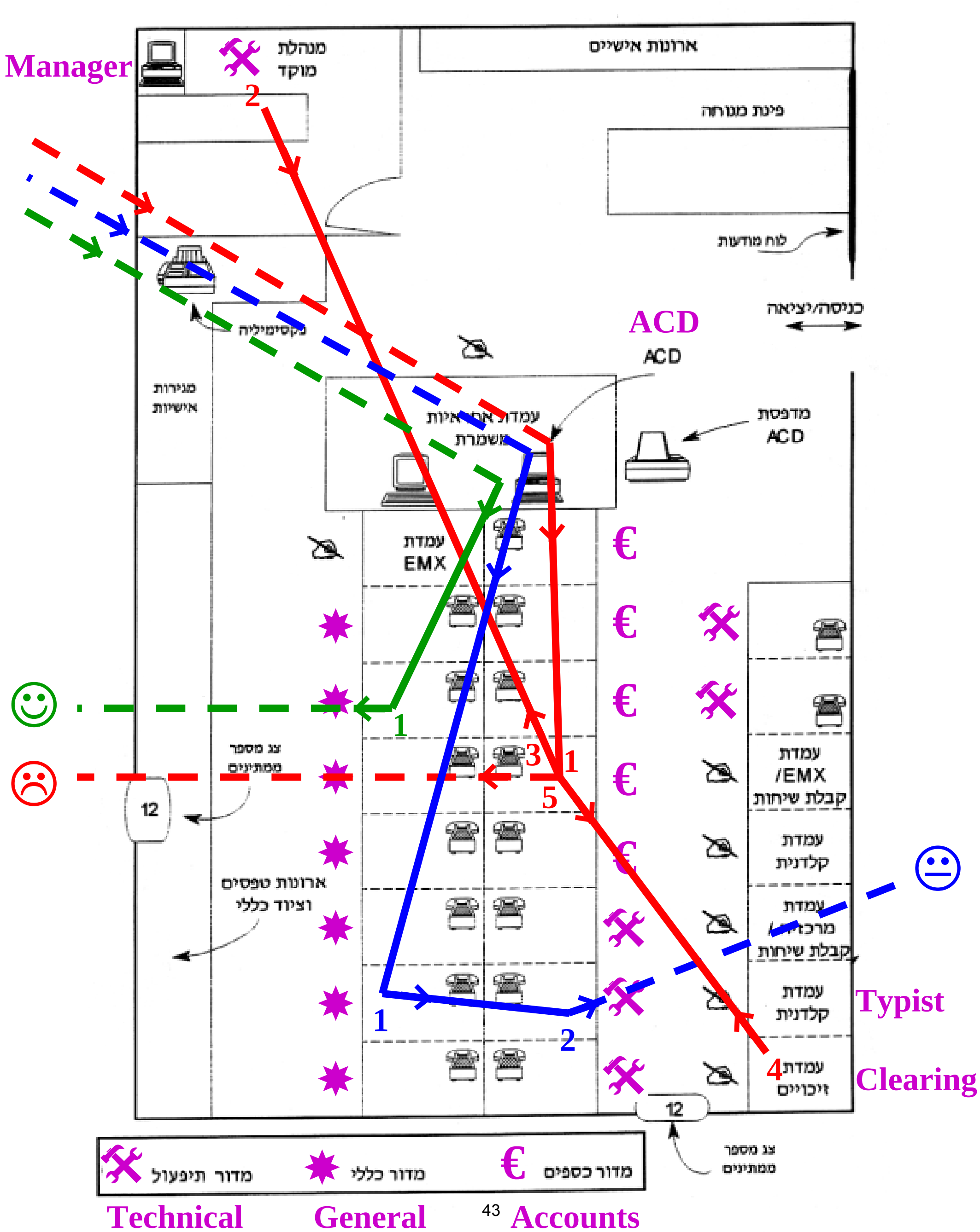
- Not Busy
- Busy
- Very Busy

Note: What can / should be done at 11:00 ?

Conclusion: Models are not always necessary but measurements are !

Conceptual Model: Call-Center Network

Schematic Chart – Pelephone Call-Center 1994
= Tele Net = Queueing Network



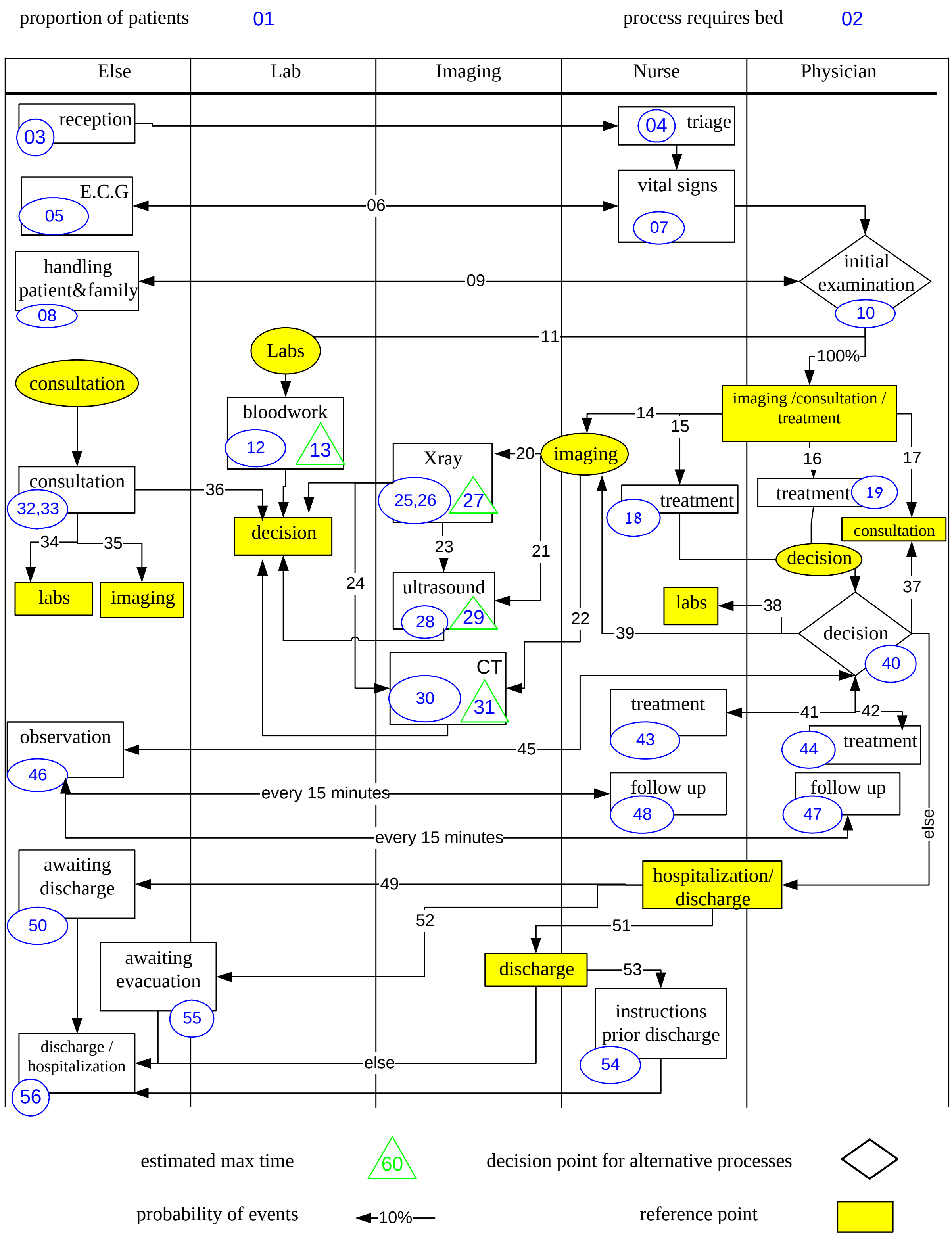
Conceptual Model: Call-Center Network

Current Status - Analysis

	Accounts Center	General Center	Technical Center
Peak days in a week	Sun, Fri	Sun	Sun
Peak days in a month	12	8-14, 2-3	10-20
Avg. applications no. in a day	4136	2476	1762
Avg. applications no. in an hour - λ_{avg}	**253.6**	**193**	**167**
Peak hours in a day	**11:00-12:00**	**10:00-11:00**	**9:00-10:00**
Avg. applications no. in peak hours - λ_{max}	**422**	**313**	**230**
Avg. waiting time (secs.)	10.9	20.0	55.9
Avg. service time (secs.)	83.5	131.3	143.2
Service index	0.88	0.87	0.72
Abandonment percentage	2.7	5.6	11.2
Avg. waiting time before abandonment (secs.)	9.7	16.8	43.2
Avg. staffing level	9.7	10.3	5.2
Target waiting time	12	25	-

Conceptual Model: Hospital Network

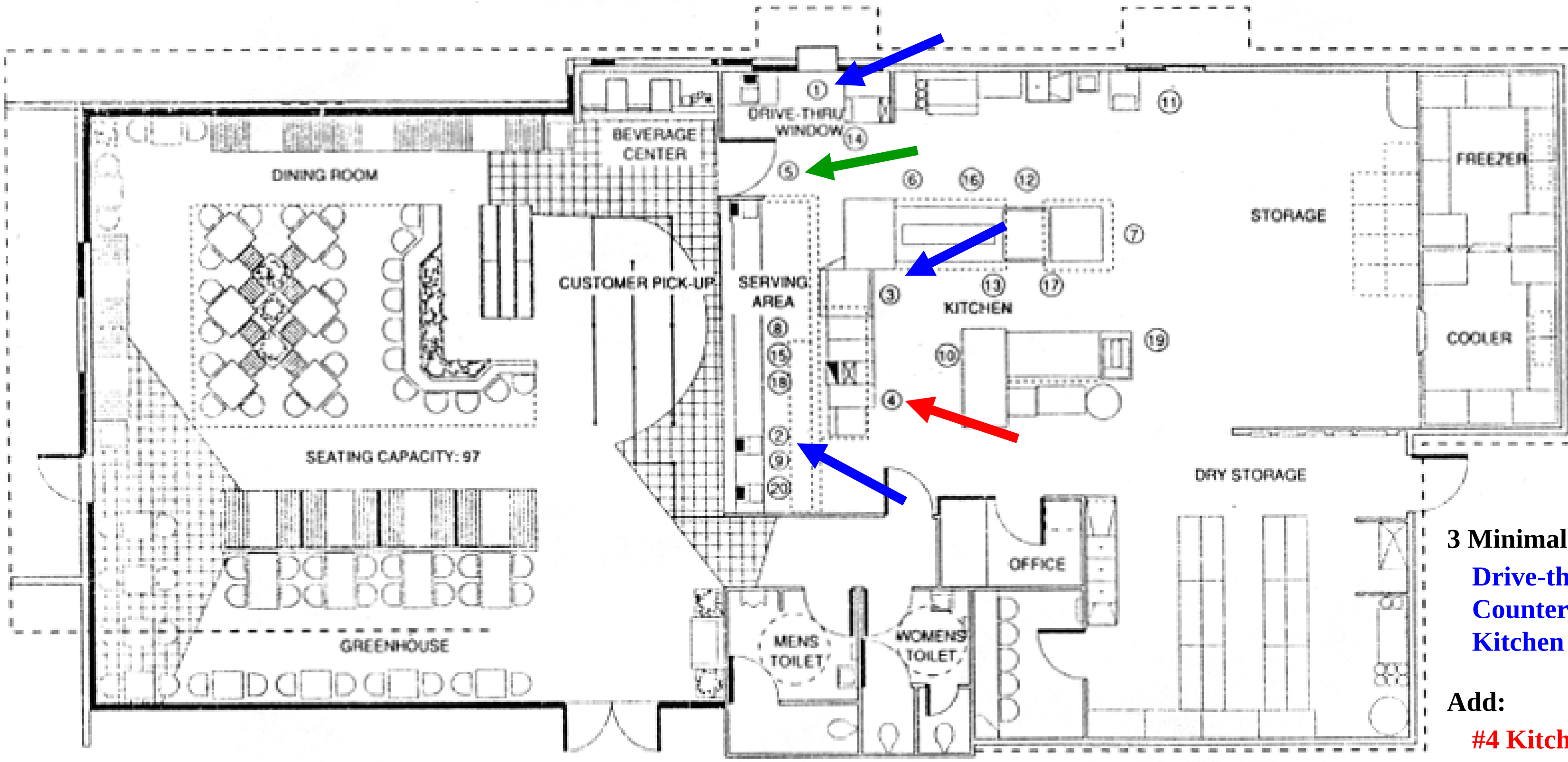
Emergency Department: Generic Flow



Conceptual Model: Burger King Bottlenecks

Bottleneck Analysis: **Short – Run Approximations**
Time – State Dependent Q-Net

TOUR F / A WORKER-PACED LINE FLOW PROCESS AND A SERVICE FACTORY 155

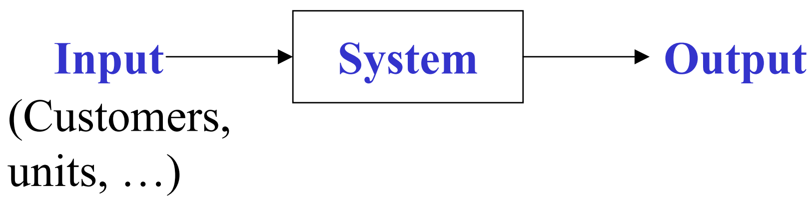


3 Minimal:
- **Drive-thru**
- **Counter**
- **Kitchen**

Add:
- **#4 Kitchen**
- **#5 Help Drive -thru**

FIGURE F1 Layout of the Noblesville Burger King. The circled numbers indicate the sequence of additions of workers to the kitchen as demand increases.

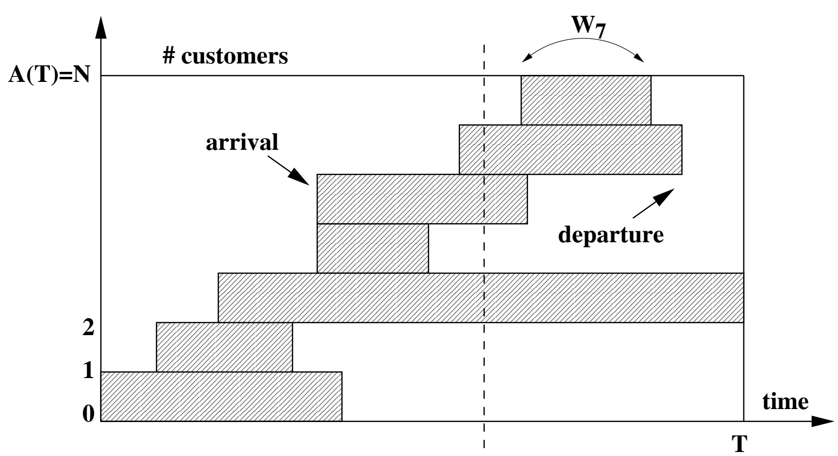
Analytical Models: Little's Law, or The First Law of Congestion



- $\boldsymbol{\lambda}$ = average arrival rate;
- $\boldsymbol{L}$ = average **number** within system;
- $\boldsymbol{W}$ = average **time** within system.

Little's Law $\quad \boxed{L = \lambda W}$

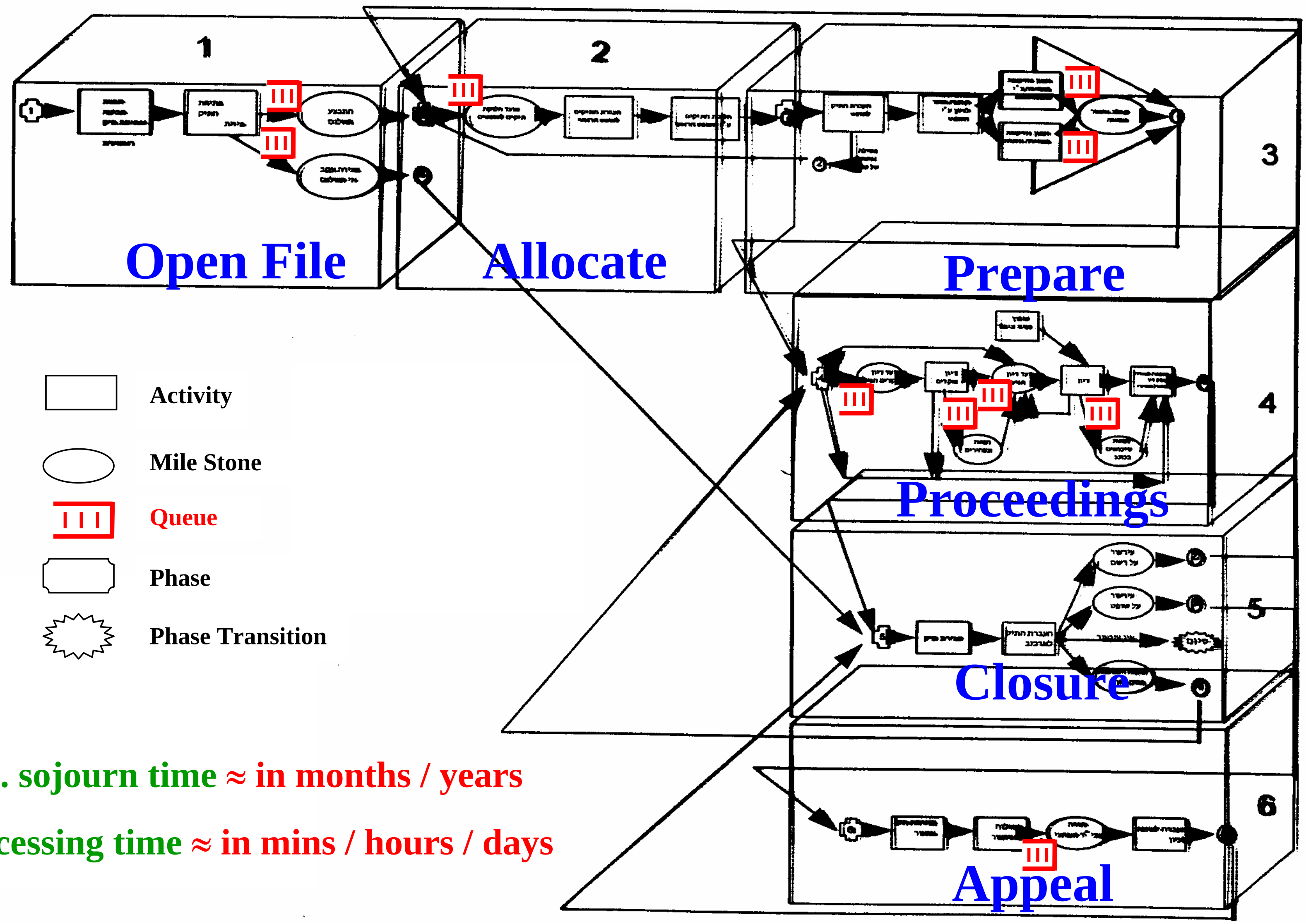
Finite-Horizon Version



Long-Run (Stochastic) Example

M/M/1: $L = \dfrac{\rho}{1-\rho} = \dfrac{\lambda}{\mu - \lambda}, \quad W = \dfrac{1}{\mu - \lambda} = \dfrac{1}{\mu}\dfrac{1}{1-\rho}.$

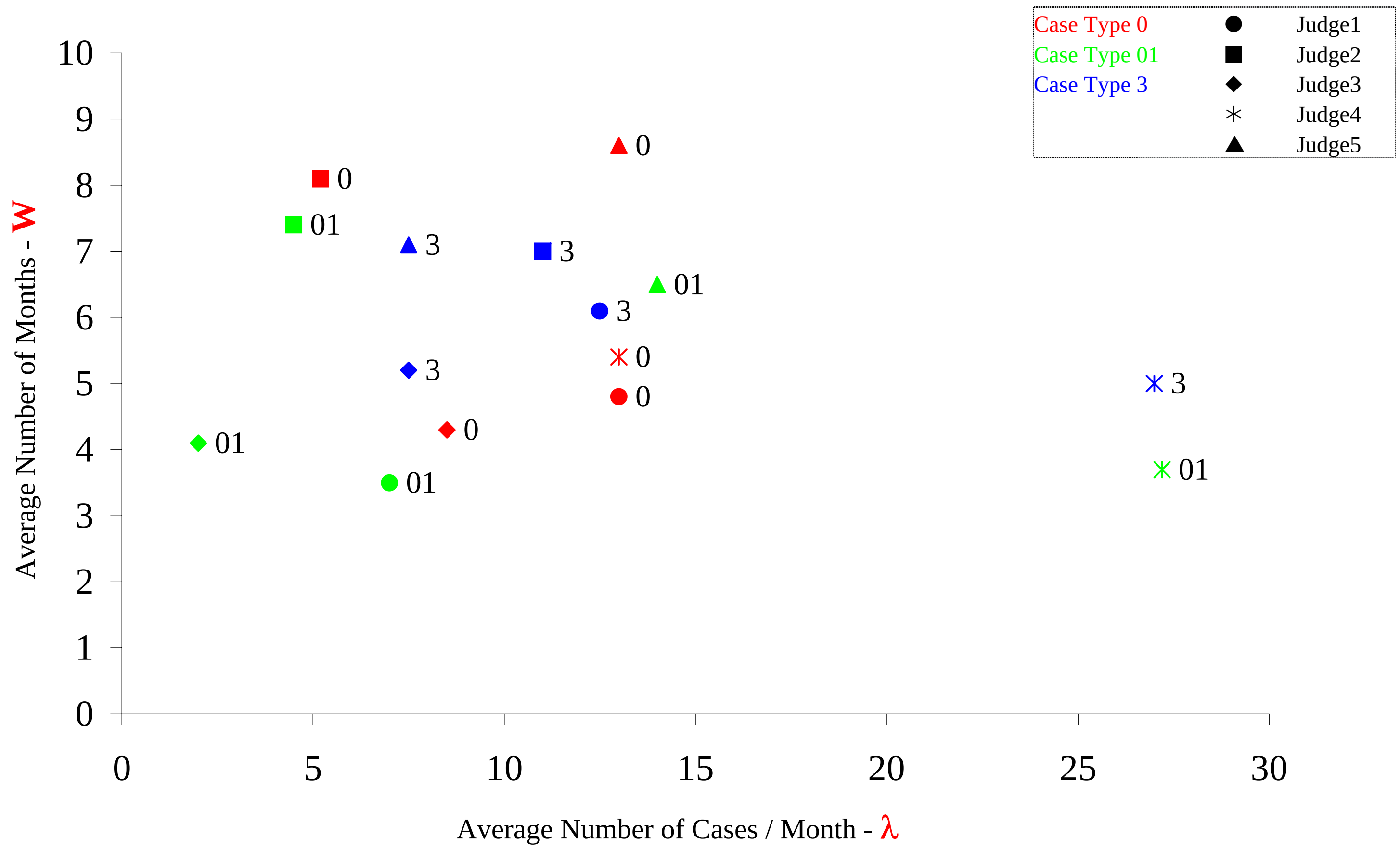
Conceptual Model: The Justice Network, or The Production of Justice



