Sergey Zeltyn January 2005

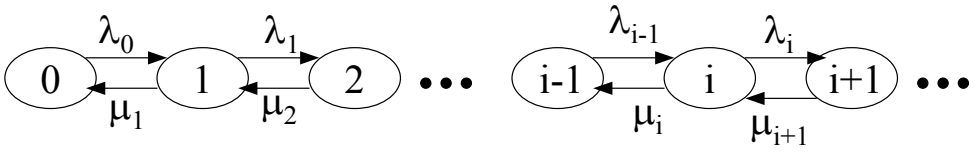
STAT 991. Service Engineering.
The Wharton School. University of Pennsylvania.

Classical Queueing Models.

Based on:

- Mandelbaum A. *Service Engineering* course, Technion.
  `http://iew3.technion.ac.il/serveng2005W`
- General knowledge on classical queueing models.
  (E.g. Wolff. *Stochastic Modelling and the Theory of Queues.*)
- 4CallCenters software: examples of output.

Birth & Death Model of a Service Station



- i – number-in-system;
- λ_i – arrival rate given i customers in system;
- μ_i – service rate given i customers in system.

Cuts at $i \leftrightarrow i+1$ yield:

$\pi_i \lambda_i = \pi_{i+1} \mu_{i+1},\ i \geq 0$, and

$$\pi_{i+1} = \frac{\lambda_i}{\mu_{i+1}}\ \pi_i = \frac{\lambda_i \lambda_{i-1}}{\mu_{i+1}\mu_i}\ \pi_{i-1} = \cdots = \frac{\lambda_0 \lambda_1 \ldots \lambda_i}{\mu_1 \mu_2 \ldots \mu_{i+1}}\ \pi_0\ .$$

Steady-state distribution exists iff

$$\sum_{i=0}^{\infty}\ \frac{\lambda_0 \ldots \lambda_i}{\mu_1 \ldots \mu_{i+1}} < \infty\ .$$

Then

$$\begin{cases} \pi_i &= \frac{\lambda_0 \ldots \lambda_{i-1}}{\mu_1 \ldots \mu_i}\ \pi_0\ ,\ i \geq 0 \\ \pi_0 &= \left[\sum_{i \geq 0}\ \frac{\lambda_0 \ldots \lambda_i}{\mu_1 \ldots \mu_{i+1}}\right]^{-1} \end{cases}$$

Additional assumptions (classical queues):

- n statistically identical servers;
- FCFS discipline – First Come First Served;
- Work conservation: a server does not go idle if there are customers in need of service;
- Customers do not abandon.

Measures of Performance

- L - number of customers at the service station;
- L_q - number of customers in the queue;
- W - sojourn time of a customer at the service station;
- W_q - waiting time of a customer in the queue.

In steady state,

$$\mathrm{E}[L] \;=\; \sum_{k\geq 0} k\cdot\pi_k \;=\; \lim_{T\to\infty}\;\frac{1}{T}\cdot\int_0^T L(t)dt\,.$$

$$\mathrm{E}[L_q] = \sum_{k=n+1}^{\infty}(k-n)\cdot\pi_k\,.$$

If λ – arrival rate to system, Little's formula implies:

$$\mathrm{E}[L] = \lambda\cdot\mathrm{E}[W]; \qquad \mathrm{E}[L_q] = \lambda\cdot\mathrm{E}[W_q]\,.$$

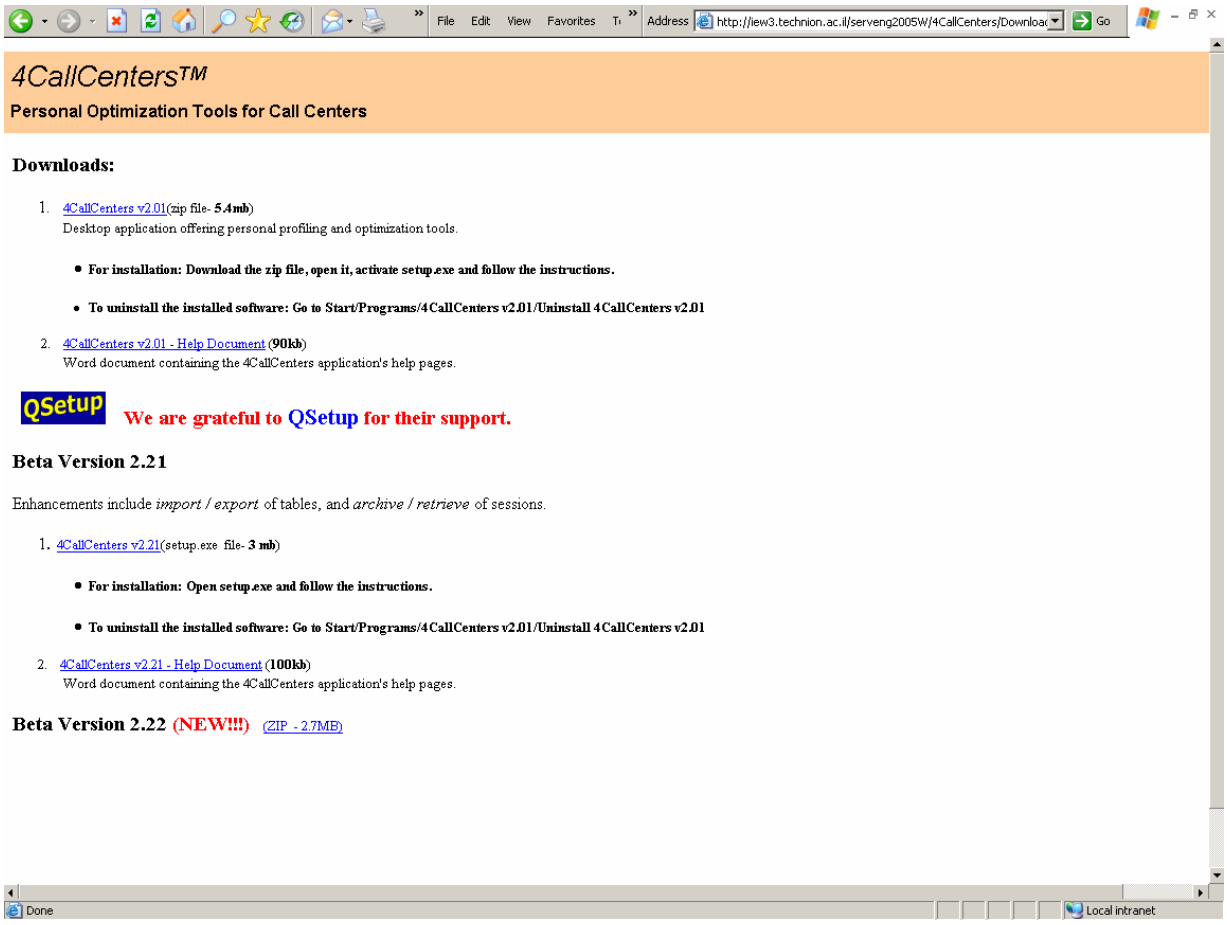
4CallCenters Software.

