

# Queueing Networks with Time Varying Rates for Modeling Call Centers

**William A. Massey**

Department of Operations Research  
and Financial Engineering,  
Princeton University

[wmassey@princeton.edu](mailto:wmassey@princeton.edu)

# **ACKNOWLEDGMENTS TO CO-AUTHORS**

Michael Fu

Robert Hampshire

Avi Mandelbaum

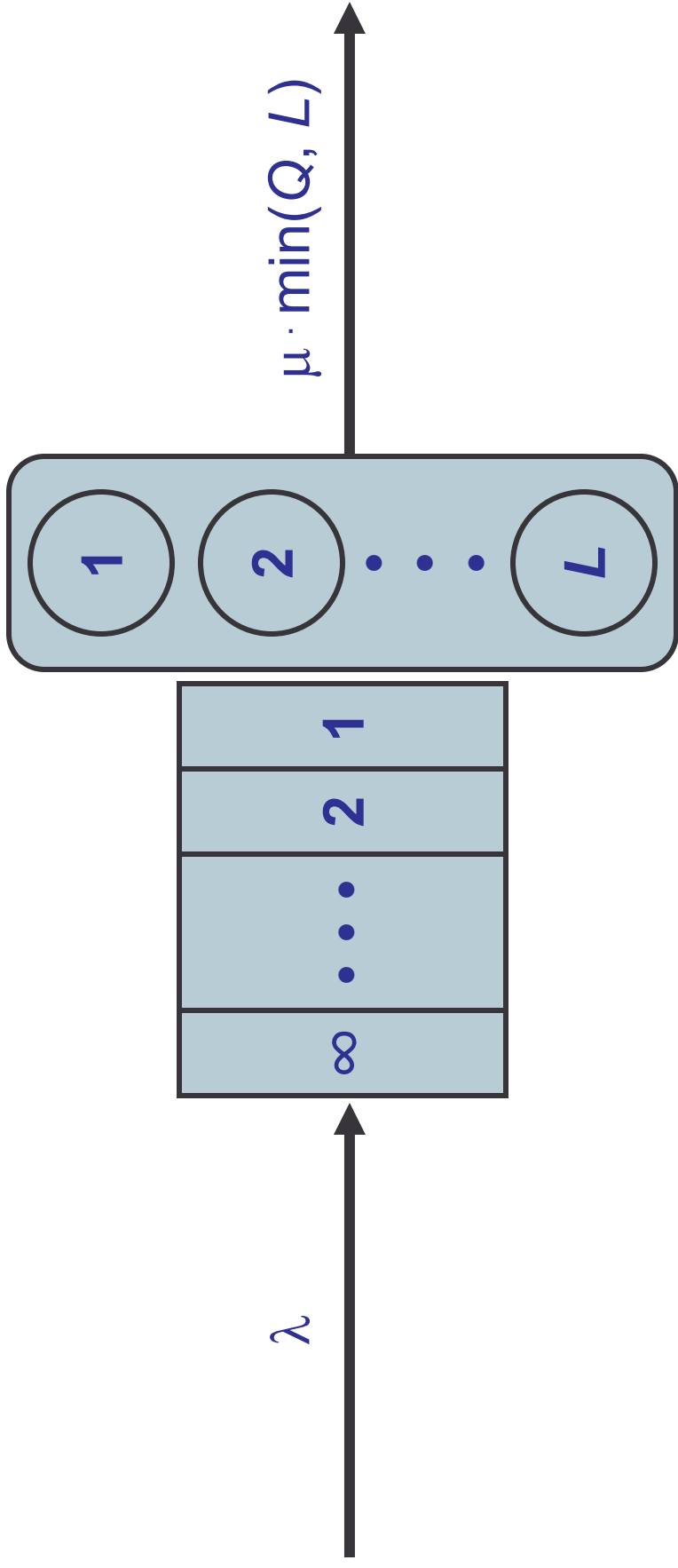
Martin Reiman

Brian Rider

Ahmed Ridley

Alexander Stolyar

# Call Centers with “Music” (Delays): The $M/M/L/\infty$ Queuing Model



We have a Poisson arrival rate  $\lambda$ , mean service time  $1/\mu$ ,  $L$  call center agents and an infinite number of telephone lines.

## THE ERLANG B FORMULA

For our purposes, the Erlang B formula is just a useful special function with a simple recursion relation (setting  $\beta_0(x) \equiv 1$ ),

$$\beta_L(x) \equiv \frac{x^L}{L!} / \sum_{\ell=0}^L \frac{x^\ell}{\ell!} \Rightarrow \beta_L(x) = \frac{x \cdot \beta_{L-1}(x)}{L + x \cdot \beta_{L-1}(x)}.$$

## THE ERLANG C FORMULA

If  $D$  is the delay for an  $M/M/L/\infty$  queue in steady state (i.e.  $q \equiv \lambda/\mu < L$ ), then we can express the probability of a non-zero delay in terms of the Erlang C formula  $\gamma_L(q)$

$$\gamma_L(q) \equiv \frac{L \cdot \beta_L(q)}{L - q + q \cdot \beta_L(q)} = P(D > 0) = P(Q \geq L).$$