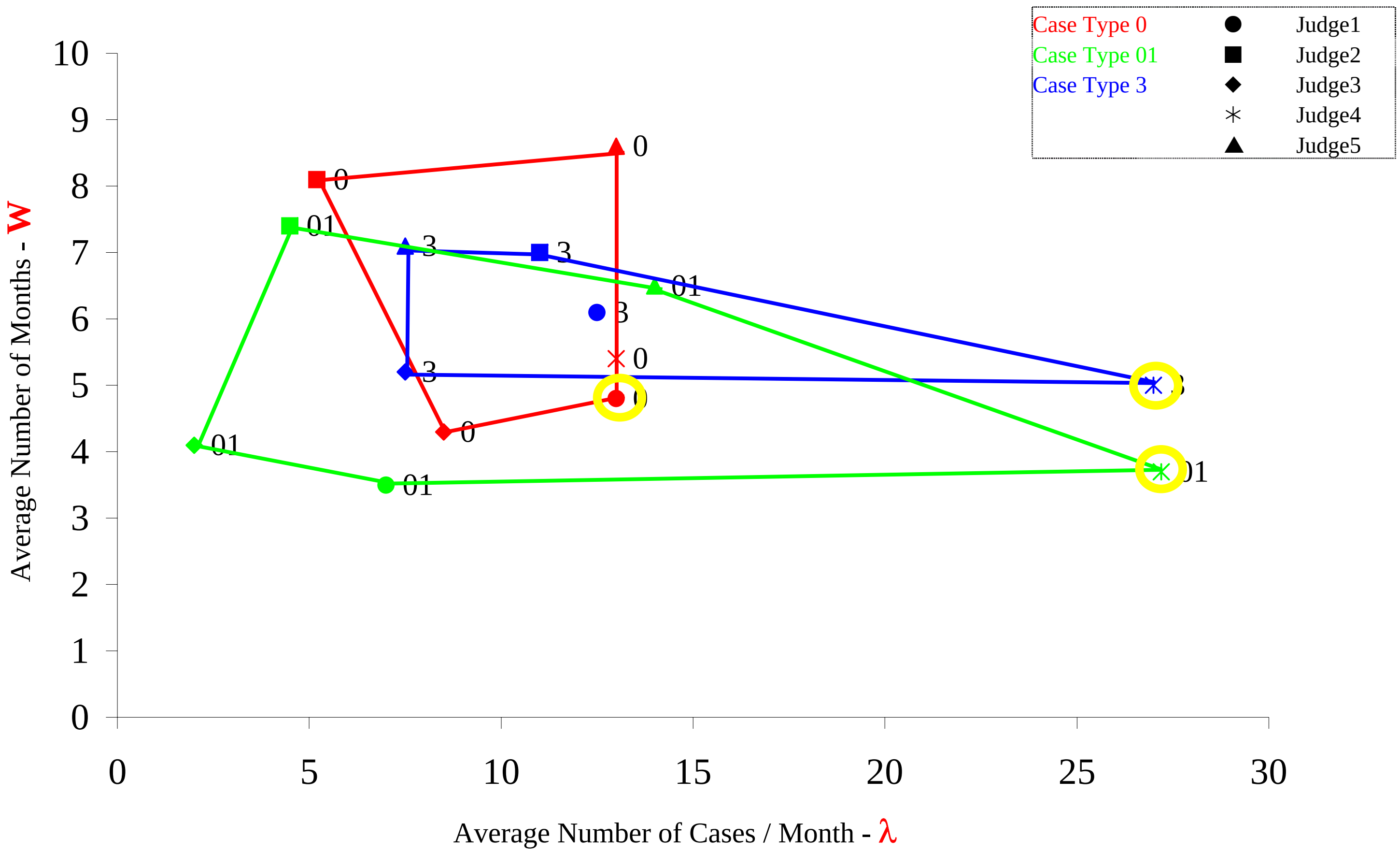
Avg. sojourn time ≈ in months / years

Processing time ≈ in mins / hours / days

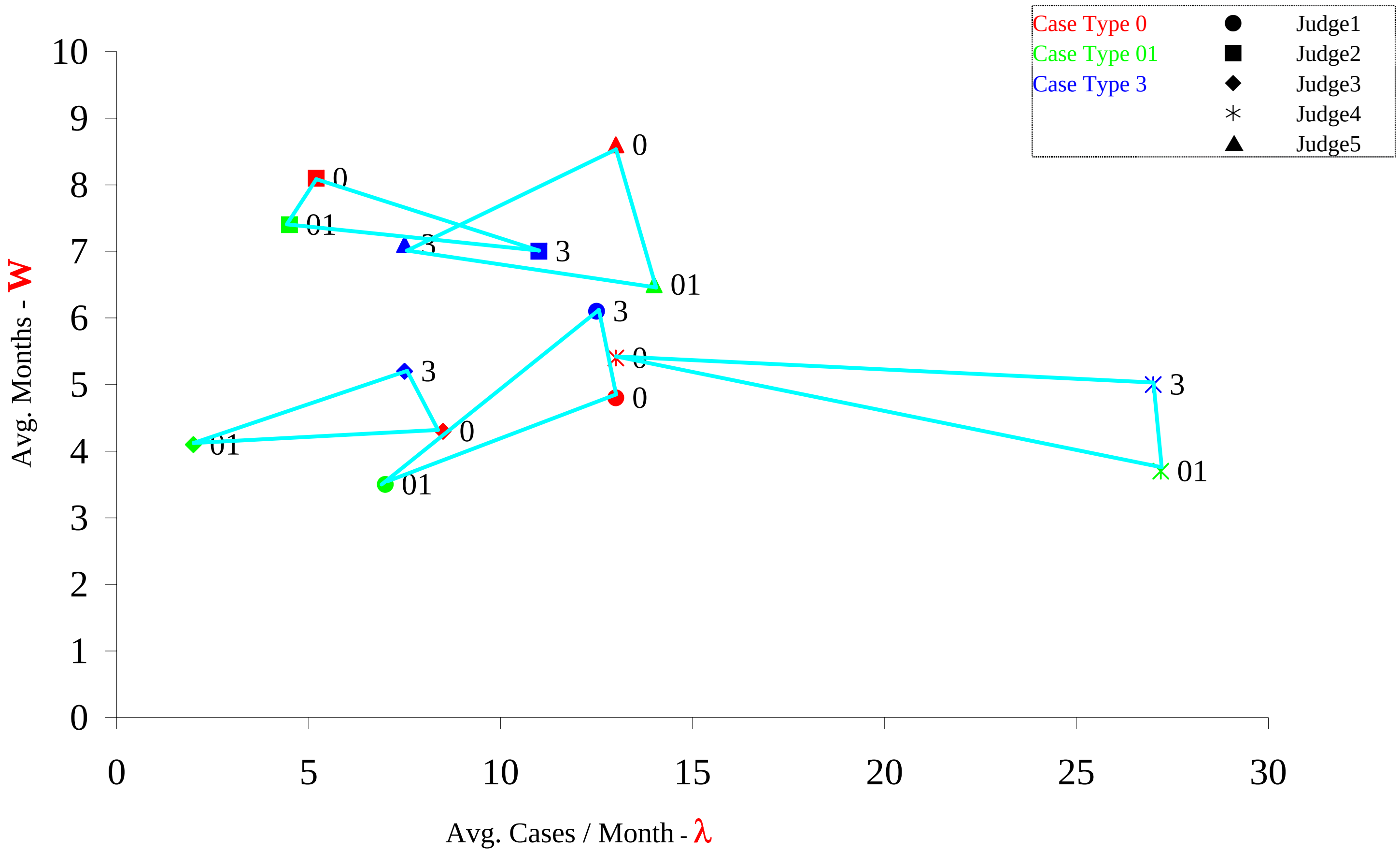
Judges: Operational Performance - Base case



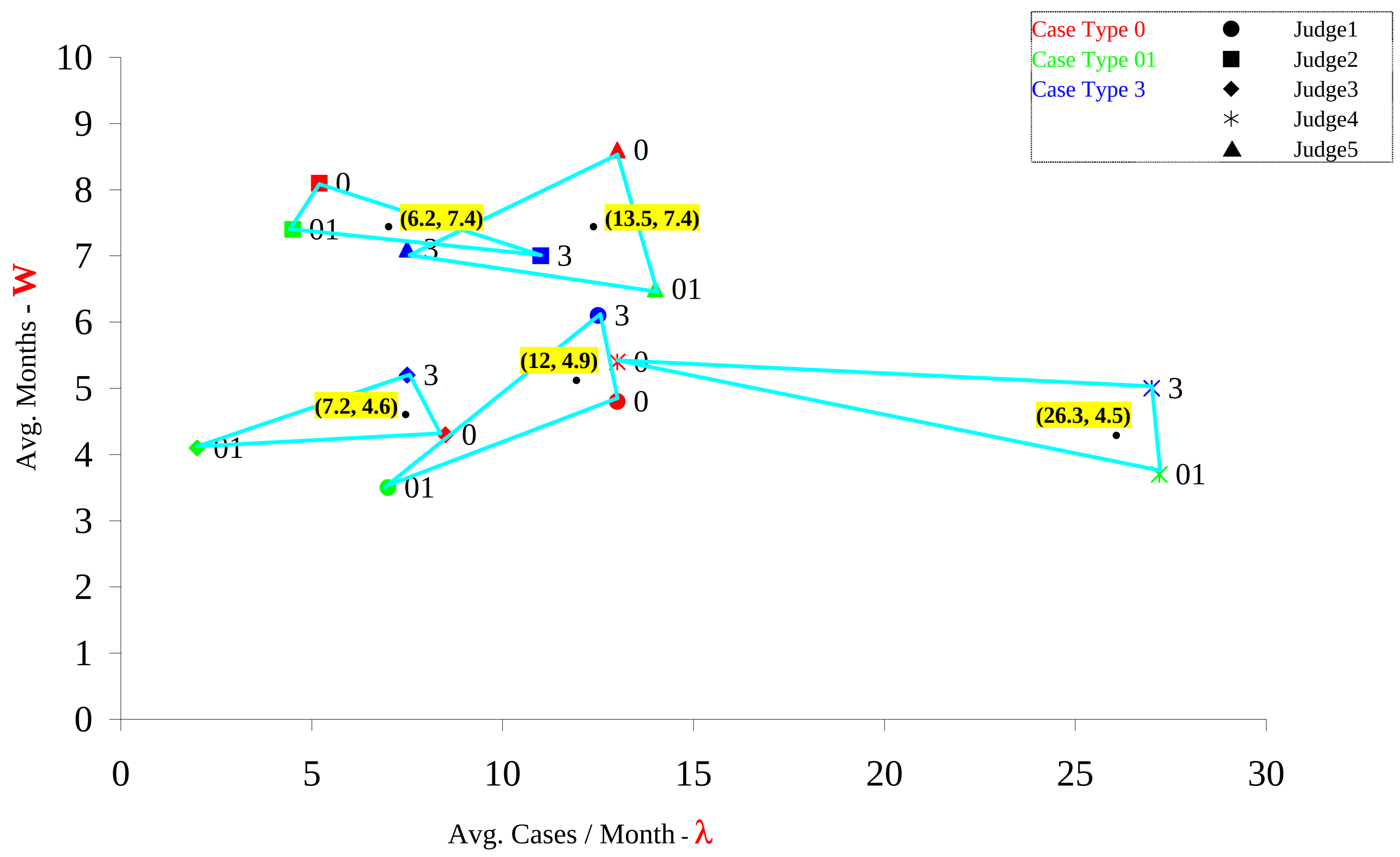
3 Case-Types: Performance by 5 Judges



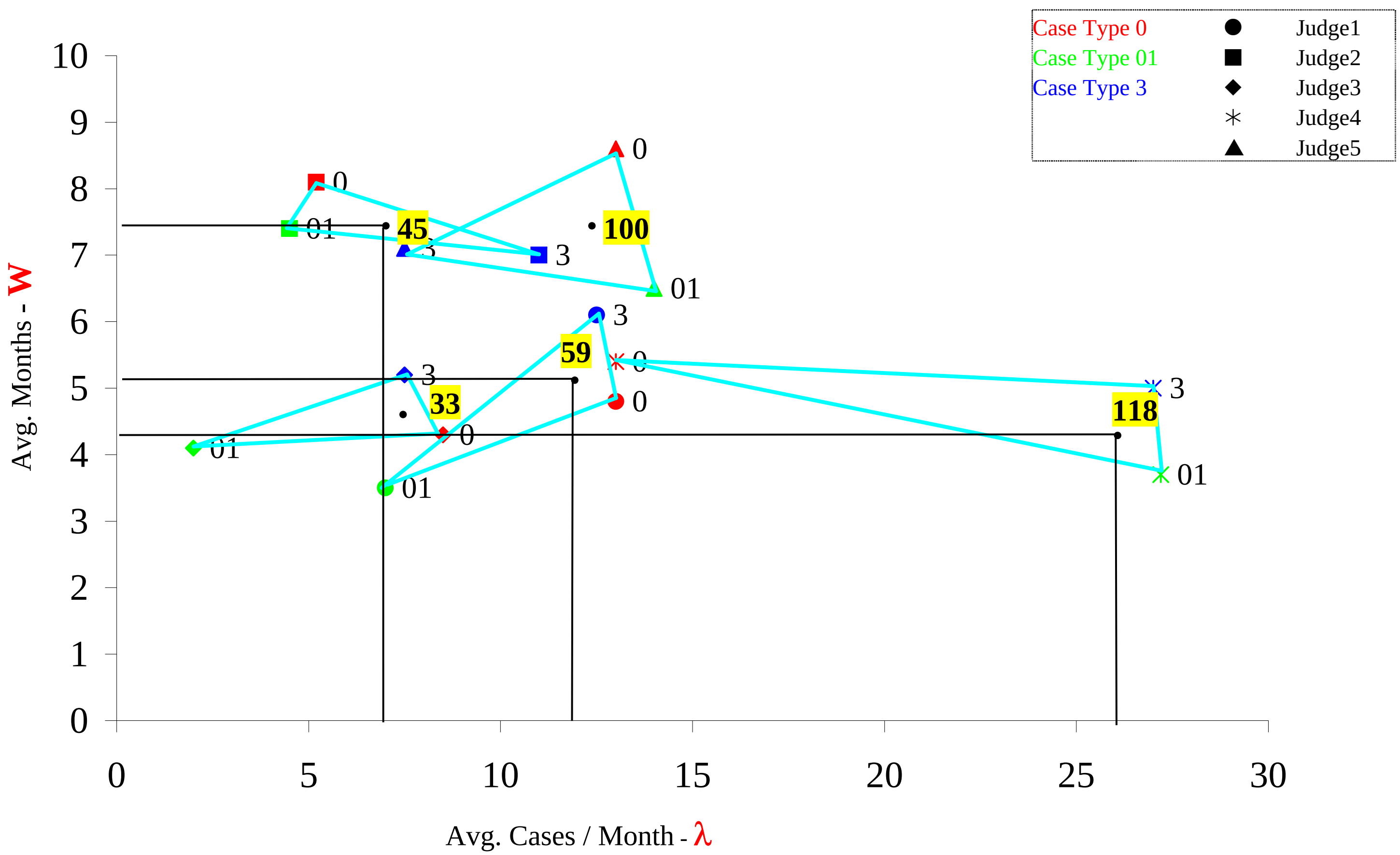
5 Judges: Performance by 3 Case-Types



Judges: Performance Analysis



Judges: Best/Worst Performance



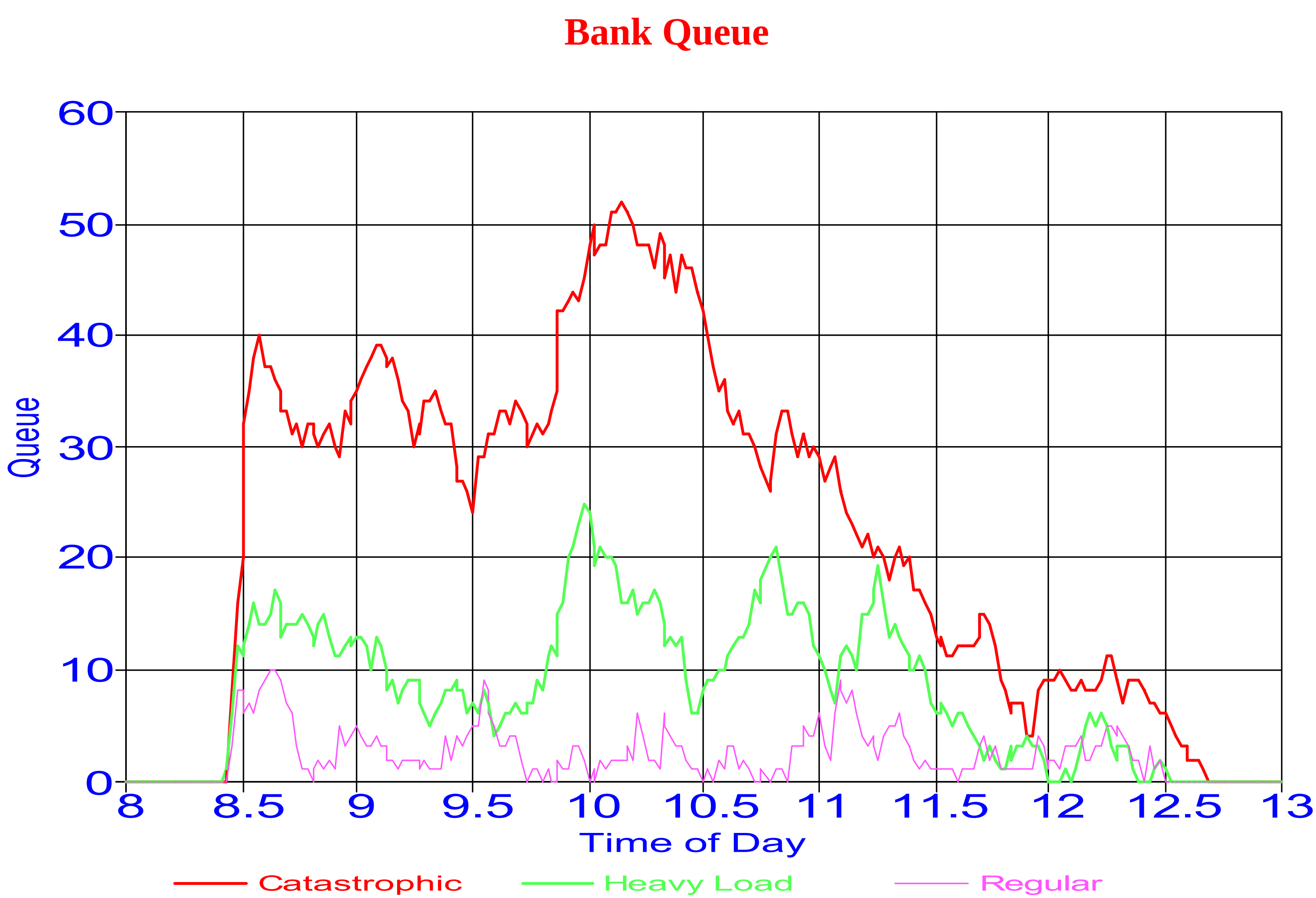
Conceptual Fluid Model

Customers/units are modeled by **fluid (continuous) flow**.

Labor-day Queueing at Niagara Falls

- Appropriate when **predictable variability** prevalent;
- Useful **first-order** models/approximations, often **suffice**;
- Rigorously justifiable via Functional Strong Laws of Large Numbers.

Empirical Fluid Model: Queue-Length at a Catastrophic/Heavy/Regular Day

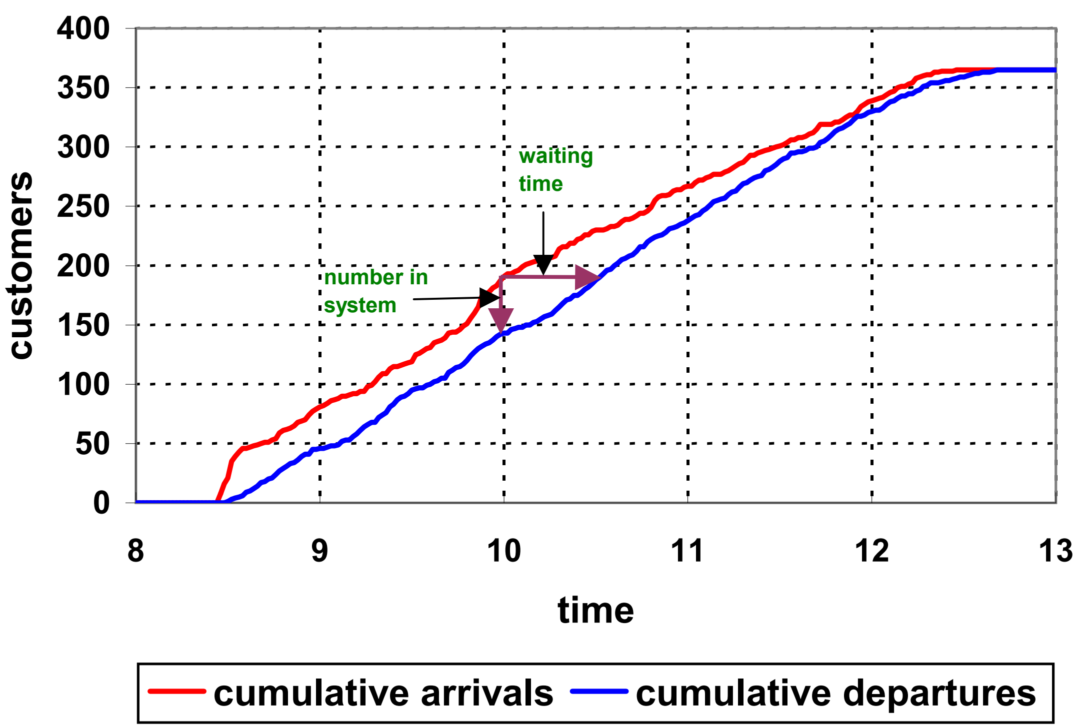


Empirical Models: Fluid, Flow

Derived directly from event-based (call-by-call) measurements. For example, an isolated service-station:

- $A(t)$ = **cumulative** # arrivals from time 0 to time t;
- $D(t)$ = **cumulative** # departures from system during $[0, t]$;
- $L(t) = A(T) - D(t)$ = # customers in system at t.

Arrivals and Departures from a Bank Branch
Face-to-Face Service



When is it possible to calculate waiting time in this way?

Mathematical Fluid Models

Differential Equations:

- $\boldsymbol{\lambda(t)}$ – **arrival rate** at time $t \in [0, T]$.
- $\boldsymbol{c(t)}$ – **maximal potential processing rate**.
- $\delta(t)$ – **effective** processing (departure) rate.
- $Q(t)$ – **total** amount in the system.

Then $Q(t)$ is **a** solution of

$$\dot{Q}(t) = \lambda(t) - \delta(t); \quad Q(0) = q_0, \quad t \in [0, T] \,.$$

In a Call Center Setting (no abandonment)

$\boldsymbol{N(t)}$ statistically-identical servers, each with service rate $\boldsymbol{\mu}$.
$\boldsymbol{c(t) = \mu N(t)}$: maximal potential processing rate.
$\delta(t) = \mu \cdot \min(N(t), Q(t))$: processing rate.

$$\dot{Q}(t) = \lambda(t) - \mu \cdot \min(N(t), Q(t)), \quad Q(0) = q_0, \quad t \in [0, T] \,.$$

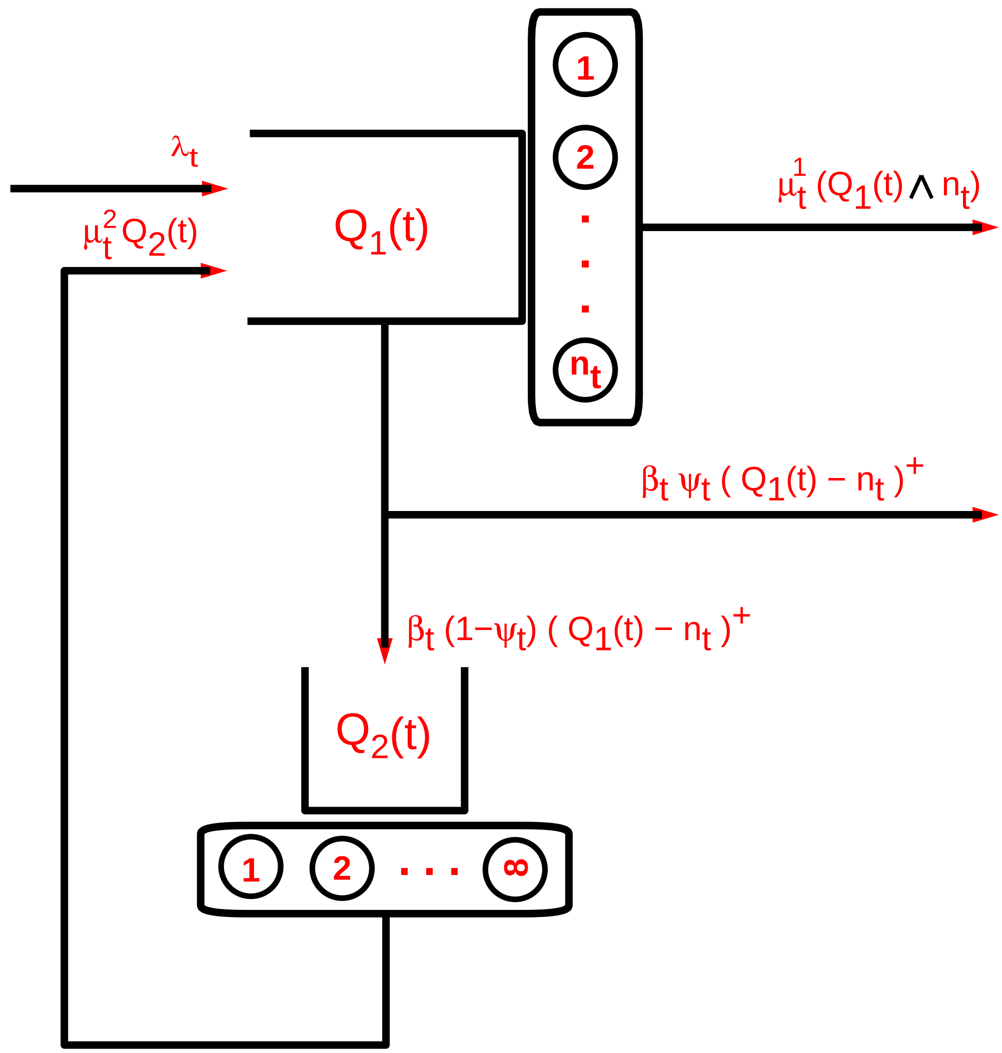
How to actually solve? Mathematics (theory, numerical), or simply: Start with $t_0 = 0$, $Q(t_0) = q_0$.
Then, for $t_n = t_{n-1} + \Delta t$:

$$Q(t_n) \;=\; Q(t_{n-1}) + \lambda(t_{n-1}) \cdot \Delta t - \mu \min(N(t_{n-1}), Q(t_{n-1})) \cdot \Delta t \,.$$

Time-Varying Queues with Abandonment and Retrials

Based on three paper with Massey, Reiman, Rider and Stolyar.

