Class 2

Flow Basics
Little's Law

Flow Basics

- Inflow, Outflow (rates)
- Capacity, Utilization (Occupancy)
- Offered Load
- Resources: Servers, Highway

Little's Law

- Little's Law (Handout): Examples and Applications
- The customer/server/manager paradigm
- Scenarios: finite horizon, periodic, steady state
- Queueing/Inventory buildup diagrams
- If time permits: Brumelle's formula, leading to K-P (but without K-P).

Recitation 2: Little's Law and Capacity Analysis.

HW 2:
Solve the problems on "Capacity Utilization and Little's Law". You can solve most of the questions already now, with mere common sense (and Little's Law: $L = \lambda W$). A capacity-analysis example will be solved in today's recitation.

Reading and "Viewing" Assignment:

Part 1: **Hall, Chapter 2, on "Observations and Measurements"**
Read Hall's Chapter 2, as a review of today's class and a preview of the next one.

Part 2: **Kaplan R. and Porter M., "How to Solve the Cost Crisis in Health", HBR, September 2011**

- The journal Harvard Business Review (HBR) is read by millions.
- Robert Kaplan, Professor at the Harvard Business School (HBS), is one of the most influential people, world-wide, in Accounting.
- Michael Porter, Professor at HBS, is one of the most influential people, world-wide, in Strategy.
- They both have joined forces in order to discover, unsurprisingly I must say, that doing what we have been advocating at the Technion (especially, collecting operational data (process maps) at the level of individual patient, e.g. via RFID systems) is what MUST be done, and will eventually be done, by every healthcare organization, in particular hospitals.

  More specifically, the theme of Kaplan and Porter, as I understand it, is as follows: Measuring Operational data is a prerequisite for understanding Financial Performance, and is well correlated with Clinical performance.

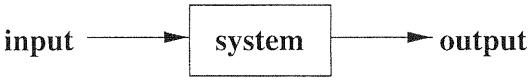
0.1

Service Engineering

Created: April 1997
Last revision: April 2007

LITTLE'S LAW

A conservation law that applies to the following general setting:



Input: Continuous flow or discrete units (examples: granules of powder measured in tons, tons of paper, number of customers, \$1000's).

System: Boundary is all that is required (very general, abstract).

Output: Same as input, call it *throughput*.

Two possible scenarios:

- System during a "cycle" (empty $\rightarrow$ empty, finite horizon);
- System in steady state/in the long run (for example, over many cycles).

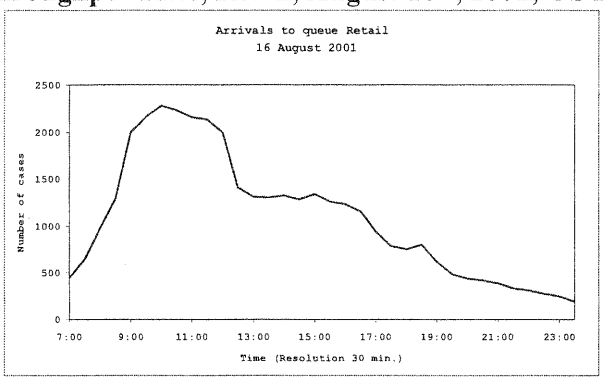
Quantities that are related via Little's law (long-run averages, or time-averages):

- λ = rate at which units *arrive*
  (= long-run average rate at which units *depart*) = *throughput-rate*, whose units are quantity/time-unit or #/time-unit;
- L = *inventory*/quantity/number in the system
  (eg. WIP: Work-In-Process, customers);
- W = time a unit spends in the system = *throughput time*
  (eg. hours) = sojourn time.

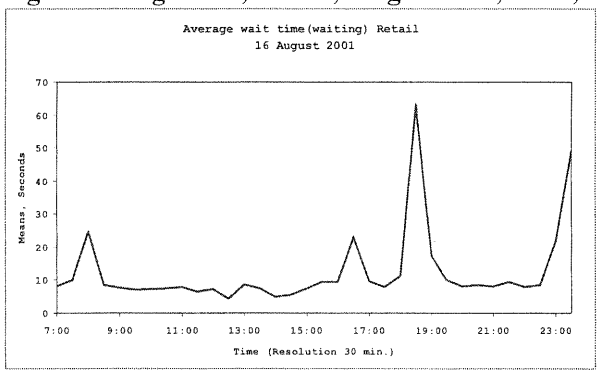
Little's Law $\boxed{L = \lambda W}$

Little's Law for Retail calls, August 16th, 2001: US Bank

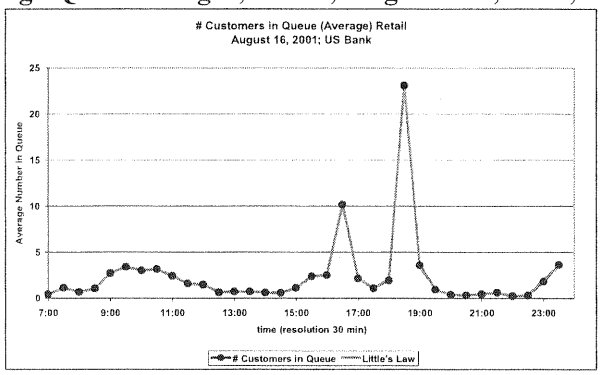
λ: Throughput Rate, Retail, August 16th, 2001; US Bank