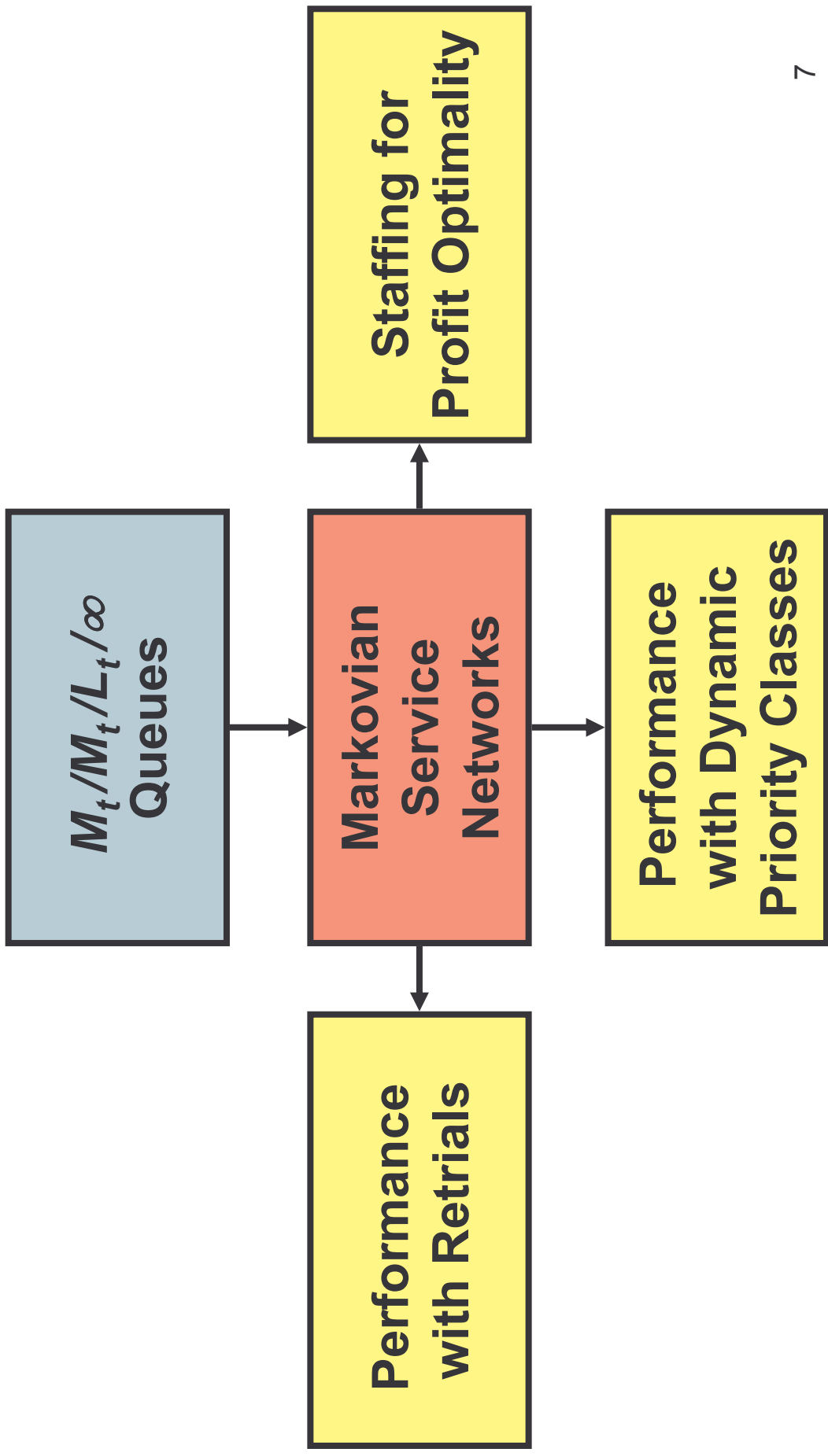
Moreover,

$$P(D > t) = \gamma_L(q) \cdot e^{-(L-q) \cdot \mu t} \quad \text{and} \quad E[D] = \frac{\gamma_L(q)}{(L-q) \cdot \mu}.$$

## **OUR RESEARCH GOAL**

Expand the family of tractable queueing models relevant to call centers.

# EXTENSIONS OF MULTI-SERVER QUEUEING MODELS



## SAMPLE PATH CONSTRUCTION FOR THE $M_t/M_t/L_t$ QUEUE

If  $\{ Q(t) \mid t \geq 0 \}$  is the  $M_t/M_t/L_t$  queueing system process, then

$$Q(t) = Q(0) + \Pi_1 \left( \int_0^t \lambda_s ds \right) - \Pi_2 \left( \int_0^t \mu_s \cdot \min(Q(s), L_s) ds \right)$$

where  $\{ \Pi_i(t) \mid t \geq 0 \}$  for  $i=1, 2$  are i.i.d. standard (rate 1) Poisson processes.

# UNIFORM ACCELERATION FOR THE $M_t/M_t/L_t$ QUEUE

Construct a scaled queueing system process  $\{Q^\eta(t) \mid t \geq 0\}$ ,  
with scale factor  $\eta$ , where

$$\begin{aligned} Q^\eta(t) &= Q^\eta(0) + \Pi_1 \left( \int_0^t \eta \lambda_s ds \right) - \Pi_2 \left( \int_0^t \eta \mu_s \cdot \min \left( \frac{1}{\eta} Q^\eta(s), L_s \right) ds \right) \\ &= Q^\eta(0) + \Pi_1 \left( \int_0^t \eta \lambda_s ds \right) - \Pi_2 \left( \int_0^t \mu_s \cdot \min \left( Q^\eta(s), \eta L_s \right) ds \right) \end{aligned}$$

If we do an asymptotic analysis for the case of  $\eta \rightarrow \infty$ , then  
we have the Halfin-Whitt scaling for multi-server queues.

# FUNCTIONAL STRONG LAW OF LARGE NUMBERS FOR THE $M_t/M_t/L_t$ QUEUE

Taking the limit as  $\eta \rightarrow \infty$  gives us

$$\frac{1}{\eta} Q^\eta \xrightarrow{\text{a.s.}} Q^{(0)} \quad (\text{u.o.c.}),$$

where  $\{ Q^{(0)}(t) \mid t \geq 0 \}$  is a dynamical system that solves the ordinary differential equation

$$\frac{d}{dt} Q^{(0)}(t) = \lambda_t - \mu_t \cdot (Q^{(0)}(t) \wedge L_t).$$

We call this the fluid limit for the  $M_t/M_t/L_t$  queue.

# FUNCTIONAL CENTRAL LIMIT THEOREM FOR THE $M_t/M_t/L_t$ QUEUE

Taking the limit as  $\eta \rightarrow \infty$  gives us

$$\sqrt{\eta} \cdot \left( \frac{1}{\eta} Q^\eta - Q^{(0)} \right) \xrightarrow{d} Q^{(1)} \quad (J_1 \text{ topology})$$

Moreover, if  $\{ t \mid Q^{(0)}(t) = L_t \}$  is a set of measure zero, then  $\{ Q^{(1)}(t) \mid t \geq 0 \}$  is a Gaussian process whose mean and autocovariance, coupled with the fluid limit, are dynamical systems.

We call this the diffusion limit (but not a heavy traffic limit) for the  $M_t/M_t/L_t$  queue.

## DIFFERENTIAL EQUATIONS FOR THE DIFFUSION LIMIT OF THE $M_t/M_t/L_t$ QUEUE

If  $\{t \mid Q^{(0)}(t) = L_t\}$  is a set of measure zero, then

$$\frac{d}{dt} E \left[ \bar{Q}^{(1)}(t) \right] = -\mu_t \cdot 1_{\{\bar{Q}^{(0)}(t) \leq L_t\}} \cdot E \left[ \bar{Q}^{(1)}(t) \right]$$

$$\begin{aligned} \frac{d}{dt} \text{Var} \left[ \bar{Q}^{(1)}(t) \right] &= -2\mu_t \cdot 1_{\{\bar{Q}^{(0)}(t) \leq L_t\}} \cdot \text{Var} \left[ \bar{Q}^{(1)}(t) \right] \\ &\quad + \lambda_t + \mu_t \cdot \left( \bar{Q}^{(0)}(t) \wedge L_t \right) \end{aligned}$$

$$\frac{d}{dt} \text{Cov} \left[ \bar{Q}^{(1)}(s), \bar{Q}^{(1)}(t) \right] = -\mu_t \cdot 1_{\{\bar{Q}^{(0)}(t) \leq L_t\}} \cdot \text{Cov} \left[ \bar{Q}^{(1)}(s), \bar{Q}^{(1)}(t) \right]$$

## MARKOVIAN SERVICE NETWORKS

We define a **Markovian service network**, as a vector-valued process  $\{ \mathbf{Q}(t) \mid t \geq 0 \}$  that is the unique solution to the equation

$$\mathbf{Q}(t) = \mathbf{Q}(0) + \sum_{i \in I} \Pi_i \left( \int_0^t \alpha_s(\mathbf{Q}(s); i) ds \right) \mathbf{v}_i$$

where the  $\{ \Pi_i \mid i \in I \}$  are an i.i.d. family of standard Poisson processes, the  $\{ \alpha(\cdot, i) \mid i \in I \}$  are time-varying, state-dependent rate functions, and the  $\{ \mathbf{v}_i \mid i \in I \}$  are transition vectors.

## UNIFORM ACCELERATION FOR MARKOVIAN SERVICE NETWORKS

Construct a scaled queueing system process  $\{ \mathbf{Q}^\eta(t) \mid t \geq 0 \}$ , with scale factor  $\eta$ , where

$$\mathbf{Q}^\eta(t) = \mathbf{Q}^\eta(0) + \sum_{i \in I} \Pi_i \left( \int_0^t \alpha_s^\eta \left( \frac{1}{\eta} \mathbf{Q}^\eta(s); i \right) ds \right) \mathbf{v}_i$$

and the time-varying, state-dependent rate functions have the following type of asymptotic behavior

$$\alpha_s^\eta(\cdot, i) = \eta \cdot \alpha_s^{(0)}(\cdot, i) + \sqrt{\eta} \cdot \alpha_s^{(1)}(\cdot, i) + o(\sqrt{\eta}).$$

# FUNCTIONAL STRONG LAW OF LARGE NUMBERS FOR MARKOVIAN SERVICE NETWORKS

Taking the limit as  $\eta \rightarrow \infty$  gives us

$$\frac{1}{\eta} \mathbf{Q}^\eta \xrightarrow{\text{a.s.}} \mathbf{Q}^{(0)} \quad (\text{u.o.c.})$$

where the fluid limit  $\{ \mathbf{Q}^{(0)}(t) \mid t \geq 0 \}$  is a dynamical system.

