

**Approximations
for
Multi-Server Queues
with
Abandonments**

Ward Whitt

IEOR Department, Columbia University

URL: <http://www.columbia.edu/~ww2040>

?

?

?

?

**Don't hesitate to
ask questions!**

From

$M/M/s+M$

To

$M/GI/s+GI$

many servers: large s

(e.g., $s \approx 100$)

non-negligible abandonments

(e.g., $P(\text{Ab}) \approx 0.05$ or higher)

Recent Prior Work

$$M/M/s+M$$

Garnett, Mandelbaum and Reiman (2002)
(QED HT limit)

$$M/M/s+GI$$

Brandt and Brandt (2002)
(exact solution)

Mandelbaum and Zeltyn (2004)

Approximations for $M/GI/s+GI$

1. Numerical Algorithm

2. ED HT Fluid Approximation

(There are papers.)

Two Applications

3. **Uncertainty** About Model Parameters
4. **Sensitivity** to Changes in Model Parameters
(There are papers.)

1. Numerical Algorithm

Approximation of

$$M/GI/s/r + GI$$

by

$$M/M/s/r + M(n)$$

exploits numerical transform inversion

M/M/s/r+M(n)

State-Dependent Abandonment Rates

abandon-time cdf: $F(t) = P(\mathbf{Time} \leq t)$

abandon-time hazard rate: $h(t) = \frac{f(t)}{1-F(t)}$

**abandon rate for customer j from the end
of the queue: $\alpha_j = h(j/\lambda)$**

Comparisons with Simulations

M/GI/s/r+GI

$\mu = 1$ mean service time

$s = 100$ servers

$\lambda = 102$ arrival rate

$r \geq 200$ extra waiting spaces (very large)

model, with mean time to abandon = 1.0
 $M/E_2/100/200 + E_2$ $M/M/100/200 + M$

<i>Performance</i>	<i>sim.</i>	<i>approx.</i>	<i>sim.</i>	<i>exact</i>
$P(W = 0)$	0.217 ± 0.0021	0.250 —	0.4092 ± 0.0013	0.4083 —
$P(A)$	0.0351 ± 0.00029	0.0381 —	0.0498 ± 0.00020	0.0499 —
$E[Q]$	11.52 ± 0.075	11.41 —	5.073 ± 0.024	5.092 —
$Var(Q)$	112.0 ± 0.71	121.9 —	44.4 ± 0.30	44.6 —
$E[W S]$	0.1115 ± 0.00071	0.1102 —	0.0489 ± 0.00023	0.0490 —
$Var(W S)$	0.0101 ± 0.000061	0.0119 —	0.00418 ± 0.000027	0.0042 —
$P(W \leq 0.1 S)$	0.510 ± 0.0030	0.528 —	0.7994 ± 0.0012	0.7986 —
$P(W \leq 0.2 S)$	0.795 ± 0.0023	0.786 —	0.9648 ± 0.00057	0.9644 —

model, mean time to abandon = 4.0

<i>Performance</i>	<i>M/M/100/300 + LN(4, 0.25)</i> <i>sim.</i>	<i>M/M/100/300 + M</i> <i>approx. numerical</i>	<i>M/M/100/300 + M</i> <i>exact numerical</i>
$P(W = 0)$	0.0096 ± 0.00082	0.0101 —	0.226 —
$P(A)$	0.0206 ± 0.00029	0.0204 —	0.0364 —
$E[Q]$	118.1 ± 0.75	117.0 —	14.84 —
$E[N]$	218.0 ± 0.75	216.9 —	113.1 —
$E[W S]$	1.154 ± 0.0073	1.144 —	0.1455 —
$E[W A]$	1.327 ± 0.0015	1.288 —	0.1429 —
$P(W \leq 0.4 S)$	0.0702 ± 0.0032	0.0710 —	0.469 —
$P(W \leq 0.4 A)$	0.000093 ± 0.0032	0.0000 —	0.449 —

*M/GI/100/200 + E₂ model with mean time to abandon = 1.0
service-time distribution*

<i>Perform.</i>	<i>D</i>	<i>E₂</i>	<i>M</i>	<i>LN(1, 1)</i>	<i>approx.</i>
$P(W = 0)$	0.180 ±0.0013	0.217 ±0.0021	0.246 ±0.0020	0.233 ±0.0021	0.250 —
$P(A)$	0.0309 ±0.0002	0.0351 ±0.00029	0.0378 ±0.0003	0.0370 ±0.00027	0.0381 —
$E[Q]$	11.08 ±0.042	11.52 ±0.075	11.75 ±0.075	11.74 ±0.063	11.41 —
$Var(Q)$	89.3 ±0.40	112.0 ±0.71	129.2 ±0.94	123.3 ±0.72	121.9 —
$E[N]$	109.9 ±0.049	109.9 ±0.092	109.9 ±0.091	110.0 ±0.72	109.5 —
$E[W S]$	0.1078 ±0.0004	0.1115 ±0.0007	0.1133 ±0.00072	0.1133 ±0.00061	0.1102 —
$Var(W S)$	0.0079 ±0.00003	0.0101 ±0.00006	0.0119 ±0.00008	0.0113 ±0.00006	0.0113 —
$P(W \leq 0.1 S)$	0.501 ±0.0018	0.510 ±0.0030	0.520 ±0.0026	0.514 ±0.0025	0.528 —
$P(W \leq 0.2 S)$	0.833 ±0.0013	0.795 ±0.0023	0.775 ±0.0023	0.780 ±0.0020	0.786 —