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

Will be extensively used in our course.

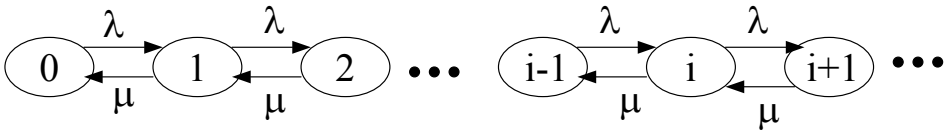
Install at

`http://iew3.technion.ac.il/serveng2005W/4CallCenters/Downloads.htm`



M/M/1 queue

- Poisson arrivals, rate λ;
- Single exponential server, rate μ, $\mathrm{E}[S] = 1/\mu$.



$\lambda_i = \lambda,\ i \geq 0; \quad \mu_i = \mu \cdot 1_{i \geq 1}$.

Cut equations: $\lambda \pi_i = \mu \pi_{i+1}\,, \qquad i \geq 0\,.$
Traffic intensity $\rho = \frac{\lambda}{\mu} < 1$ (assumed for stability).

Steady-state distribution $\mathrm{Geom}(p = 1-\rho)$ (from 0):

$$\pi_i = (1-\rho)\rho^i,\ i \geq 0.$$

Properties:

- Sojourn time is exponentially distributed:

$$W \sim \exp\left(\text{mean} = \frac{1}{\mu(1-\rho)} = \frac{1}{\mu}\left[1 + \frac{\rho}{1-\rho}\right]\right).$$

  Proof: Via moment generating functions.

  According to PASTA

  $$W \stackrel{d}{=} \sum_{i=1}^{N} X_i\,, \quad X_i \sim \exp(\mu) \text{ i.i.d.}\,, \quad N \stackrel{d}{=} \mathrm{Geom}(1-\rho) \text{ (from 1)}.$$

Moment generating function:

$$\begin{aligned}
\phi_W(t) &\triangleq \mathrm{E}\left[\exp\{tW\}\right] = \mathrm{E}\left[\exp\{t\cdot\sum_{i=1}^{N} X_i\}\right] \\
&= \mathrm{E}\left[\mathrm{E}\left[\exp\{t\cdot\sum_{i=1}^{N} X_i\}\middle| N\right]\right] \\
&= \text{(Moment generating function of Erlang r.v.)} \\
&= \mathrm{E}\left[\left(\frac{\mu}{\mu-t}\right)^N\right] = \sum_{k=1}^{\infty}(1-\rho)\rho^{k-1}\left(\frac{\mu}{\mu-t}\right)^k \\
&= \frac{\mu(1-\rho)}{\mu-t}\cdot\sum_{k=0}^{\infty}\left(\frac{\mu\rho}{\mu-t}\right)^k = \frac{\mu(1-\rho)}{\mu(1-\rho)-t} \\
&= \phi_{\exp(\mu(1-\rho))}(t)\,.
\end{aligned}$$

- Delay probability (PASTA)

$$\mathrm{P}\{W_q > 0\} = \rho\,.$$

- Waiting time in queue, given delay, is exp:

$$\frac{W_q}{1/\mu} \stackrel{d}{=} \begin{cases} 0 & \text{wp } 1-\rho \\ \exp\left(\text{mean} = \frac{1}{1-\rho}\right) & \text{wp } \rho \end{cases}$$