W: Average Waiting Time, Retail, August 16th, 2001; US Bank



L: Average Queue Length, Retail, August 16th, 2001; US Bank

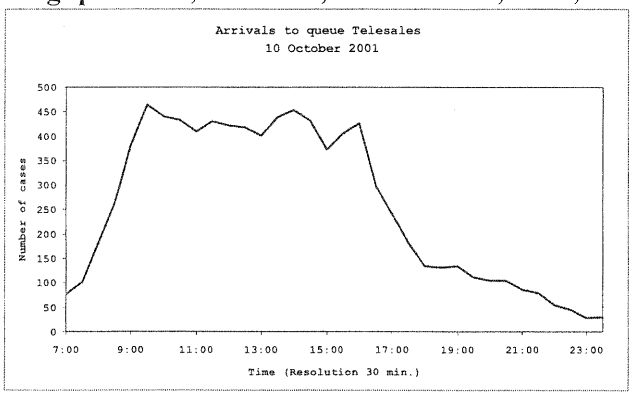


Time	7:00	7:30	8:00	8:30	9:00	9:30	10:00	10:30	11:00	11:30	12:00	12:30	13:00	13:30	14:00	14:30	15:00
λ	443	639	987	1291	1998	2166	2278	2231	2158	2135	2000	1408	1311	1303	1323	1285	1340
W	1.7	3.2	1.2	1.5	2.4	2.8	2.4	2.6	2.0	1.3	1.3	0.8	1.0	1.0	0.8	0.8	1.5
λ * W	0.42	1.14	0.68	1.06	2.72	3.42	3.01	3.18	2.44	1.55	1.47	0.64	0.72	0.72	0.62	0.59	1.09
L	0.42	1.14	0.68	1.06	2.72	3.40	3.02	3.17	2.41	1.59	1.48	0.64	0.72	0.72	0.62	0.57	1.11

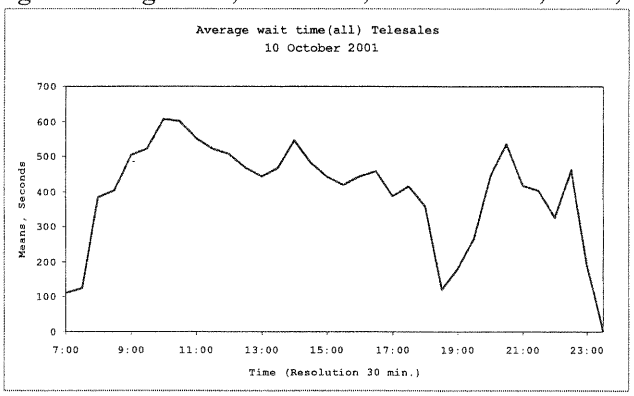
Time	15:30	16:00	16:30	17:00	17:30	18:00	18:30	19:00	19:30	20:00	20:30	21:00	21:30	22:00	22:30	23:00	23:30
λ	1258	1235	1157	942	788	752	803	619	485	437	421	386	336	311	274	251	193
W	3.5	3.6	15.8	4.2	2.4	4.9	51.9	10.0	3.5	1.7	1.3	2.1	3.3	1.4	2.0	14.3	32.6
λ * W	2.422	2.45	10.2	2.173	1.06	2.05	23.16	3.43	0.95	0.41	0.314	0.44	0.62	0.24	0.30	2.00	3.50
L	2.37	2.49	10.17	2.16	1.07	1.94	23.11	3.59	0.95	0.40	0.31	0.45	0.62	0.24	0.30	1.83	3.63

Little's Law for Telesales calls, October 10th, 2001: US Bank

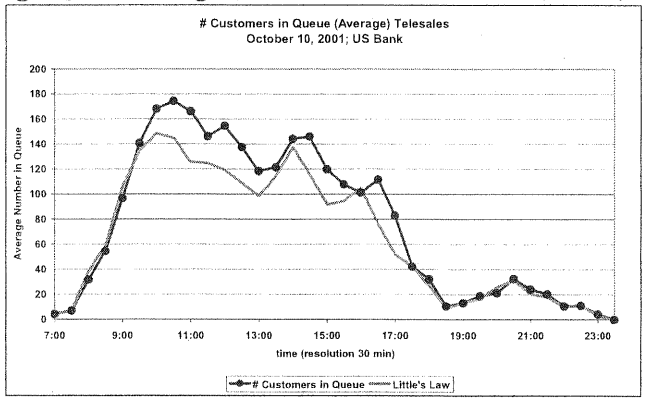
λ: Throughput Rate, Telesales, October 10th, 2001; US Bank