Call Center: a Multiserver Queue with Abandonment and Retrials



Primitives: Time-Varying Predictability

$\boldsymbol{\lambda_t}$ exogenous arrival rate;
e.g., continuously changing, sudden peak.

$\boldsymbol{\mu_t^1}$ service rate;
e.g., change in nature of work or fatigue.

$\boldsymbol{n_t}$ number of servers;
e.g., in response to predictably varying workload.

$\boldsymbol{Q_1(t)}$ number of customers in call center (queue+service).

$\boldsymbol{\beta_t}$ abandonment rate while waiting;
e.g., in response to IVR discouragement at predictable overloading.

$\boldsymbol{\psi_t}$ probability of no retrial.

$\boldsymbol{\mu_t^2}$ retrial rate;
if constant, $1/\mu^2$ – average time to retry.

$\boldsymbol{Q_2(t)}$ number of customers that will retry.

In our examples, we vary λ_t only, other primitives are constant.

Fluid Model

Replacing random processes by their rates yields

$$Q^{(0)}(t) = (Q_1^{(0)}(t),\ Q_2^{(0)}(t))$$

Solution to nonlinear differential balance equations

$$\begin{aligned} \frac{d}{dt} Q_1^{(0)}(t) &= \lambda_t - \mu_t^1 \left(Q_1^{(0)}(t) \wedge n_t\right) \\ &\quad + \mu_t^2 Q_2^{(0)}(t) - \beta_t \left(Q_1^{(0)}(t) - n_t\right)^+ \\ \frac{d}{dt} Q_2^{(0)}(t) &= \beta_1 (1 - \psi_t)(Q_1^{(0)}(t) - n_t)^+ \\ &\quad - \mu_t^2 Q_2^{(0)}(t) \end{aligned}$$

Justification: **Functional Strong Law of Large Numbers**,
with $\lambda_t \to \eta\lambda_t,\ n_t \to \eta n_t$.

As $\eta \uparrow \infty$,

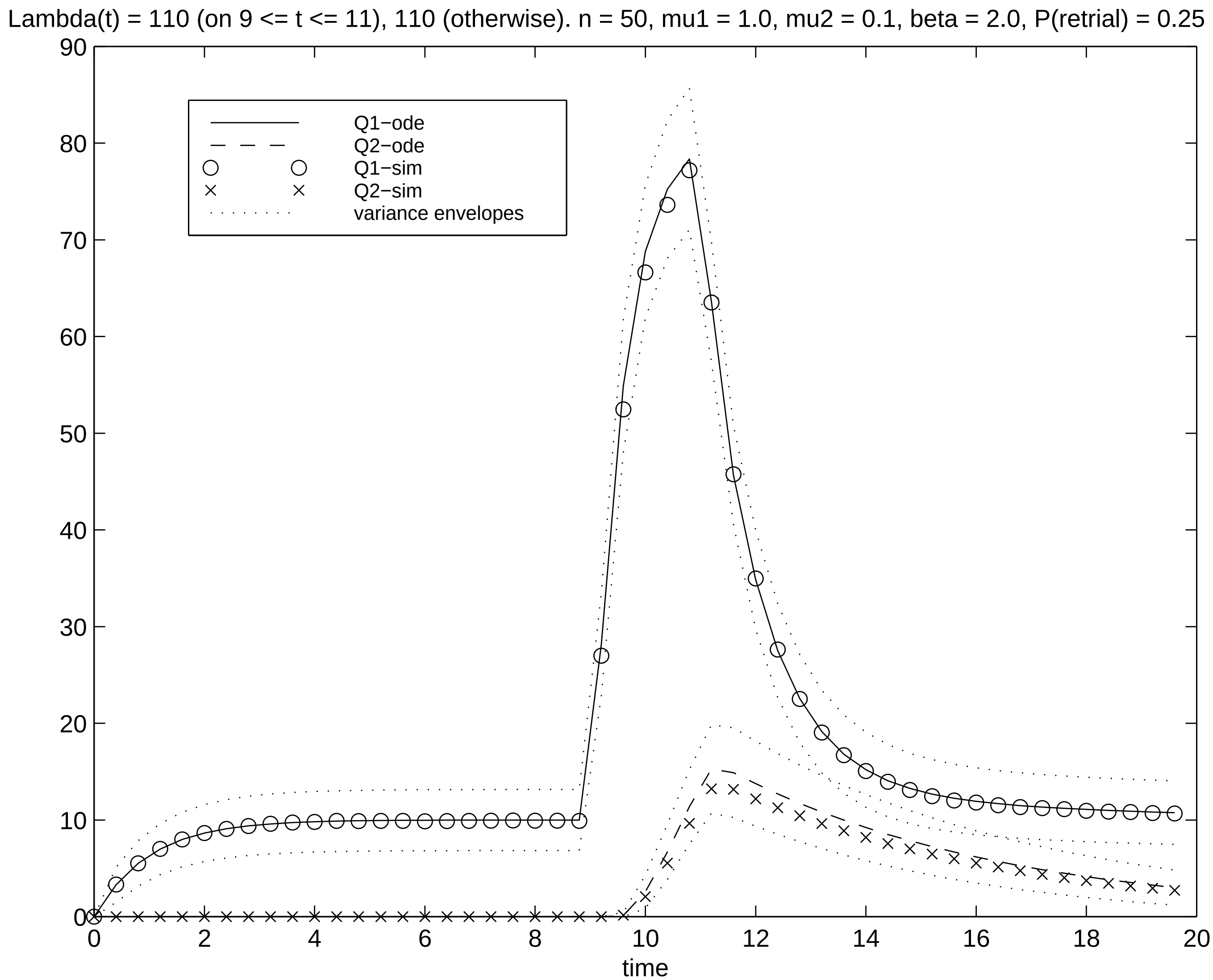
$$\frac{1}{\eta} Q^\eta(t) \to Q^{(0)}(t)\,, \quad \text{uniformly on compacts, a.s.}$$

given convergence at $t = 0$

Sudden Rush Hour

$$n \quad = \quad \text{50 servers;} \qquad \mu = 1$$

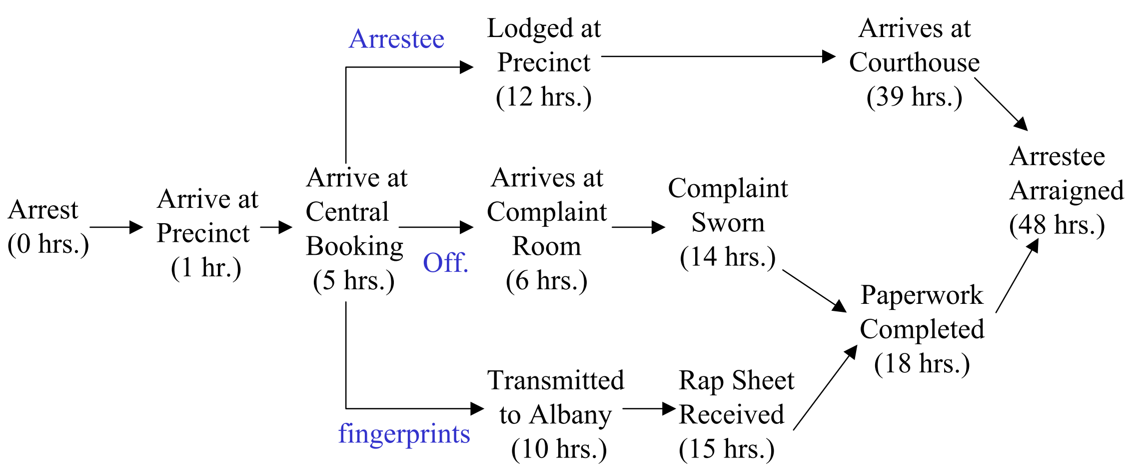
$$\lambda_t \quad = \quad \mathbf{110} \quad \text{for } 9 \leq t \leq 11, \quad \lambda_t = \mathbf{10} \quad \text{otherwise}$$



Stochastic Framework: DS PERT/CPM

DS = Dynamic Stochastic (Fork-Join, Split-Match)
PERT = **P**rogram **E**valuation and **R**eview **T**echnique
CPM = **C**ritical **P**ath **M**ethod
Operations Research in Project Management: Standard Successful.

New-York Arrest-to-Arraignment System
(Larson et al., 1993)



CRM – task times are deterministic/averages (standard).
S-PERT (**S**tochastic PERT) – task times random variables.
DS-PERT/CPM – multi-project (dynamic) environment, with tasks processed at dedicated service stations.

- **Capacity analysis:** Can we do it? (LP)
- **Response-time** analysis: How long will it take? (S-Nets)
- **What if**: Can we do better? (Sensitivity, Parametric)
- **Optimality**: What is the best one can do?

Stochastic Model of a Basic Service Station

Building blocks:

- Arrivals
- Service durations (times)
- Customers' (im)patience.
- Customers' returns (during service process, after service)

First study these building blocks one-by-one:

- Empirical analysis, which motivates
- Theoretical model(s).

Then integrate building blocks, via protocols, into (Basic) Models:

- Erlang-B/C (Arrivals, Services)
- Erlang-A (+ Abandonment), Erlang-R (+ Returns).

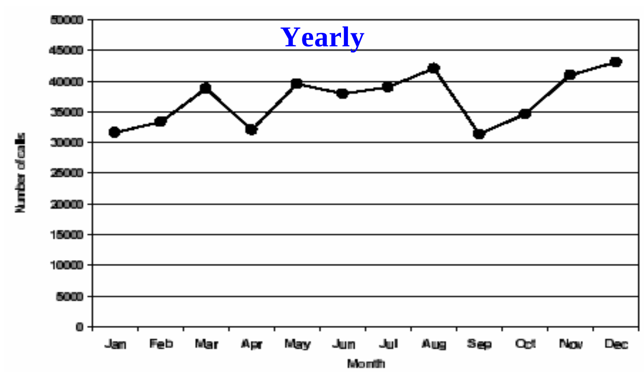
The models support, for example,

- Staffing Workforce, for which Basic Models are already useful; and beyond:
- Routing Customers
- Scheduling Servers
- Matching Customers-Needs with Servers-Skills (SBR).

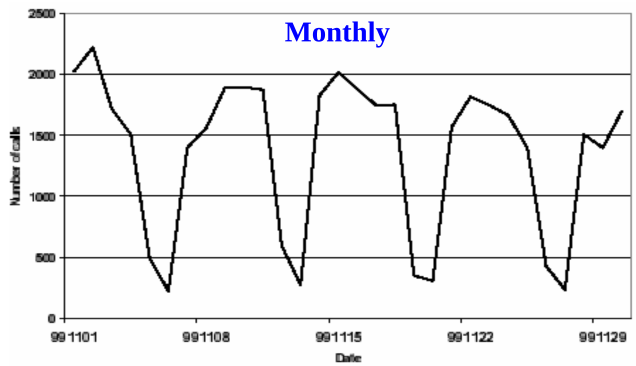
Arrivals to Service

Arrivals to a Call Center (1999): Time Scale

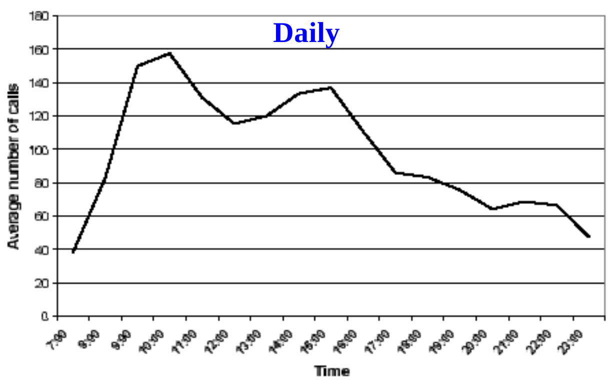
Strategic



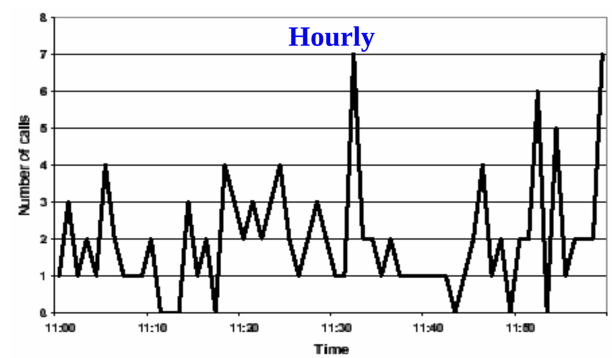
Tactical



Operational



Stochastic



Arrivals Process, in 1976

(E. S. Buffa, M. J. Cosgrove, and B. J. Luce,
"An Integrated Work Shift Scheduling System")

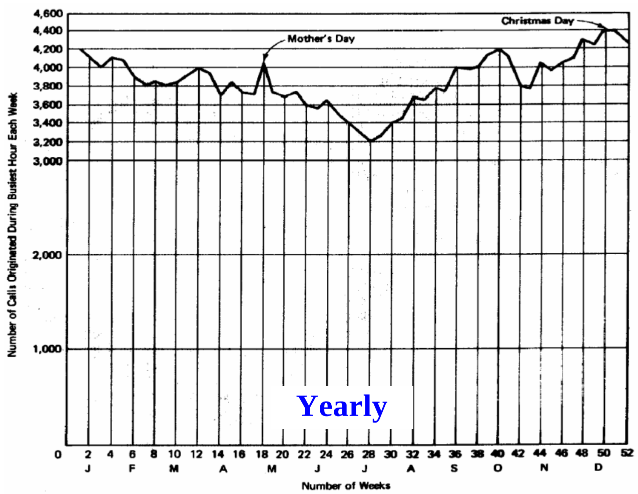


Figure 1 Typical distribution of calls during the busiest hour for each week during a year.

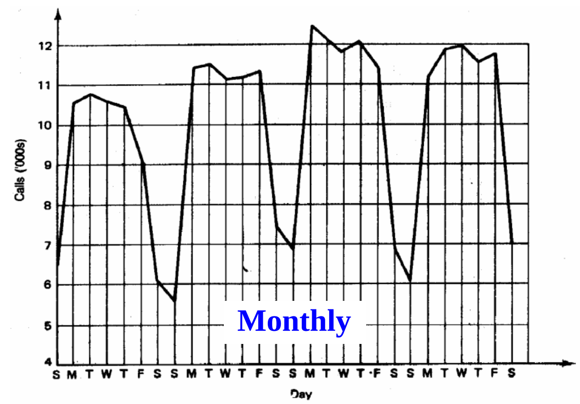


Figure 2 Daily call load for Long Beach, January 1972.

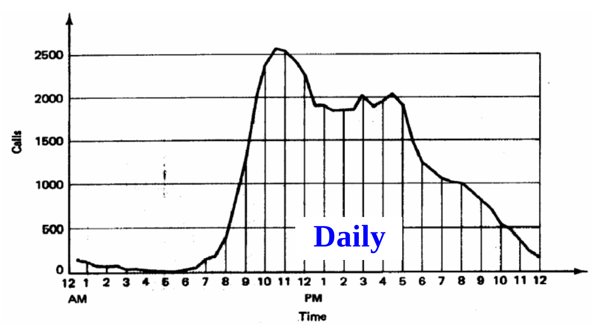


Figure 3 Typical half-hourly call distribution (Bundy D A).

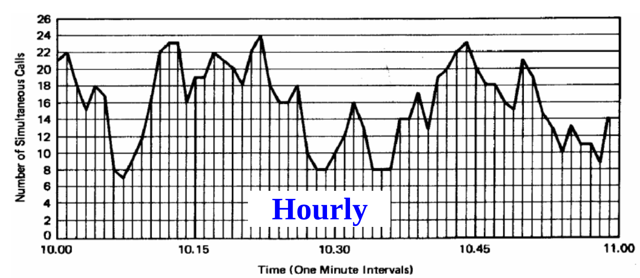


Figure 4 Typical intrahour distribution of calls, 10:00–11:00 A.M.

Q-Science: Predictable Variability

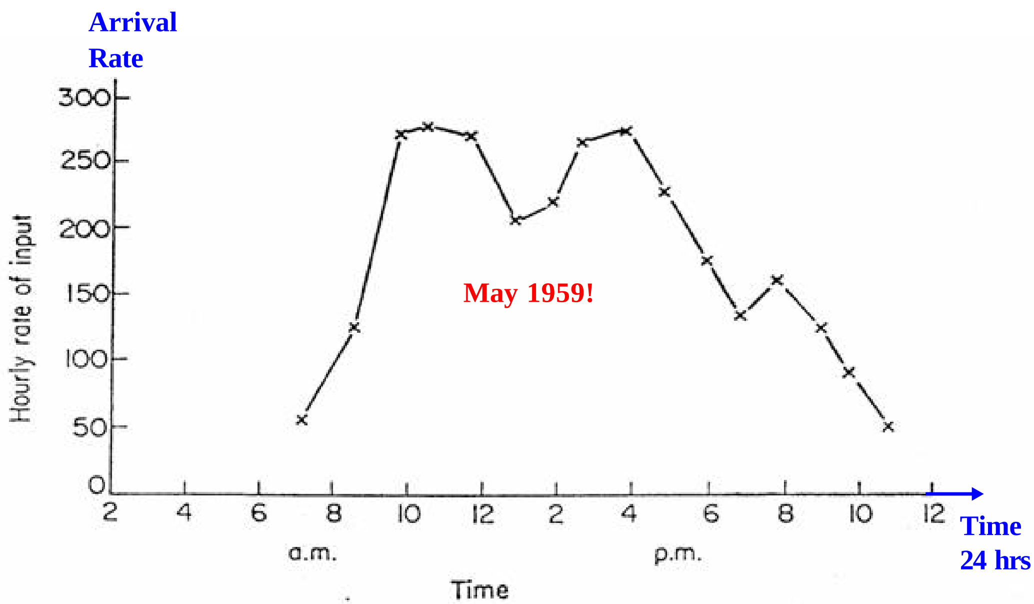
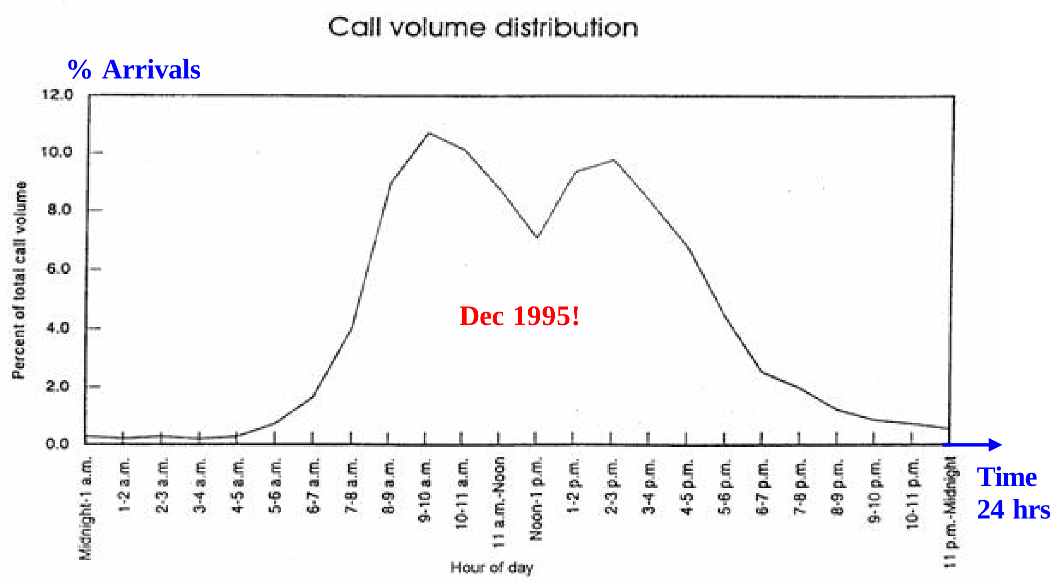


Fig. 15.1 The variation in the hourly input rates of reservations calls during a typical day (in May 1959)

(Lee A.M., Applied Q-Th)

1995 Help Desk and Customer Support Practices Report

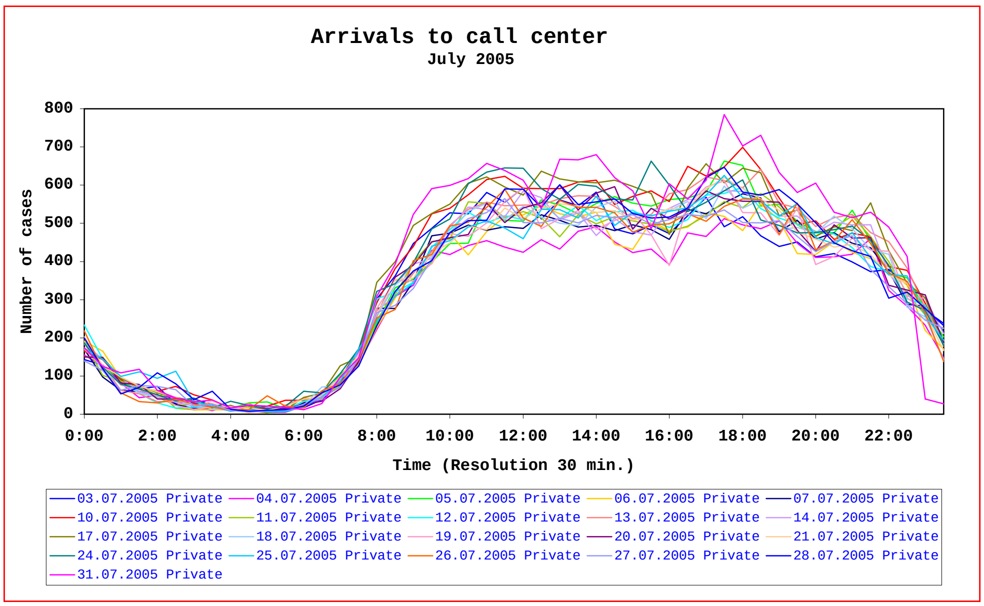


Number of respondents = 522

(Help Desk Institute)

Arrivals to Service: Poisson Processes

Weekday Arrival Rates (Israeli CC, MOCCA)



- Arrivals over short (but not too short) intervals (15, 30 min) are close to homogeneous **Poisson**, with **over-dispersion**.
- Arrivals over the day are (over-dispersed) **non-homogeneous** Poisson.

Practice: model as Poisson with piecewise-constant arrival rates.

Poisson Phenomena:

- **PASTA** = **P**oisson **A**rrivals **S**ee **T**ime **A**verages;
- **Biased sampling:** Why is the service time we encounter upon arrival longer than a "typical" service time?

Arrivals to Service: Forecasting

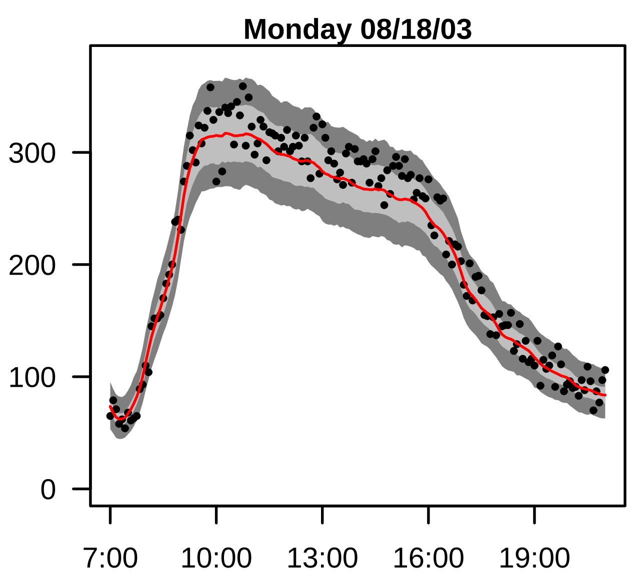
How to **predict** Poisson arrival rates? **Time Series** models.
Days are divided into **time intervals** over which arrival rates are assumed **constant**.

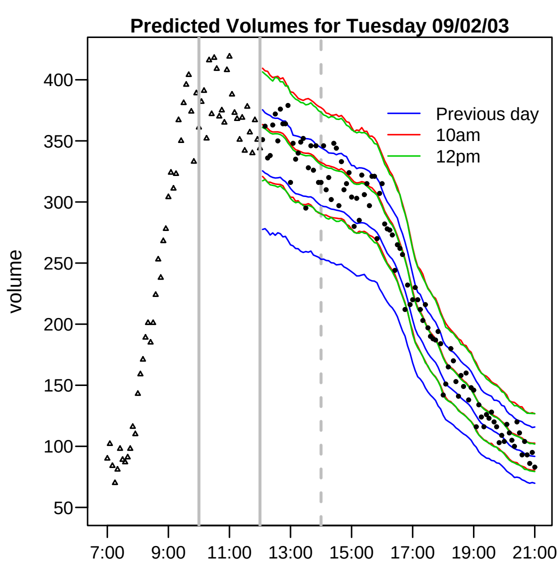
Standard Resolutions: 15 min, 30 min, 1 hour.