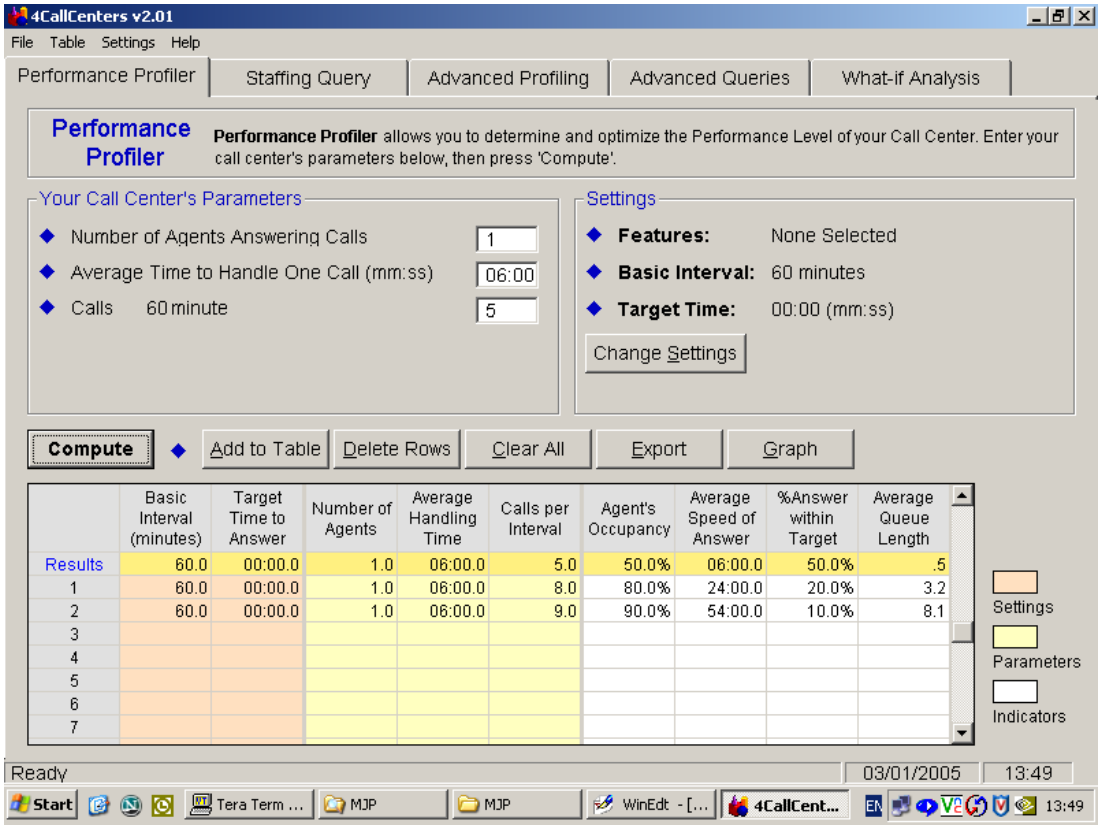
- Number-in-system

$$\mathrm{E}[L] = \frac{\rho}{1-\rho}\,; \qquad \mathrm{E}[L_q] = \frac{\rho^2}{1-\rho}\,.$$

- Server's utilization (occupancy) is $\rho = \lambda/\mu$.
  (Little's formula, system = server.)

- Departure process in steady state is Poisson (λ)
  (Burke theorem) – important in queueing networks.

M/M/1. 4CallCenters output



Note large waiting times:

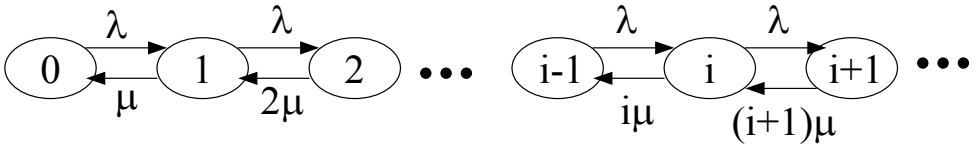
$\mathrm{E}[S]$ for $\rho = 50\%$, $\quad 9 \cdot \mathrm{E}[S]$ for $\rho = 90\%$, $\quad 19 \cdot \mathrm{E}[S]$ for $\rho = 95\%$.

4CallCenters: performance measures.

- Average Speed of Answer $= \mathrm{E}[W_q]$
  (will be different in queues with abandonment);
- %Answer within Target $= \mathrm{P}\{W_q < T\}$;
- Average Queue Length $= \mathrm{E}[L_q]$.

M/M/∞ queue

- Poisson arrivals, rate λ;
- Infinite number of exponential servers, rate μ.



$\lambda_i = \lambda, \quad i \geq 0; \qquad \mu_i = i \cdot \mu, \quad i > 0.$

Cut equations:

$$\lambda \pi_i = (i+1) \cdot \mu \pi_{i+1}\,, \qquad i \geq 0\,.$$

Always stable.

Steady-state distribution is Poisson:

$$\pi_i = e^{-R} \cdot \frac{R^i}{i!}\ ,\ i \geq 0\,,$$

where $R = \dfrac{\lambda}{\mu}$ is the **offered load** (measured in Erlangs).

$$\mathrm{E}[L] \ = \ \mathrm{E}(\#\text{ busy servers}) \ = \ \lambda \cdot \frac{1}{\mu} \ = \ R\,.$$

(Little's formula, system = service.)

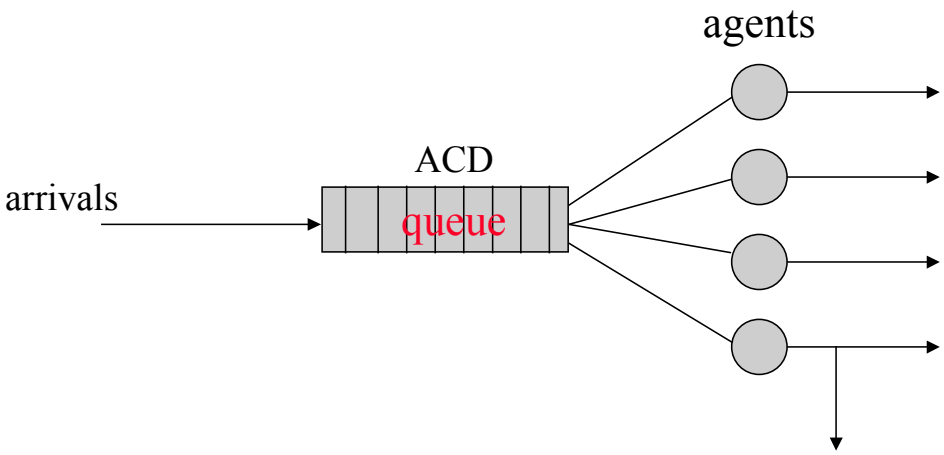
Very *useful*: ∞-server models provide bounds (next lecture: queues with abandonment).

Results above valid for M/G/∞ – generally distributed service times.