W: Average Waiting Time, Telesales, October 10th, 2001; US Bank



L: Average Queue Length, Telesales, October 10th, 2001; US Bank



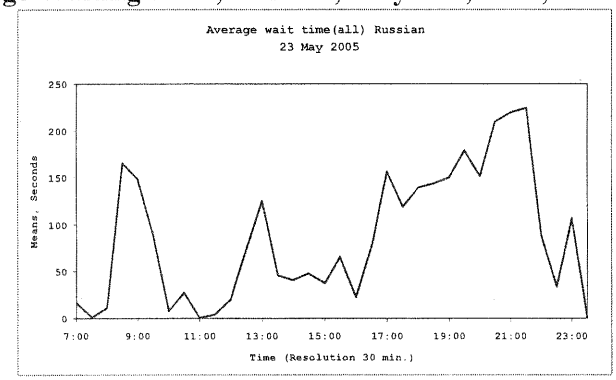
Time	7:00	7:30	8:00	8:30	9:00	9:30	10:00	10:30	11:00	11:30	12:00	12:30	13:00	13:30	14:00	14:30	15:00
λ	76	102	182	262	379	464	440	433	410	431	422	418	401	439	453	432	373
W	109.8	123.8	383.5	403.7	503.5	522.5	607.9	602.1	552.4	521.1	508.6	468.8	442.1	467.3	545.9	483.1	442.1
λ * W	4.63	7.01	38.77	58.76	106.01	134.69	148.60	144.84	125.82	124.77	119.23	108.86	98.48	113.98	137.39	115.93	91.61
L	4.28	6.91	31.73	54.36	96.50	140.70	168.10	174.34	166.14	146.13	154.48	137.47	118.29	121.44	144.07	146.01	119.83

Time	15:30	16:00	16:30	17:00	17:30	18:00	18:30	19:00	19:30	20:00	20:30	21:00	21:30	22:00	22:30	23:00	23:30
λ	405	427	298	242	182	134	132	134	112	105	105	87	80	55	45	28	30
W	419.2	442.2	458.8	387.9	415.1	357.1	121.6	179.8	267.9	445.7	536.0	416.9	403.9	326.0	463.6	187.3	0.9
λ * W	94.31	104.89	75.96	52.15	41.97	26.58	8.92	13.38	16.67	26.00	31.27	20.15	17.95	9.96	11.59	2.91	0.02
L	107.86	101.22	111.60	82.93	42.23	32.32	10.57	13.24	18.67	21.07	32.50	24.10	20.33	10.69	11.13	4.35	0.02

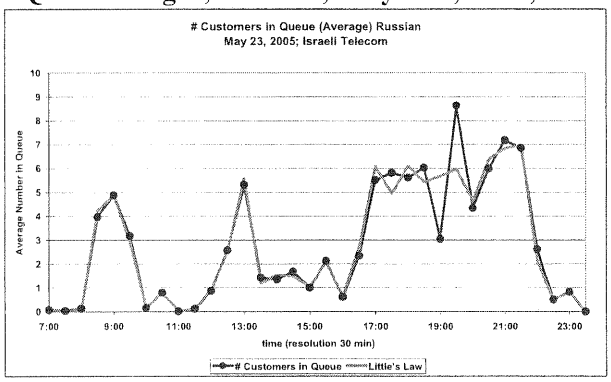
Little's Law for Russian calls, May 23rd, 2005: Israeli Telecom

λ: Throughput Rate, Russian, May 23rd, 2005; Israeli Telecom

W: Average Waiting Time, Russian, May 23rd, 2005; Israeli Telecom



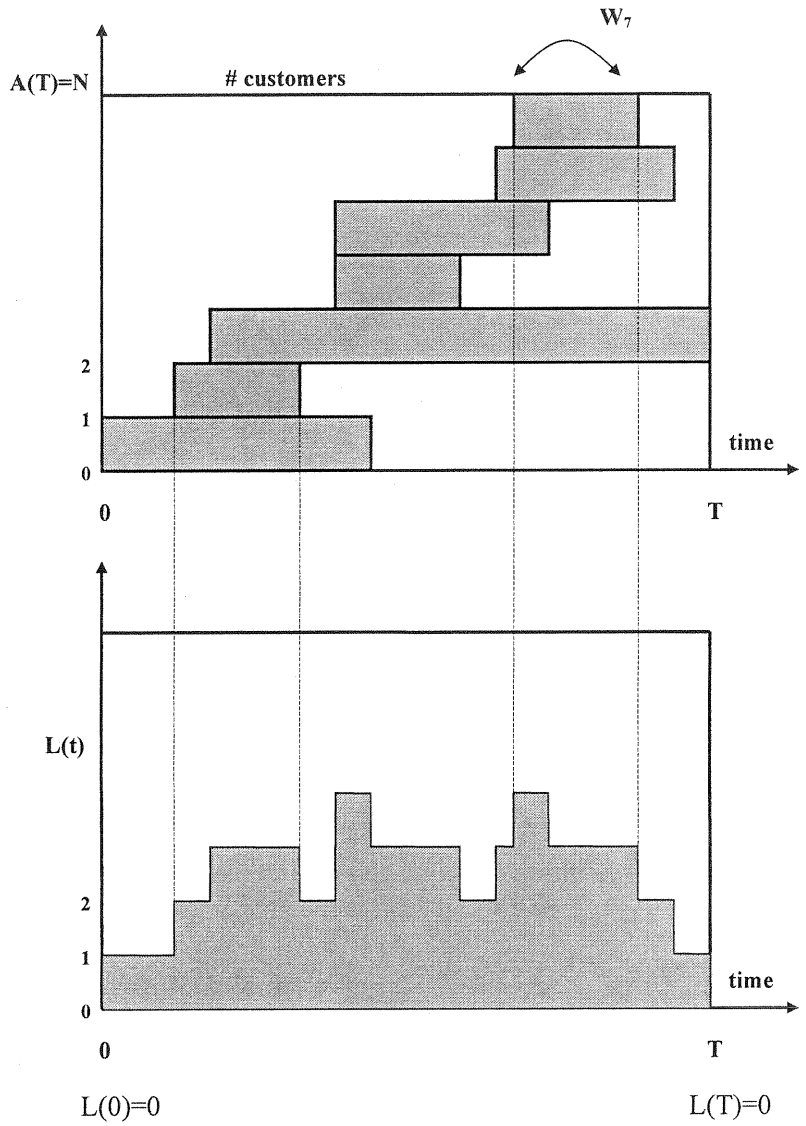
L: Average Queue Length, Russian, May 23rd, 2005; Israeli Telecom