N_{jk} = number of arrivals on day j during interval k.
Assume K time intervals and J days overall.

- **One-day-ahead** prediction:
  $N_{1\cdot}, \ldots, N_{j-1,\cdot}$ known. Predict $N_{j1}, \ldots, N_{jK}$.
- **Several days (weeks) ahead** prediction.
- **Within-day** prediction.

Forecast Accuracy (U.S. Bank, Weinberg)





Service Times (Durations)

http://iew3.technion.ac.il/serveng/Lectures/ServiceFull.pdf

Why Significant? +1 second of 1000 agents costs $500K yearly.

Why Interesting?
Must accurately **Model, Estimate, Predict, Analyze**:

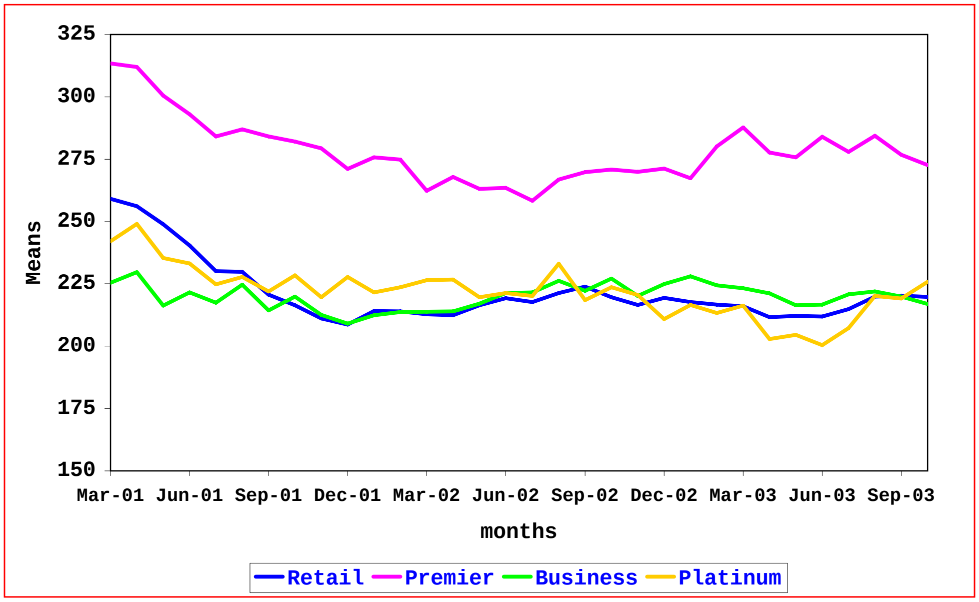
- Resolution: Sec's (phone)? min's (email)? hr's (hospital)
- Parameter, Distribution (Static) or Process (Dynamic)?
- Does it include after-call work?
- Does it include interruptions?
  - Whisper time, hold time, phones during face-to-face,...
- Does is account for return services?

How affected by covariates?

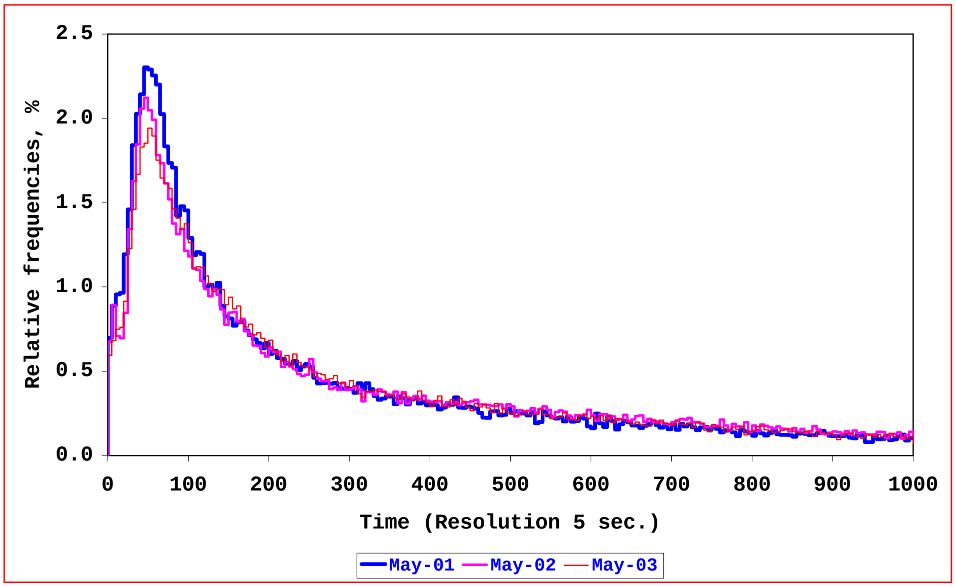
- Experience and Skill of agents (Learning Curve)
- Type of Customer: Service Type, VIP Status
- Time-of-Day: Congestion-Level
- Human Factor: Incentives, pending workload, fatigue

Service Times: Trends and Stability

Average Customer Service Time, Weekdays (MOCCA)



USBank Service-Time Histograms for Telesales (MOCCA)

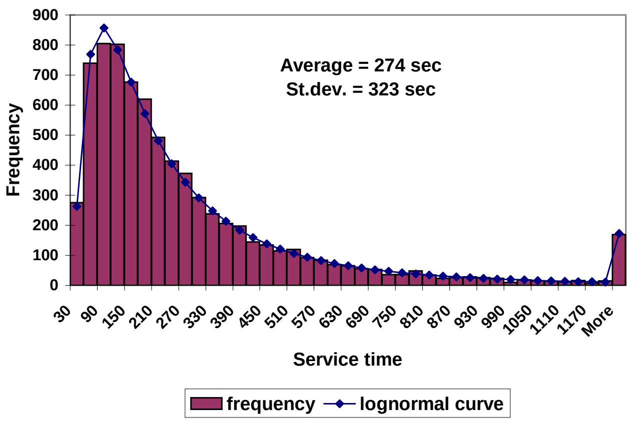


Service Times: Static Models, or Averages Do Not Tell the Whole Story

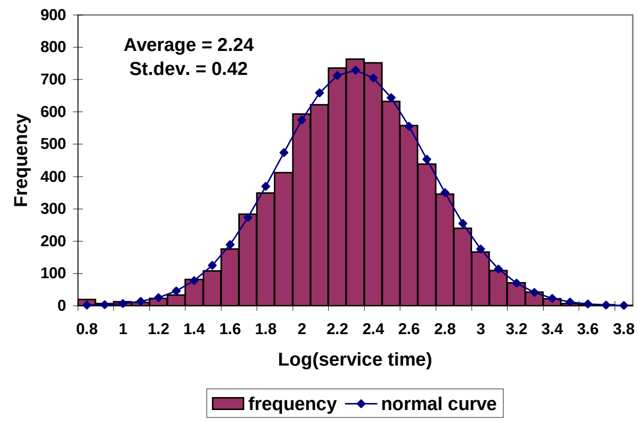
Distributions: Parametric (Exponential, Lognormal), Semi-Parametric (Phase-Type), Non-Parametric (Empirical).

Lognormal Service Times in an Israeli Bank

Histogram

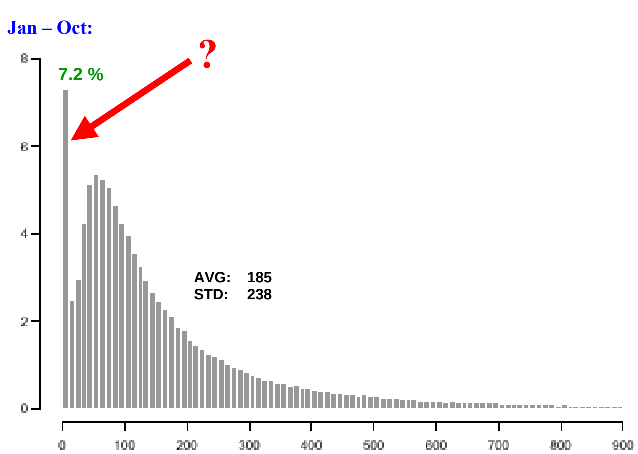


Histogram in Logarithmic Scale

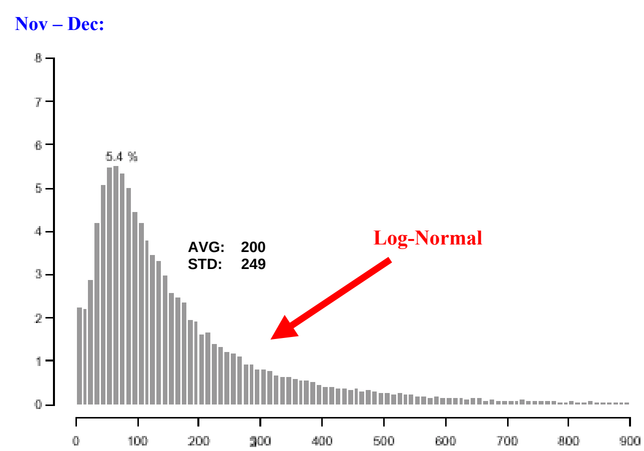


A Typical Call Center?

January-October

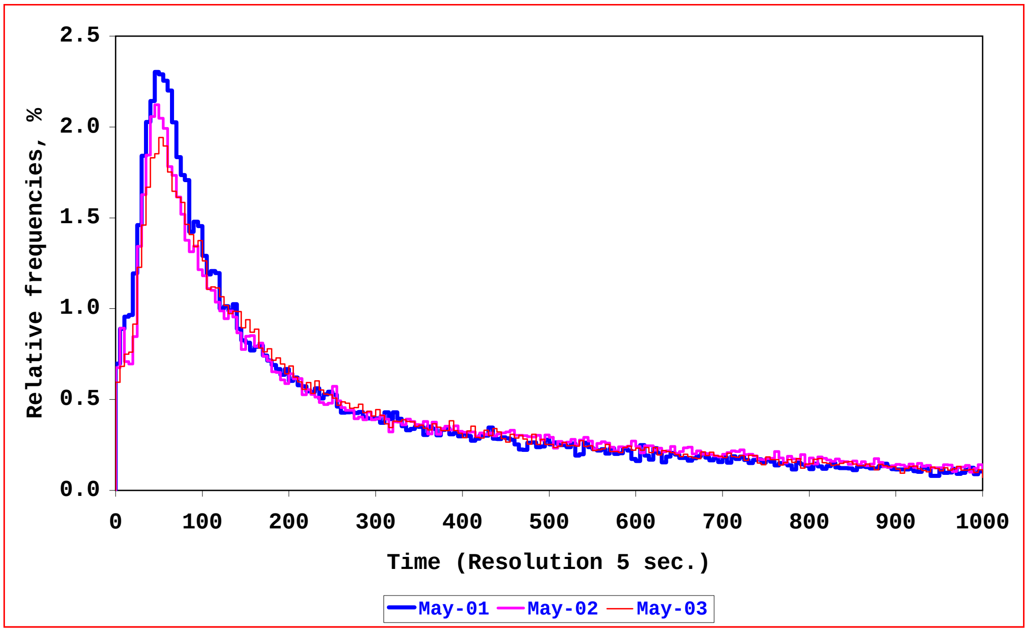


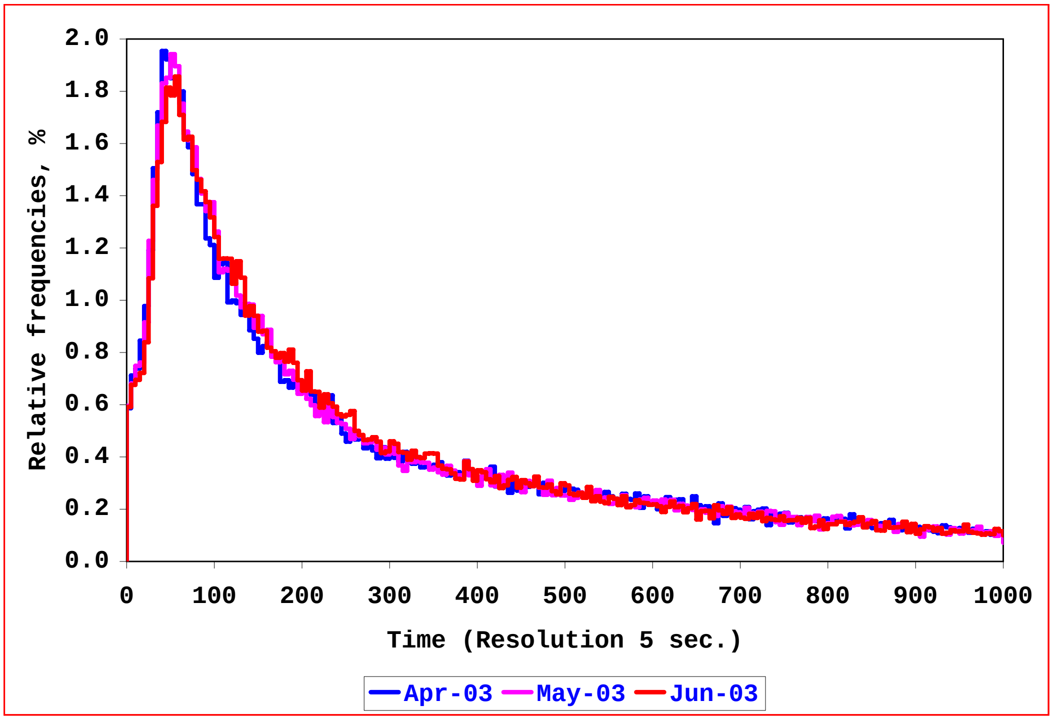
November-December



Service Times: 5 Sec's Resolution

USBank. Service-Time Histograms for Telesales (MOCCA)





Local Municipalities

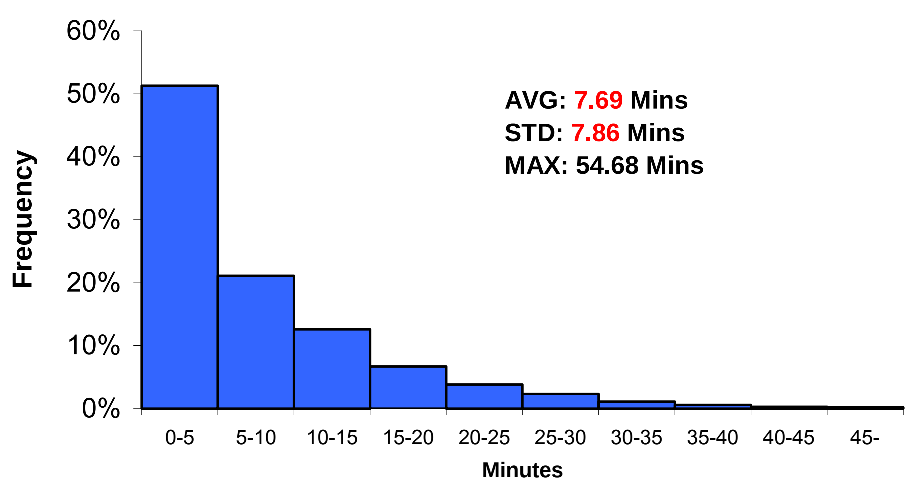
Department	Station No.	Total Customers	Avg. Arrival Rate (1/Hr)	Avg. Service Time (Mins)	STD (Mins)	Maximal Service Time (Mins)	Utilization	Avg. Waiting Time (Mins)
Water	N/A	187	1.8 ± 0.2	**8.87** ± 1.0	**8.15**	54.68	13.3%	4.76
Tellers	N/A	1328	12.6 ± 0.5	**8.82** ± 0.4	**8.55**	49.37	30.8%	7.73
Cashier	N/A	757	7.2 ± 0.4	**6.64** ± 0.4	**6.94**	29.95	79.7%	3.89
Manager	N/A	190	1.8 ± 0.2	**7.99** ± 1.0	**8.44**	38.97	24.1%	9.16
Discounts	N/A	317	3.0 ± 0.3	**4.59** ± 0.4	**4.54**	36.72	23.1%	3.65

Water	1	57	N/A	**7.80** ± 1.70	**7.61**	31.28	6.5%	N/A
	2	130	N/A	**9.34** ± 1.20	**8.37**	54.68	19.3%	N/A
Tellers	3	336	N/A	**9.04** ± 0.80	**8.93**	49.05	48.2%	N/A
	4	208	N/A	**9.93** ± 1.00	**8.82**	49.12	33.0%	N/A
	5	417	N/A	**8.97** ± 0.70	**8.55**	49.37	59.4%	N/A
	6	144	N/A	**9.53** ± 1.20	**8.75**	41.70	21.8%	N/A
	7	156	N/A	**8.03** ± 1.10	**7.96**	35.27	19.8%	N/A
	8	67	N/A	**3.74** ± 0.70	**3.58**	21.03	4.0%	N/A
Cashier	9	757	N/A	**6.64** ± 0.40	**6.94**	29.95	79.7%	N/A
Manager	10	190	N/A	**1.99** ± 1.00	**8.44**	38.97	24.1%	N/A
Discounts	11	317	N/A	**4.59** ± 0.40	**4.54**	36.72	23.1%	N/A

*Service time ranges given with 90% confidence.

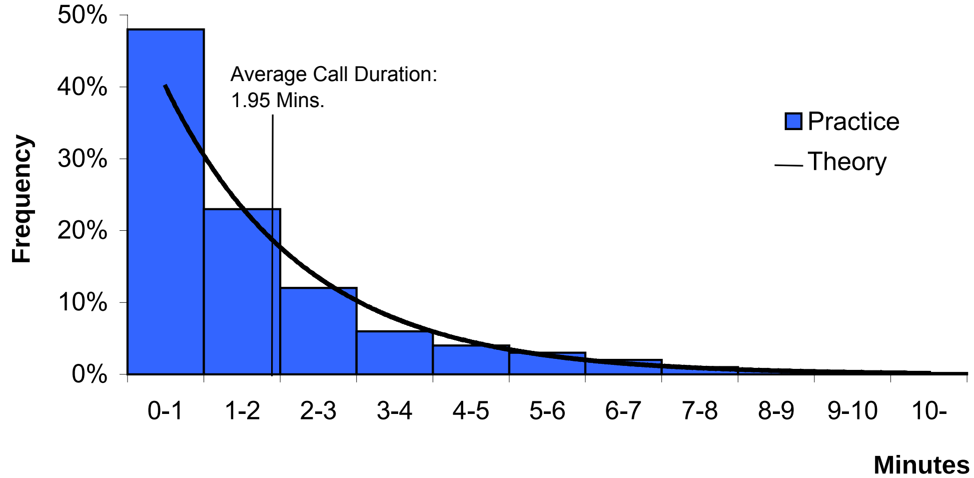
Service Time Histogram – Overall:

Range	Frequency
0-5	51.3
5-10	21.1
10-15	12.6
15-20	6.7
20-25	3.8
25-30	2.3
30-35	1.1
35-40	0.6
40-45	0.3
45-	0.2

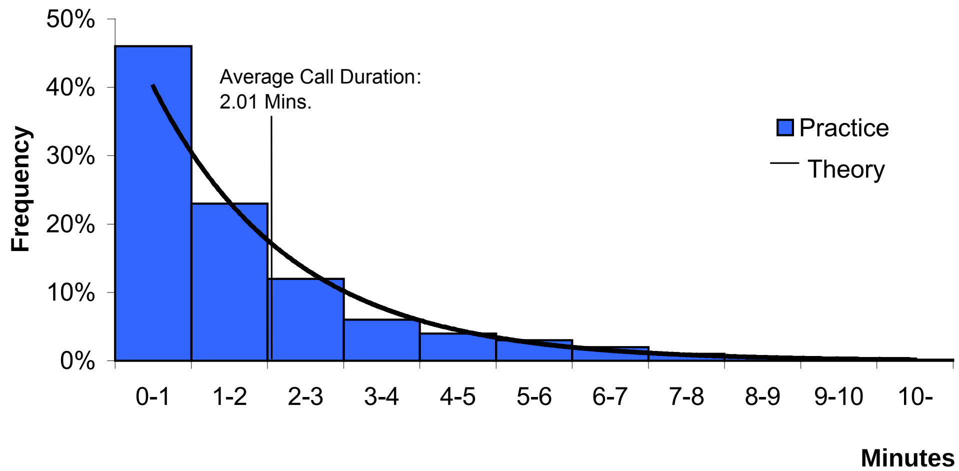


Service Times: Exponential (Phone Calls)

Call-Duration Frequency - North:



Call-Duration Frequency – Central:

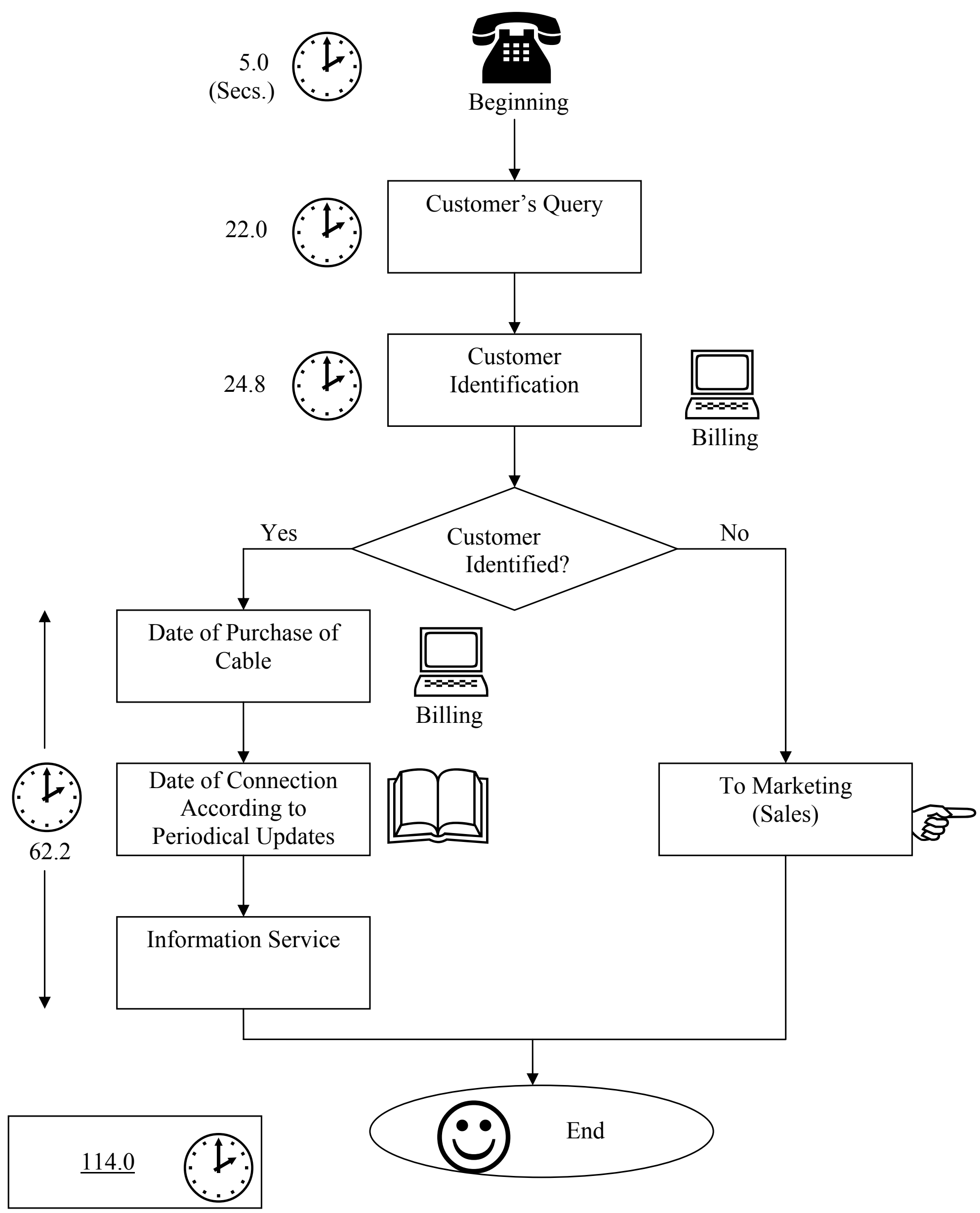


Q. How to recognize “Exponential” when you "see" one?

A. Geometric Approximation.

Service Times: Phase-Type Model

Late Connections



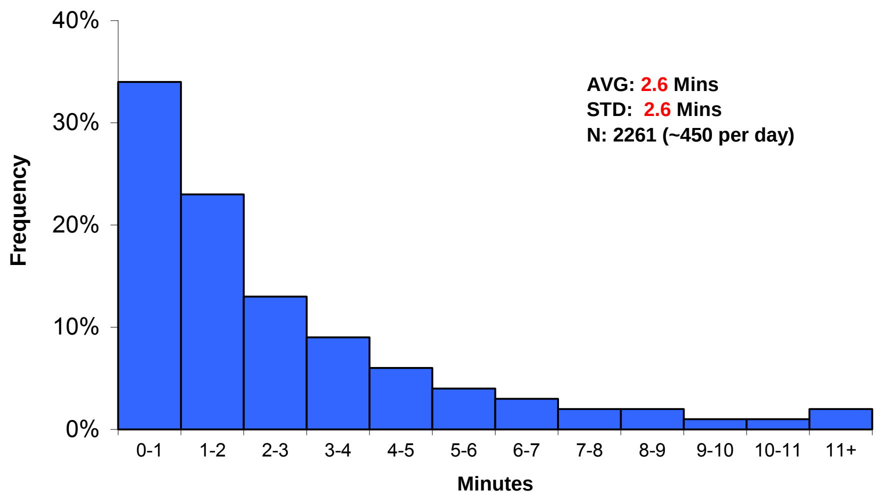
? Where does human-service start / end (recall 144)?
"Average" picture.