2. ED Fluid Approximation

Approximation for

$$G/GI/s/r+GI$$

in

Efficiency-Driven (ED) Regime

Many-Server Heavy-Traffic Regimes

s large

QD

$$\rho < 1$$

$$P(W > 0) \approx 0$$

$$P(Ab) \approx 0$$

QED

$$\rho \approx 1$$

$$0 < P(W > 0) < 1$$

$$P(Ab) \approx 0$$

ED

$$\rho > 1$$

$$P(W > 0) \approx 1$$

$$0 < P(Ab) < 1$$

Halfin and Whitt (1981), Mandelbaum

G/GI/s/r+GI

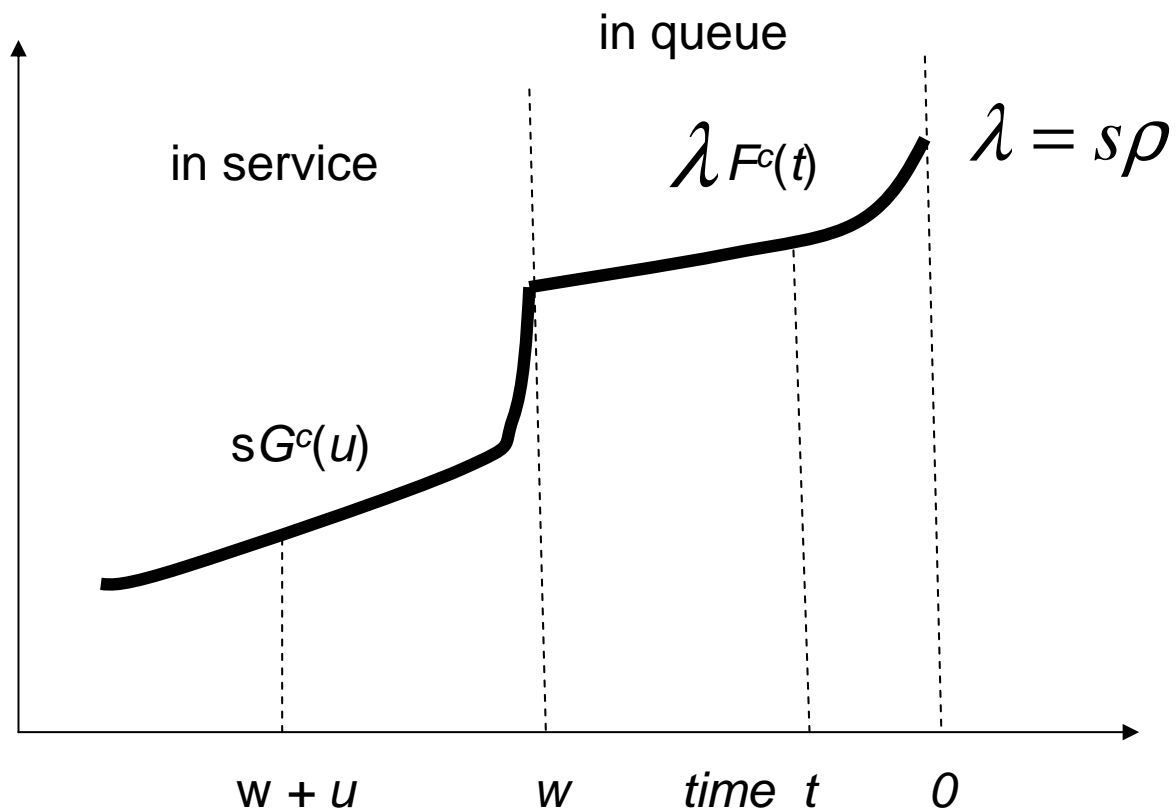
Model Elements

service-time cdf: **G** (mean 1)

abandon-time cdf: **F**

traffic intensity: ρ

Equilibrium in the ED Regime



3. Uncertainty About the Model Parameters

“Staffing a Call Center
with Uncertain Arrival Rate
and Absenteeism”