Time	7:00	7:30	8:00	8:30	9:00	9:30	10:00	10:30	11:00	11:30	12:00	12:30	13:00	13:30	14:00	14:30	15:00
λ	12	12	22	46	59	59	36	52	43	56	81	61	80	46	67	56	50
W	16.9	1.3	11.4	166.0	148.9	88.7	7.9	27.4	0.7	3.9	20.3	74.9	125.4	46.3	41.4	47.9	37.7
λ * W	0.11	0.01	0.14	4.24	4.88	2.91	0.16	0.79	0.02	0.12	0.91	2.54	5.57	1.18	1.54	1.49	1.05
L	0.08	0.04	0.14	3.97	4.88	3.18	0.16	0.79	0.02	0.12	0.88	2.57	5.32	1.44	1.36	1.67	1.00

Time	15:30	16:00	16:30	17:00	17:30	18:00	18:30	19:00	19:30	20:00	20:30	21:00	21:30	22:00	22:30	23:00	23:30
λ	57	52	62	70	75	79	68	68	60	55	55	56	56	43	27	14	6
W	65.7	22.4	79.2	156.6	118.6	139.3	143.6	150.2	179.2	151.3	209.5	219.5	224.7	88.4	33.7	107.0	0.7
λ * W	2.08	0.65	2.73	6.09	4.94	6.11	5.42	5.68	5.97	4.62	6.40	6.83	6.99	2.11	0.51	0.83	0.00
L	2.13	0.62	2.34	5.51	5.82	5.61	6.03	3.04	8.63	4.34	5.99	7.18	6.85	2.61	0.51	0.83	0.00

6



Note: Vertical cut = number of customers in the system.

S = Shaded area (units: customer × hours), measures total waiting.

$W = \dfrac{S}{N}$, divides waiting among customers
(the **customer's** view).

$L = \dfrac{S}{T}$, divides waiting over time
(the **manager's** view).

$\lambda = \dfrac{N}{T}$, implies $L = \lambda \cdot W$
(which adds the **server's** view).

5.2 Cars flow over a single-loop detector, that can measure Occupancy = % time there is a car above the detector;
Flow = avg. # cars per hour.

System = Detector

L = Occupancy (E [Indicator])

λ = Flow

W = $\frac{\ell}{V}$ time to traverse one detector

where V = Velocity, ℓ = av. car length.

By Little's Law:

$$\boxed{\text{Occupancy} = \frac{\text{Flow} \times \text{car-length}}{\text{Velocity}}} \times 100\%$$

Note: Occupancy = Density × car-length.

5.3 Empirically, transportation flow reveals the following "flow vs. occupancy" relation ("flow vs. density" would look the same):

From "Causes and Cures of Highway Congestion", Chao Chen, Zhanfeng Jia and Pravin Varaiya, 2001

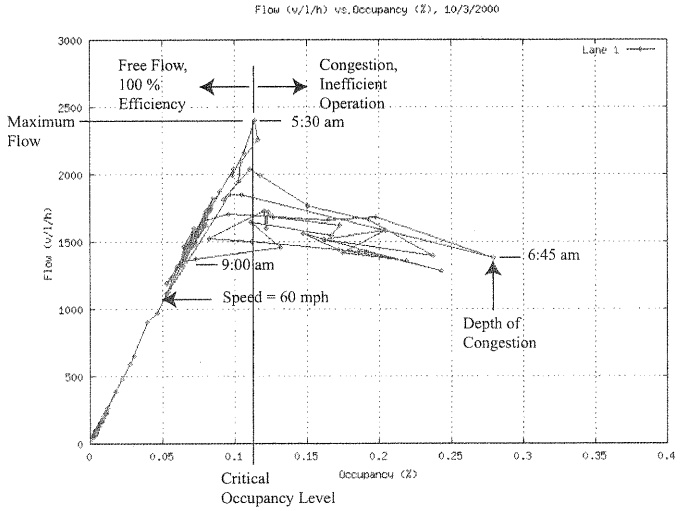


Figure 6: Flow vs. occupancy on a section at postmile 37.18 on I-10W, midnight to noon on October 3, 2000.