# FUNCTIONAL CENTRAL LIMIT THEOREM FOR MARKOVIAN SERVICE NETWORKS

Taking the limit as  $\eta \rightarrow \infty$  gives us

$$\sqrt{\eta} \cdot \left( \frac{1}{\eta} \mathbf{Q}^\eta - \mathbf{Q}^{(0)} \right) \xrightarrow{d} \mathbf{Q}^{(1)} \quad (J_1 \text{ topology})$$

Moreover, if a derived set of time points from the fluid limit has measure zero, then diffusion limit  $\{ \mathbf{Q}^{(1)}(t) \mid t \geq 0 \}$  is a Gaussian process whose mean and covariance matrices, coupled with the fluid limit, are dynamical systems.

## STRONG APPROXIMATION FOR STANDARD POISSON PROCESSES

If  $\{ \Pi(t) \mid t \geq 0 \}$  is a standard Poisson process, then we can construct a standard Brownian motion  $\{ B(t) \mid t \geq 0 \}$  such that

$$E \left[ \sup_{t \geq 0} \frac{|\Pi(t) - t - B(t)|}{\log(2 \vee t)} \right] < \infty.$$

This result is due to Kolmos, Major and Tusnady.

# PROPERTIES OF STANDARD BROWNIAN MOTION

Rotational Invariance:

$$\{B(t) | t \geq 0\} \stackrel{d}{=} \{-B(t) | t \geq 0\}$$

Reflection Principle:

$$\sup_{0 \leq s \leq t} B(s) \stackrel{d}{=} |B(t)|$$

Self-Similarity:

$$\{B(\eta t) / \sqrt{\eta} | t \geq 0\} \stackrel{d}{=} \{B(t) | t \geq 0\}$$

# ALMOST SURE CONVERGENCE LIMIT FOR POISSON PROCESSES

$$\frac{1}{\eta} \sup_{0 \leq s \leq t} |B(\eta s)| \xrightarrow{\text{a.s.}} 0$$

$$E \left[ \sup_{t \geq 0} \frac{|\Pi(t) - t - B(t)|}{\log(2 \vee t)} \right] = E \left[ \sup_{t \geq 0} \frac{|\Pi(\eta t) - \eta t - B(\eta t)|/\eta}{\log(2 \vee \eta t)/\eta} \right]$$

$$\boxed{\frac{1}{\eta} \sup_{0 \leq s \leq t} |\Pi(\eta s) - \eta s| \xrightarrow{\text{a.s.}} 0}$$

# CONVERGENCE IN PROBABILITY LIMITS FOR POISSON PROCESSES

$$\left\{ \left( \Pi^\eta(t), \hat{B}(t) \right) \middle| t \geq 0 \right\} \stackrel{d}{=} \left\{ \left( \Pi(t), B(\eta t) / \sqrt{\eta} \right) \middle| t \geq 0 \right\}$$

$$E \left[ \sup_{t \geq 0} \frac{|\Pi(t) - t - B(t)|}{\log(2 \vee t)} \right] = E \left[ \sup_{t \geq 0} \frac{\left| \Pi^\eta(t) - \eta t \right| / \sqrt{\eta} - \hat{B}(t)}{\log(2 \vee \eta t) / \sqrt{\eta}} \right]$$

$$\frac{1}{\eta} \sup_{0 \leq s \leq t} |\Pi^\eta(\eta s) - \eta s| \xrightarrow{p} 0$$

$$\sup_{0 \leq s \leq t} \frac{1}{\sqrt{\eta}} \left| \Pi(\eta s) - \eta s \right| - \hat{B}(s) \xrightarrow{p} 0$$

$$\boxed{\frac{1}{\eta} \mathbf{Q}^\eta \xrightarrow{a.s.} \mathbf{Q}^{(0)}}$$

 $\Downarrow$ 

$$\frac{1}{\eta} \hat{\mathbf{Q}}^\eta \xrightarrow{p} \mathbf{Q}^{(0)}$$

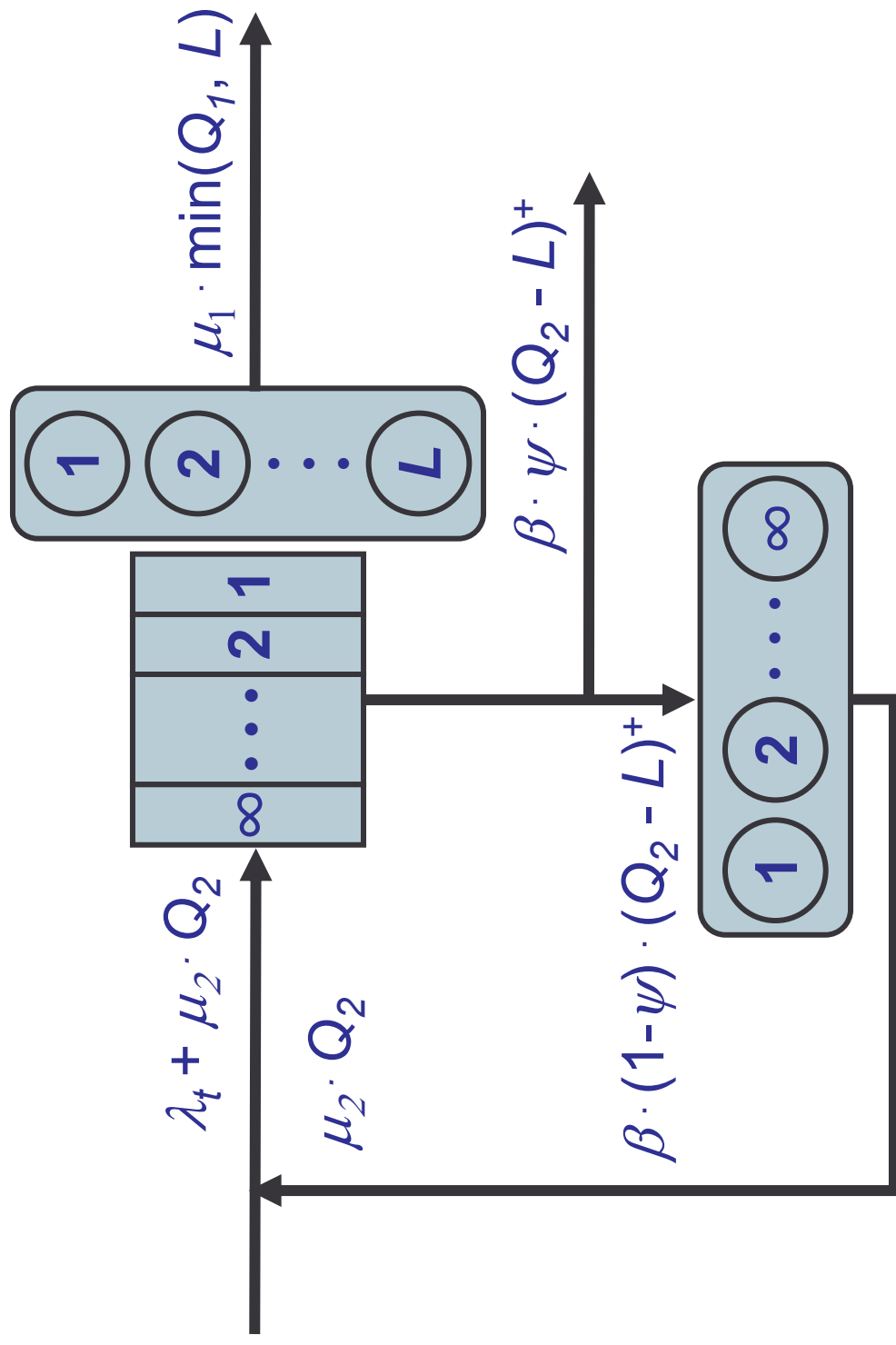
 $\Downarrow$ 

$$\sqrt{\eta} \left( \frac{1}{\eta} \hat{\mathbf{Q}}^\eta - \mathbf{Q}^{(0)} \right) \xrightarrow{p} \mathbf{Q}^{(1)}$$