Service Times: Exponential, Phase-Type

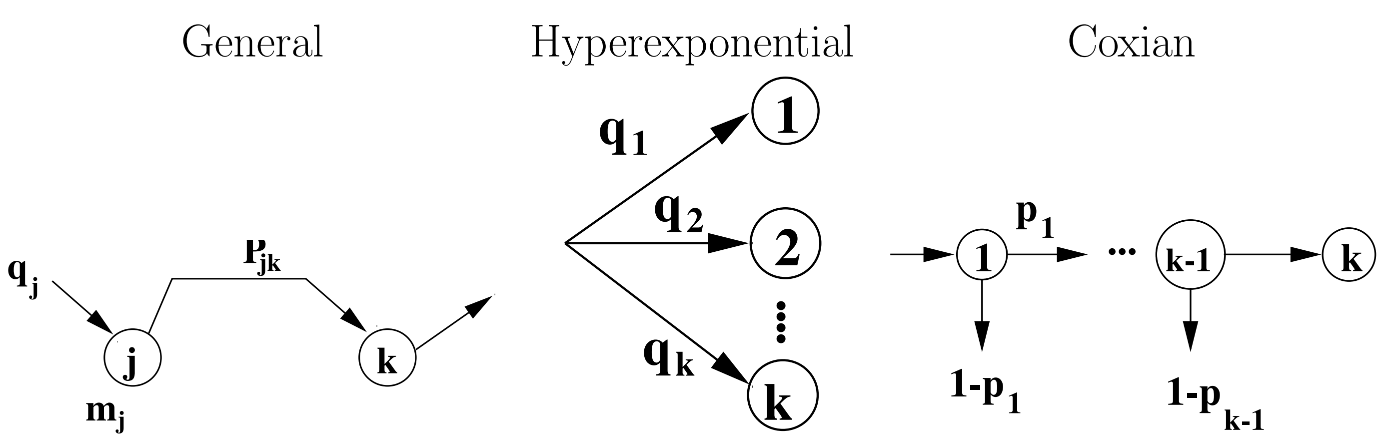
Static Model: Exponential Duration

Face-to-Face Services in a Government Office

Service Times Histogram:

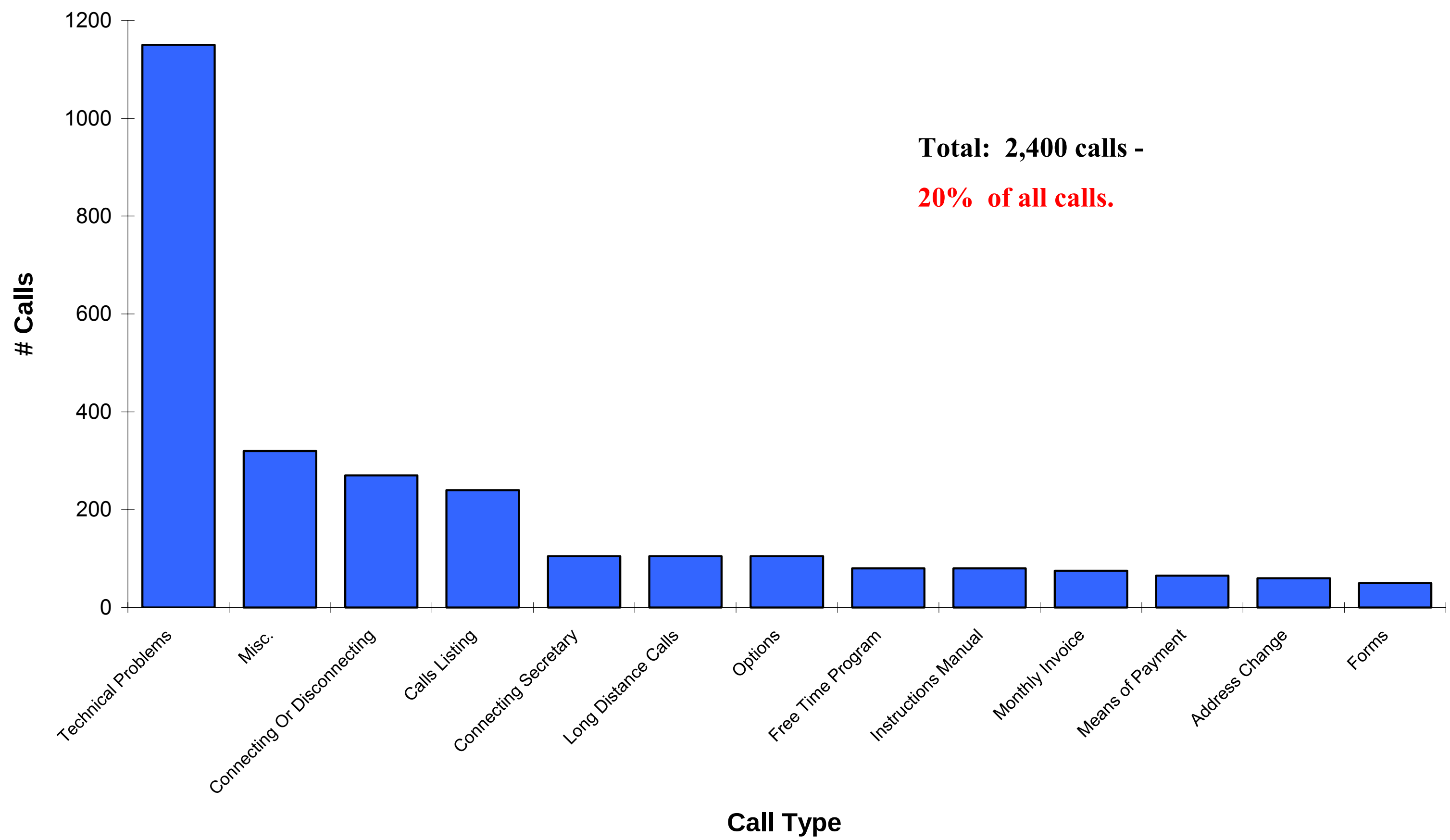


Dynamic Model: Phase-Type Duration



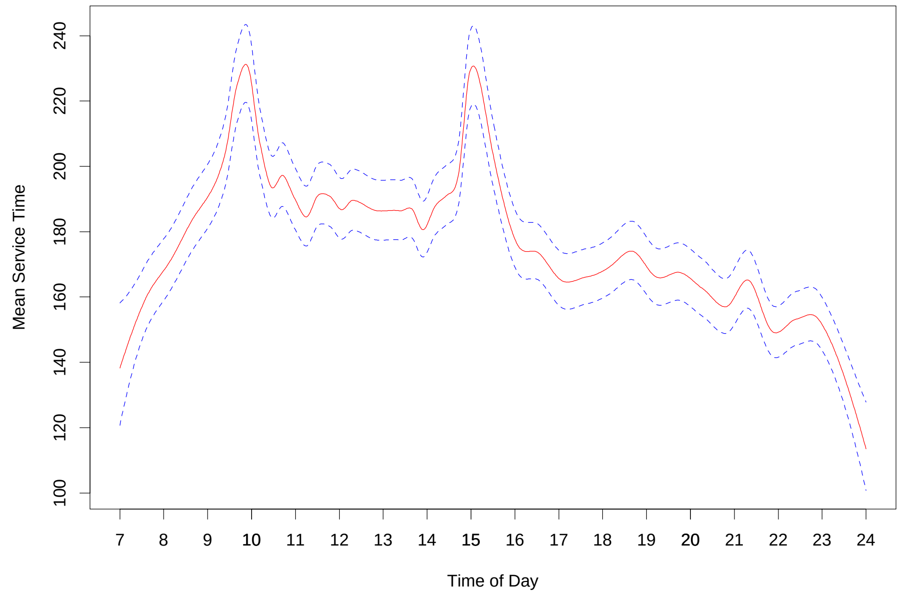
Service Times: Returns

Bank Classification of "Continued – Calls"

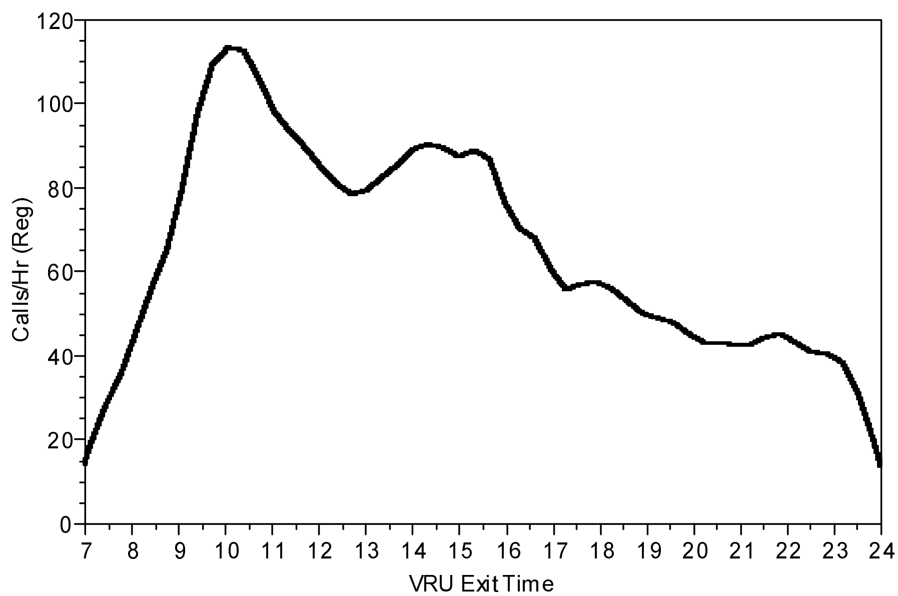


Service Times: The Human Factor, or Why Longest During Peak Loads?

Mean-Service-Time (Regular) vs. Time-of-Day (95% CI) (n=42613)



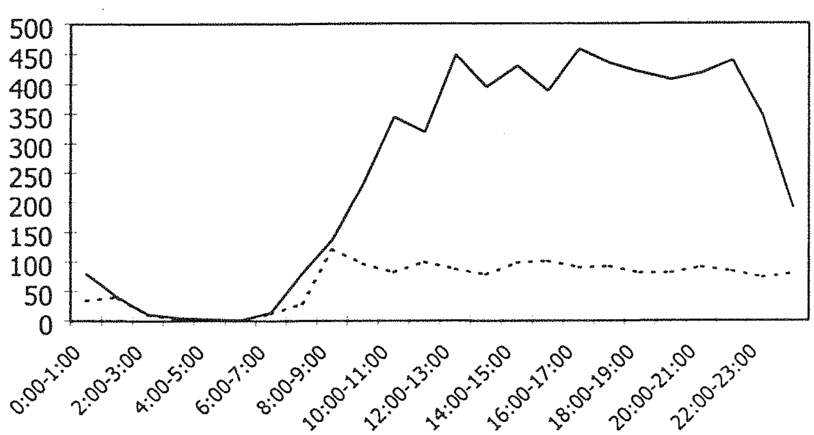
Arrivals to Queue or Service - Regular Calls (Inhomogeneous Poisson)



Customers' (Im)Patience

Marketing Campaign at a Call Center

Average wait 376 sec, 24% calls **answered**



Abandonment **Important and Interesting**

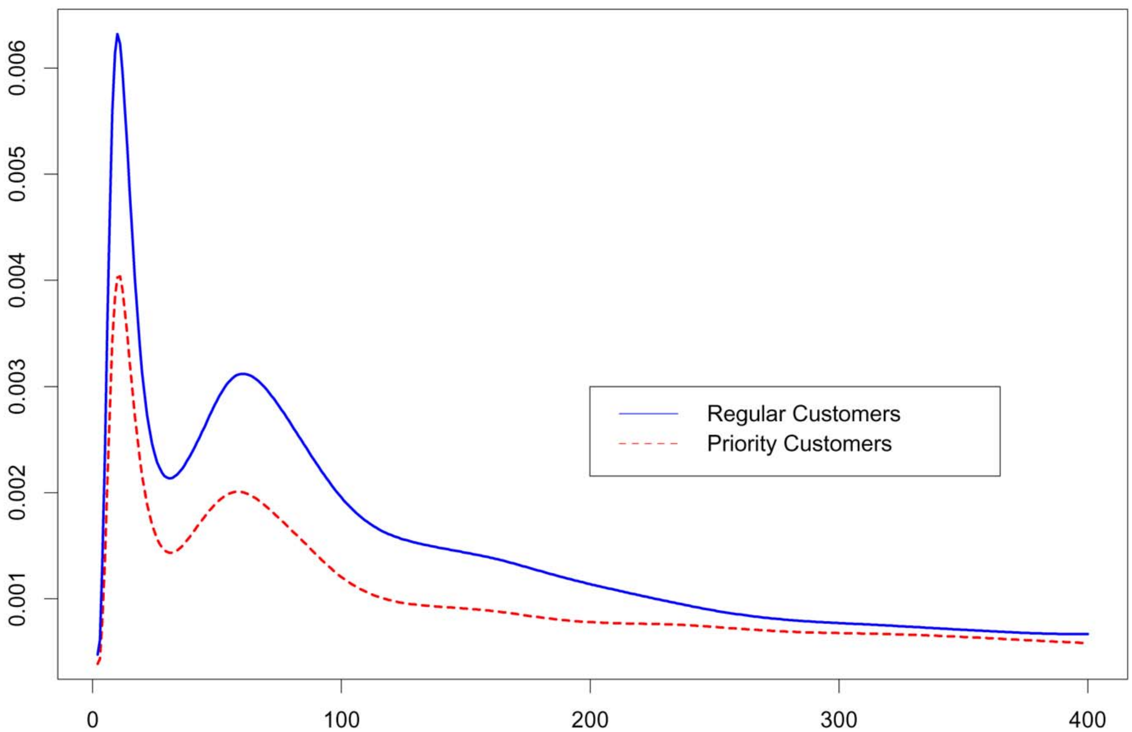
- One of two customer-subjective performance measures (2$^{\text{nd}}$=Redials)
- Poor service level (future losses)
- Lost business (present losses)
- 1-800 costs (present gains; out-of-pocket vs. alternative)
- Self-selection: the "fittest survive" and wait less (much less)
- Accurate Robust models (vs. distorted instability-prone)
- Beyond Operations/OR: Psychology, Marketing, Statistics
- Beyond Telephony: VRU/IVR (Opt-Out-Rates), Internet (over 60%), Hospitals ED (LWBS).

Understanding (Im)Patience

- **Observing** (Im)Patiecne – Heterogeneity:
  Under a single roof, the fraction abandoning varies
  from 6% to 40%, depending on the type of service/customer.
- **Describing** (Im)Patience Dynamically:
  Irritation proportional to Hazard Rate (Palm's Law).
- **Managing** (Im)Patience:
  - VIP vs. Regulars: who is more "Patient"?
  - What are we actually measuring?
  - (Im)Patience Index:
    "How long Expect to wait" relative to
    "How long Willing to wait".
- **Estimating** (Im)Patience: Censored Sampling.
- **Modeling** (Im)Patience:
  - The "Wait" Cycle:
    Expecting, Willing, Required, Actual, Perceived, etc.
    The case of the Experienced & Rational customer.
  - (Nash) Equilibrium Models.

Palm's Law of Irritation (1943-53): $\propto$ Hazard-Rate of (Im)Patience Distribution

Small Israeli Bank (1999):
Regular over **Priority (VIP)** Customers



Hazard-Rate function of $\tau \geq 0$ (absolutely continuous):

$$h(t) \ = \frac{g(t)}{1 - G(t)}\,,$$

$g =$ Density function of τ,
$G =$ Distribution function of τ.

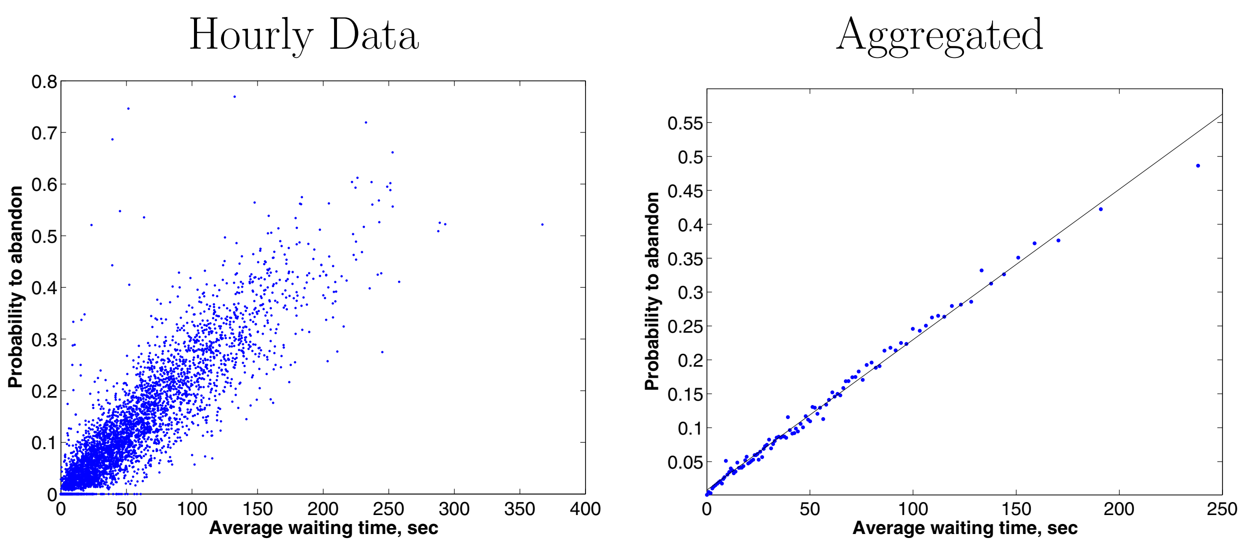
Intuition: $\mathrm{P}\{\tau \leq t + \Delta | \tau > t\} \approx h(t) \cdot \Delta$.

$P\{Ab\} \propto \mathrm{E}[W_q]$

Claim: (Im)Patience that is exp(θ) implies

$$\mathrm{P}\{\mathrm{Ab}\} \;=\; \theta \cdot \mathrm{E}[W_q]\,.$$

Small Israeli Bank: 1999 Data



The graphs are based on 4158 hour intervals.

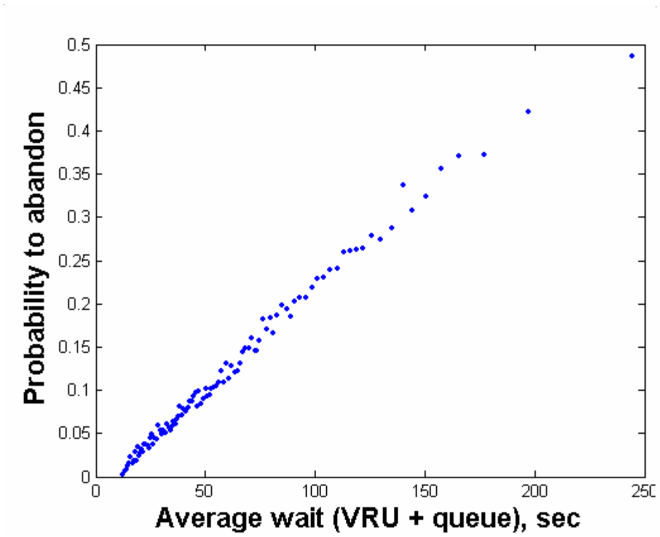
Regression $\Rightarrow$ average patience $(1/\theta) \approx \dfrac{250}{0.56} \approx 446$ sec.

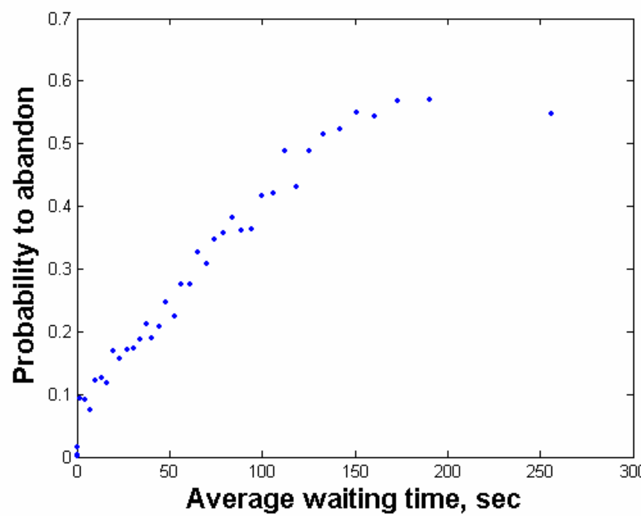
But (im)patience at this bank is **not** exponential ! ?

Moreover,

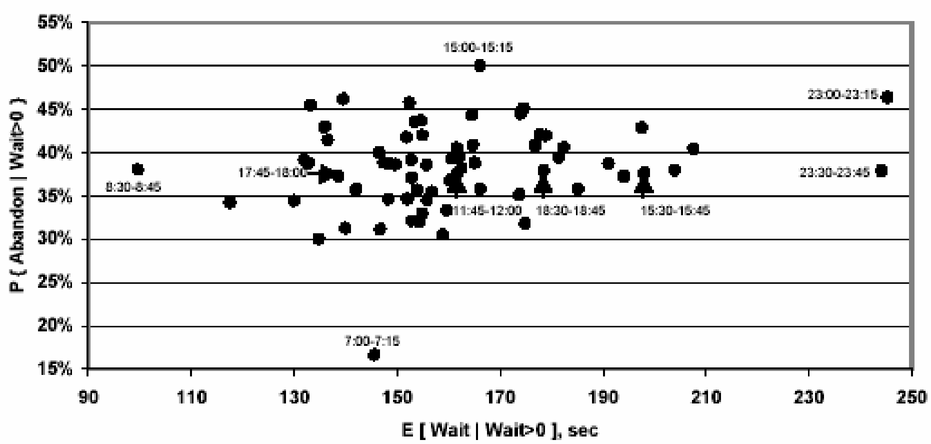
Queueing Science: Human Behavior

Delayed Abandons (IVR) | **Balking (New Customers)**

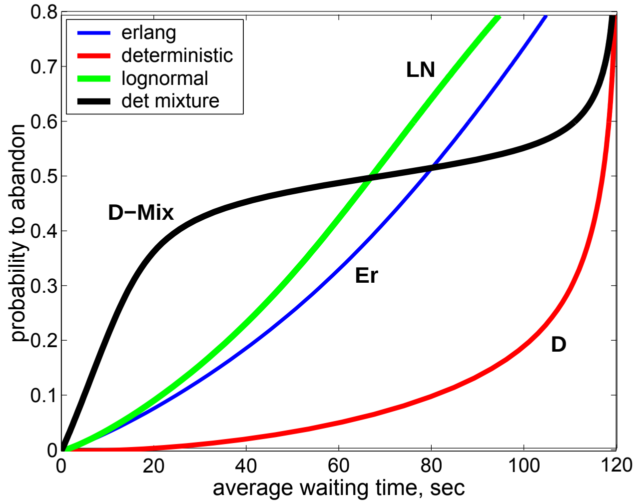




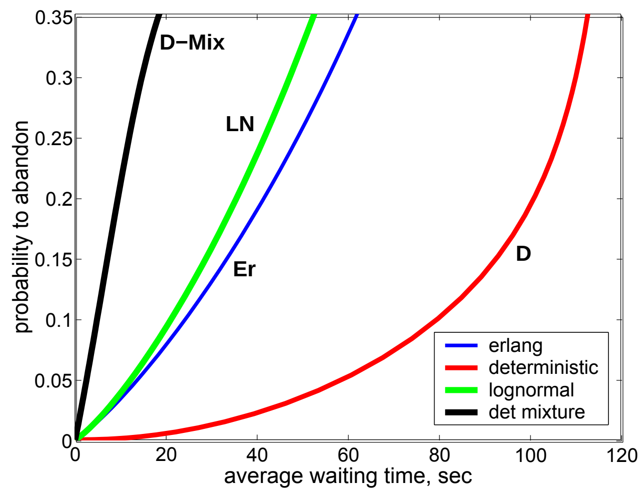
Learning (Internet Customers)



Examples of non-linear relations



moderate loads



Patience distributions:

- **D**: Deterministic: 2 minutes exactly;
- **Er**: Erlang with two exp(mean=1) phases;
- **LN**: Lognormal, both average and standard deviation equal to 2;
- **D-Mix**: 50-50% mixture of two constants: 0.2 and 3.8.