Random Arrival Rate Λ

Random Number of Servers Γ_s

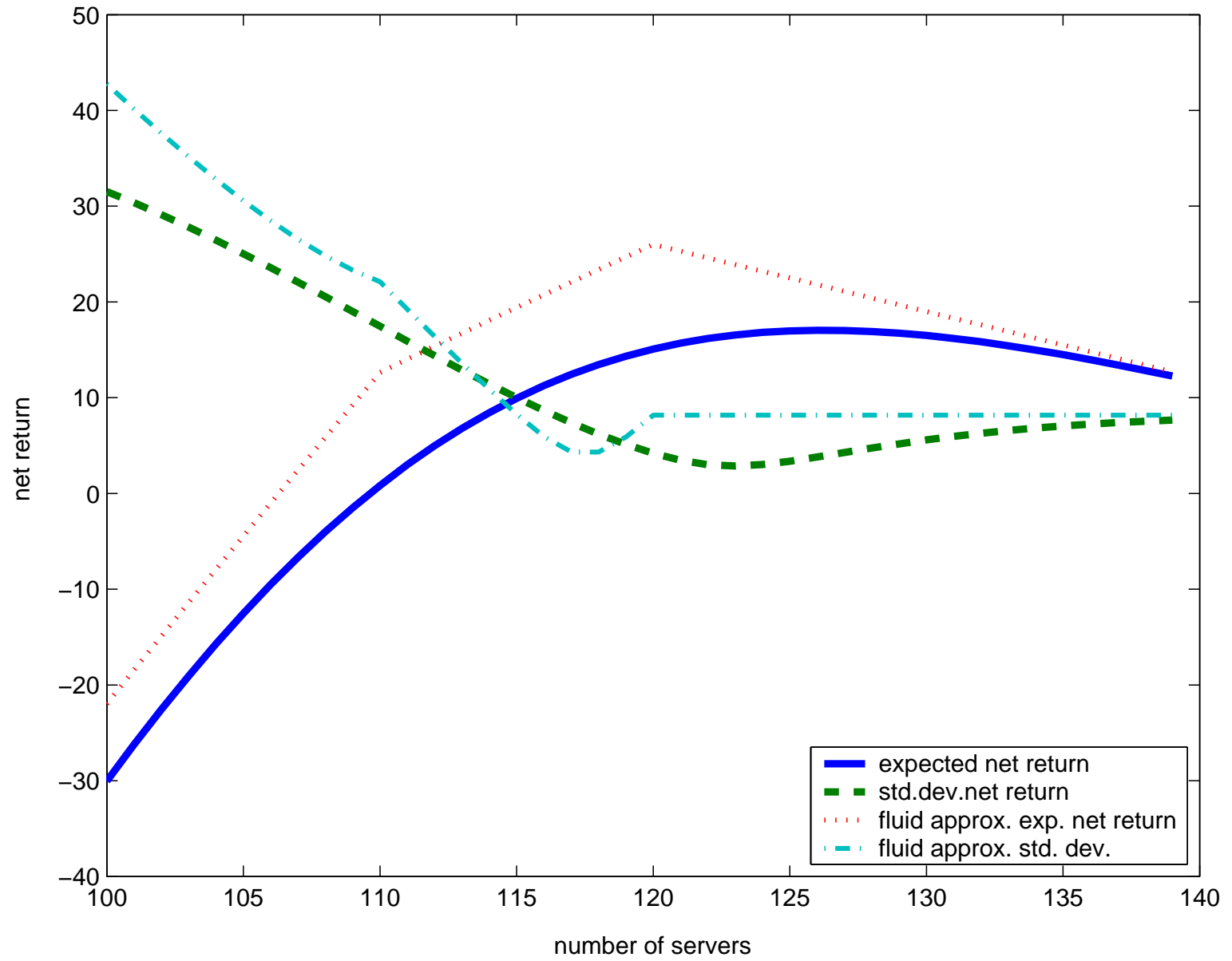
Revenue

$$R(s) = r_t T(s) - c_s \Gamma s - c_a L(s) - c_w \Lambda W(s)$$

perf. measure	notation	fluid approx.
throughput	$T(s)$	$\Lambda \wedge \Gamma s$
loss rate	$L(s)$	$(\Lambda - \Gamma s)^+$
waiting rate	$\Lambda W(s)$	$(\Lambda - \Gamma s)^+ / f(0)$