 $\Downarrow$ 

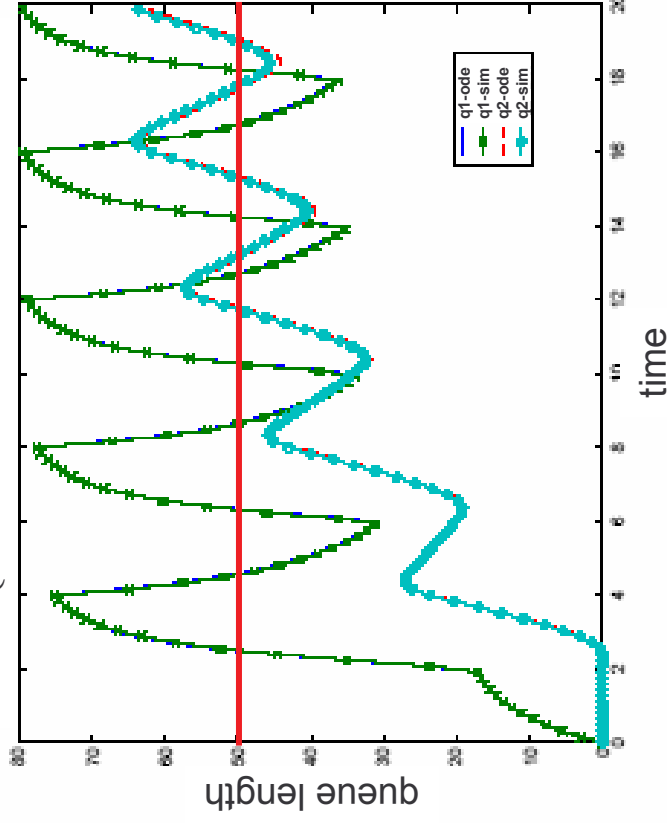
$$\boxed{\sqrt{\eta} \left( \frac{1}{\eta} \mathbf{Q}^\eta - \mathbf{Q}^{(0)} \right) \xrightarrow{d} \mathbf{Q}^{(1)}}$$

# MULTISERVER QUEUE WITH ABANDONMENT AND RETRIALS

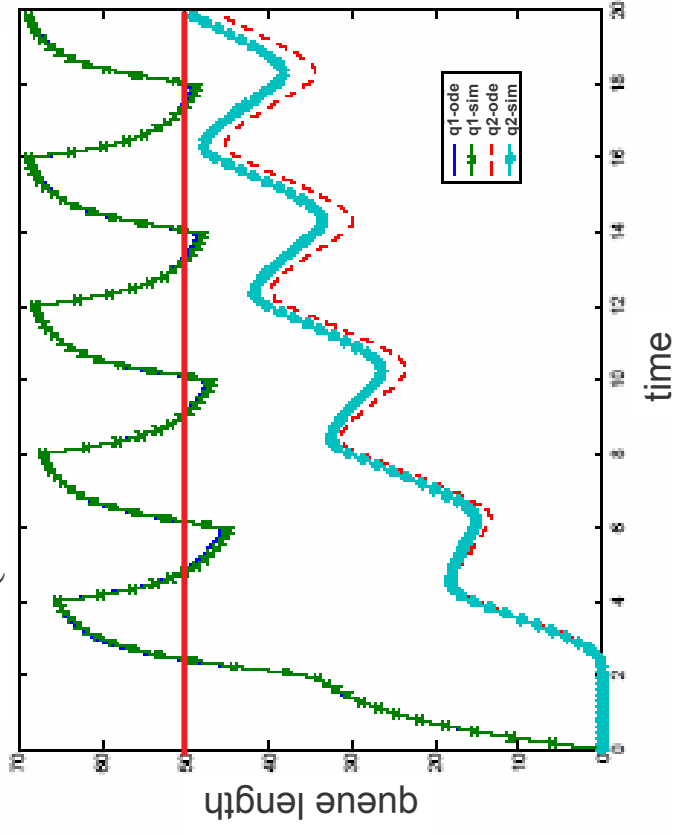


# SAMPLE MEAN VS. FLUID APPROXIMATION FOR THE QUEUEING SYSTEM

$$\lambda(t) = \begin{cases} 20 & \text{if } t \in [0, 2) \cup [4, 6) \cup \dots, \\ 100 & \text{otherwise.} \end{cases}$$



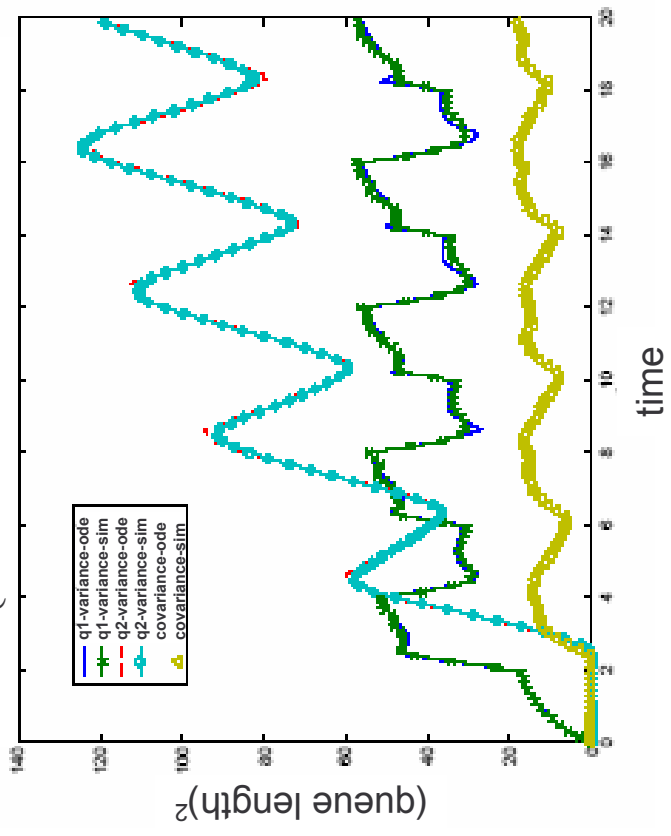
$$\lambda(t) = \begin{cases} 40 & \text{if } t \in [0, 2) \cup [4, 6) \cup \dots, \\ 80 & \text{otherwise.} \end{cases}$$



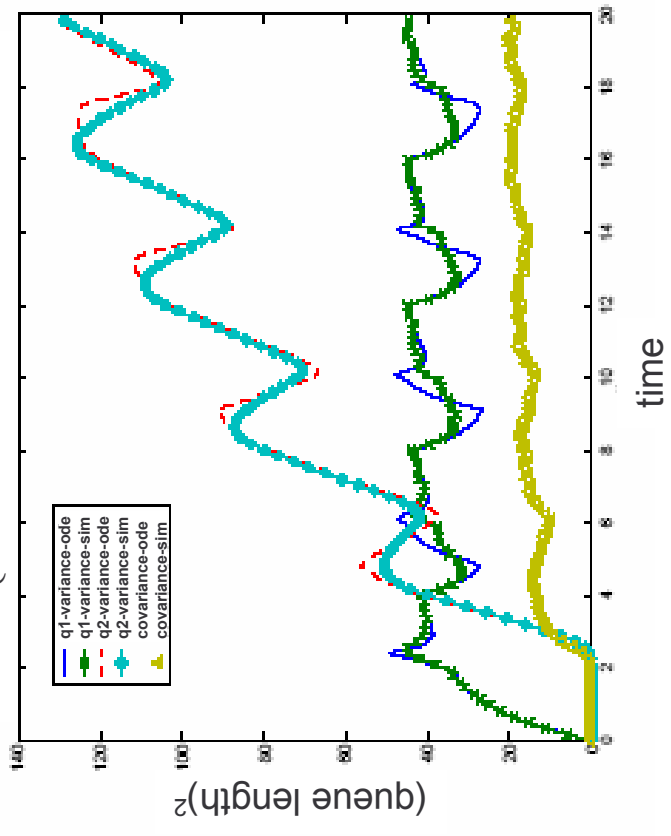
$$L = 50, 1/\mu_1 = 1.0, 1/\mu_2 = 5.0, 1/\beta = 0.5, \psi = 0.5$$