10

The **critical occupancy** is the occupancy-level beyond which congestion starts building up.

Note: For each point on the curve, the slope of the line connecting it with the origin is proportional (equal) to the velocity; indeed:

$$\frac{\text{Flow}}{\text{Occupancy}} = \frac{\text{Velocity}}{\text{Car-lenght}}; \quad \frac{\text{Flow}}{\text{Density}} = \text{Velocity}$$

This explains the (almost) straight line to the left of the critical occupancy: its slope is the congestion-free velocity (60 miles/hr in California highways).

Note: with a single-loop detector covering N **lanes**, and assuming that traffic is evenly divided among the lanes (though typically this is not the case), the Occupancy should be calculated by using Flow/N, instead of merely Flow.

6. **Abandonment:** Calls arrive at a call center at rate α. A fraction P_{ab} of them abandons due to impatience. Individual abandonment rate is θ.

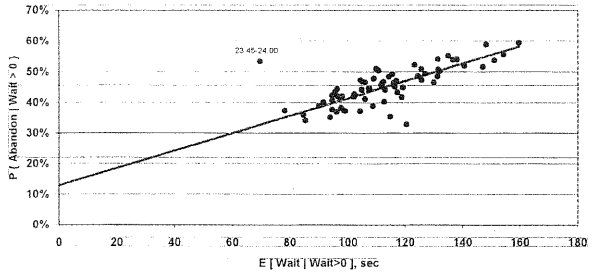
   Let L_q, W_q denote, respectively, the average number of customers waiting to be served, and the average queueing time (waiting for service). Then

   $$\alpha \cdot P_{ab} = \theta \cdot L_q.$$

   But $L_q = \alpha W_q$, hence

   $$P_{ab} = \theta \cdot W_q.$$

   Thus, the abandonment rate is proportional to the average waiting time. This has been confirmed empirically for new (potential) customers. Indeed, (P_{ab}, W_q) were observed and scatterplotted. The slope (via regression) can be used to estimate customers' (average) patience.



The data is from a bank call center. Each point corresponds to a 15-minute period of a day (Sunday to Thursday), starting at 7:00am. ending at midnight. and averaged over the whole year of 1999.

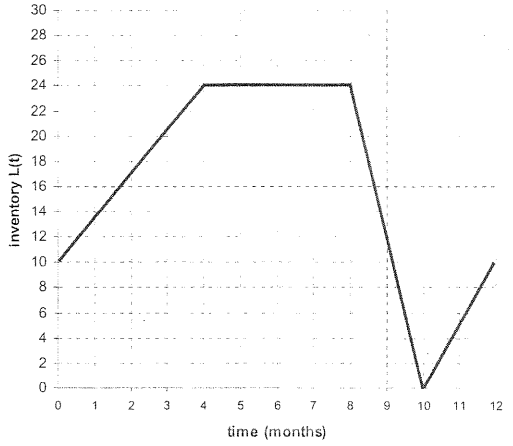
- Why a positive y-intercept?
- What about *experienced* customers?

7. **Loan Application Flow** from *Managing Business Process Flows.* by R.Anupindi, S.Chopra, S.Deshmukh, J.Van Mieghem, E.Zemel, Chapter 3. (In Recitation.)

8. **Process Flow**: A supermarket receives from suppliers 300 tons of fish over the course of a full year, which averages out to 25 tons per month. The average quantity of fish held in freezer storage is 16.5 tons.
   On average, how long does a ton of fish remain in freezer storage between the time it is received and the time it is sent to the sales department?

   $W = L/\lambda = 16.5/25 = 0.66$ months, on average, is the period that a ton of fish spends in the freezer.
   How does one get $L = 16.5$? This comes out of the following inventory build-up diagram by calculating the area below the graph:

Inventory/Queue Build-up Diagram.

