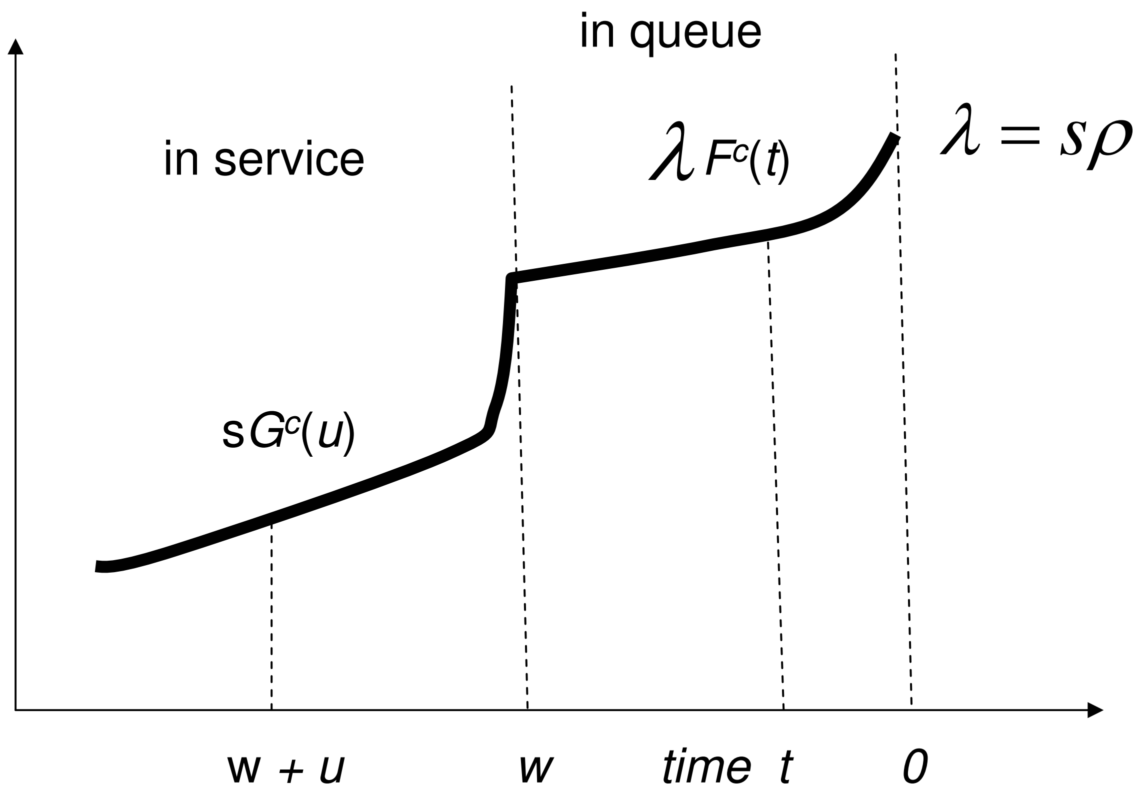
# Fluid Approximations for Many-Server Queues with Abandonment and their Applications