# SAMPLE COVARIANCE VS. DIFFUSION COVARIANCE FOR THE QUEUEING SYSTEM

$$\lambda(t) = \begin{cases} 20 & \text{if } t \in [0, 2) \cup [4, 6) \cup \dots, \\ 100 & \text{otherwise.} \end{cases}$$



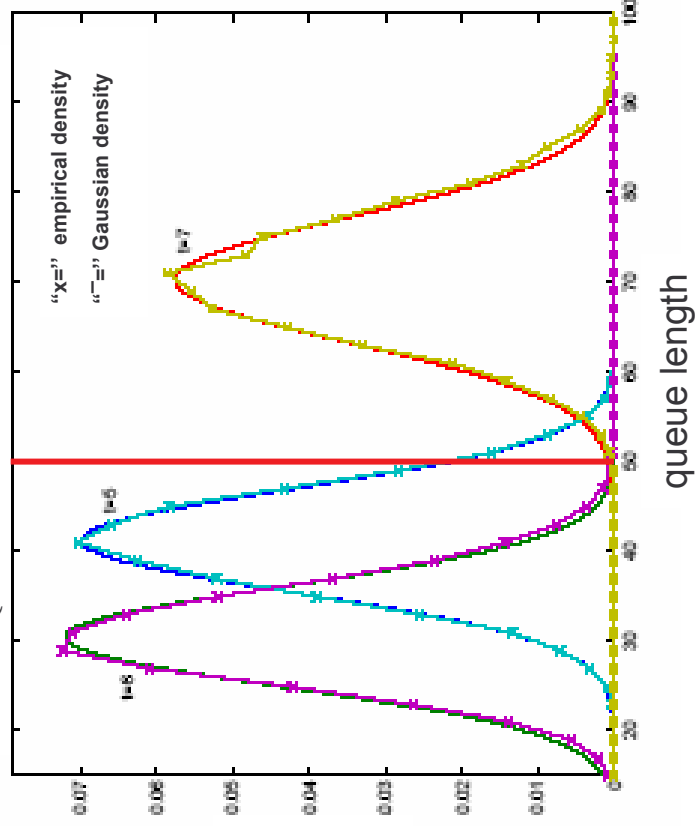
$$\lambda(t) = \begin{cases} 40 & \text{if } t \in [0, 2) \cup [4, 6) \cup \dots, \\ 80 & \text{otherwise.} \end{cases}$$



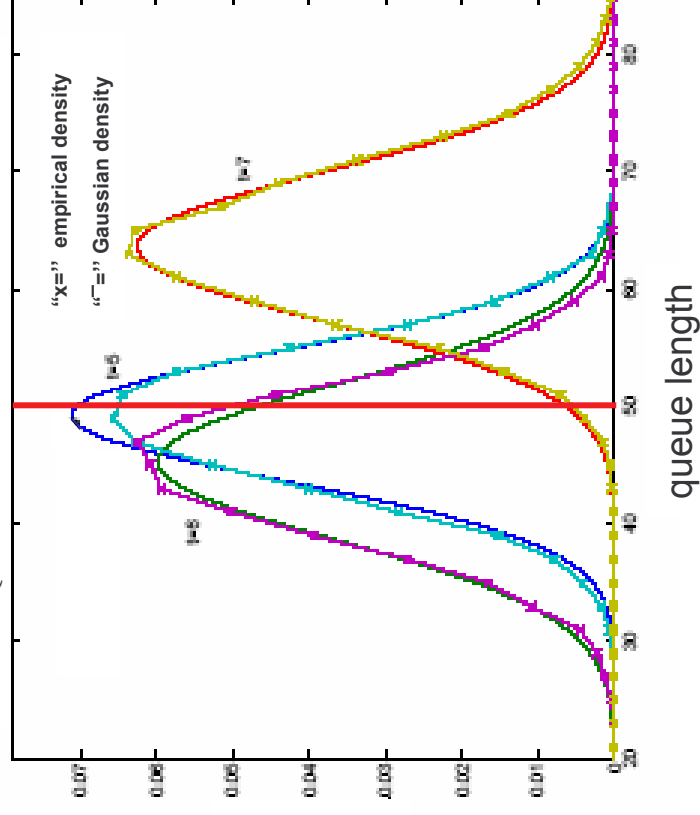
$$L = 50, 1/\mu_1 = 1.0, 1/\mu_2 = 5.0, 1/\beta = 0.5, \psi = 0.5$$

# SAMPLE DENSITY VS. GAUSSIAN APPROXIMATION FOR THE QUEUEING SYSTEM

$$\lambda(t) = \begin{cases} 20 & \text{if } t \in [0, 2) \cup [4, 6) \cup \dots, \\ 100 & \text{otherwise.} \end{cases}$$



$$\lambda(t) = \begin{cases} 40 & \text{if } t \in [0, 2) \cup [4, 6) \cup \dots, \\ 80 & \text{otherwise.} \end{cases}$$



$$L = 50, 1/\mu_1 = 1.0, 1/\mu_2 = 5.0, 1/\beta = 0.5, \psi = 0.5$$

## VIRTUAL DELAY FOR THE $M_t/M_t/L_t/\infty$ /FCFS QUEUE

Construct an associated queueing system process with arrival rate

$$\hat{\lambda}_t = \begin{cases} \lambda_t & \text{if } t \leq \tau, \\ 0 & \text{otherwise.} \end{cases}$$

The virtual delay for the original multi-server queueing model of a customer arriving at time  $\tau$  is  $W(\tau) - \tau$  where

$$W(\tau) \equiv \inf \{ t \geq \tau \mid \hat{Q}(t) \leq L_t - 1 \}.$$

The uniformly accelerated version of the virtual delay is

$$W^\eta(\tau) \equiv \inf \{ t \geq \tau \mid \hat{Q}^\eta(t) \leq \eta L_t - 1 \}.$$

## VIRTUAL DELAY FLUID LIMIT

Taking the limit as  $\eta \rightarrow \infty$  gives us

$$W^\eta \xrightarrow{\text{a.s.}} W^{(0)},$$

where

$$W^{(0)}(\tau) \equiv \inf \left\{ t \geq \tau \mid \hat{Q}^{(0)}(t) \leq L_t - 1 \right\},$$