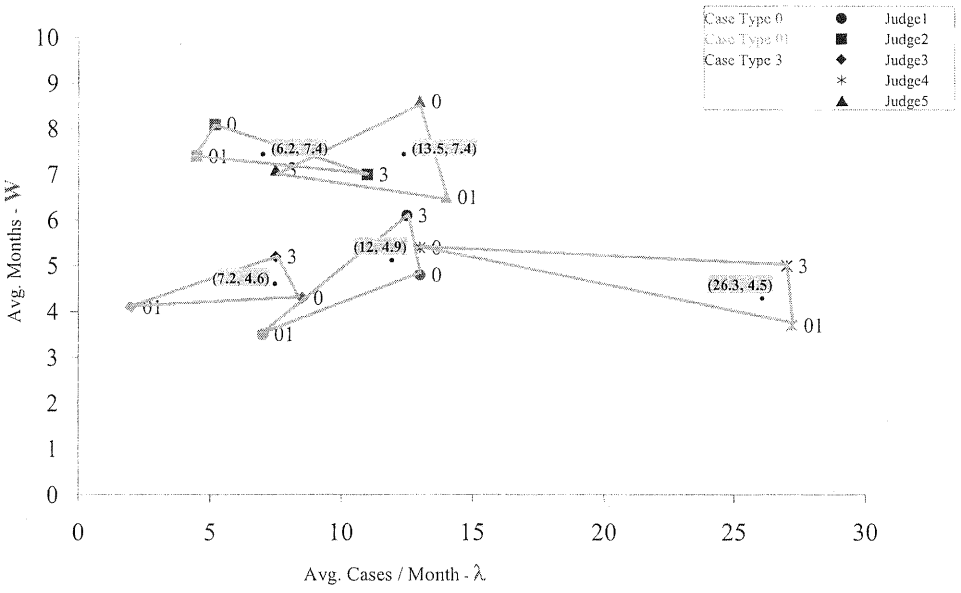


$$17 \times \tfrac{4}{12} + 24 \times \tfrac{4}{12} + 12 \times \tfrac{2}{12} + 5 \times \tfrac{2}{12} =$$
$$\tfrac{17}{3} + 8 + 2 + \tfrac{5}{6} = 16.5.$$

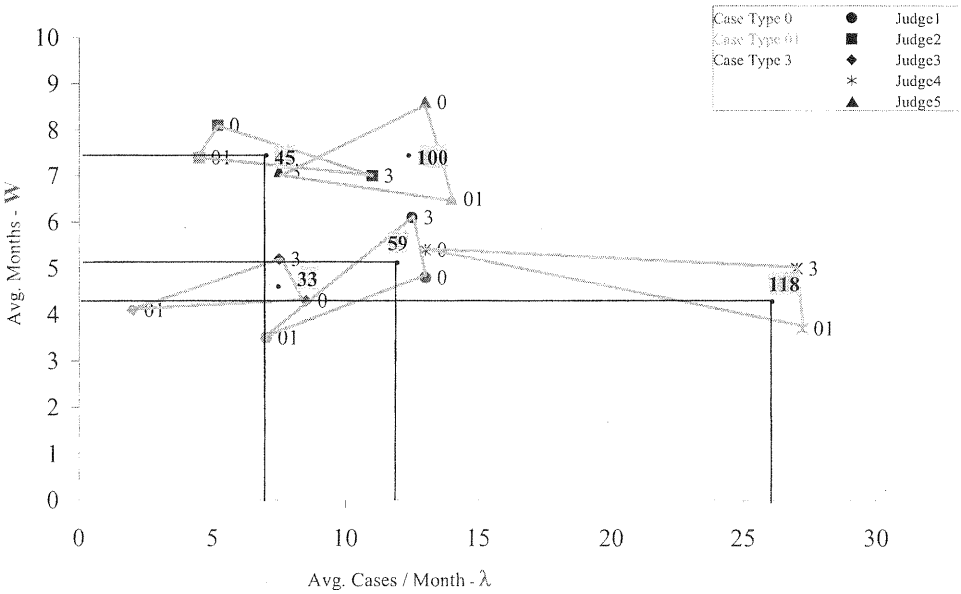
15. **Little's Law in the "Production of Justice".**

- 5 Judges "process" 3 types of files.
- System = "drawer" of a Judge.

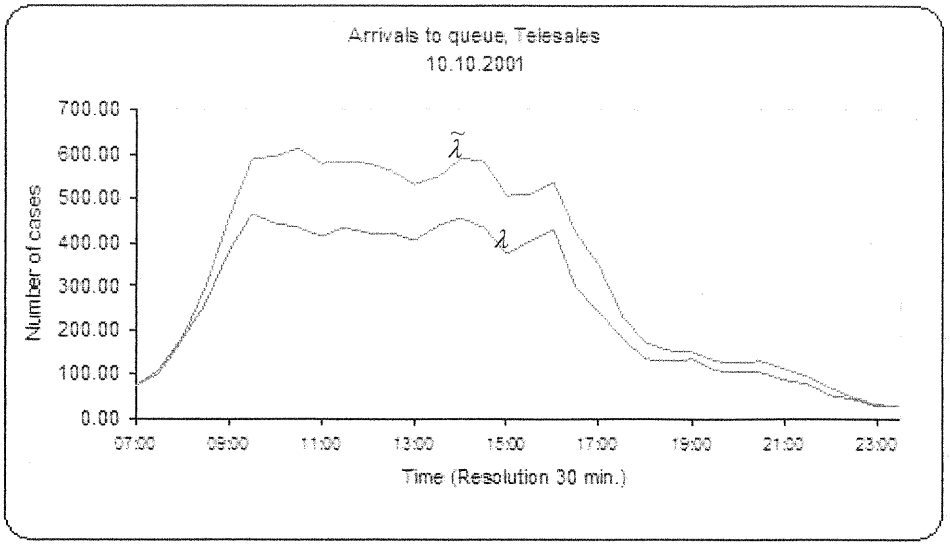
Judges: Performance Analysis (λ, W)