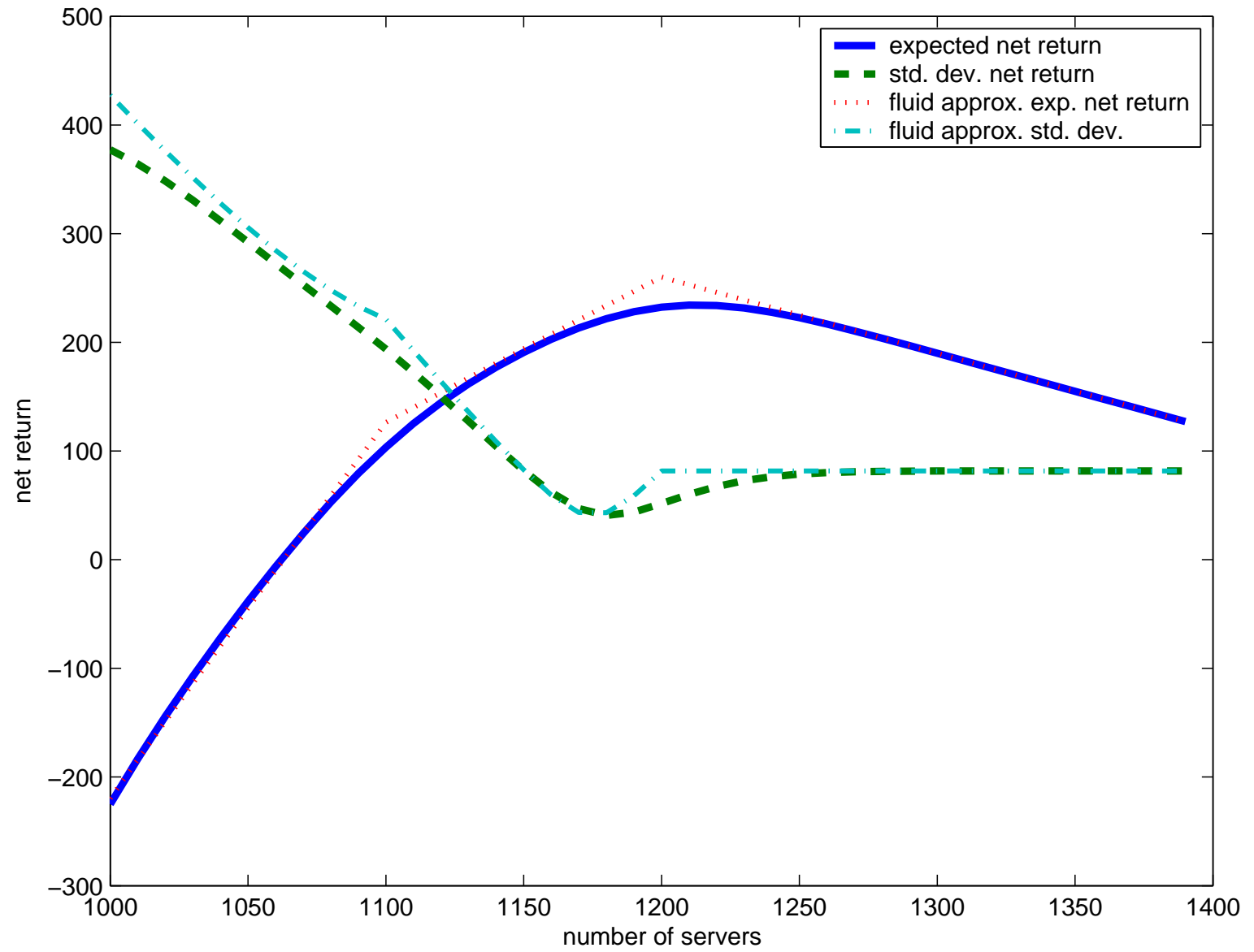
Example 1.

$\Lambda = 100, 110$ or 120
each with probability $1/3$



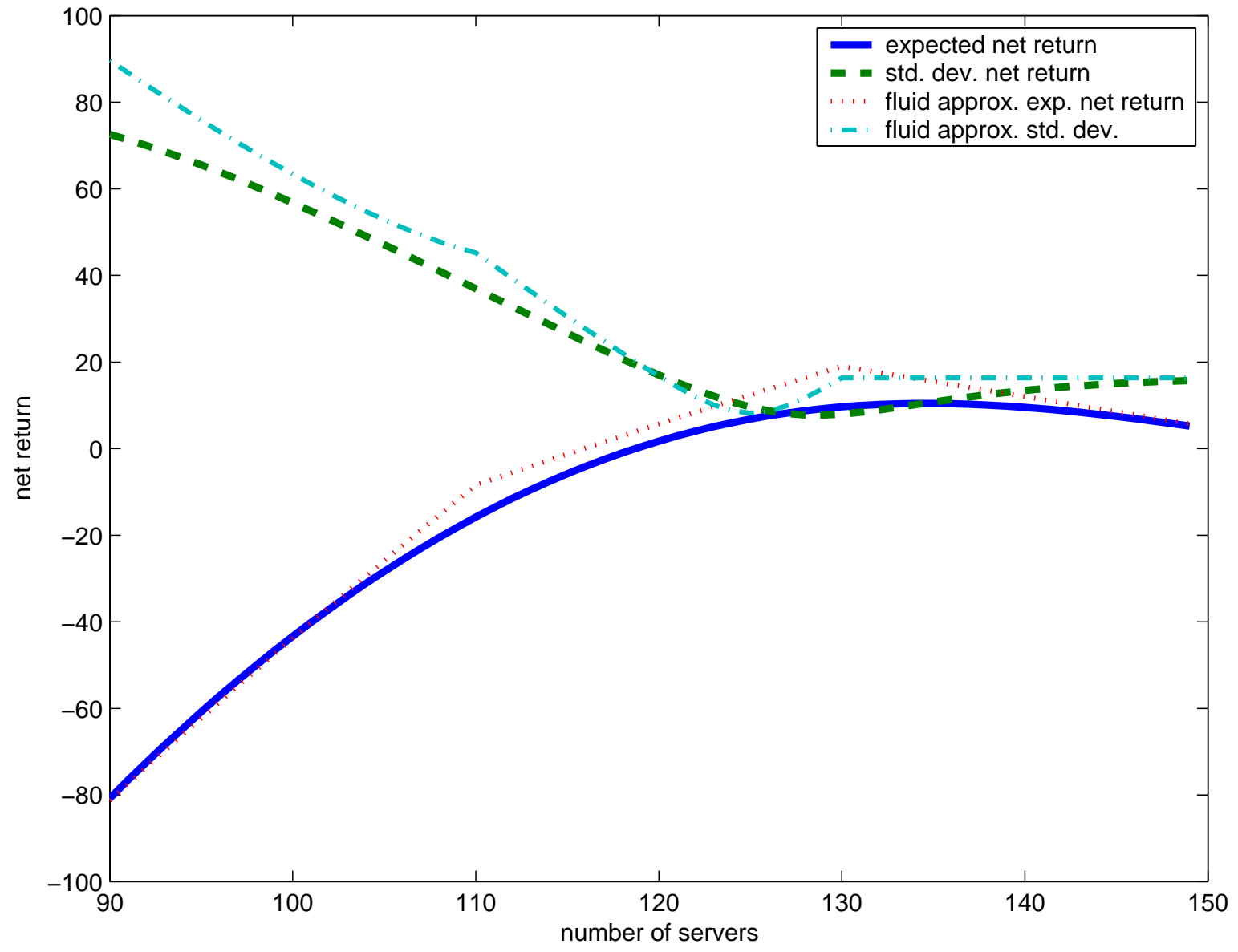
Example 2.