and

$$\hat{Q}^{(0)}(t) = Q^{(0)}(\tau) - \int_\tau^t \mu_s L_s ds.$$

## VIRTUAL DELAY DIFFUSION LIMIT

Taking the limit as  $\eta \rightarrow \infty$  gives us

$$\sqrt{\eta} \cdot \left( W^\eta - W^{(0)} \right) \xrightarrow{d} W^{(1)}$$

where

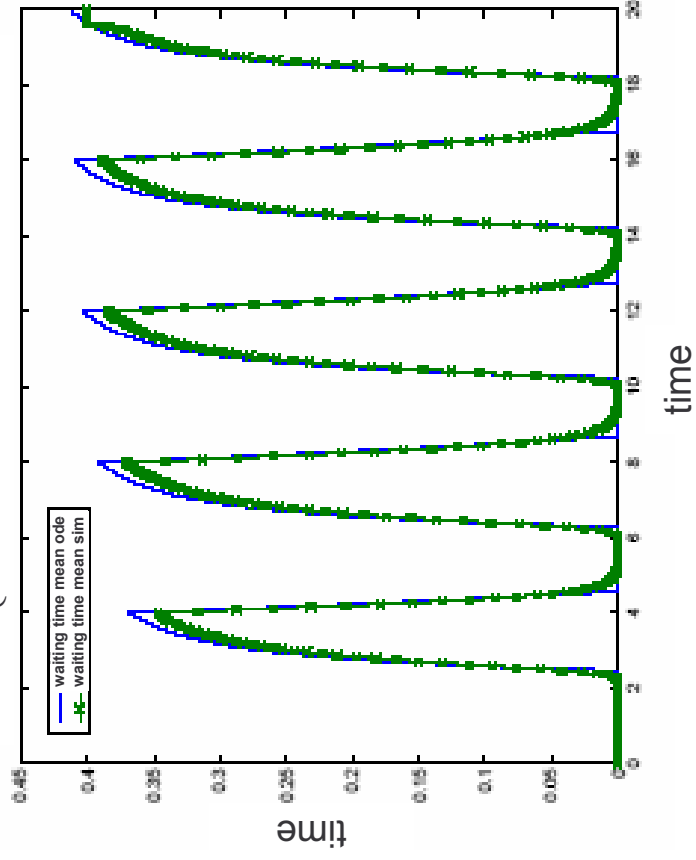
$$W^{(1)}(\tau) = \frac{\hat{Q}^{(1)}(M) \left( W^{(0)}(\tau) \right)}{\mu_{W^{(0)}(\tau)} \cdot L_{W^{(0)}(\tau)}}$$

and when the diffusion limit for the queueing system process is Gaussian, we have

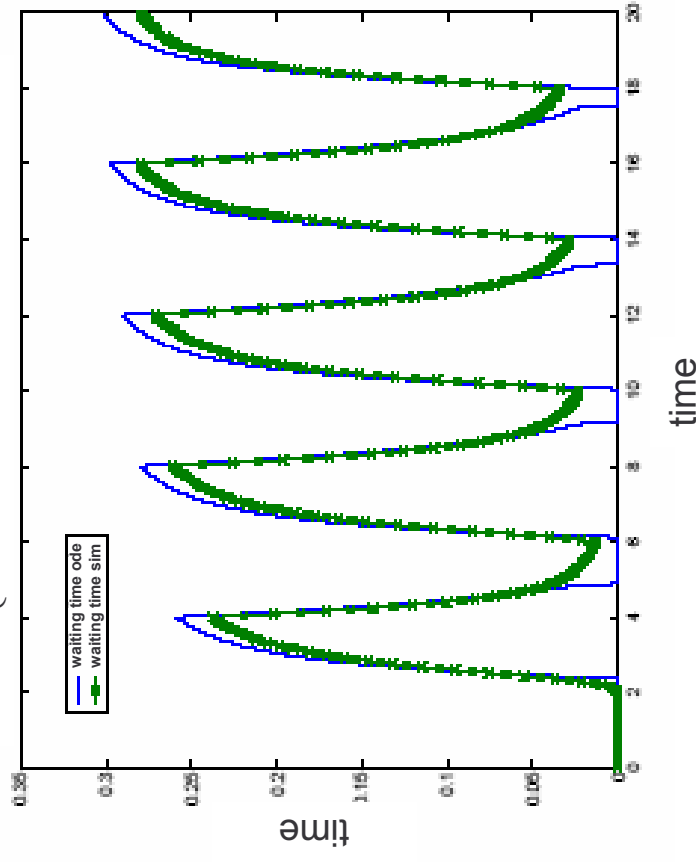
$$\text{Var} \left[ \hat{Q}^{(1)} \left( W^{(0)}(\tau) \right) \right] = \text{Var} \left[ \hat{Q}^{(1)}(\tau) \right] + \int_{W^{(0)}(\tau)}^{\tau} \mu_s L_s ds.$$

# SAMPLE MEAN VS. FLUID APPROXIMATION FOR THE VIRTUAL DELAY

$$\lambda(t) = \begin{cases} 20 & \text{if } t \in [0, 2) \cup [4, 6) \cup \dots, \\ 100 & \text{otherwise.} \end{cases}$$



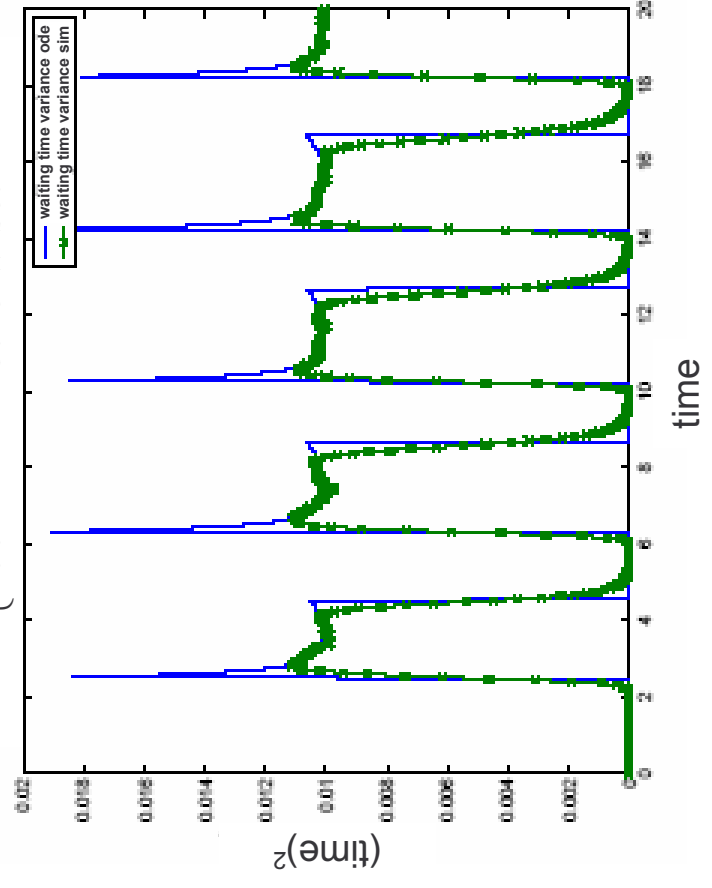
$$\lambda(t) = \begin{cases} 40 & \text{if } t \in [0, 2) \cup [4, 6) \cup \dots, \\ 80 & \text{otherwise.} \end{cases}$$



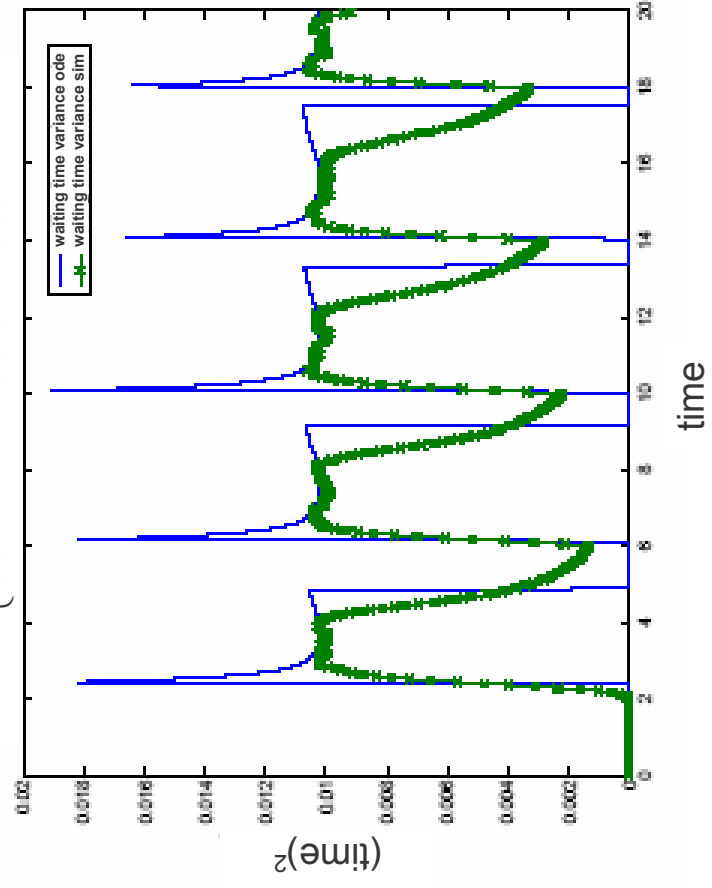
$$L = 50, 1/\mu_1 = 1.0, 1/\mu_2 = 5.0, 1/\beta = 0.5, \psi = 0.5$$