A Patience Index

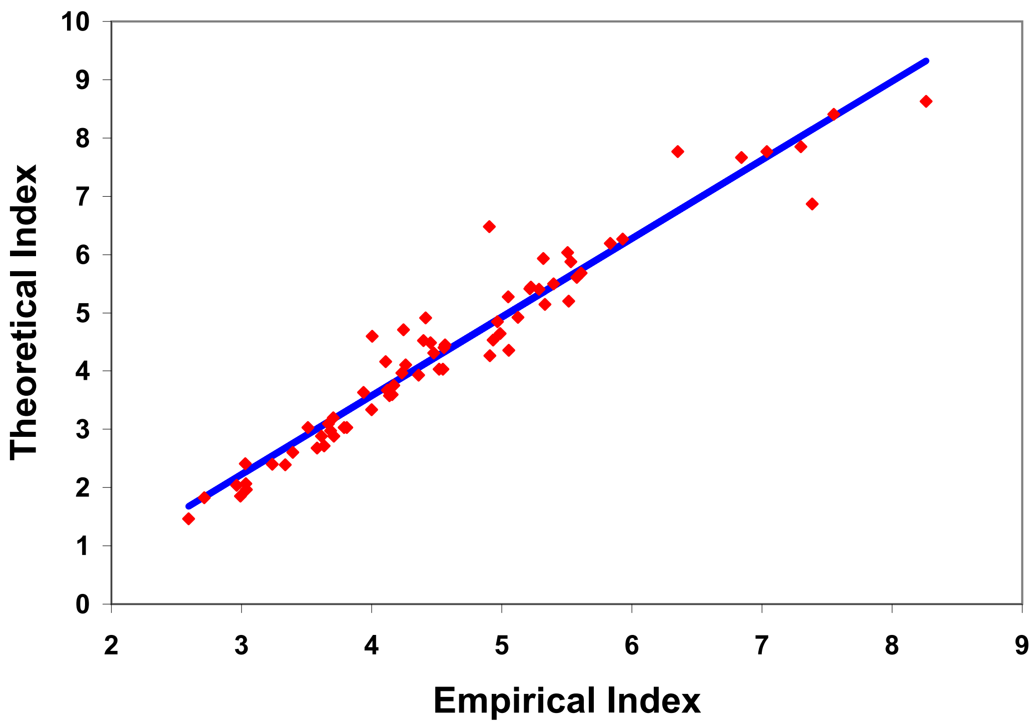
How to quantify (im)patience?

$$\textbf{Theoretical}\text{ Patience Index } = \frac{\text{Willing to Wait}}{\text{Expected to Wait}}.$$

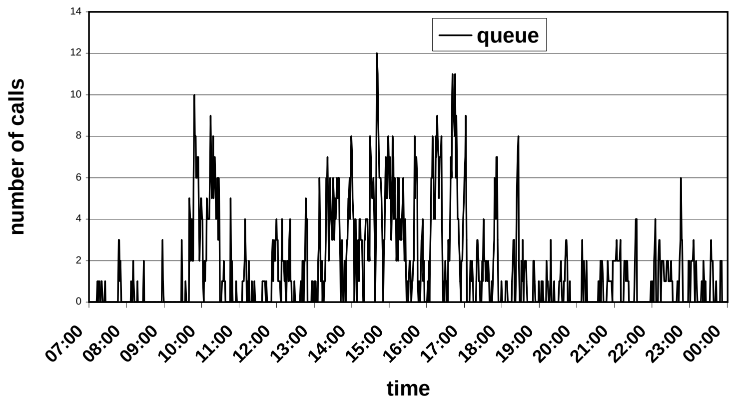
How to measure? Calculate? Assume **Experienced** customers. Then, a simple (but not too simple) model suggests the easy-to-measure:

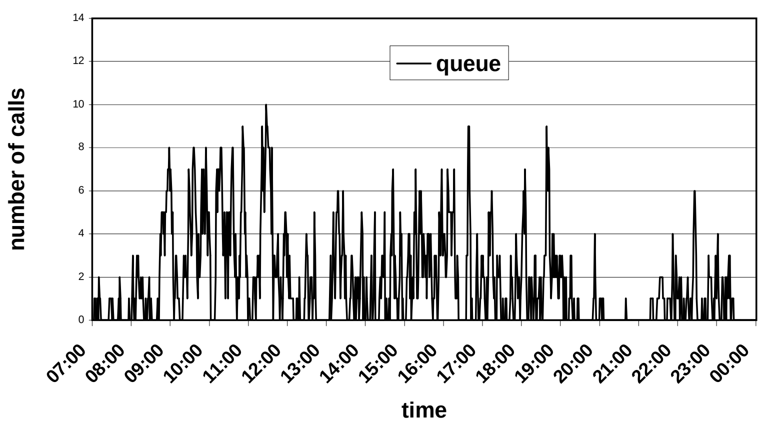
$$\textbf{Empirical}\text{ Patience Index } \triangleq \frac{\%\text{ Served}}{\%\text{ Abandoned}}.$$

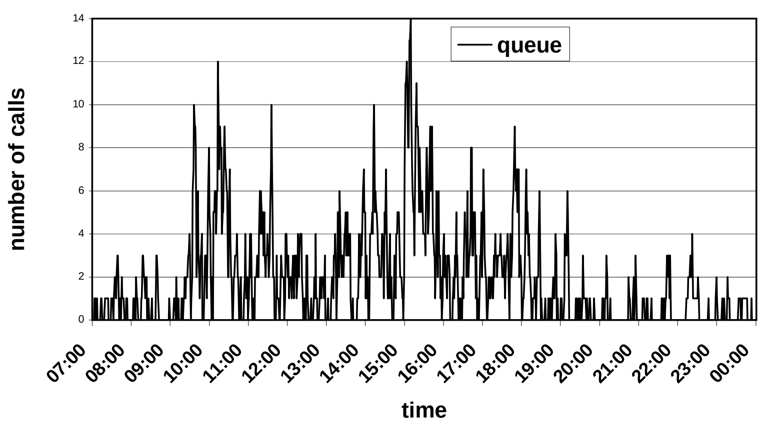
Patience index – Empirical vs. Theoretical



Queues = Integrating the Building Blocks

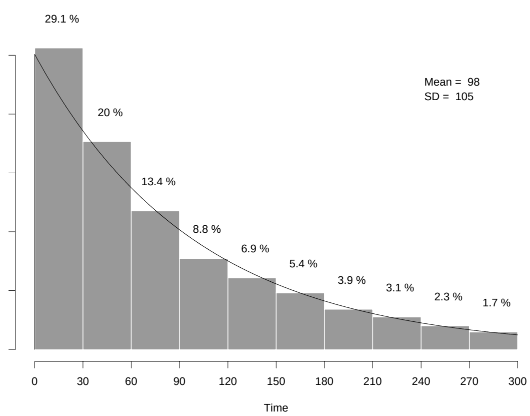


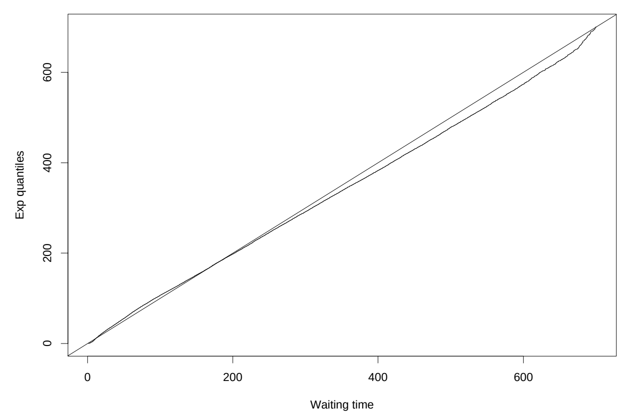




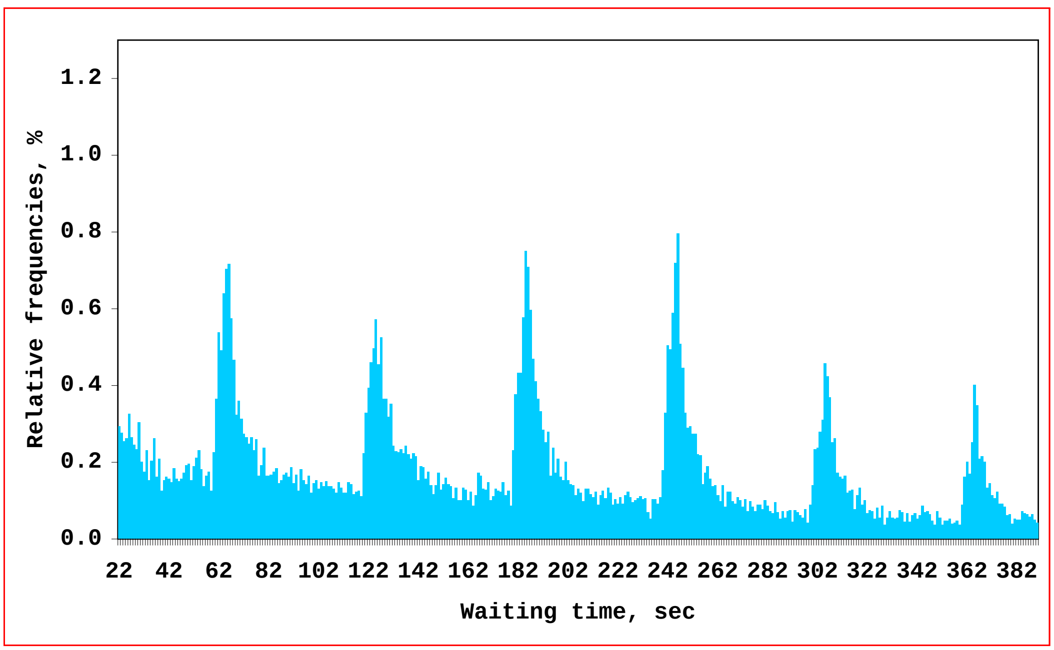
Delays = Integrating the Building Blocks

Exponential Delays:
Small Call Center of an Israeli Bank (1999)





Delays:
Medium-Size Call Center of an Israeli Bank (2006)

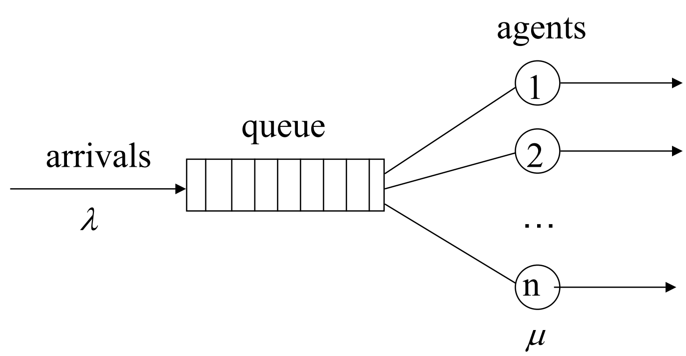


Basic (Markovian) Queueing Models of a Basic Service Station

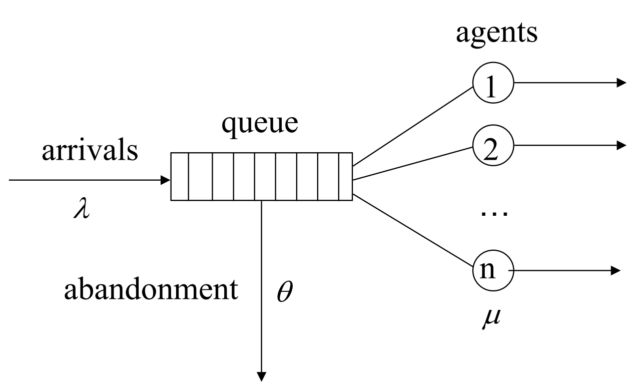
Poisson arrivals, **Exponential** service times, **Exponential** (im)patience.

Mathematical Framework: Markov Jump-Processes (Birth&Death).

M/M/n (Erlang-C) Queue



M/M/n+M (Palm/Erlang-A) Queue



Additional Markovian Models: Balking, Trunks; Retrials.

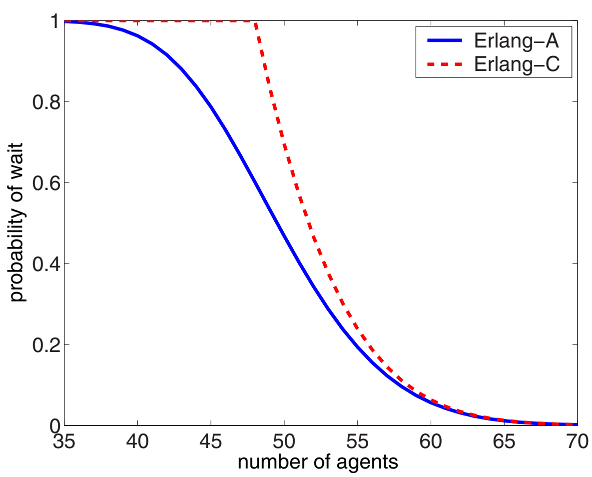
Applications: Performance Analysis, Design (EOS), Staffing.

"The Fittest Survive" and Wait Less - Much Less!

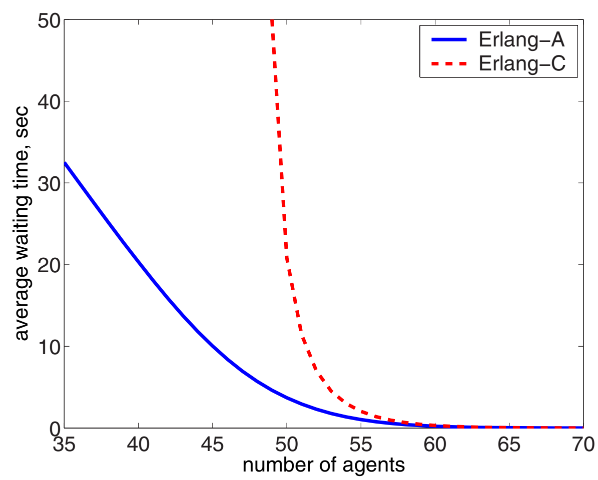
Erlang-A vs. Erlang-C

48 calls per min, 1 min average service time,
2 min average patience

probability of wait
vs. number of agents



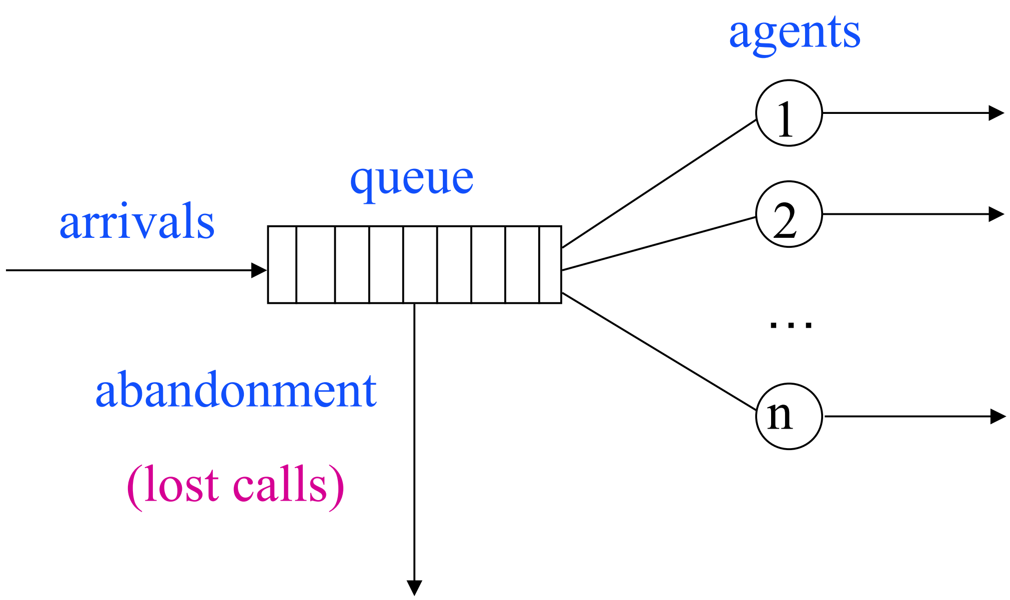
average wait
vs. number of agents



If 50 agents:

	M/M/n	M/M/n+M	M/M/n, $\lambda \downarrow 3.1\%$
Fraction abandoning	–	3.1%	-
Average waiting time	20.8 sec	3.7 sec	8.8 sec
Waiting time's 90-th percentile	58.1 sec	12.5 sec	28.2 sec
Average queue length	17	3	7
Agents' utilization	96%	93%	93%

Modelling (Im)Patience: Time Willing vs. Time Required to Wait



- **(Im)Patience Time** $\tau \sim G$:
  Time a customer **willing to wait** for service.

- **Offered Wait** V:
  Time a customer **required to wait** for service;
  in other words, waiting-time of an infinitely-patient customer.

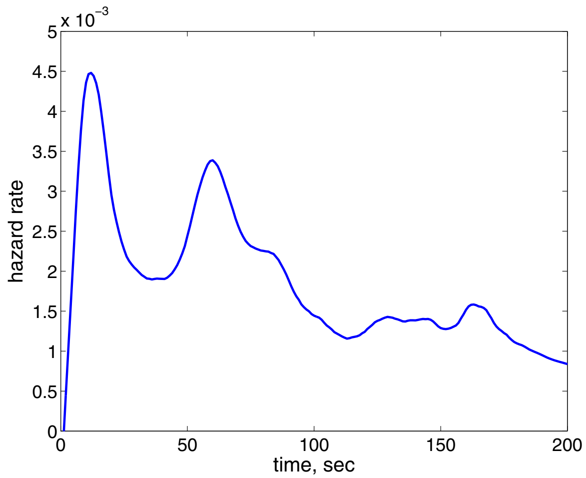
- If $\tau \leq V$, customer **Abandons**;
  otherwise, customer **Served**;

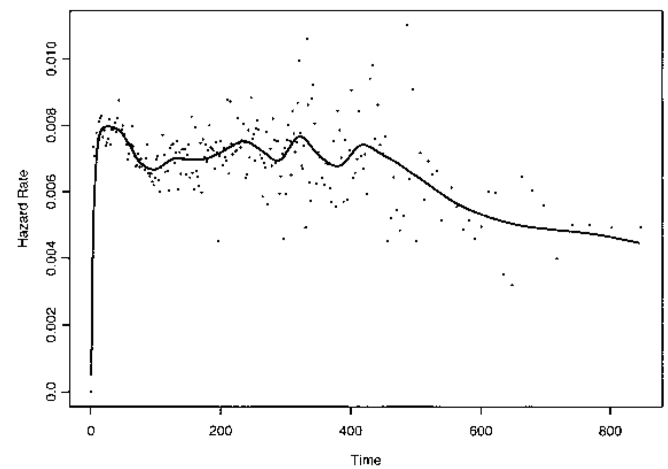
- **Actual wait** $W = \min(\tau, V)$.

Call Center Data: Hazard Rates (Un-Censored)

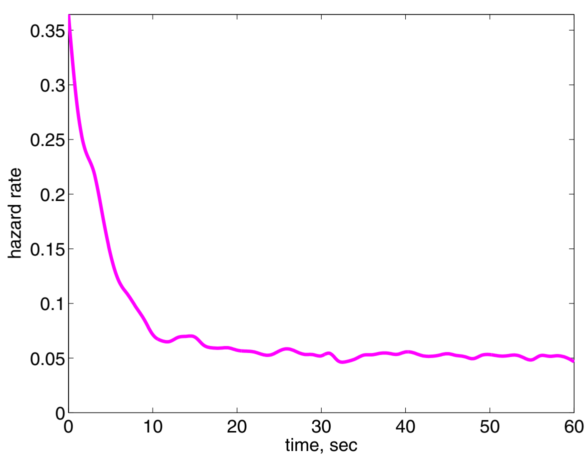
(Im)Patience Time | Required/Offered Wait

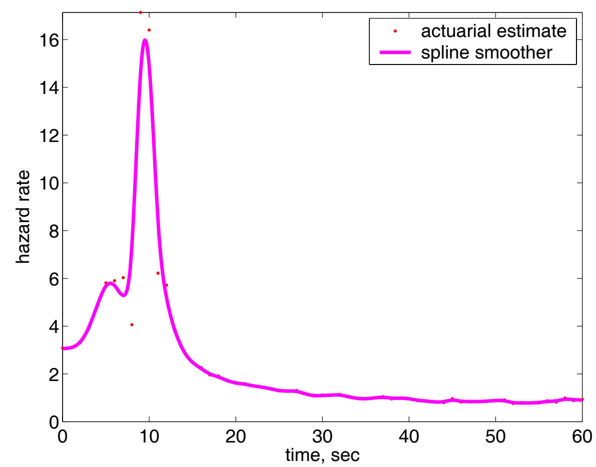
Israel





U.S.





Predicting Performance

Model **Primitives** (eg. Erlang-A):

- Arrivals to service (eg. Poisson)
- (Im)Patience while waiting τ (eg. Exp)
- Service times (eg. Exp)
- Number of Agents.

Model **Output**: ***Offered-Wait V***

Operational Performance Measure calculable in terms of (τ, V).

- eg. Average Wait = $E[\min\{\tau, V\}]$
- eg. % Abandonment = $P\{\tau < V\}$

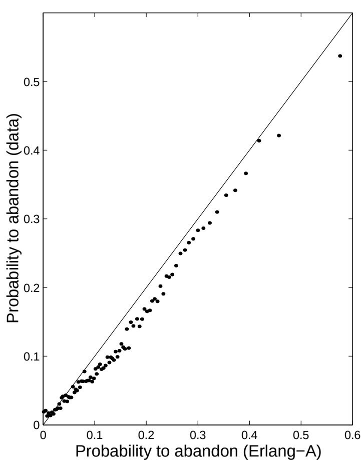
Applications:

- **Performance Analysis**
- **Design, Phenomena** (Pooling, Economies of Scale)
- **Staffing – How Many Agents** (FTE's = Full-Time-Equivalent's)
- Note: Control requires model-refinements - later, in SBR.

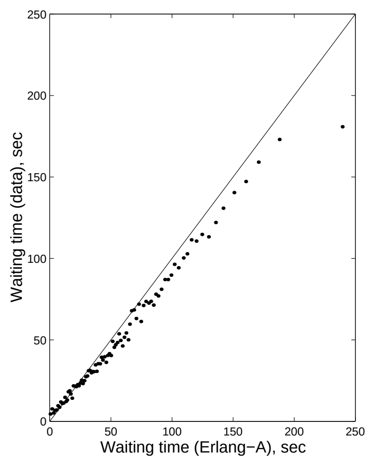
Erlang-A: A Simple Model at the Service of Complex Realities

- Small Israeli bank (10 agents);
- Data-Based Estimation of Patienc (P{Ab}/E[W_q]);
- Graph: Actual Performance vs. Erlang-A Predictions (aggregation of 40 similar hours).

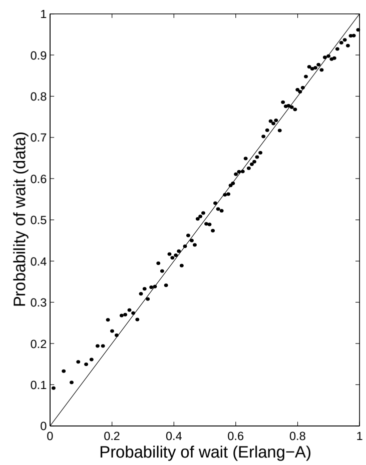
P{Ab}



E[W_q]



P{$W_q > 0$}

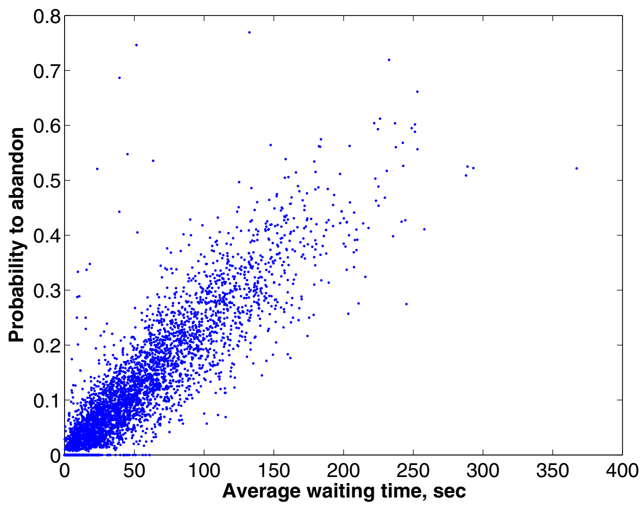


- **Question:** Why Erlang-A works? indeed, all its underlying assumptions fail (Arrivals, Services, Impatience)
- **Towards a Theoretical Answer:** Robustness and Limitations, via Asymptotic (QED) Analysis.
- **Practical Significance:** Asymptotic results applicable in small systems (eg. healthcare).