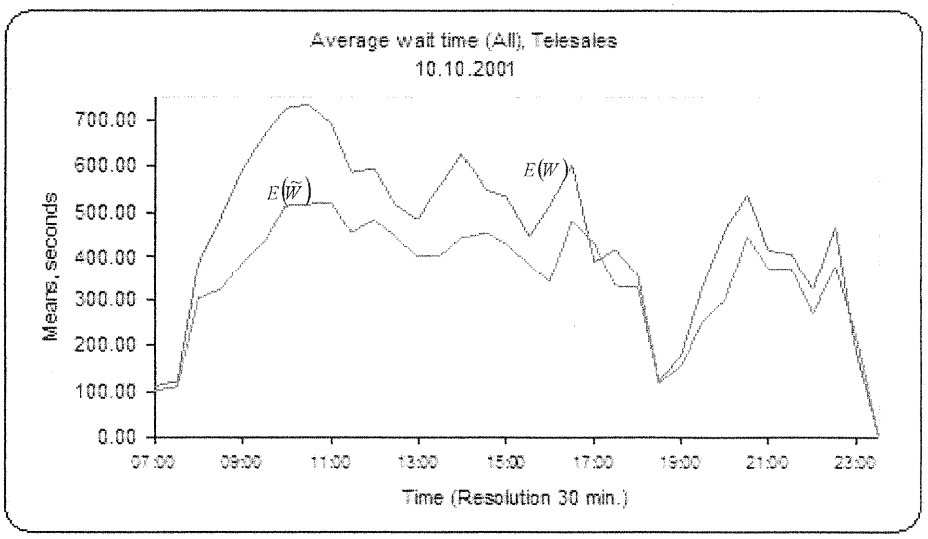


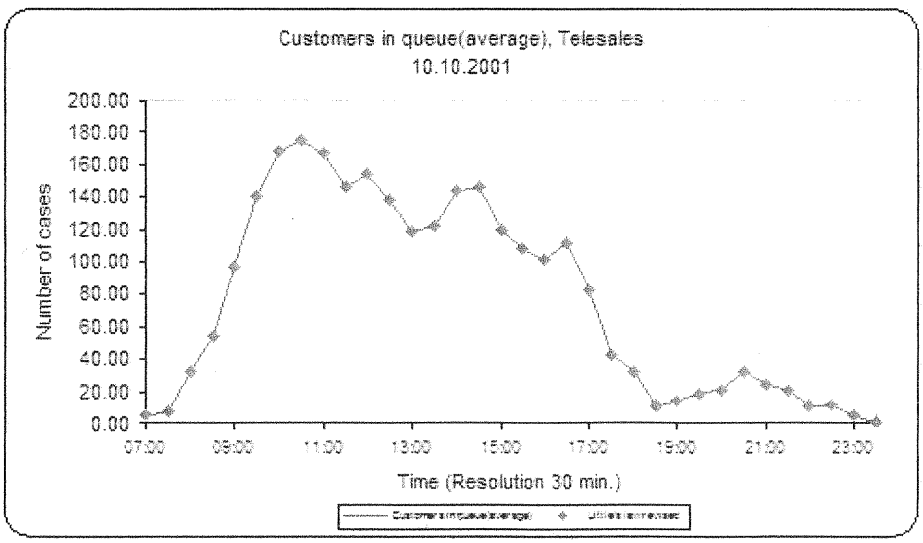
Judges: Performance Analysis (L)



14







Note: Recall that waiting-times are not truncated to 30 minutes, the latter being SEEStat standard.

Claim (Abin Koren, 2011)

Let $t_1 < t_2$.

a. If $V(t_1) = V(t_2)$ then on the interval $[t_1, t_2)$, $L = \lambda \cdot W$.

b. The absolute difference between L and $\lambda \cdot W$ is given by: $\lambda W - L = \frac{V(t_2) - V(t_1)}{t_2 - t_1}$

Little's Law: Finite-Horizon (SEEStat)

Explanation (for 'b' only as it is a generalization of 'a'):

Recall that in an interval which starts with an empty system and ends with an empty system, $L = \lambda \cdot E(W)$.

Let us force this condition on some interval $[t_1, t_2)$ by treating all entities of Type 3 from above (i.e. entities which arrive before t_1) as if they arrived exactly at time t_1; and treating all entities of Type 2 from above (i.e. entities which depart after t_2) as if they depart exactly at time t_2).