Ward Whitt

IEOR Department, Columbia University

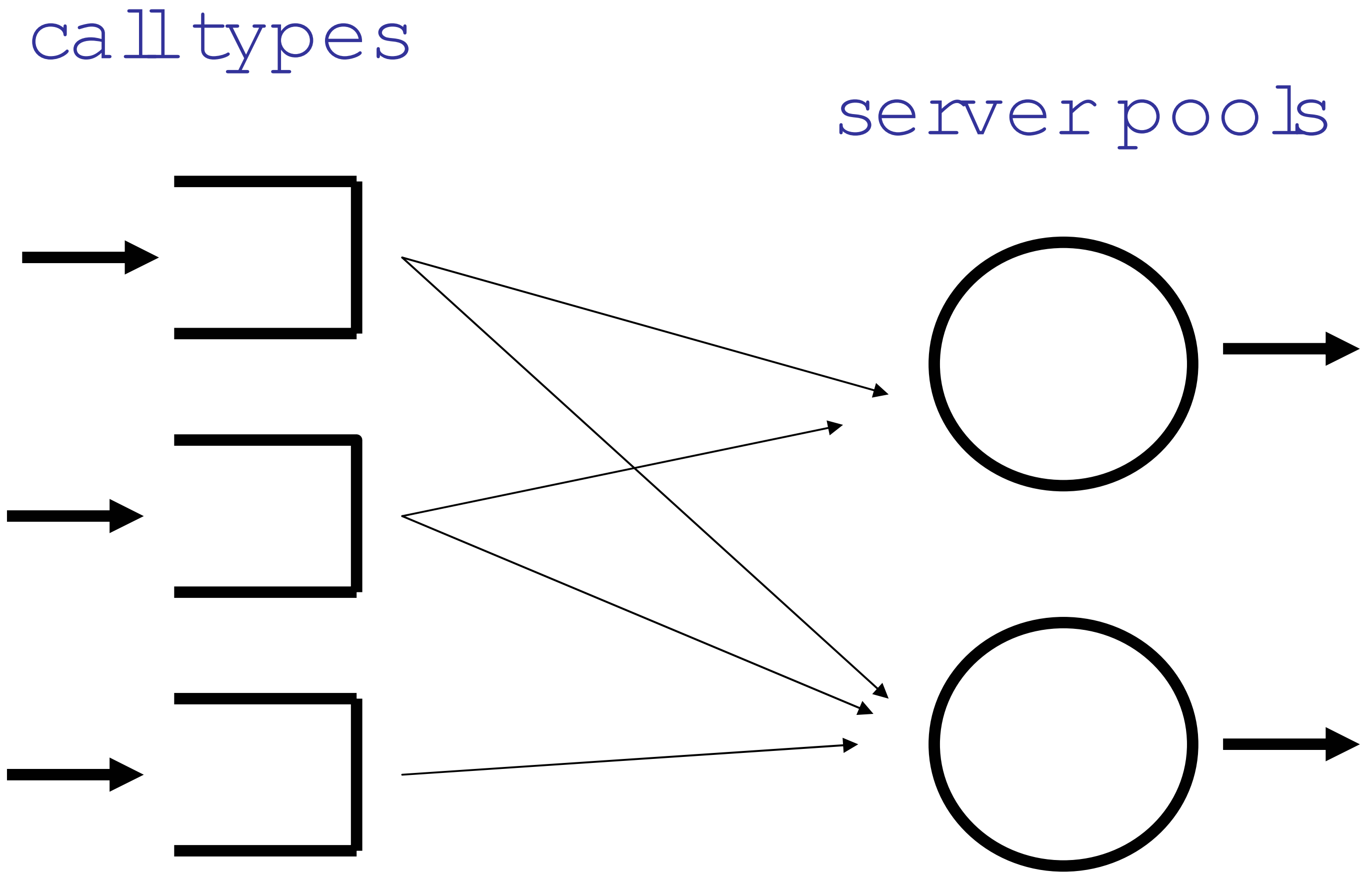
URL: http://www.columbia.edu/∼ww2040

# Equilibrium in the ED Regime



# Many Servers

# skill-based routing



**Here only**

# M/GI/s+GI

How is system behavior affected by the two non-exponential distributions?

# Deterministic Fluid Approximation

for

M/GI/s+GI

with large s

# Applications

1. **Delay Announcements**

2. **Uncertainty About Model Parameters**

3. **Sensitivity to Changes in Model Parameters**

4. **Time-Varying Arrivals**

# M/GI/s+GI

## Model Elements

service-time cdf: G (mean 1)

abandon-time cdf: F

arrival rate: $\lambda$

traffic intensity: $\rho = \lambda / s$

# State in M/GI/s+GI Model

## Time Plus Number

# State in M/GI/s+GI Model

B(t,x) - number of servers busy
for less than or equal to time x
at time t

Q(t,x) - number of customers in queue
waiting for less than or equal to time x
at time t