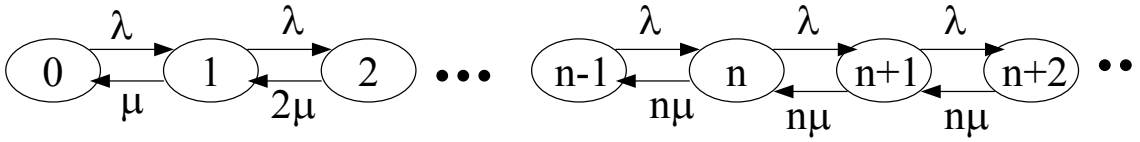
M/M/n (Erlang-C) queue

- Poisson arrivals, rate λ;
- n exponential servers, rate μ.

Widely used in call centers.



Transition-rate diagram



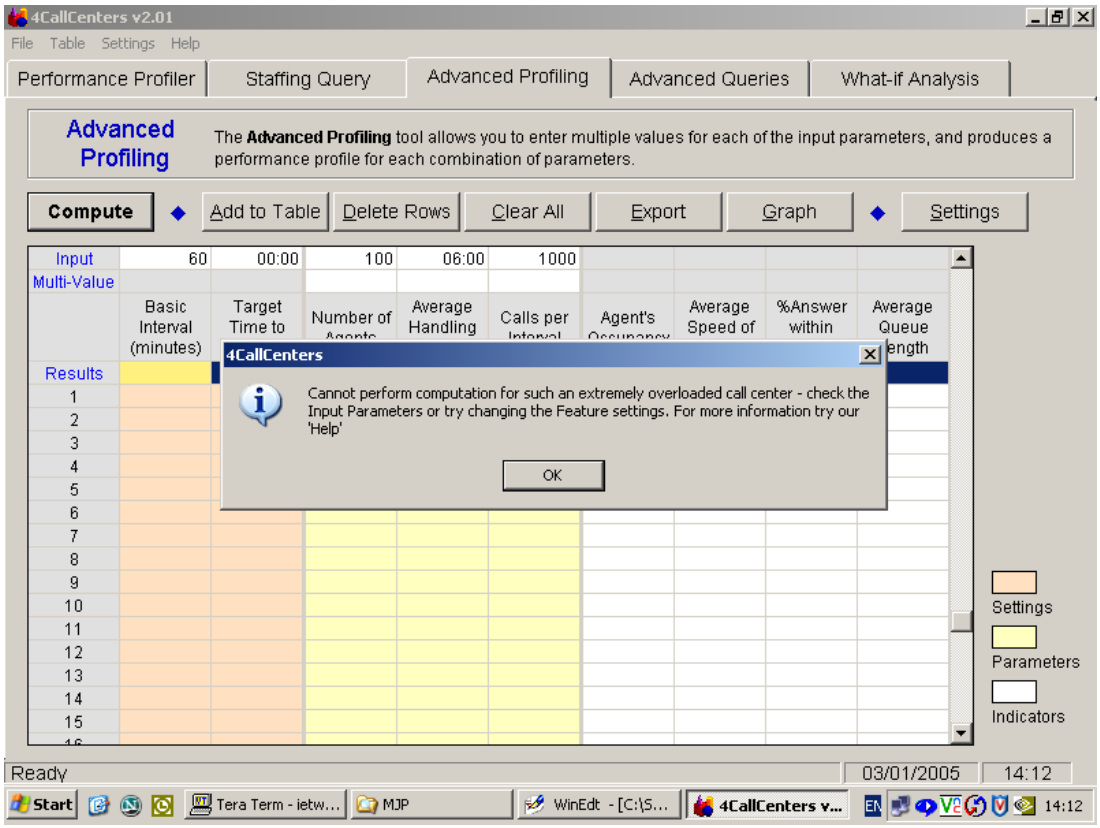
$$\begin{aligned} \lambda_j &= \lambda, \qquad j \geq 0, \\ \mu_j &= (j \wedge n)\mu, \quad j \geq 1. \end{aligned}$$

Agents' utilization

$$\rho = \frac{\lambda}{n\mu}.$$

Assume $\rho < 1$ ($R < n$) to ensure stability (as in M/M/1).

4CallCenters output: Instability, $\rho \geq 1$



Steady-state distribution:

$$\begin{aligned}
\pi_i &= \frac{R^i}{i!}\,\pi_0, && i \leq n, \\
&= \frac{n^n \rho^i}{n!}\,\pi_0, && i \geq n, \\
\pi_0 &= \left[\sum_{j=0}^{n-1} \frac{R^j}{j!} + \frac{R^n}{n!(1-\rho)}\right]^{-1},
\end{aligned}$$

where $R = \dfrac{\lambda}{\mu}$ is the **offered load**.

Erlang-C Formula (1917):
Delay probability:

$$\mathrm{P}\{W > 0\} \stackrel{\Delta}{=} E_{2,n} = \sum_{i \geq n} \pi_i = \frac{R^n}{n!} \frac{1}{1-\rho} \cdot \pi_0 \,.$$

Erlang-C computation: recursion, see Erlang-B below.

Number-in-queue:

$$\mathrm{P}\{L_q = i\} = E_{2,n} \cdot (1-\rho)\rho^i \,, \qquad i > 0,$$

or

$$L_q = \begin{cases} 0 & \text{wp } 1 - E_{2,n} \\ \mathrm{Geom}(1-\rho) & \text{wp } E_{2,n} \end{cases}$$

Waiting time distribution:

$$\frac{W_q}{1/\mu} = \begin{cases} 0 & \text{wp } 1 - E_{2,n} \\ \exp\left(\text{mean} = \frac{1}{n} \cdot \frac{1}{1-\rho}\right) & \text{wp } E_{2,n} \end{cases}$$

Compare with M/M/1!