$\Lambda = 1000, 1100$ or 1200
each with probability $1/3$



Example 3.

$\Lambda = 90, 110$ or 130
each with probability $1/3$



4. Sensitivity

“Sensitivity of Performance
in the $M/M/s+M$ Model
to Changes in the Model Parameters”

λ arrival rate

μ service rate

s number of servers

θ abandonment rate

Elasticities

$$f(\lambda) \equiv P(W > 0)$$

$$\mathcal{E}(f, \lambda) \equiv \frac{\frac{df(\lambda)}{d\lambda}}{\frac{f(\lambda)}{\lambda}} = \frac{\lambda f'(\lambda)}{f(\lambda)}$$

“percentage change in $f(\lambda)$ caused by small percentage change in λ ”

$$\delta\% \text{ in } \lambda \Rightarrow \mathcal{E}(f, \lambda)\delta\% \text{ change in } f(\lambda)$$

Finite-Difference Approximation

$$f'(\lambda) \approx \frac{f(\lambda+h) - f(\lambda)}{h}$$

(use numerical algorithm)

Principal Conclusion

Sensitivity of performance
in the $M/M/s+M$ model
to changes in the abandonment rate (θ)
is much less than
to changes in the arrival rate (λ)
and the other parameters.

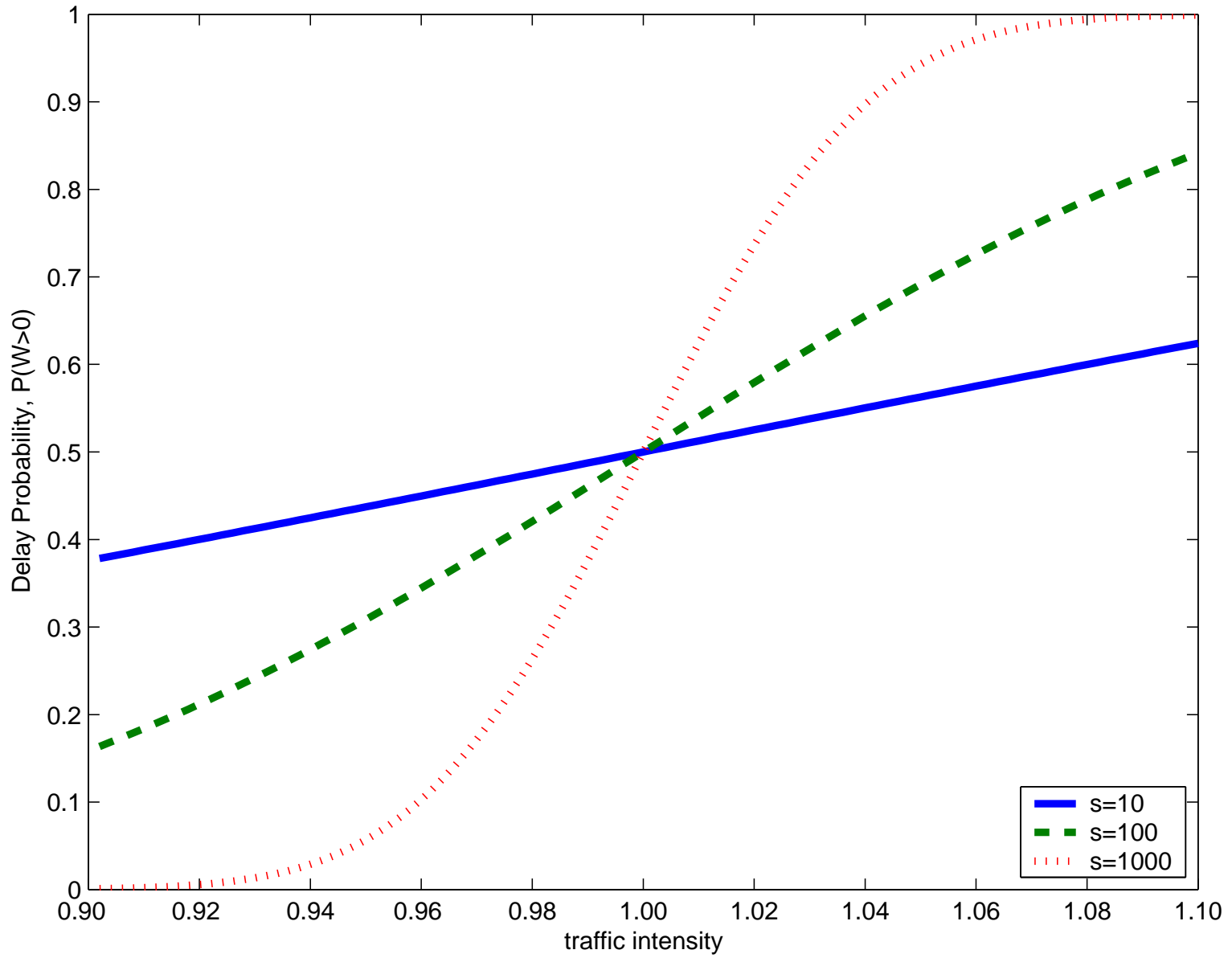
Insights from QED Limits

$$\mathcal{E}(f, \lambda) = O(\sqrt{s}) \quad \text{as } s \rightarrow \infty$$

$$\mathcal{E}(f, \theta) = O(1) \quad \text{as } s \rightarrow \infty$$

$$\mathcal{E}(f, \lambda) = -\mathcal{E}(f, \lambda^{-1})$$

$$\mathcal{E}(f, \lambda) \sim -\mathcal{E}(f, \mu) \sim -\mathcal{E}(f, s)$$



Numerical Examples

parameters		performance measures					
θ	$\lambda = s$	$P(W > 0)$	$P(Ab)$	EN	$SD(Q)$	$SD(N)$	$SD(W)$
10.0	10	0.320	0.186	8.33	0.53	2.08	0.041
	100	0.262	0.0605	94.6	1.50	6.8	0.014
	1000	0.247	0.0192	982.8	4.58	21.9	0.0045
1.0	10	0.542	0.125	10.0	1.96	3.16	0.183
	100	0.513	0.0399	100.0	5.95	10.0	0.058
	1000	0.504	0.0126	1000.	18.6	31.6	0.0185
0.1	10	0.779	0.0605	15.4	6.4	7.1	0.63