# Many-Server Heavy-Traffic Limit

$$s \to \infty$$

**and**

$$\lambda \to \infty$$

**traffic intensity:** $\rho = \lambda/s$ **fixed**

# Many-Server Heavy-Traffic Limit

$$\frac{B_s(t,x)}{s} \rightarrow B(t,x) = \int_0^x b(t,y)\, dy$$

$$\frac{Q_s(t,x)}{s} \rightarrow Q(t,x) = \int_0^x q(t,y)\, dy$$

**as** $s \rightarrow \infty$

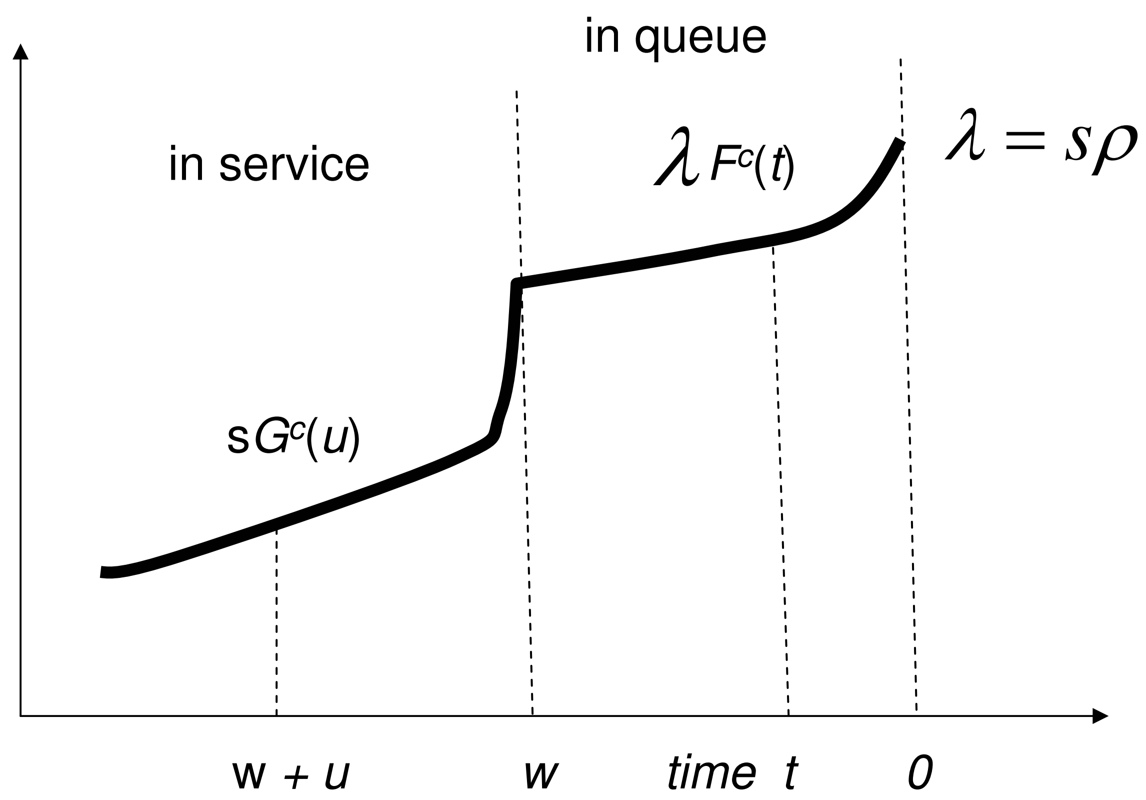
# Many-Server Heavy-Traffic Regimes

**s large**

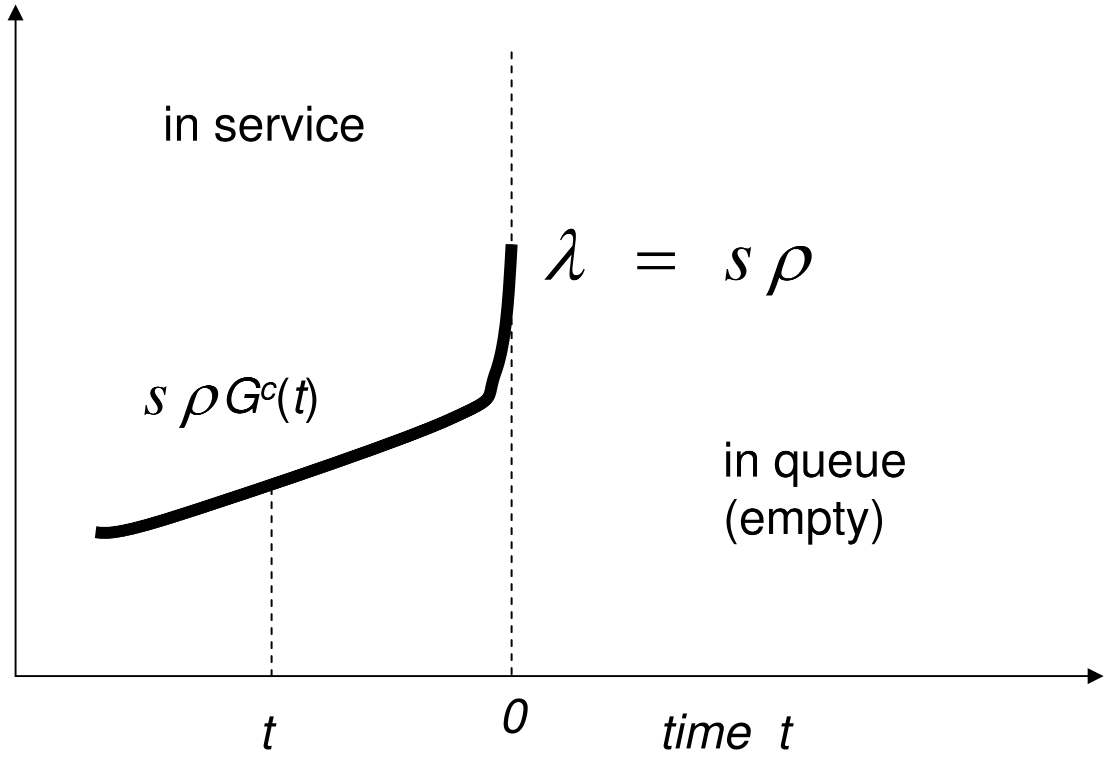
| QD | QED | ED |
|---|---|---|
| $\rho < 1$ | $\rho \approx 1$ | $\rho > 1$ |
| $P(W > 0) \approx 0$ | $0 < P(W > 0) < 1$ | $P(W > 0) \approx 1$ |
| $P(Ab) \approx 0$ | $P(Ab) \approx 0$ | $0 < P(Ab) < 1$ |