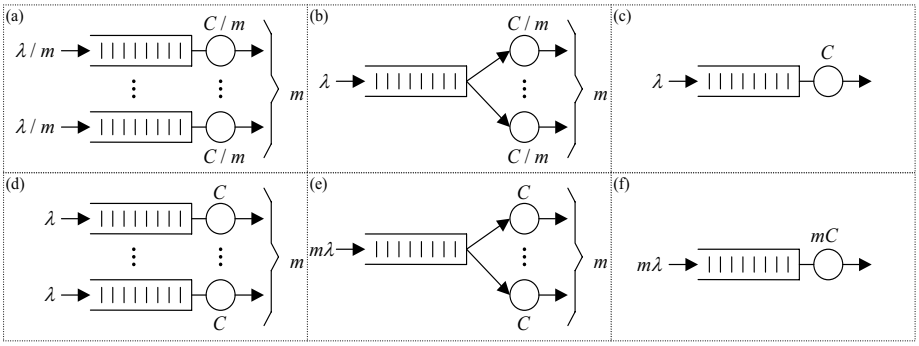
Departure process: Poisson(λ) in steady-state.
Proof via reversibility.

M/M/n: derivation of waiting-time distribution

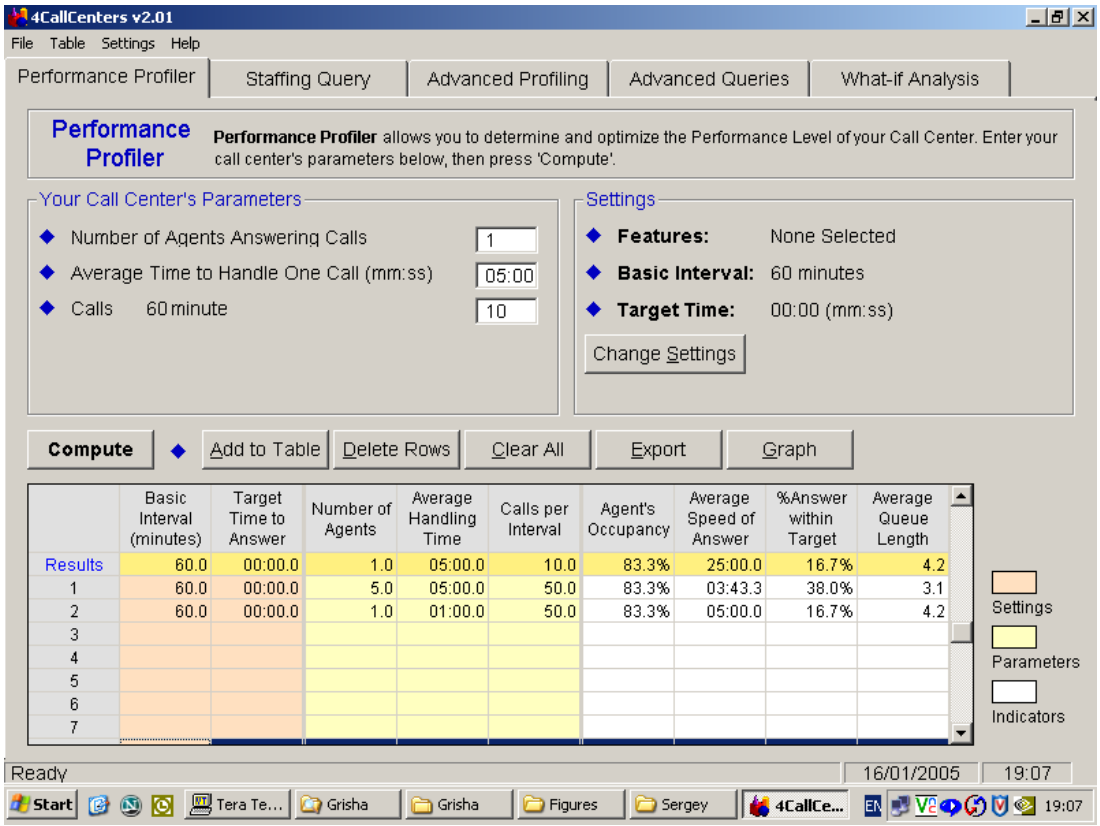
$$
\begin{aligned}
P\{W_q > t\} &= \sum_{k=1}^{\infty} \mathrm{P}\{L_q = k-1\} \cdot \mathrm{P}\{E_k > t\} \\
&\quad (\text{where } E_k \sim \text{Erlang}(k, n\mu)) \\
&= E_{2,n} \cdot \sum_{k=1}^{\infty} \left[(1-\rho)\rho^{k-1} \cdot \int_t^{\infty} \frac{n\mu(n\mu x)^{k-1}}{(k-1)!} e^{-n\mu x} dx \right] \\
&= E_{2,n} \cdot n\mu(1-\rho) \cdot \int_t^{\infty} \left(e^{-n\mu x} \cdot \sum_{k=1}^{\infty} \frac{(n\mu\rho x)^{k-1}}{(k-1)!} \right) dx \\
&= E_{2,n} \cdot n\mu(1-\rho) \cdot \int_t^{\infty} e^{-n\mu(1-\rho)x} dx \\
&= E_{2,n} \cdot e^{-n\mu(1-\rho)t}
\end{aligned}
$$

Pooling

Example: Kleinrock, L. Vol.II, Chapter 5 (1976)