	100	0.766	0.0192	117.2	20.0	22.2	0.198
	1000	0.762	0.00605	1055.	62.8	69.9	0.063

Several performance measures in the Erlang A model, as a function of the abandonment rate, θ and the number of servers, s , when $\lambda = s$ and $\mu = 1$.

scaled performance measures

$P(W > 0)$	$\sqrt{s}P(Ab)$	$(EN - s)/\sqrt{s}$	$SD(Q)/\sqrt{s}$	$SD(N)/\sqrt{s}$	$\sqrt{s}SD(W)$
0.320	0.59	-0.53	0.168	0.66	0.13
0.262	0.61	-0.54	0.150	0.68	0.14
0.247	0.61	-0.54	0.145	0.69	0.14
0.542	0.40	0.00	0.62	1.00	0.58
0.513	0.40	0.00	0.60	1.00	0.58
0.504	0.40	0.00	0.59	1.00	0.59
0.779	0.19	1.71	0.20	2.2	2.0
0.766	0.19	1.72	0.20	2.2	2.0
0.762	0.19	1.74	0.20	2.1	2.0

Scaled versions of the performance measures in the previous table.

parameters		performance measures					
θ	$\lambda = s$	$P(W > 0)$	$P(Ab)$	$EQ\&EW$	EN	$SD(Q)$	$SD(W)$
10.0	10	-0.22	0.103	-0.90	-0.04	-0.57	-0.74
	100	-0.33	0.119	-0.88	-0.013	-0.63	-0.66
	1000	-0.37	0.120	-0.88	-0.004	-0.65	-0.66
1.0	10	-0.21	0.25	-0.75	-0.125	-0.54	-0.58
	100	-0.24	0.25	-0.75	-0.04	-0.56	-0.57
	1000	-0.23	0.37	-0.64	-0.011	-0.42	-0.43
0.1	10	-0.108	0.38	-0.62	-0.26	-0.49	-0.51
	100	-0.116	0.41	-0.62	-0.11	-0.50	-0.50
	1000	-0.119	0.38	-0.62	-0.04	-0.50	-0.50

The abandonment-rate elasticities, $\mathcal{E}(f, \theta)$, of several performance measures (the f) in the same setting.

parameters		performance measures					
θ	$\lambda = s$	$P(W > 0)$	$P(Ab)$	EQ	$SD(Q)$	EW	$SD(W)$
10.0	10	0.73	0.76	1.08	0.54	0.76	0.32
	100	0.77	0.88	0.98	0.48	0.88	0.40
	1000	0.80	0.92	0.95	0.46	0.92	0.44
1.0	10	0.73	1.04	1.36	0.70	1.04	0.44
	100	0.78	1.19	1.29	0.62	1.19	0.54
	1000	0.80	1.23	1.27	0.60	1.23	0.57
0.1	10	0.73	2.02	2.34	1.23	2.02	1.01
	100	0.78	2.17	2.27	1.15	2.17	1.09
	1000	0.80	2.22	2.25	1.13	2.24	1.11