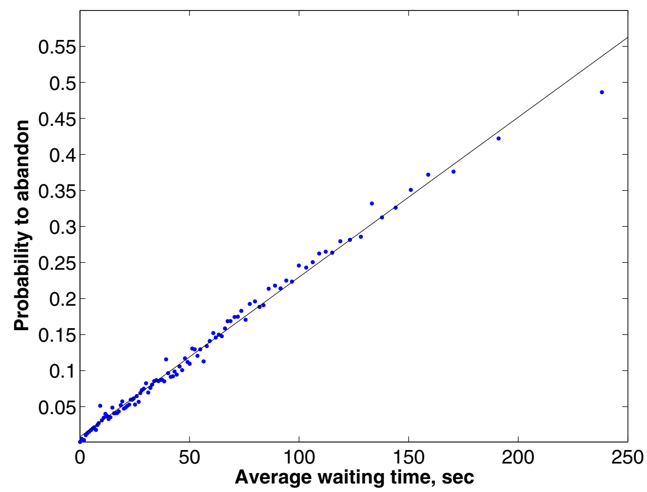
Queueing Science: In Support of Erlang-A

Israeli Bank: Yearly Data

Hourly Data



Aggregated



Data: P{Ab} $\propto$ **E**$[W_q]$.

Theory: P{Ab} $= \theta \cdot$ **E**$[W_q]$, **if** (Im)Patience = Exp(θ).

Proof: Let λ = Arrival Rate. Then, by Conservation & Little:

$$\lambda \cdot \mathrm{P}\{\mathrm{Ab}\} \;=\; \theta \cdot \mathrm{E}[L_q] \;=\; \theta \cdot \lambda \cdot \mathrm{E}[W_q], \quad \text{q.e.d.}$$

Recipe: Use Erlang-A, with $\hat{\theta} = \mathrm{P}\widehat{\{\mathrm{Ab}\}}/\mathrm{E}[\widehat{W_q}]$ (slope above).

But (Im)Patience is **not** Exponentially distributed !?

Queueing Science: via Data & Theory, Linearity Robust.
Service Engineering: via Theory & Simulations, often-enough,

- Reality $\approx$ M/G/n + G $\approx$ Erlang-A, in which $\theta = g(0)$;
- P{Ab} $\approx$ $g(0) \cdot \mathrm{E}[W_q]$, hence recipe prevails, often enough.

4CallCenters: Personal Tool for Workforce Management

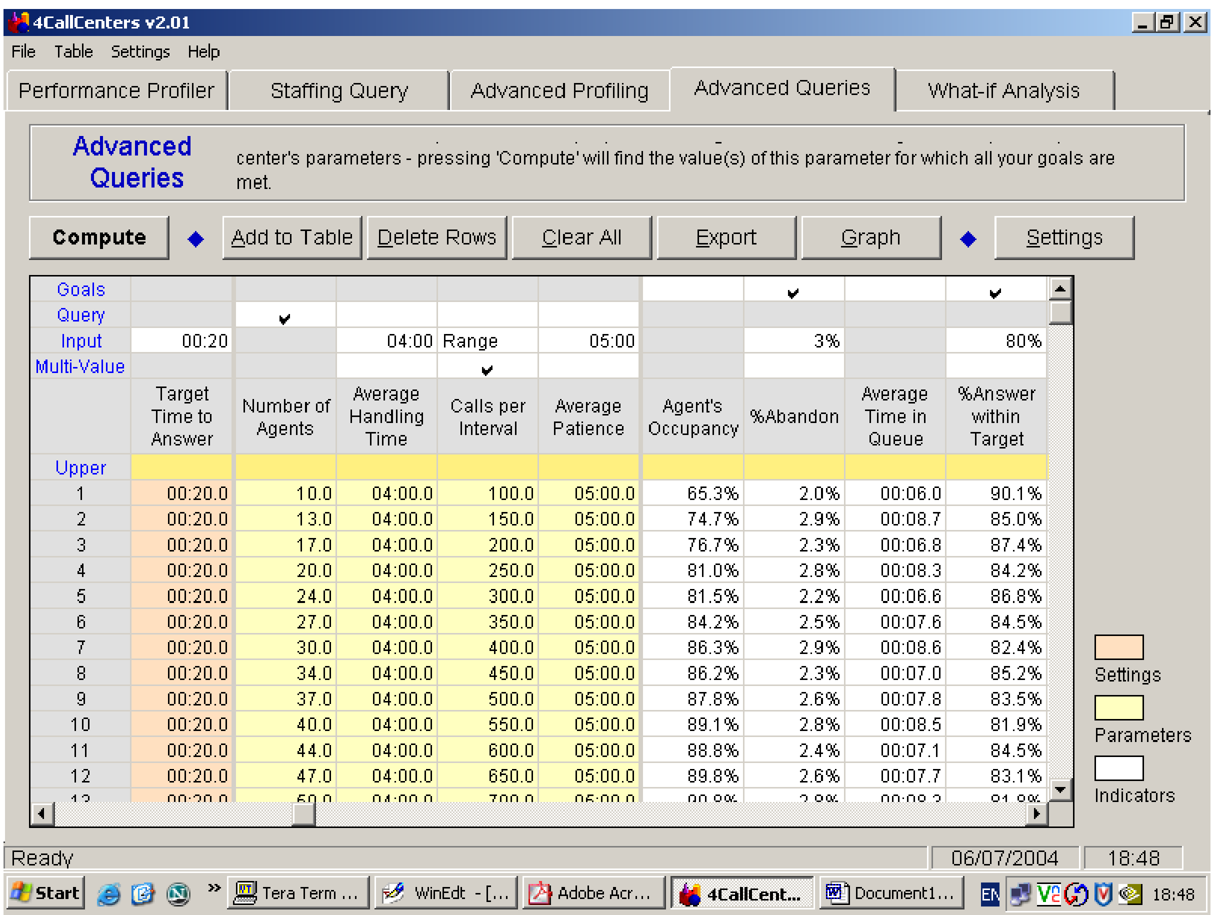
Calculations based on the M.Sc. thesis of Ofer Garnett.

Is extensively used in Service Engineering.

Install at

http://ie.technion.ac.il/serveng/4CallCenters/Downloads.htm

4CallCenters: Output Example



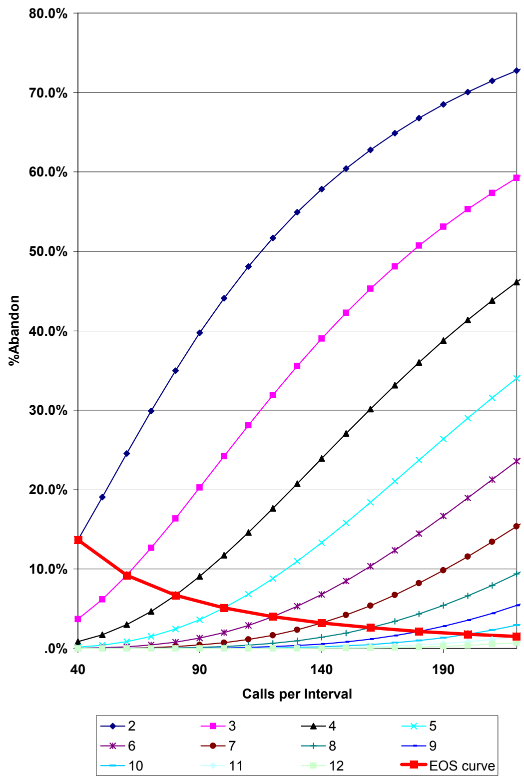
	Target Time to Answer	Number of Agents	Average Handling Time	Calls per Interval	Average Patience	Agent's Occupancy	%Abandon	Average Time in Queue	%Answer within Target
Goals							✔		✔
Query		✔							
Input	00:20		04:00	Range	05:00		3%		80%
Multi-Value				✔					
Upper									
1	00:20.0	10.0	04:00.0	100.0	05:00.0	65.3%	2.0%	00:06.0	90.1%
2	00:20.0	13.0	04:00.0	150.0	05:00.0	74.7%	2.9%	00:08.7	85.0%
3	00:20.0	17.0	04:00.0	200.0	05:00.0	76.7%	2.3%	00:06.8	87.4%
4	00:20.0	20.0	04:00.0	250.0	05:00.0	81.0%	2.8%	00:08.3	84.2%
5	00:20.0	24.0	04:00.0	300.0	05:00.0	81.5%	2.2%	00:06.6	86.8%
6	00:20.0	27.0	04:00.0	350.0	05:00.0	84.2%	2.5%	00:07.6	84.5%
7	00:20.0	30.0	04:00.0	400.0	05:00.0	86.3%	2.9%	00:08.6	82.4%
8	00:20.0	34.0	04:00.0	450.0	05:00.0	86.2%	2.3%	00:07.0	85.2%
9	00:20.0	37.0	04:00.0	500.0	05:00.0	87.8%	2.6%	00:07.8	83.5%
10	00:20.0	40.0	04:00.0	550.0	05:00.0	89.1%	2.8%	00:08.5	81.9%
11	00:20.0	44.0	04:00.0	600.0	05:00.0	88.8%	2.4%	00:07.1	84.5%
12	00:20.0	47.0	04:00.0	650.0	05:00.0	89.8%	2.6%	00:07.7	83.1%

4CallCenters: Congestion Curves

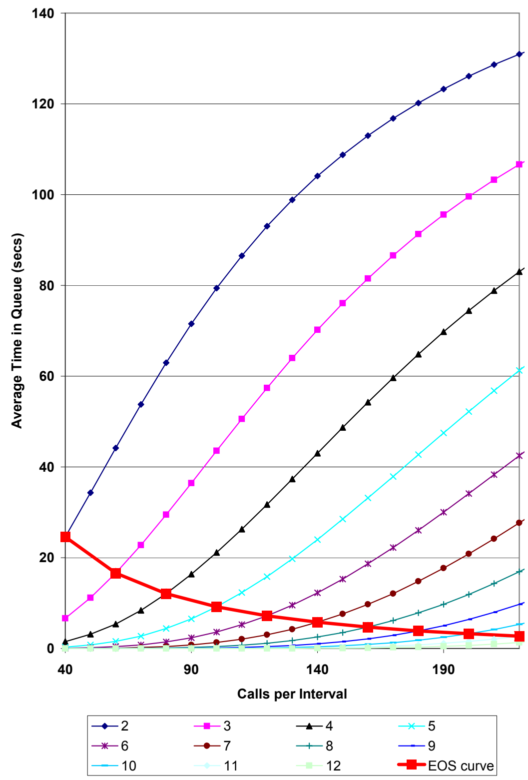
Vary input parameters of Erlang-A and display output (performance measures) in a table or graphically.

Example: $1/\mu = 2$ minutes, $1/\theta = 3$ minutes;
λ varies from 40 to 230 calls per hour, in steps of 10;
n varies from 2 to 12.

Probability to abandon



Average wait



Red curve: offered load per server fixed.
EOS (Economies-Of-Scale) observed.
Why the two graphs are similar?

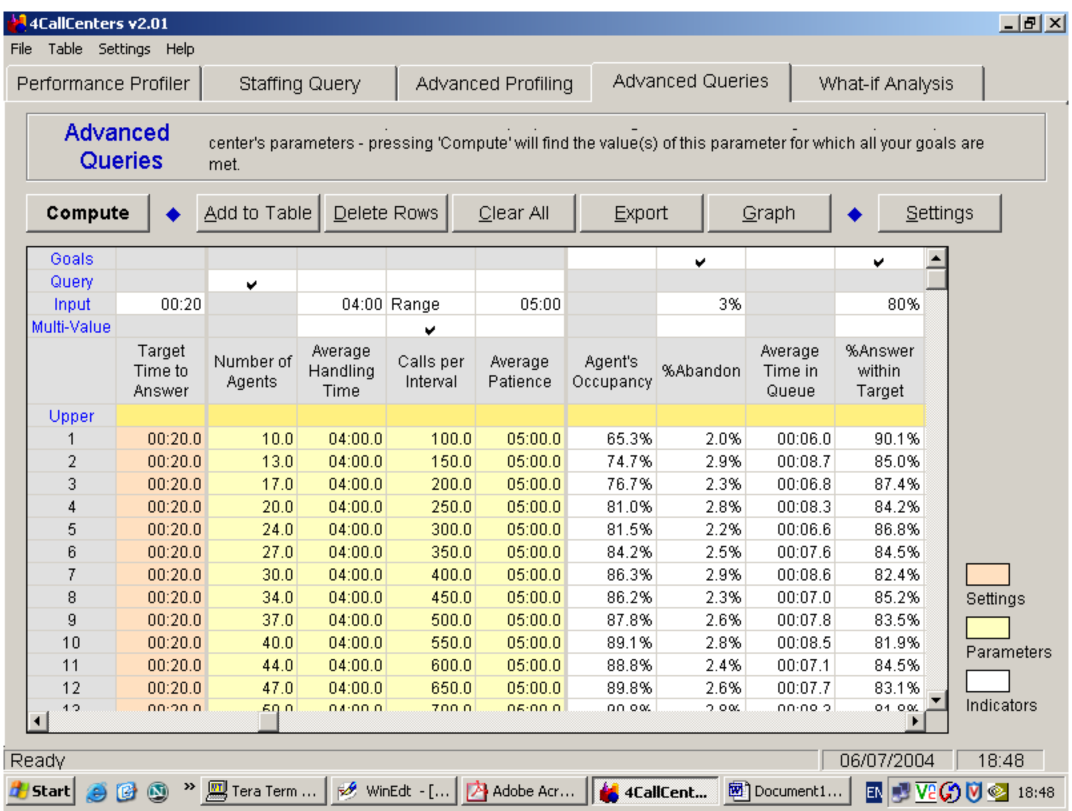
4CallCenters:
Advanced Staffing Queries

Set multiple performance goals.

Example: $1/\mu = 4$ minutes, $1/\theta = 5$ minutes;
λ varies from 100 to 1200, in steps of 50.

Performance targets:
$\mathrm{P}\{\mathrm{Ab}\} \leq 3\%$; $\quad \mathrm{P}\{W_q < 20 \text{ sec}; \ \mathrm{Sr}\} \geq 0.8$.

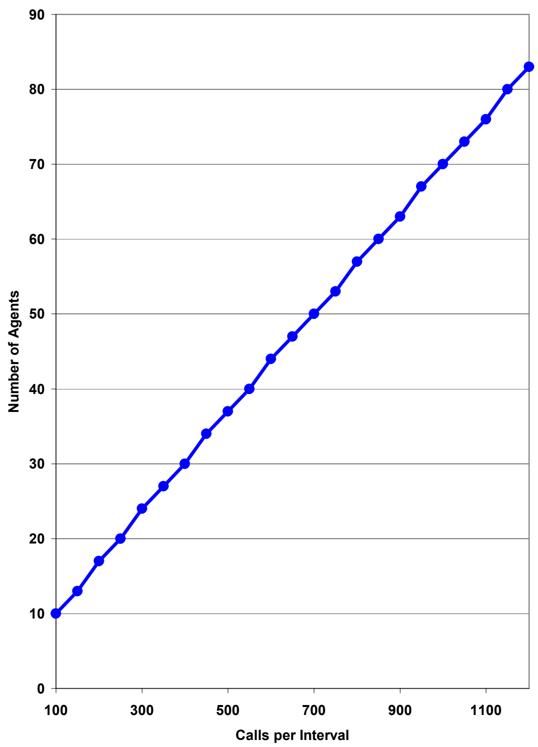
4CallCenters output

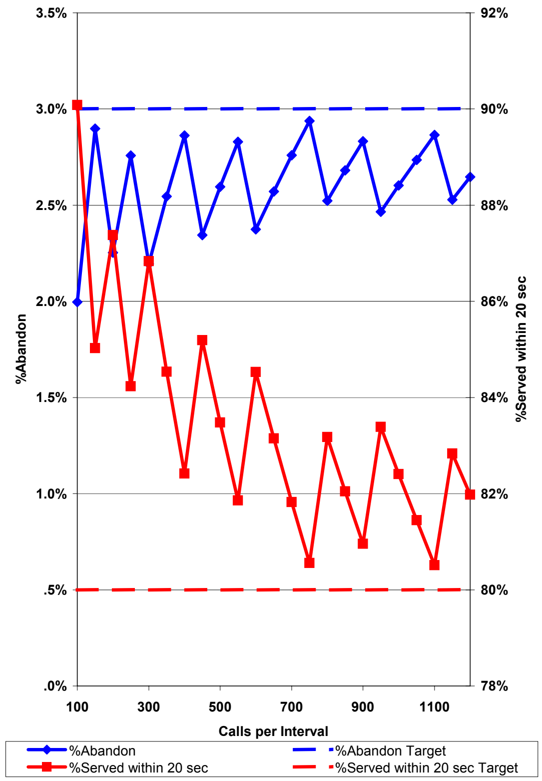


Advanced Staffing Queries II

Recommended staffing level

Target performance measures





EOS: 10 agents needed for 100 calls per hour but only 83 for 1200 calls per hour.

Call Centers:
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

Operational Regimes in Many-Server Queues

The **Quality-Efficiency Tradeoff** in services (call centers).

Offered Load: $R = \lambda \times \mathrm{E}[S]$ Erlangs, namely minutes of work (=service) that arrive per minute.

Efficiency-Driven (ED):

$$n \approx R - \gamma R, \qquad \gamma > 0.$$

Understaffing with respect to the offered load.

Quality-Driven (QD):

$$n \approx R + \delta R, \qquad \delta > 0.$$

Overstaffing with respect to the offered load.

Quality and Efficiency-Driven (QED):

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

The **Square-Root Staffing Rule:**

- Introduced by **Erlang**, already in 1924!
- Rigorized by **Halfin-Whitt**, only in 1981 (Erlang-C);
- Above version: with Garnett, Reiman, Zeltyn (Erlang-A/G).

Operational Regimes: Rules-of-Thumb

Assume that **offered load** R is not small ($\lambda \to \infty$).

ED regime:

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

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

QD regime:

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

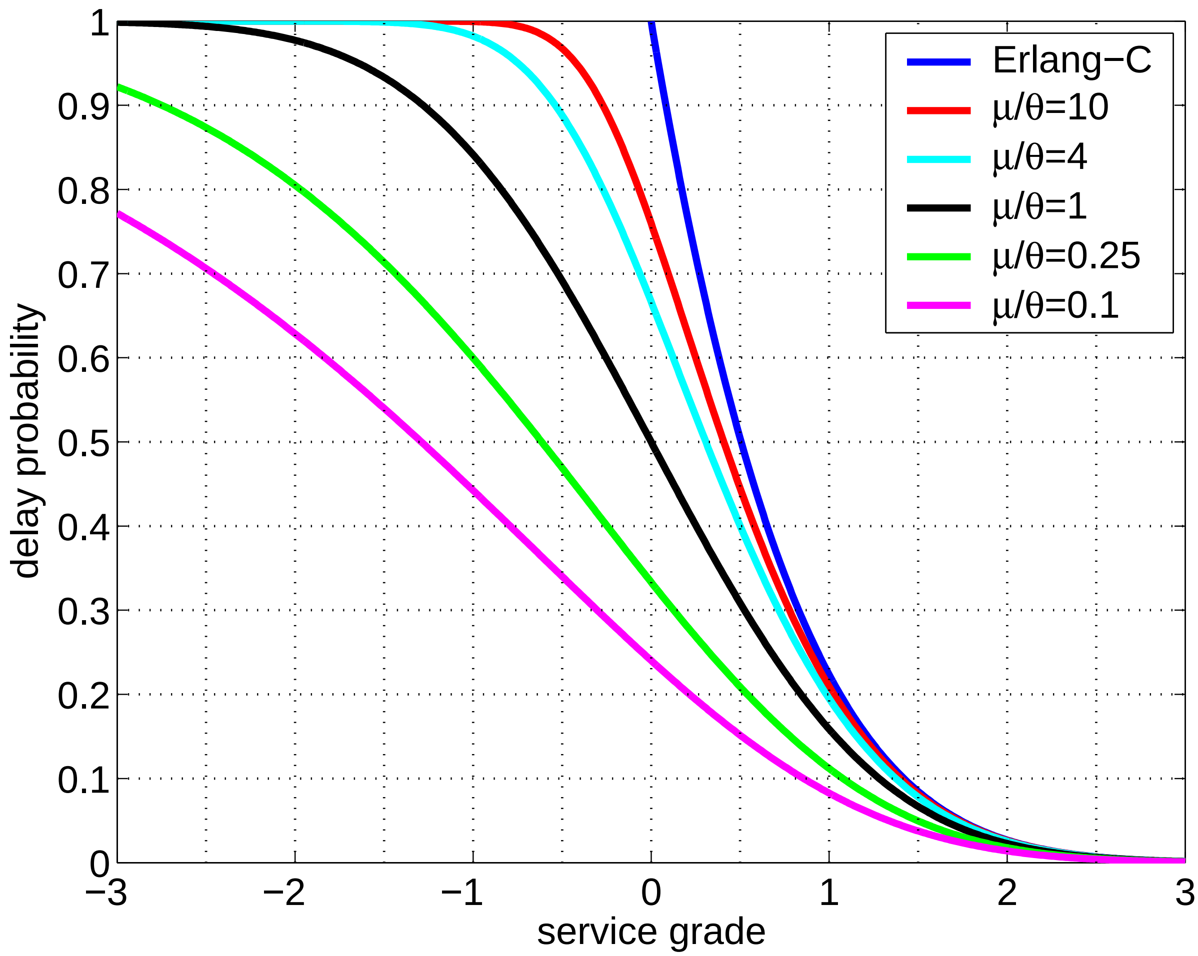
Essentially **no** delays.

QED regime:

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

- %Delayed between **25% and 75%**;
- %Abandoned is 1-5%;
- Average wait is one-order less than average service-time (seconds vs. minutes).

The QED Regime in Erlang-A: Delay Probability



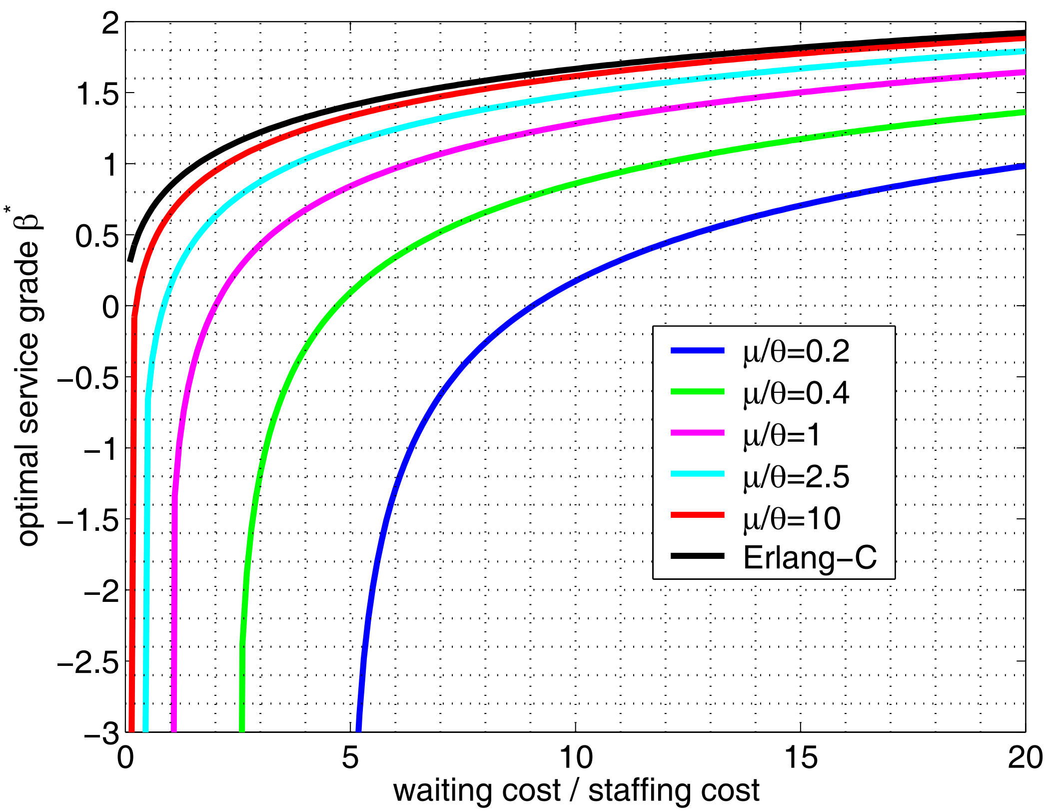
Note. **Erlang-C** is the limit of**Erlang-A**, as patience increases indefinitely.

Dimensioning Erlang-A: Optimal QoS

Cost $= c \cdot n + d \cdot \lambda \mathrm{E}[W_q]$.
(Abandonment cost can be accommodated via $P\{Ab\} = \theta E[W_q]$.)