**Halfin and Whitt (1981), Mandelbaum, Reiman, . . .**

# Equilibrium in the ED Regime



# Underloaded Equilibrium



$M/GI/100/200+GI$ *model with* $\lambda = 120$ *and* $E[T] = 1.0$

| | $E_2$ *time-to-abandon cdf* *service cdf* | | | $LN(1,4)$ *time-to-abandon cdf* *service cdf* | | |
|---|---|---|---|---|---|---|
| Perf. Meas. | $E_2$ | $LN(1,4)$ | approx. | $E_2$ | $LN(1,4)$ | approx. |
| $P(A_s)$ | 0.1665 | 0.1668 | 0.1667 | 0.168 | 0.170 | 0.1667 |
| $E[Q_s]$ | 40.3 | 39.6 | 41.1 | 14.5 | 14.5 | 14.6 |
| | ±0.06 | ±0.10 | – | ±0.02 | ±0.04 | – |
| $Var(Q_s)$ | 140 | 222 | 0.00 | 61 | 82 | 0.00 |
| | ±0.7 | ±1.1 | – | ±0.2 | ±0.3 | – |
| $SCV(Q_s)$ | 0.09 | 0.14 | 0.00 | 0.29 | 0.39 | 0.00 |
| $P(W_s = 0)$ | 0.0005 | 0.007 | 0.000 | 0.032 | 0.07 | 0.00 |
| $E[W_s \mid S_s]$ | 0.353 | 0.343 | 0.365 | 0.126 | 0.125 | 0.131 |
| | ±0.0005 | ±0.0010 | – | ±0.0002 | ±0.0004 | – |

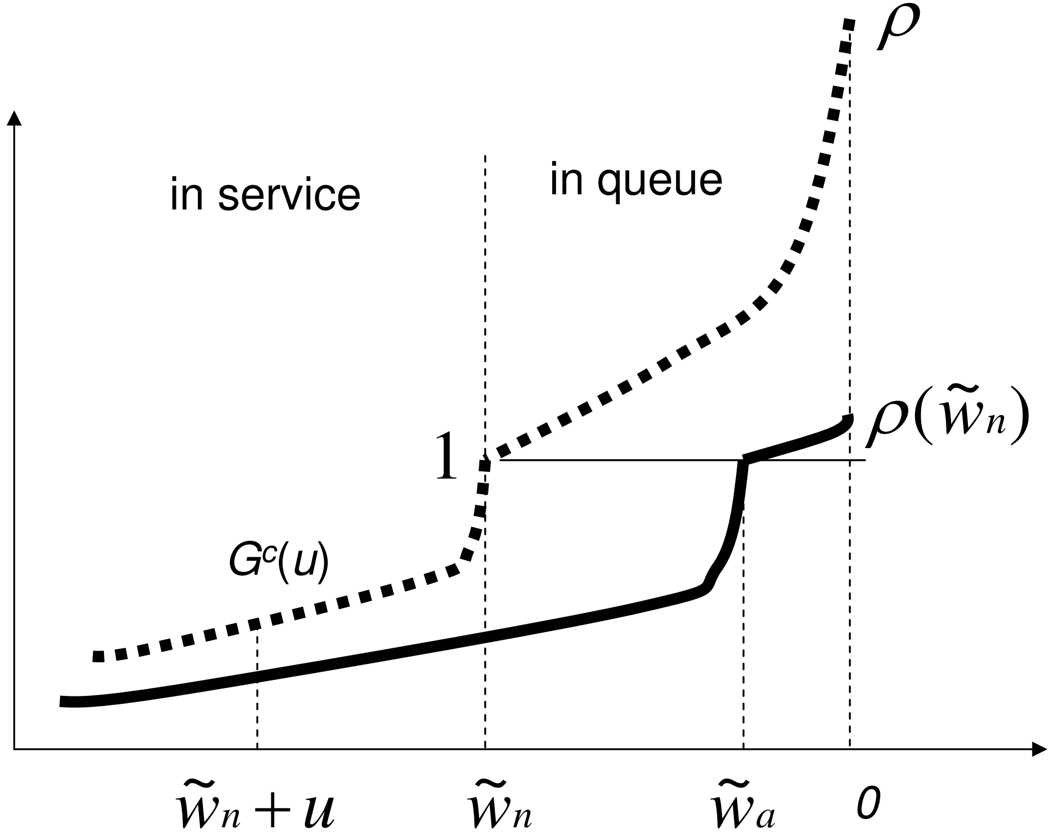
# Application 1.