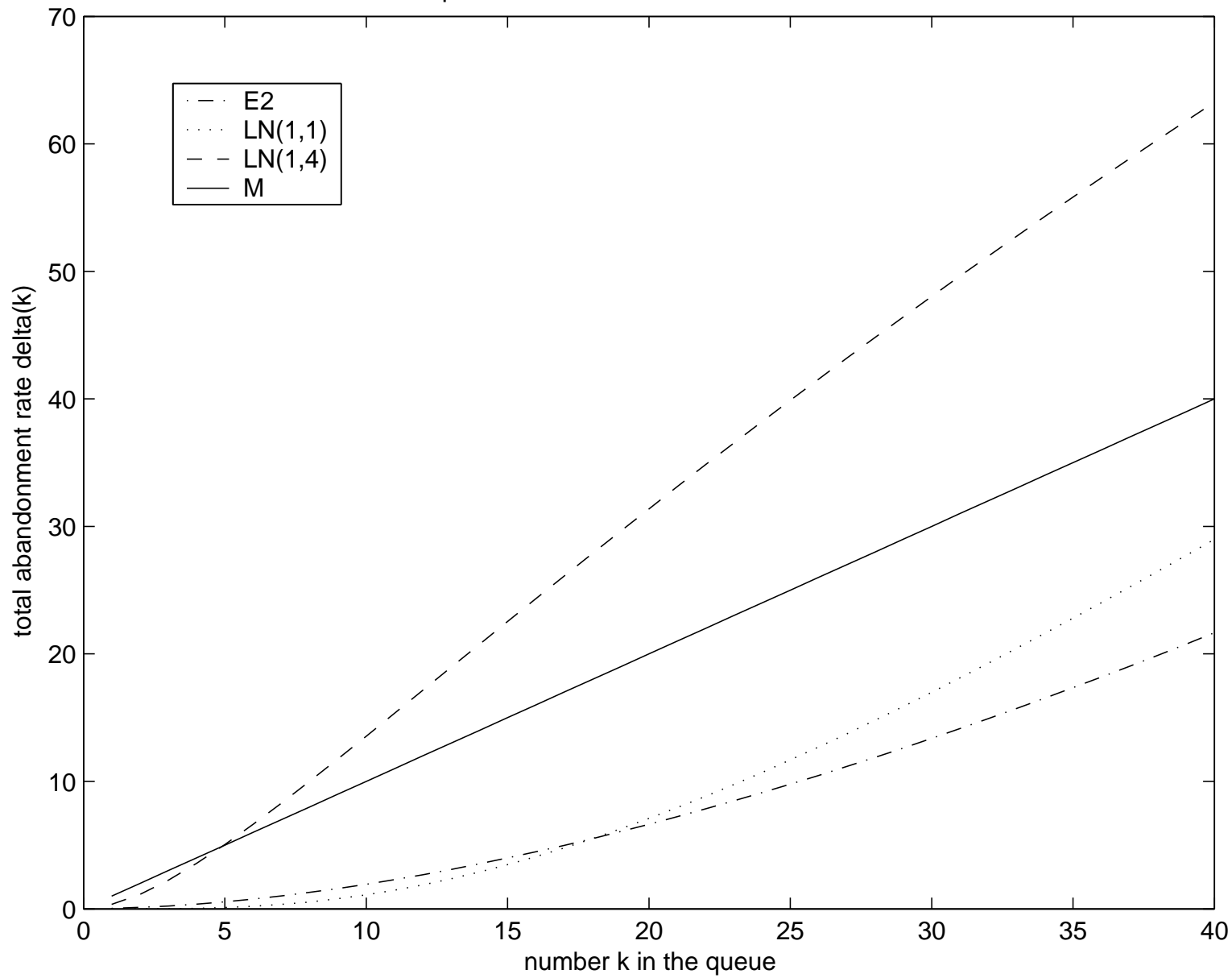
The arrival-rate elasticities, $\mathcal{E}(f, \lambda)$, of several performance measures in the same setting. The arrival-rate elasticities have been scaled by dividing by \sqrt{s} .

Done At Last

*M/GI/100/200 + LN(1, 1) model with mean time to abandon = 1.0
service-time distribution*

<i>Perform.</i>	E_2	M	$LN(1, 1)$	$LN(1, 4)$	<i>approx.</i>
$P(W = 0)$	0.211 ± 0.0013	0.242 ± 0.0026	0.229 ± 0.0015	0.286 ± 0.0020	0.247 —
$P(A)$	0.0348 ± 0.0002	0.0376 ± 0.0003	0.0366 ± 0.0002	0.0425 ± 0.0002	0.0379 —
$E[Q]$	11.40 ± 0.039	11.42 ± 0.071	11.44 ± 0.051	11.55 ± 0.048	11.02 —
$Var(Q)$	102.7 ± 0.39	115.6 ± 0.46	110.6 ± 0.43	137.6 ± 0.49	107.2 —
$E[N]$	109.9 ± 0.053	109.6 ± 0.092	109.7 ± 0.062	109.2 ± 0.071	109.1 —
$E[W S]$	0.1097 ± 0.0004	0.1094 ± 0.00067	0.1098 ± 0.0005	0.1096 ± 0.0004	0.1058 —
$Var(W S)$	0.0091 ± 0.00003	0.0104 ± 0.00004	0.0099 ± 0.00004	0.0126 ± 0.00005	0.0097 —
$P(W \leq 0.1 S)$	0.502 ± 0.0016	0.518 ± 0.0028	0.511 ± 0.0021	0.542 ± 0.0020	0.527 —
$P(W \leq 0.2 S)$	0.807 ± 0.0011	0.792 ± 0.0018	0.797 ± 0.0016	0.773 ± 0.0011	0.807 —