4CallCenters output



	1		**2**		**3**
	$n \times$ M/M/1	$\overset{\text{pooling}}{\longrightarrow}$	M/M/n	$\overset{\text{technology}}{\longrightarrow}$	M/M/1
	λ, μ		$n\lambda, \mu$		$n\lambda, n\mu$
$\mathrm{P}\{W_q > 0\}$	ρ		$E_{2,n}$		ρ
$\mathrm{E}[W_q]$	$\frac{1}{\mu} \cdot \frac{\rho}{1-\rho}$		$\frac{1}{\mu} \cdot \frac{E_{2,n}}{n(1-\rho)}$		$\frac{1}{n\mu} \cdot \frac{\rho}{1-\rho}$
$\mathrm{E}[S]$	$\frac{1}{\mu}$		$\frac{1}{\mu}$		$\frac{1}{n\mu}$
$\mathrm{E}[W]$	$\frac{1}{\mu} \cdot \frac{1}{1-\rho}$		$\frac{1}{\mu} \cdot \left[\frac{E_{2,n}}{n(1-\rho)} + 1\right]$		$\frac{1}{n\mu} \cdot \frac{1}{1-\rho}$

Statement: $1 - \rho \;<\; 1 - E_{2,n} \;<\; n(1-\rho)\,.$

Proof: Consider M/M/n.

$$
\begin{aligned}
1 - \rho &= \mathrm{P}\{\text{server } i \text{ idle}\}, \quad \text{for } \; i = 1, \ldots, n\,; \\
1 - E_{2,n} &= \mathrm{P}\{\text{at least one server idle}\} \;=\; \mathrm{P}\left\{ \bigcup_{i=1}^{n} \{i \text{ idle}\} \right\} \\
n(1-\rho) &= \sum_{i=1}^{n} \mathrm{P}\{\text{server } i \text{ idle}\}
\end{aligned}
$$

Conclusions

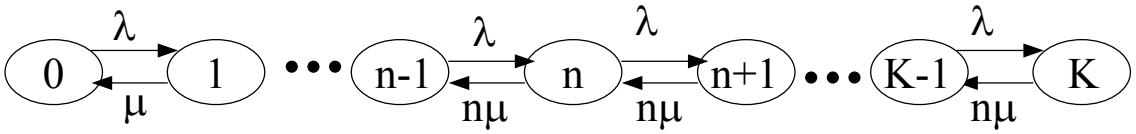
1 $\longrightarrow$ 2 : Pooling yields $\mathrm{E}[W_q]$ decrease by more than factor n;

1 $\longrightarrow$ 3 : Fast server yields $\mathrm{E}[W]$ and $\mathrm{E}[W_q]$ decrease by factor n;

2 $\longrightarrow$ 3 : Fast server better for $\mathrm{E}[W]$;
Pooling better for $\mathrm{E}[W_q]$.

M/M/n/K queue

- Poisson arrivals, rate λ;
- n exponential servers, rate μ;
- K trunks ($K \geq n$);
- If all trunks busy, arriving customer blocked (busy signal).

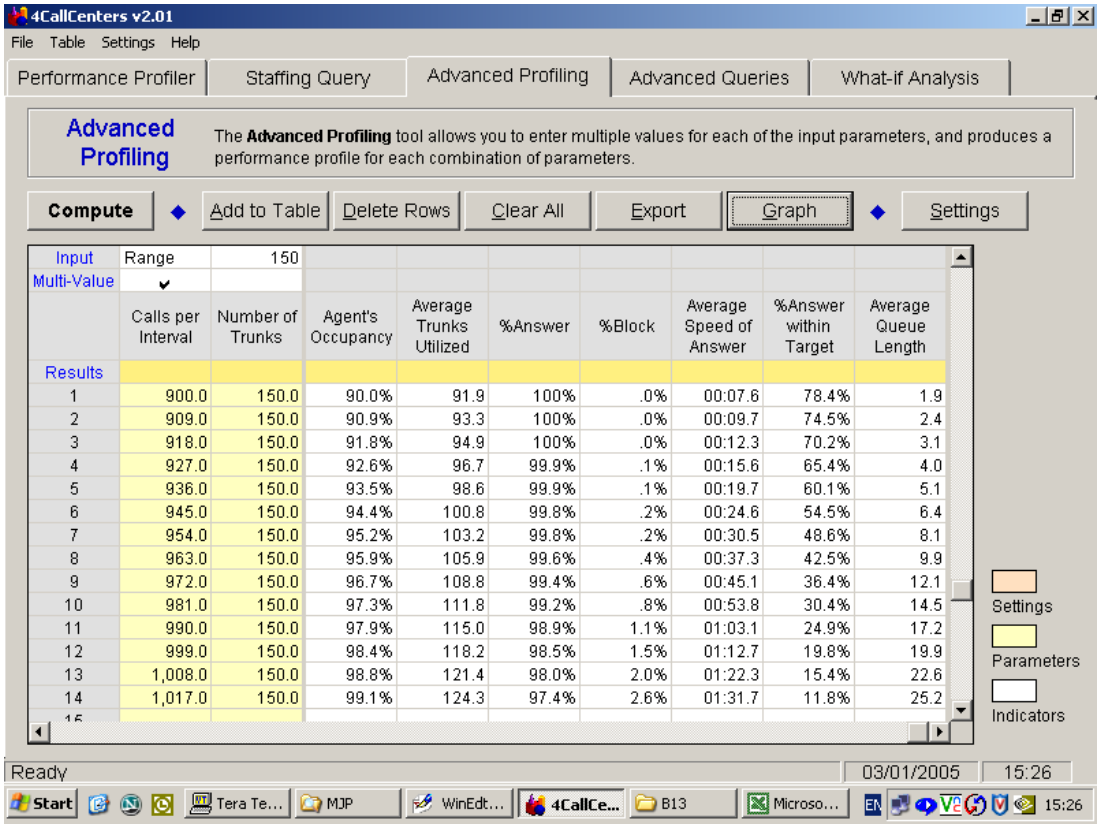


$$\lambda_j = \lambda, \qquad 0 \leq j \leq K-1,$$

$$\mu_j = (j \wedge n)\mu, \quad 1 \leq j \leq K.$$

Formulae straightforward but cumbersome (simply truncate M/M/n). Always reaches steady state.

4CallCenters output.