## Delay Announcements

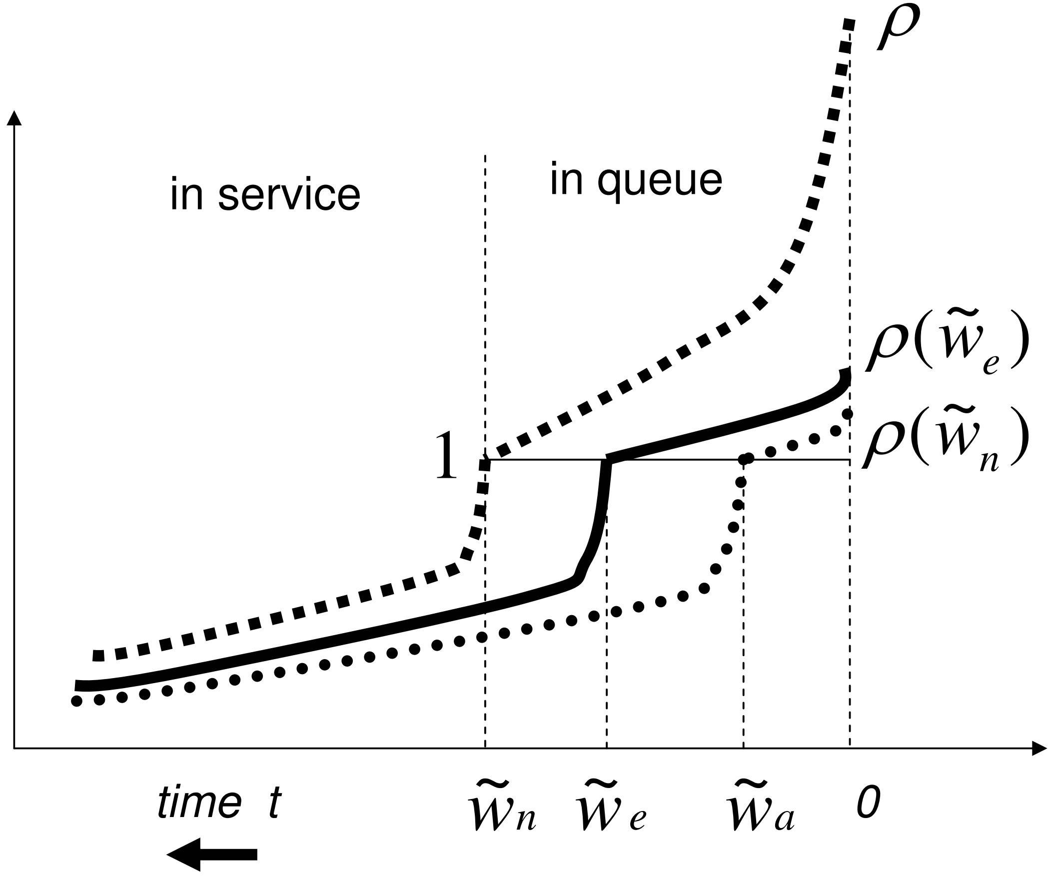
"The Impact of Delay Announcements
in Many-Server Queues
with Abandonment"