Comparison of four abandon-time distributions



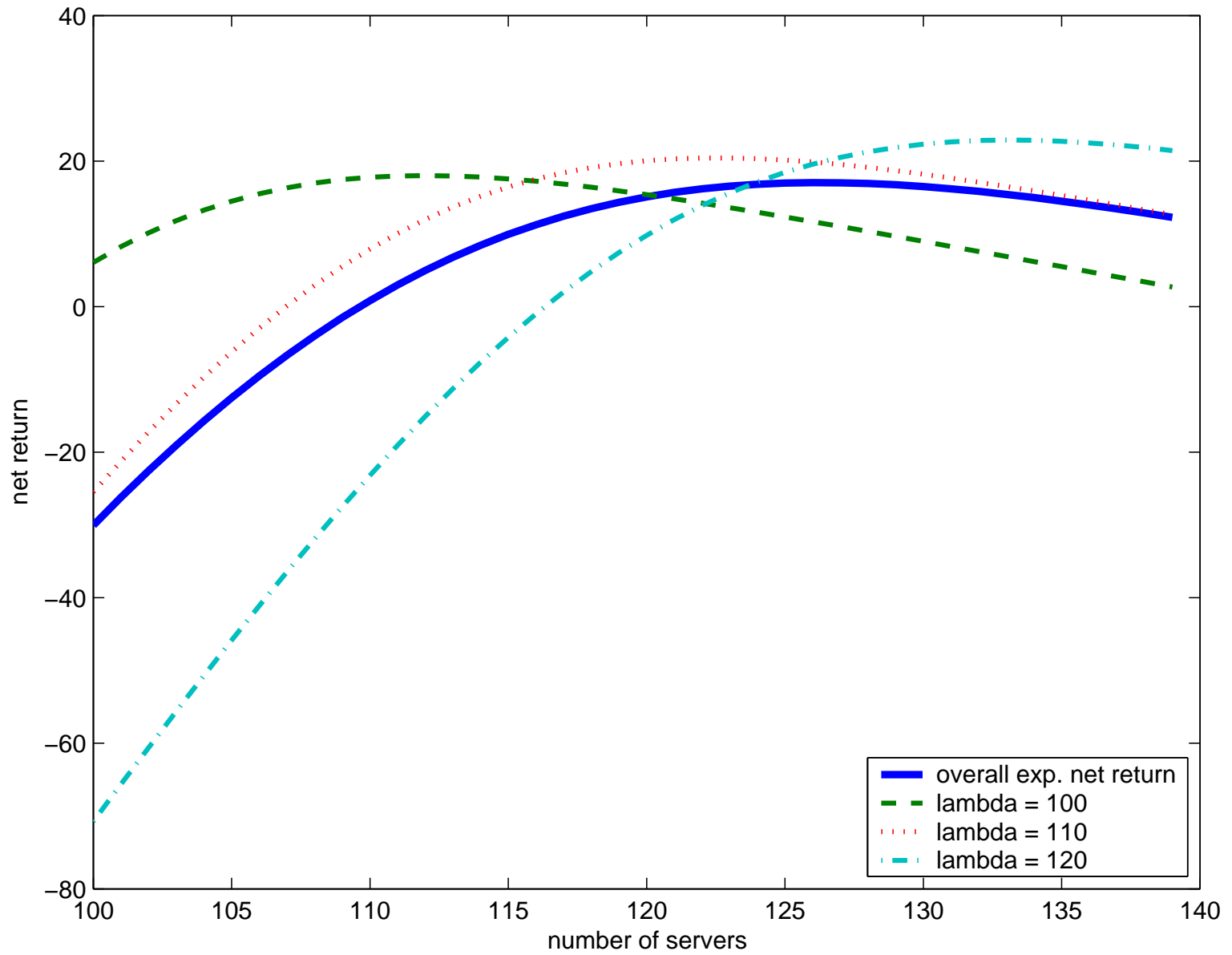
The approximate total abandonment rate δ_k for four time-to-abandon distributions when $\lambda = 100$.

Example 1a.

$\Lambda = 100, 110$ or 120

each with probability $1/3$

Overall versus conditional expected returns



$$\hat{w}_s(z) \equiv E[e^{-z(W)} \mathbf{1}_{\{S\}}] = \sum_{k=1}^r p_{s+k-1}^a \Gamma_k \hat{e}_k(z) ,$$

where p_{s+k-1}^a is the probability an arrival finds $s+k-1$ customers in the system, $\hat{e}_k(z)$ is the transform of the time until customer $s+k$ receives service, i.e.,

$$\hat{e}_k(z) \equiv \prod_{j=1}^k \left(\frac{m_{k,j}^{-1}}{m_{k,j}^{-1} + z} \right) ,$$

and Γ_k is the probability that customer $s+k$ eventually receives service, i.e.,

$$\Gamma_k = (1 - \gamma_{k,1})(1 - \gamma_{k,2}) \dots (1 - \gamma_{k,k}) ,$$

with

$$\gamma_{k,j} \approx \frac{\alpha_j}{s\mu + (\delta_k - \delta_{j-1})} ,$$

$$m_{k,j} \approx \frac{1}{s\mu + (\delta_k - \delta_{j-1})} .$$