Use `Change Settings` $\Longrightarrow$ `Features` $\Longrightarrow$ `Trunks`.

Note new indicators:
`Average Trunks Utilized` and `%Blocked`.

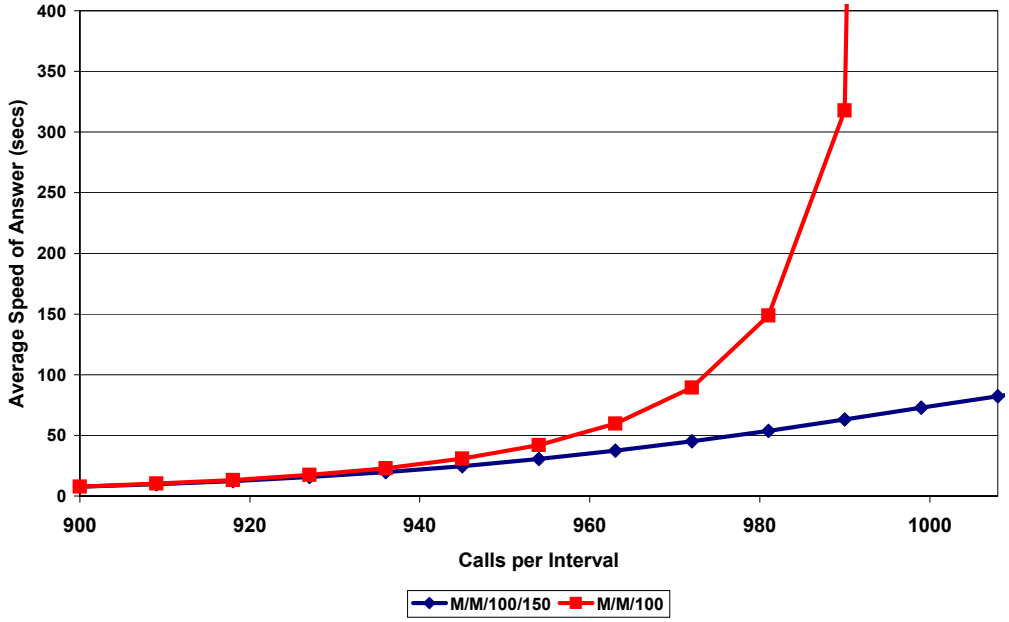
4CallCenters: Advanced Profiling

Arrival rate varied from 900 to 1017 per hour, in step 9.

Excel interface: graphs and spreadsheets.

M/M/n/K vs. Erlang-C

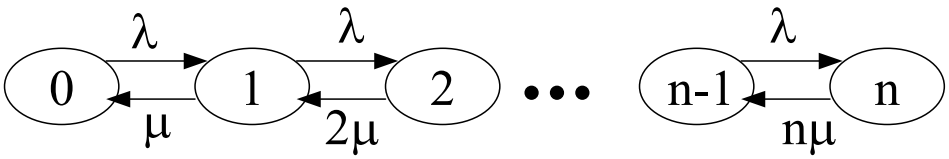
Average service time = 6 min, 100 agents, 150 trunks



Similar performance for light loads.

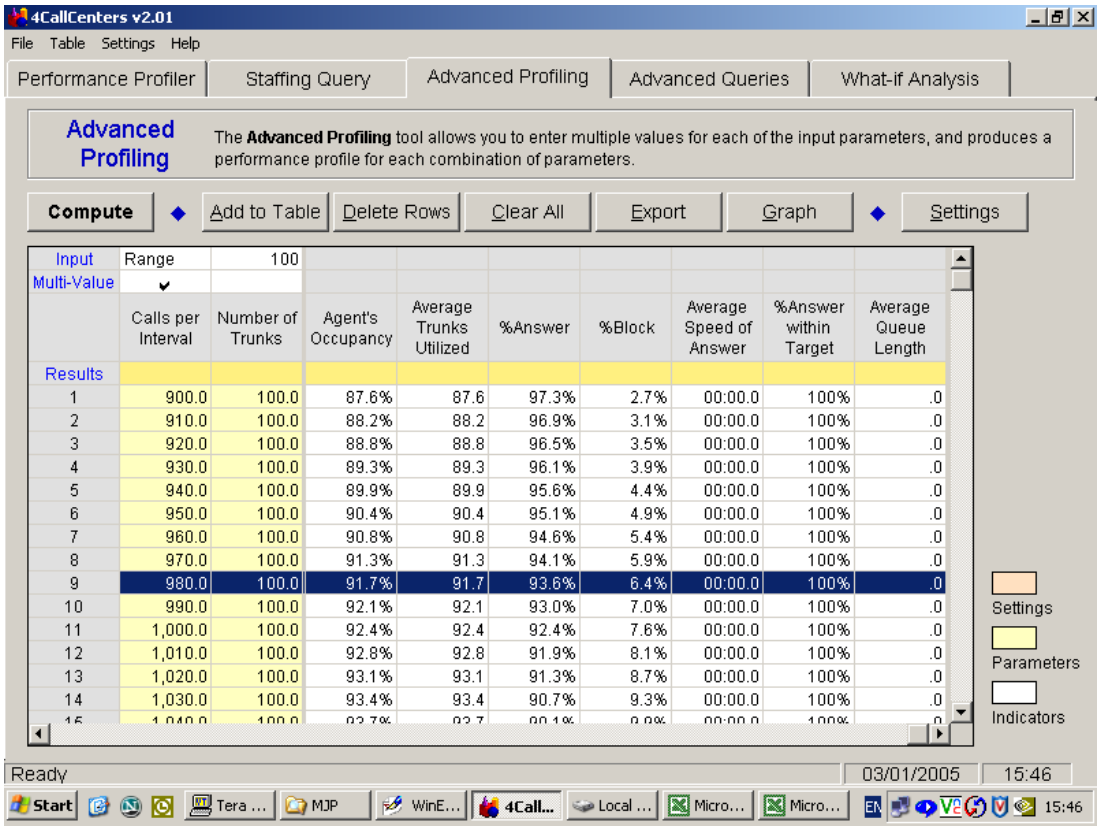
Erlang-C "explodes" as $\rho = \dfrac{\lambda}{n\mu} \uparrow 1$.