Joint work with
Mor Armony and Nahum Shimkin

# Direct Response to Delay Announcement



# An Equilibrium Delay Announcement



# Application 2.

## Uncertainty About the Model Parameters

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

**Random Arrival Rate $\Lambda$**

**Random Number of Servers $\Gamma s$**

# Revenue

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

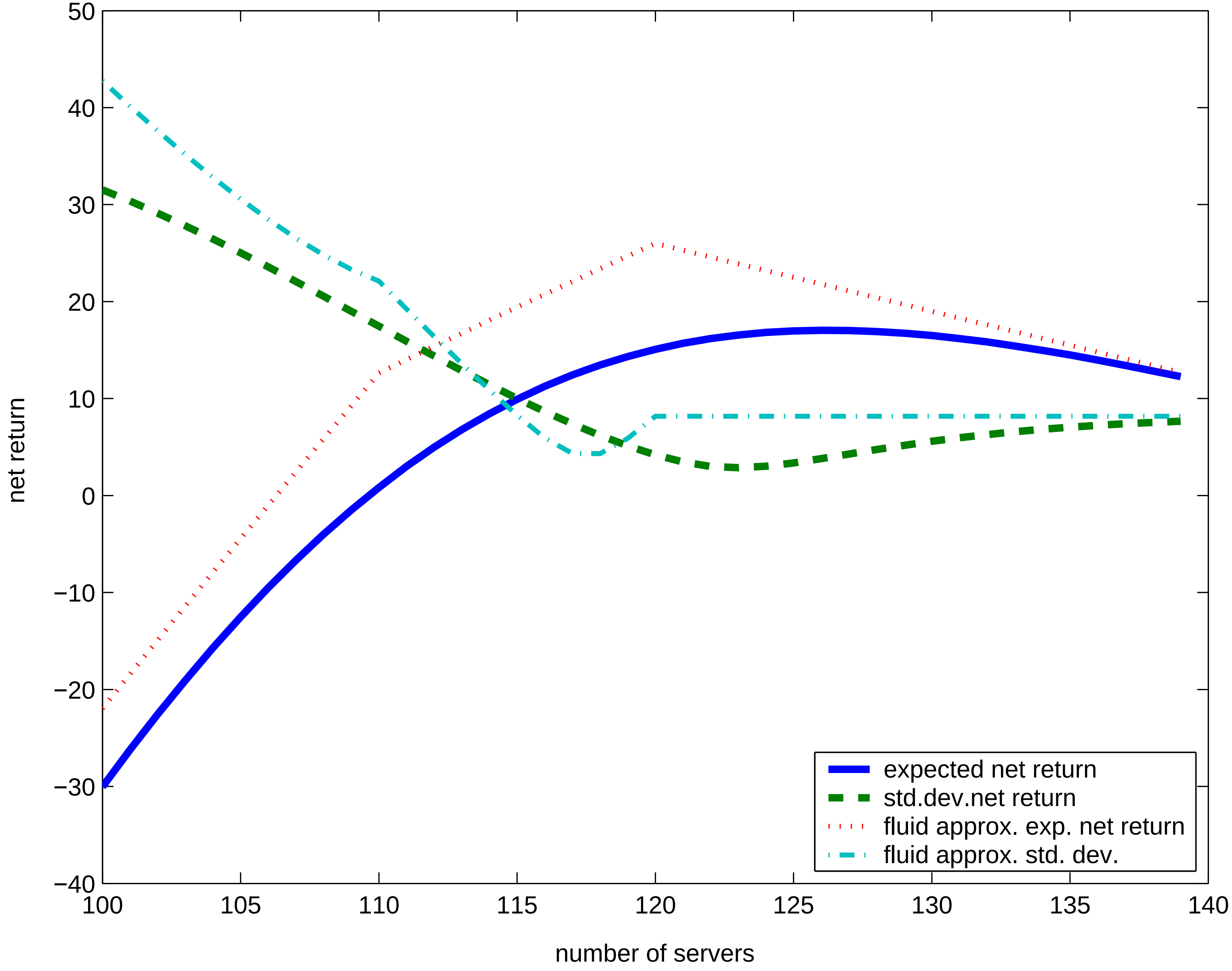
| perf. measure | notation | fluid approx. |
|---|---|---|
| throughput | T(s) | $\Lambda \wedge \Gamma s$ |
| loss rate | L(s) | $(\Lambda - \Gamma s)^+$ |
| waiting rate | $\Lambda$ W(s) | $(\Lambda - \Gamma s)^+ / f(0)$ |