# SAMPLE VARIANCE VS. DIFFUSION VARIANCE FOR THE VIRTUAL DELAY

$$\lambda(t) = \begin{cases} 20 & \text{if } t \in [0, 2) \cup [4, 6) \cup \dots, \\ 100 & \text{otherwise.} \end{cases}$$



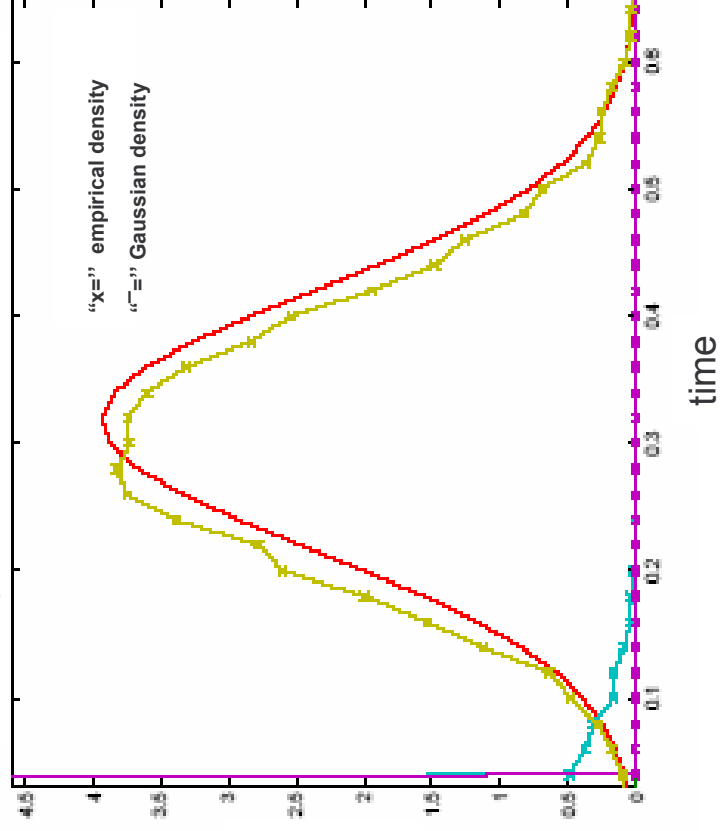
$$\lambda(t) = \begin{cases} 40 & \text{if } t \in [0, 2) \cup [4, 6) \cup \dots, \\ 80 & \text{otherwise.} \end{cases}$$



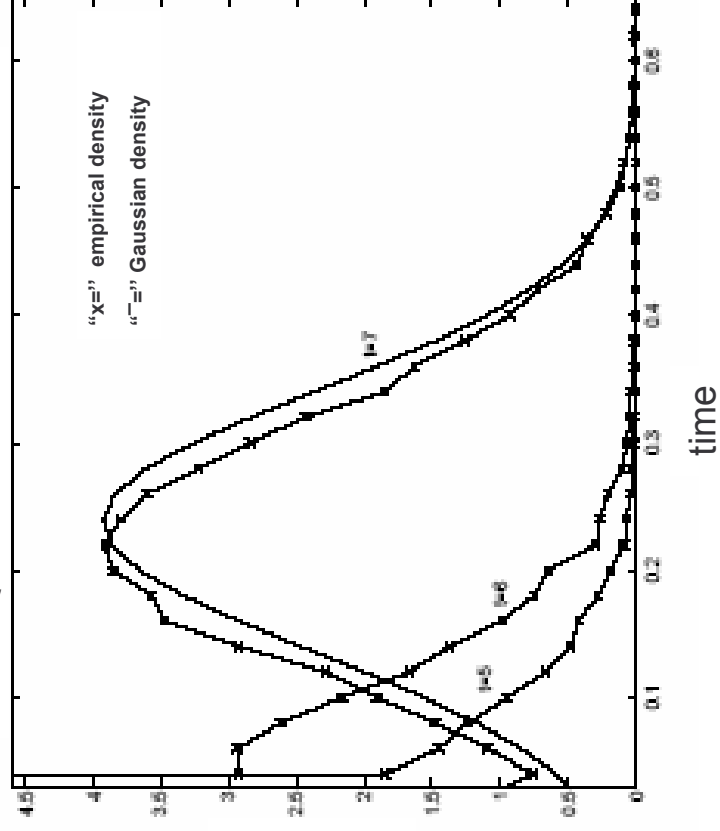
$$L = 50, 1/\mu_1 = 1.0, 1/\mu_2 = 5.0, 1/\beta = 0.5, \psi = 0.5$$

# SAMPLE DENSITY VS. GAUSSIAN APPROXIMATION FOR THE VIRTUAL DELAY

$$\lambda(t) = \begin{cases} 20 & \text{if } t \in [0, 2) \cup [4, 6) \cup \dots, \\ 100 & \text{otherwise.} \end{cases}$$