Optimal staffing level:

$$n^* \approx R + \beta^*(r;s)\sqrt{R}, \qquad r = d/c, \quad s = \sqrt{\mu/\theta}\,,$$



- $r < \theta/\mu$ implies that "close-the-gate" is optimal.
- $r \leq 20 \;\Rightarrow\; \beta^* < 2; \quad r \leq 500 \;\Rightarrow\; \beta^* < 3$!
- **Remarkable** accuracy and robustness, via numerical tests.

Non-Parametric Queueing Models: A Basic Service Station

Assumptions:

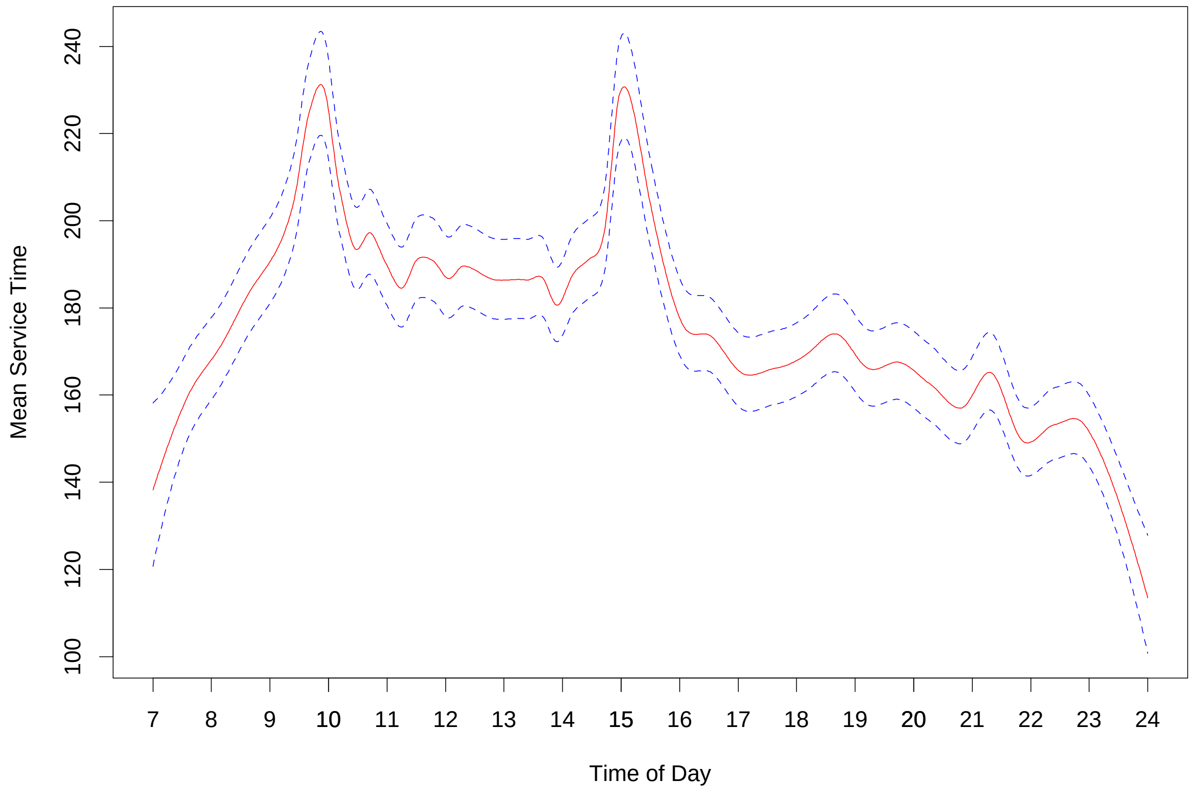
- Non-Poisson (Renewal) Arrivals;
- Non-Exponential i.i.d. Service Times;
- Non-Exponential i.i.d. (Im)Patience.

Analysis:

- Intractable Models, hence resort to Approximations;
- Single- and Moderately-Few Servers in Heavy-Traffic; (Many-Server Models with General Service Times is still a Theory in the Making);
- Steady-State Analysis;
- Two-Moment Theory: Means and Coefficients-of-Variations;
- Priorities;
- Optimal Scheduling of Customer Classes: The $c\mu$-Rule, and Relatives.

Interdependence of the Building Blocks

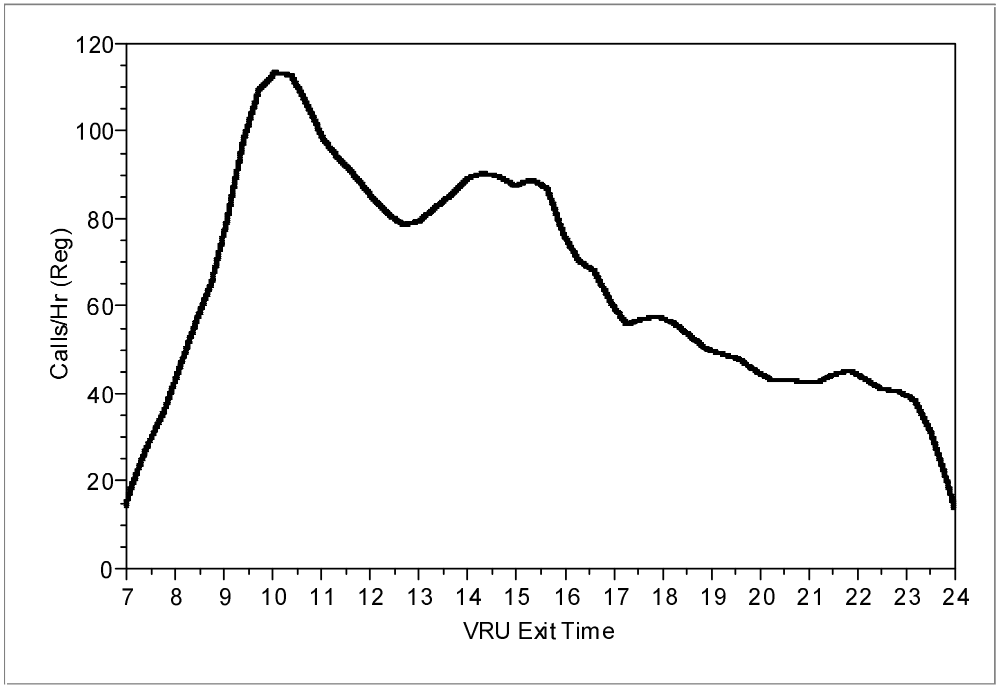
Figure 12: Mean Service Time (Regular) vs. Time-of-day (95% CI) ($n =$ 42613)



Arrival Rates: Longest Services at Peak Loads

Arrivals: Inhomogeneous Poisson

Figure 1: Arrivals (to queue or service) – "Regular" Calls



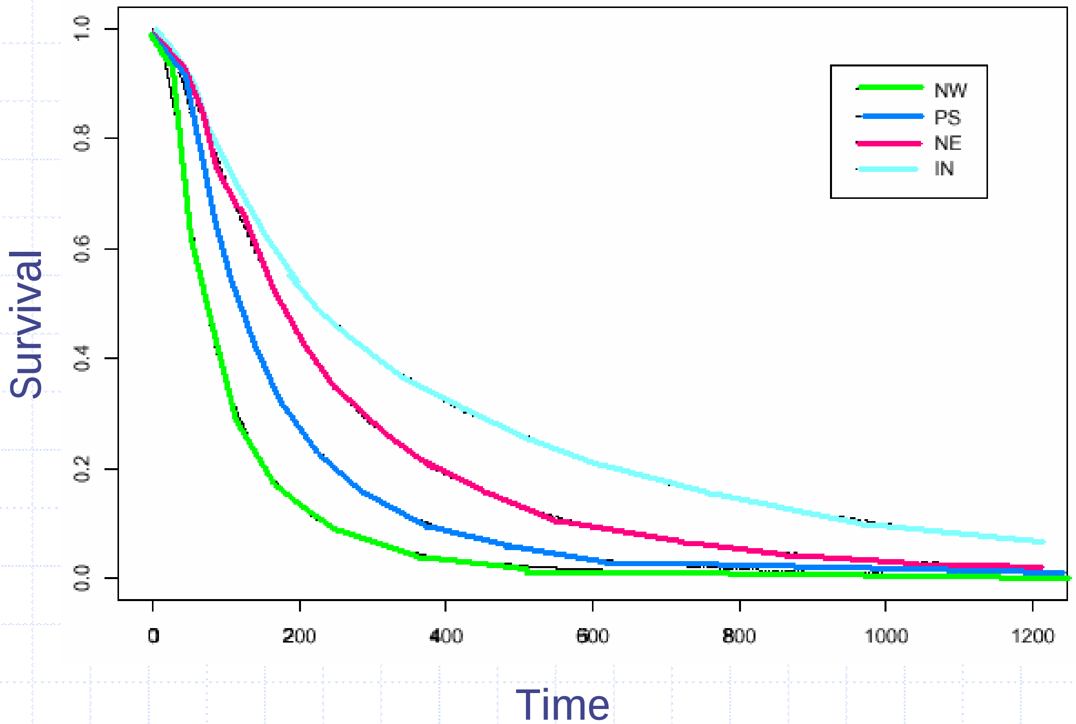
Service Time

	Overall	Regular service	New customers	Internet	Stock
Mean	188	181	111	381	269
SD	240	207	154	485	320
Med	114	117	64	196	169

84

Service Time

Survival curve, by Types



Means (In Seconds)

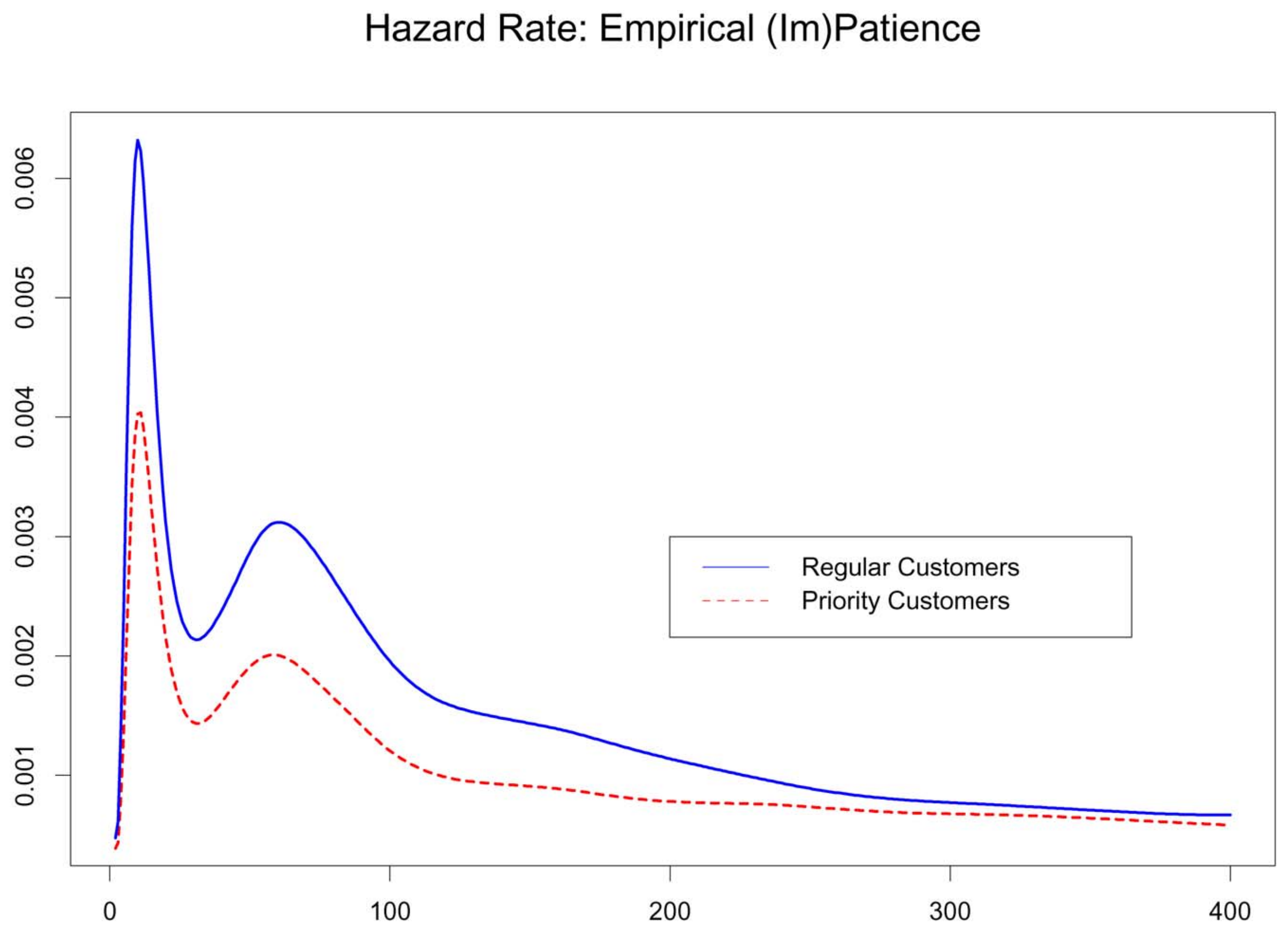
NW (New) = 111

PS (Regular) = 181

NE (Stocks) = 269

IN (Internet) = 381

(Im)Patience: Regulars vs. VIP



BONUS SUPPLEMENT: E-TAILING'S FUTURE
e.b GEN
www.businessweek.com
BusinessWeek
OCTOBER 23, 2000
A PUBLICATION OF THE McGRAW-HILL COMPANIES
Mutual Funds
How to avoid a big tax bill
Wall Street
Will tech's slide keep spreading?
Dot-coms
The search for new business models
DOT COM
Managed Care
Employers seek a new solution
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
#BXBBGDD****CAR-RT SORT**B083
#06032865631763#J010201 018489
52/INDUSTRIAL 0830
ENGINEERING LIBRARY 103
PO BOX 830657
BIRMINGHAM AL 35283-0657
AOL Keyword: BW

Customer Relationships Management

NationsBank's Design of the Service Encounter
Examples of Specifications:
Assignable Grade Of Service

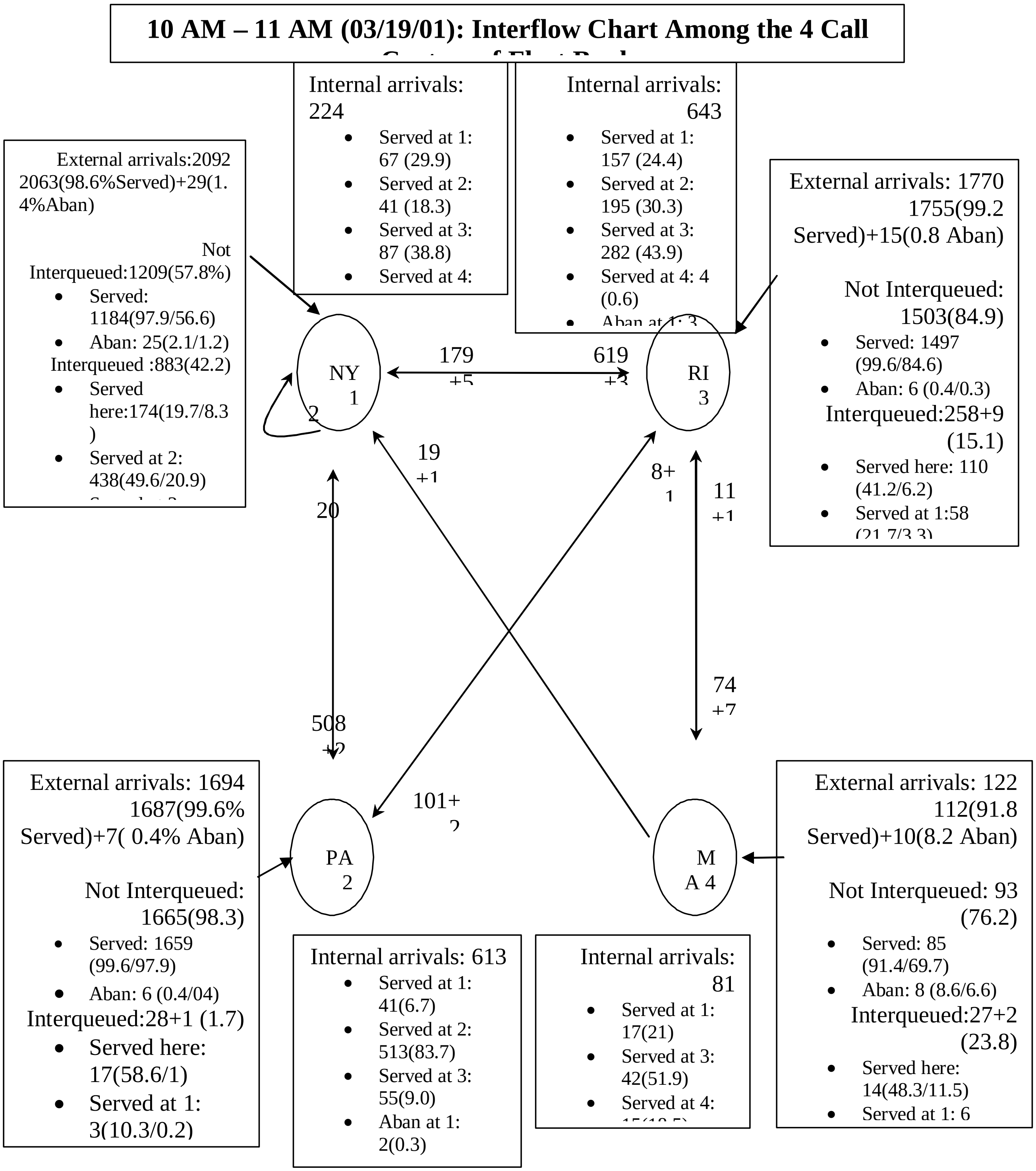
	RG1	RG2	RG3
VRU Target	70% of calls	85% of calls	90% of calls
Abandonment rate	< 1%	< 5%	< 9%
Speed of Answer	100% in 2 rings	80% in 20 seconds	50% in 20 seconds
Average Talk Time	no limit	4 min. average	2 min. average
Rep. Training	universal	product experts	basic product
Rep. Personalization	request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

NationsBank CRM: Relationship Groups:

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

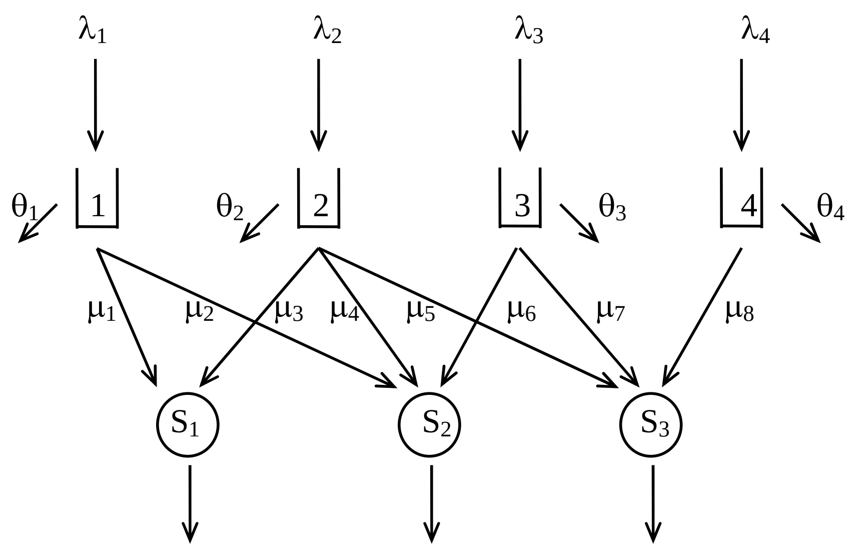
CRM = Customer **Revenue** Management

Distributed Call Center (U.S. Bank)



Skills-Base Routing: Operational Complexities

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



Here the λ's designate arrival rates, the μ's service rates, the θ's abandonment rates, and the S's are the number of servers in each server-pool.

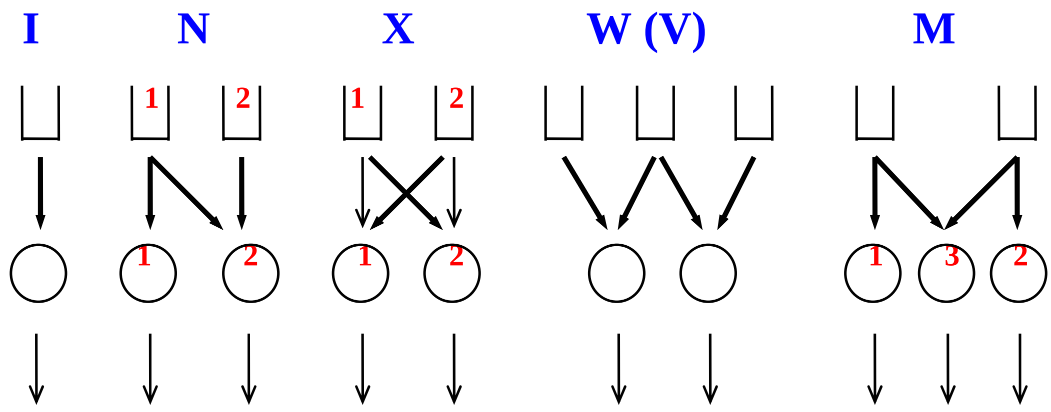
Skills-Based Design:

- **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 μ_6 customers/hour.

Customers of type 3 arrive randomly at rate λ_3 customers/hour, equipped with an impatience rate of θ_3.

Some Canonical Designs - Animation



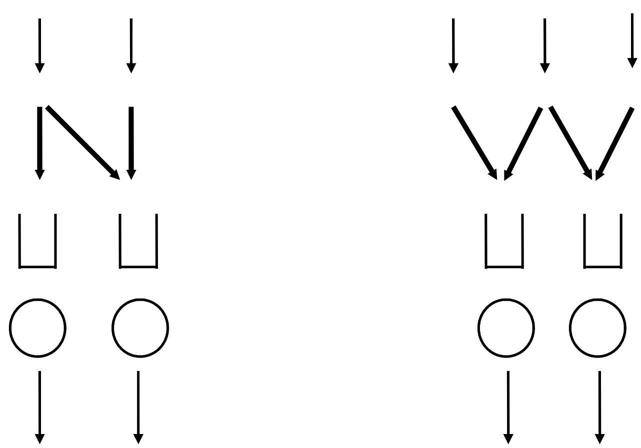
I – dedicated (specialized) agents

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)

SBR: Where are We?

Still a **challenge**, both theoretically and practically.

- "Exact" analysis of Markovian models (but mostly "queue-less"), by Koole et al.
- The ED-regime is relatively-well covered, in conventional heavy-traffic a-la Stolyar's (control) and the fluid-models of Harrison et al (staffing + control, accommodating also non-parametric models with "time-varying randomness").
- Control in the QED-regime is "theoretically-covered" by Atar et al. (exponential service-times).
- Staffing + Control in the QED-regime covers special cases: Gurvich, Armony; Dai, Tezcan; Gurvich, Whitt; ...

Still plenty to do.

Interesting and Significant Additional Topics

- Stochastic Service **Networks**:
  - Classical Markovian: Jackson and Gordon-Newell, Kelly/BCMP Networks;
  - Non-Parametric Network Approximations (QNA, SBR).
- Service Quality (Psychology, Marketing);
- Additional Significant Service Sectors: Healthcare, Hospitality, Retail, Professional Services (Consulting), ...; e-health, e-retail, e-·, ...;
- Convergence of Services and Manufacturing: After-Sale or Field Support (life-time customer-value);
- Service Supply-Chains;
- New-Service Development (or Service-Engineering in Germany);
- Design and Management of the Customer-System Interface: Multi-Media Channels; Appointments; Pricing; ...
- Revenue Management (Finite Horizon, Call Centers, ...)

Call Centers = Q's w/ Impatient Customers 15 Years History, or "A Modelling Gallery"

1. Kella, Meilijson: Practice ⇒ Abandonment important
2. Shimkin, Zohar: No data ⇒ Rational patience in Equilibrium
3. Carmon, Zakay: Cost of waiting ⇒ Psychological models
4. Garnett, Reiman; Zeltyn: Palm/Erlang-A to replace Erlang-C/B as the standard Steady-state model
5. Massey, Reiman, Rider, Stolyar: Predictable variability ⇒ Fluid models, Diffusion refinements
6. Ritov; Sakov, Zeltyn: Finally Data ⇒ Empirical models
7. Brown, Gans, Haipeng, Zhao: Statistics ⇒ Queueing Science
8. Atar, Reiman, Shaikhet: Skills-based routing ⇒ Control models
9. Nakibly, Meilijson, Pollatchek: Prediction of waiting ⇒ Online Models and Real-Time Simulation
10. Garnett: Practice ⇒ 4CallCenters.com
11. Zeltyn: Queueing Science ⇒ Empirically-Based Theory
12. Borst, Reiman; Zeltyn: Dimensioning M/M/N+G
13. Momcilovic: Non-Parametric (G/GI/N+GI) QED Q's
14. Jennings; Feldman, Massey, Whitt: Time-stable performance (ISA)