## Example 1.

**M/M/s+M model**

$\Lambda = 100,\ 110$ **or** $120$

**each with probability** 1/3

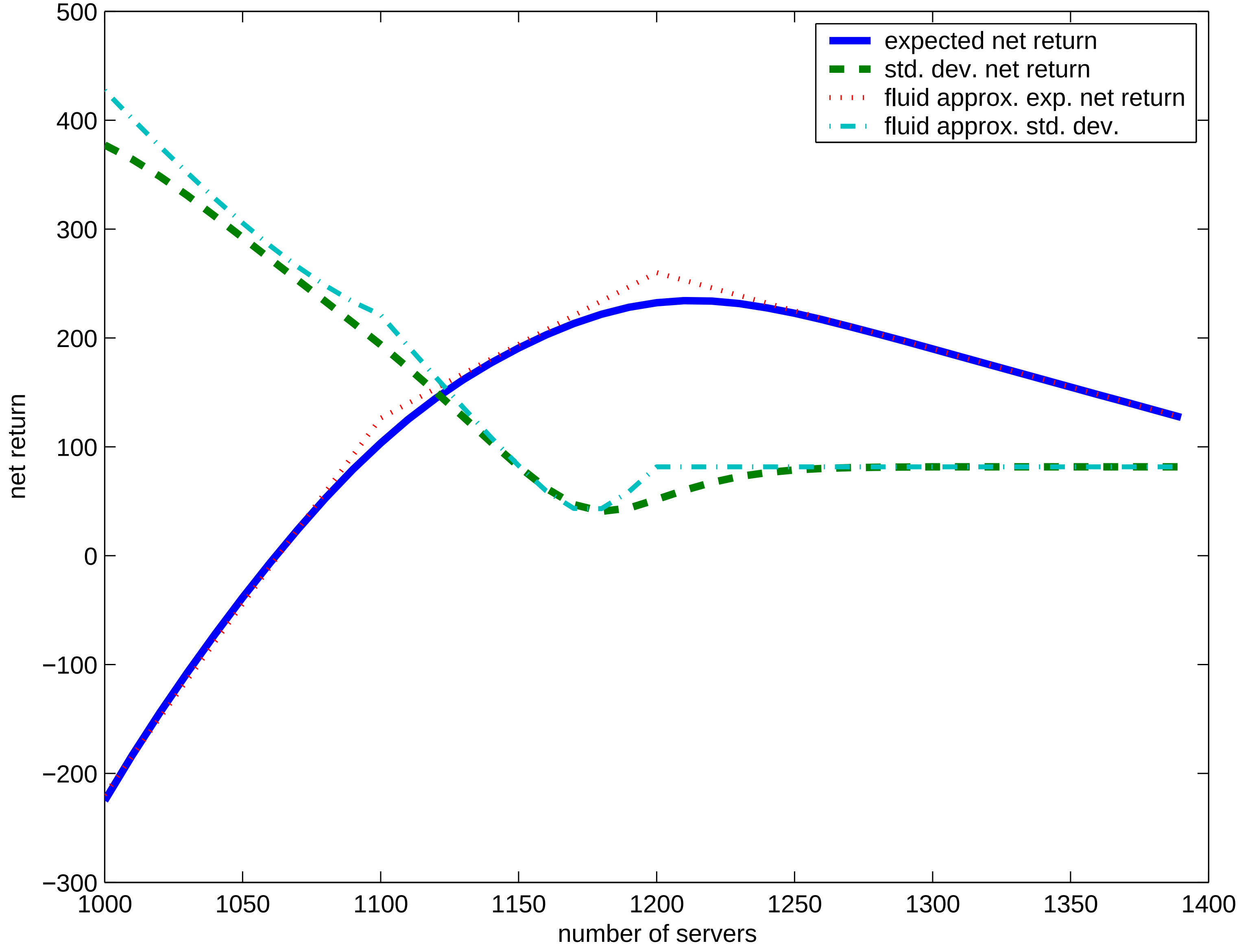


50
40
30
20
10
0
−10
−20
−30
−40
net return
100
105
110
115
120
125
130
135
140
number of servers
expected net return
std.dev.net return
fluid approx. exp. net return
fluid approx. std. dev.