We now compute L according to Little's Law:

$$L = \underbrace{\frac{\#\,entities\ that\ spent\ some\ time\ in\ the\ system}{t_2 - t_1}}_{\lambda} \cdot \underbrace{\frac{\sum sojourn\ times\ enclosed\ within [t_1, t_2)}{\#\,entities\ that\ spent\ some\ time\ in\ the\ system}}_{E(W)}$$

Next, recall how SEEStat computes the arrival rate and average waiting time and finds their product:

$$\lambda \cdot E(W) = \frac{\#\,Arrivals\ during [t_1, t_2)}{t_2 - t_1} \cdot \frac{\sum sojourn\ times\ of\ entities\ to\ arrive\ during [t_1, t_2)}{\#\,Arrivals\ during [t_1, t_2)}.$$

But:

$$\sum sojourn\ times\ enclosed\ within [t_1, t_2) =$$
$$\sum sojourn\ times\ of\ entities\ to\ arrive\ during [t_1, t_2) + V(t_1) - V(t_2)$$

Hence:

$$L - \lambda \cdot E(W) = \frac{V(t_1) - V(t_2)}{t_2 - t_1}$$

Let us return to the example above:

$$\lambda = \frac{4}{7};\ E(W) = \frac{2+4+5+3}{4} = \frac{14}{4} \rightarrow \lambda \cdot E(W) = 2$$
$$L = \frac{6}{7} \cdot \frac{3+7+2+4+3+2}{6} = \frac{21}{7} = 3$$

And indeed:

$$\frac{V(t_2) - V(t_1)}{t_2 - t_1} = \frac{5-12}{7} = -1.$$

$$\Delta L = \frac{V(18:30) - V(18:00)}{18:30 - 18:00} = \frac{189 - 5}{1800} = 0.102 \ .$$

This matches the difference presented previously, when putting L against $\lambda \cdot W$.
Our analysis may well explain why, on this day, we see that Little's Law 'works properly'.

Finally, we examine an interval where we previously found Little's Law does NOT work: Consider October 10th 2001 in Telesales of US Bank call center, during 11:00-11:30. It is possible to compute the remaining work at both the beginning and the end of the interval. We get (note how loaded the call center is):

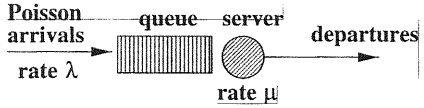
$$V(11:00) = 170,670 \text{ sec},$$
$$V(11:30) = 154,634 \text{ sec}.$$

Therefore:

$$\lambda \cdot W - L = \frac{154,634 - 170,670}{1800} = \frac{-16036}{1800} = -8.91.$$

Comparing this result to the measured difference in the corresponding graph above (note that you should compare it to the difference found in the graph with untruncated waiting times), we get the exact same difference.

Stochastic example: M/M/1



Model

Birth-and-death process, birth rate λ, death rate μ.

Assumption

$\rho = \frac{\lambda}{\mu} < 1$, answers existence of stationary (limit) distribution π:

$$\begin{aligned} \pi_k &= (1-\rho)\rho^k, \quad k = 0, 1, 2, \ldots \text{ (geometric distribution).} \\ L &= \sum_{k=0}^{\infty} k\pi_k = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu - \lambda}\,. \end{aligned}$$

Little: $W = \frac{1}{\lambda}\, L = \frac{1}{\mu - \lambda} = \frac{1}{\mu}\, \frac{1}{1-\rho}$.

Check out:

$$\begin{aligned} W &= \text{(PASTA)} = \sum_{k=0}^{\infty} E[\text{sojourn time}/k \text{ customers in system}]\, \pi_k \\ &= \text{(memoryless property)} = \sum_{k=0}^{\infty} \left[\frac{k}{\mu} + \frac{1}{\mu}\right] \pi_k \\ &= \frac{1}{\mu} + \frac{1}{\mu}\, L = \cdots = \frac{1}{\mu}\, \frac{1}{1-\rho}\,. \end{aligned}$$

System = queue: $L_q = \lambda\, W_q$, $W_q = W - \frac{1}{\mu} = \frac{1}{\mu}\, \frac{\rho}{1-\rho}$.

L_q – queue-length,
W_q – waiting-time.

System = server:

$L = \lambda \cdot \frac{1}{\mu}$;
$L = \rho$ = probability that the system is not empty (customer waits)
= proportion of time when the server is busy *(traffic intensity)*.

Stochastic Model (à la Serfozo[1])

$\{(A_n, D_n),\ n \geq 1\}$ random variables; limits are a.s. (with probability 1)

e.g. $\lambda = \lim_{t \uparrow \infty} \frac{1}{t} A(t)$ a.s. ; $\frac{1}{T} \int_0^T L(t)dt \to L$ a.s., as $T \uparrow \infty$, etc.

"**Periodic**" System (Serfozo, pg. 17)

A system is *periodically empty* if there exist strictly increasing random times $\tau_n \uparrow \infty$, such that

1. $\tau_n \sim \tau_{n+1}$ i.e. $\lim_{n \uparrow \infty} \frac{\tau_{n+1}}{\tau_n} = 1$ a.s. (implied, for example, by $\tau_n/n \to c$).

2. For all n, there exists $t \in [\tau_n, \tau_{n+1})$ such that $A(t) = D(t)$, i.e. $L(t) = 0$.

Theorem. *If a system is periodically empty, the existence of any two positive limits out of (L, λ, W) implies existence of the third, as well as the relation $L = \lambda W$.*

Typical application: τ_n starts a "cycle" (eg. empty system; state 7), which gives rise to a regenerative structure (eg. Markovian).

[1] Introduction To Stochastic Networks, Springer 1999, Chapter 5