M/M/n/n (Erlang-B) queue



$$\lambda_i \equiv \lambda, \qquad 0 \le i \le n-1\,,$$

$$\mu_i = i \cdot \mu, \qquad 1 \le i \le n\,.$$

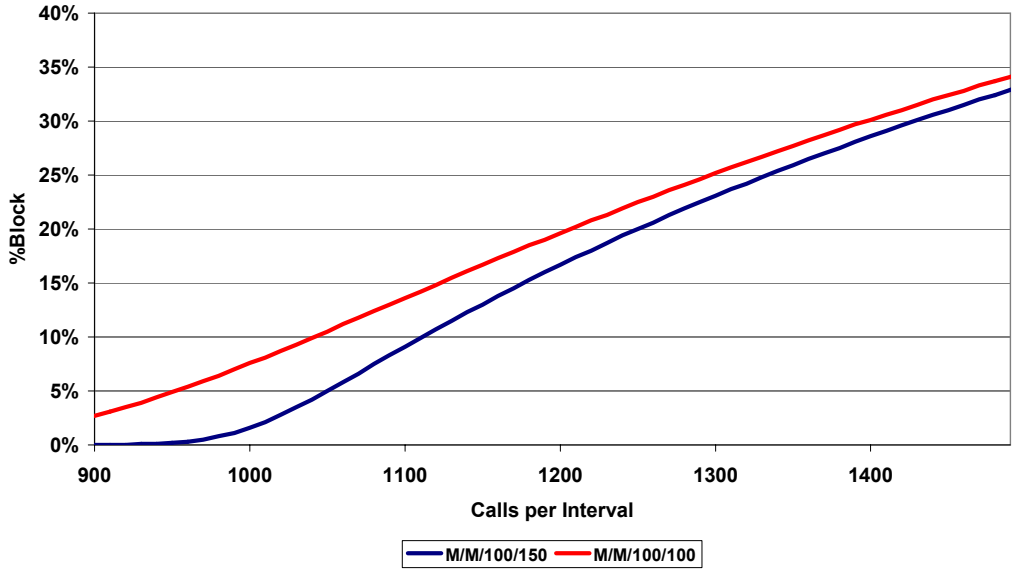


No queue ⇒ no wait.

$$\pi_i = \frac{R^i}{i!} \Big/ \sum_{j=0}^{n} \frac{R^j}{j!}\,, \qquad 0 \le i \le n.$$

M/M/n/K vs. Erlang-B

Average service time = 6 min, 100 agents



Moderate load: additional trunks prevent blocking.
Heavy load: % blocking $\approx 1 - 1/\rho$ (*"fluid limit"*).

Erlang-B Formula (1917):
Loss probability

$$E_{1,n} = \pi_n = \frac{R^n}{n!} \Big/ \sum_{j=0}^{n} \frac{R^j}{j!} \tag{2}$$

Follows from PASTA.
(2) valid for M/G/n/n! (Generally distributed service time.)

$\lambda\pi_n$ – rate of lost customers,
$\lambda(1 - \pi_n)$ – effective throughput.

Erlang-B computation: via recursion

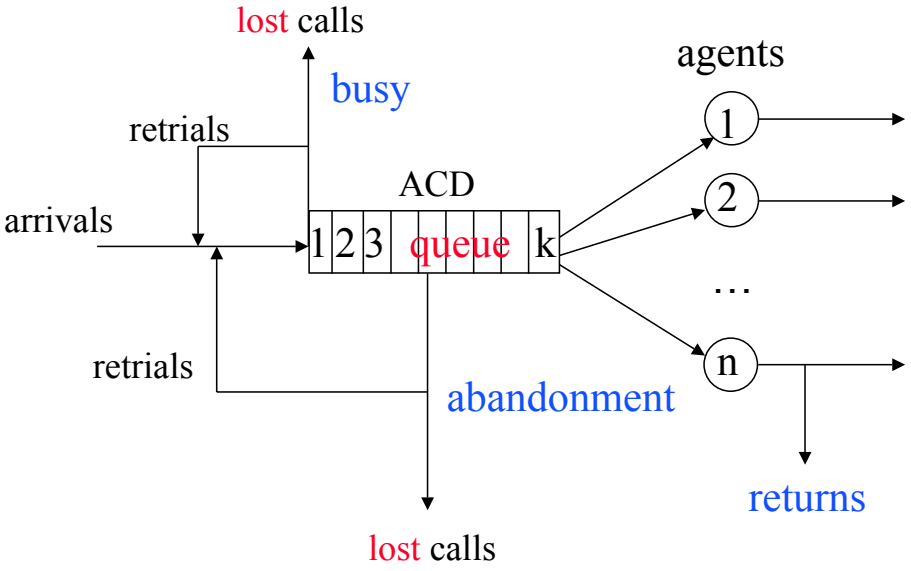
$$E_{1,n} \;=\; \frac{RE_{1,n-1}}{n + RE_{1,n-1}} \;=\; \frac{\rho E_{1,n-1}}{1 + \rho E_{1,n-1}} \qquad E_{1,0} = 1\,.$$

Note:

$$E_{1,n} \;=\; \frac{(n-R)E_{2,n}}{n - RE_{2,n}}\;; \qquad E_{2,n} = \frac{E_{1,n}}{(1-\rho) + \rho E_{1,n}}\;;$$

$$E_{2,n} \;>\; E_{1,n},\ \text{as expected: why?}$$

Schematic representation of a telephone call center



How to model Abandonment?