**Example 2.**

$\Lambda = 1000,\ 1100$ **or** $1200$

**each with probability** $1/3$

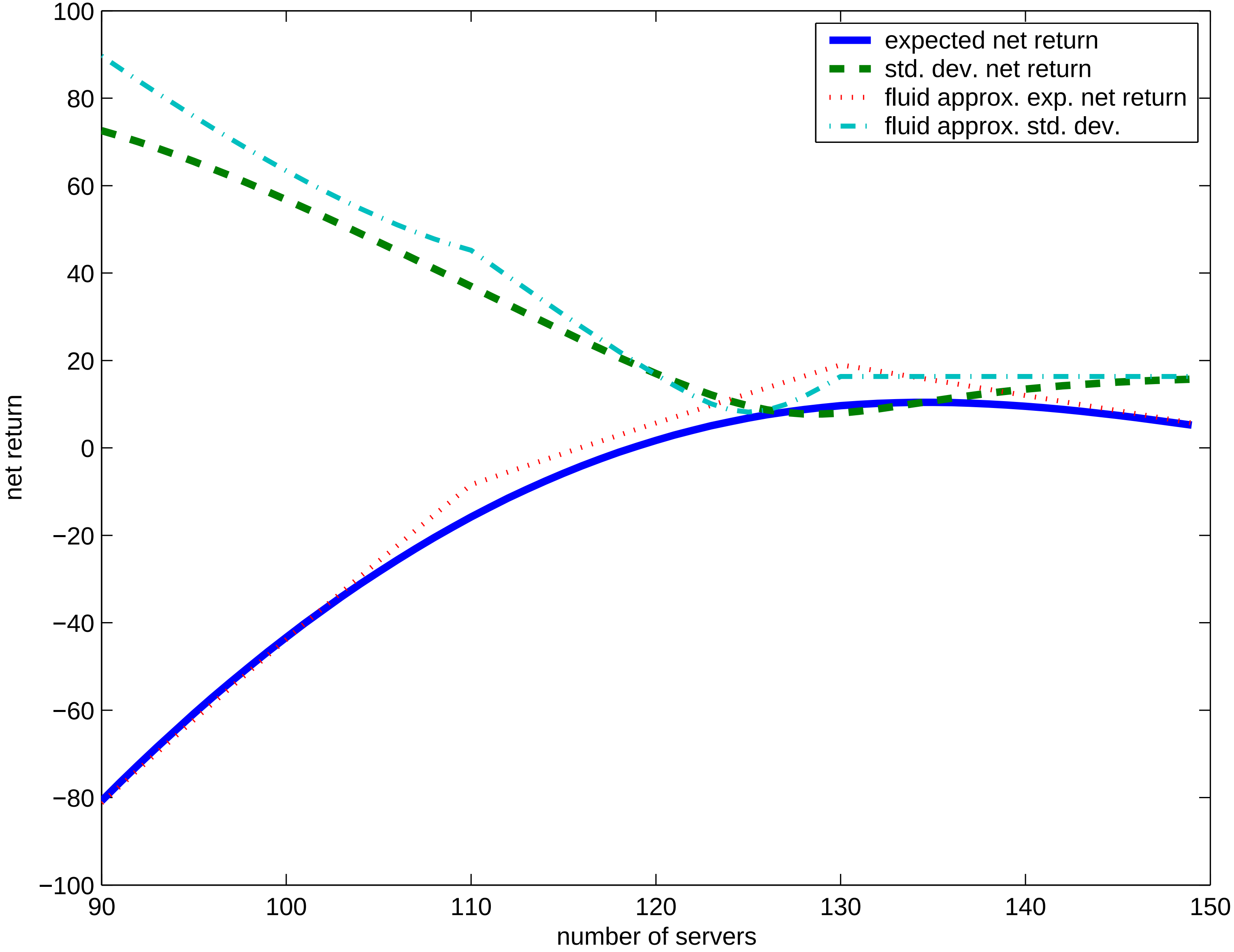


500
400
300
200
100
0
−100
−200
−300
1000
1050
1100
1150
1200
1250
1300
1350
1400
net return
number of servers
expected net return
std. dev. net return
fluid approx. exp. net return
fluid approx. std. dev.

**Example 3.**

$\Lambda = 90,\ 110$ **or** $130$

**each with probability** $1/3$


expected net return
std. dev. net return
fluid approx. exp. net return
fluid approx. std. dev.
net return
number of servers
100
80
60
40
20
0
−20
−40
−60
−80
−100
90
100
110
120
130
140
150

# Summary

## Fluid Approximation for M/GI/s+GI

♠ See impact of non-exponential distributions

♠ Has useful applications

◇ Delay announcements

◇ Model parameter uncertainty

# References


1. WW, Fluid models for multi-server queues with abandonments. *Operations Research*, forthcoming. Available at http://columbia.edu/∼ww2040.

2. Mor Armony, Nahum Shimkin and WW. The impact of delay announcements in many-server queues with abandonments. Working paper, 2005. Available at http://columbia.edu/∼ww2040.

3. WW, Staffing a call center with uncertain arrival rate and absenteeism. *Production and Operations Management*, forthcoming. Available at http://columbia.edu/∼ww2040.

4. N. G. Duffield and WW. Control and recovery from rare congestion events in a large multi-server system. *Queueing Systems* 26 (1997) 69–104.