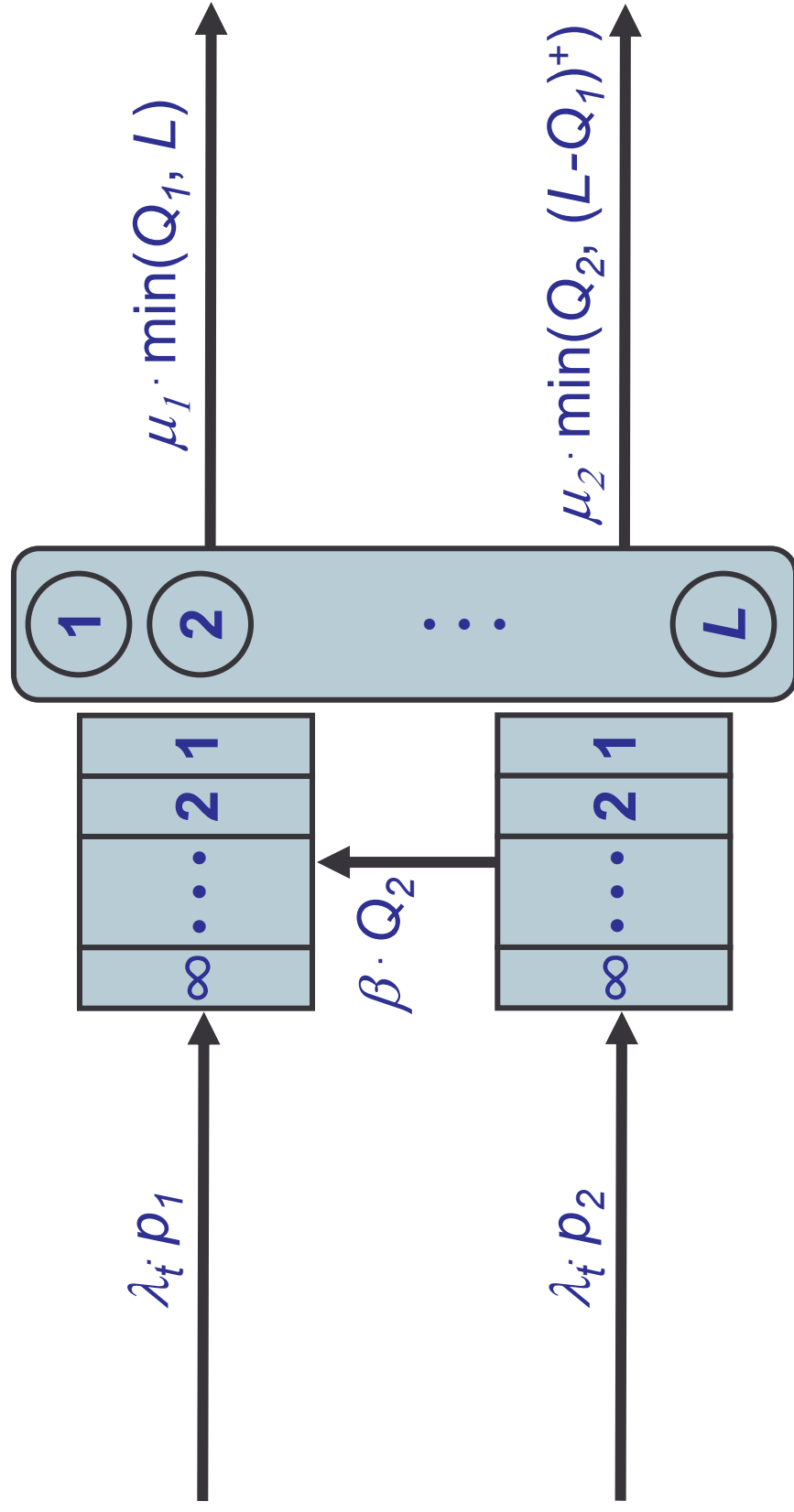


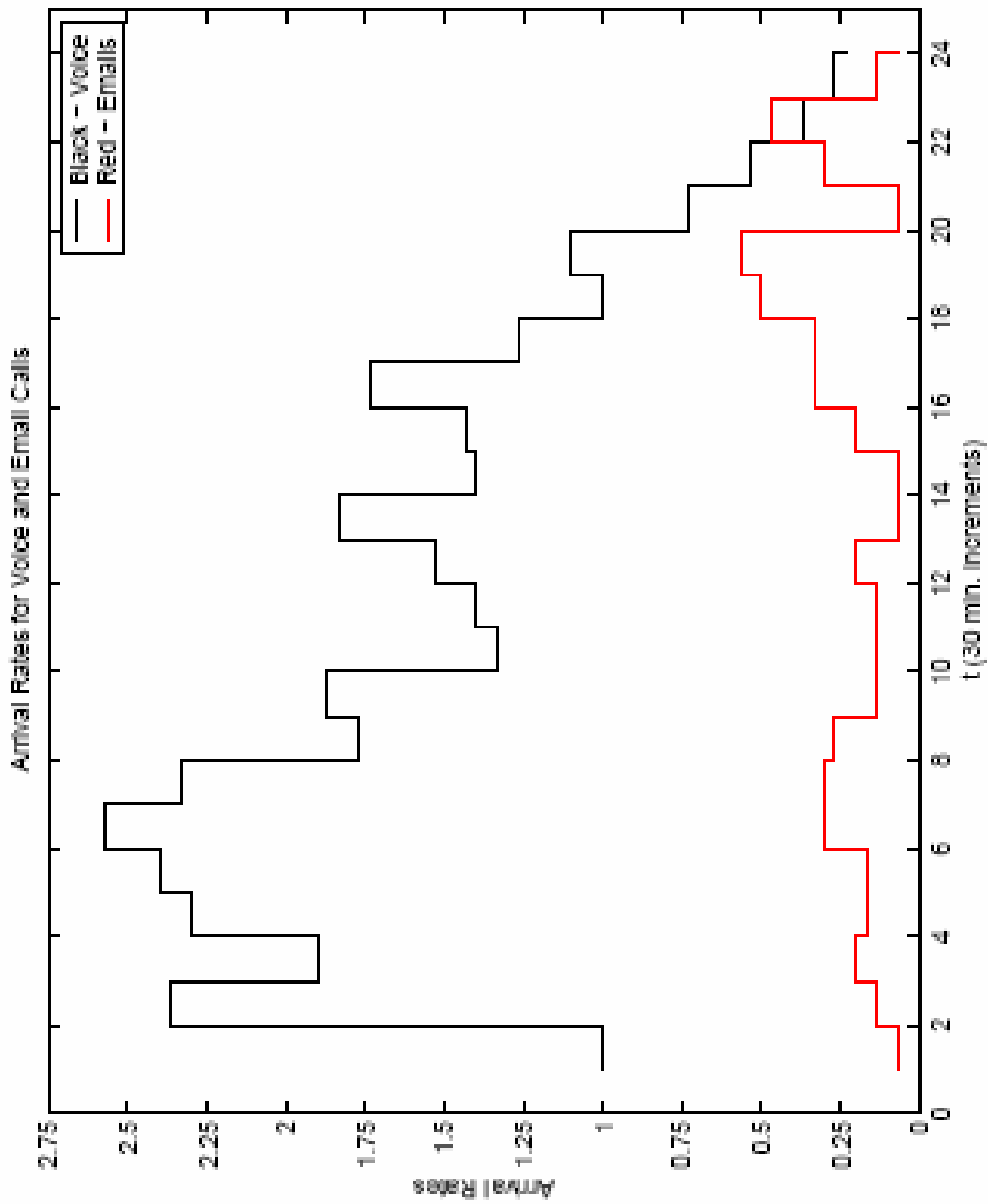
$$\lambda(t) = \begin{cases} 40 & \text{if } t \in [0, 2) \cup [4, 6) \cup \dots, \\ 80 & \text{otherwise.} \end{cases}$$

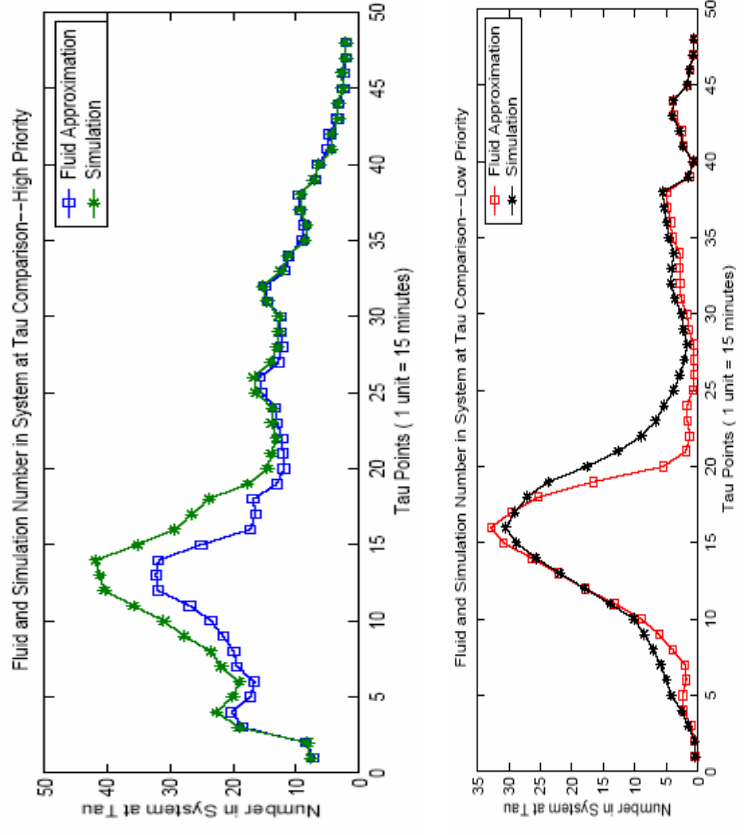
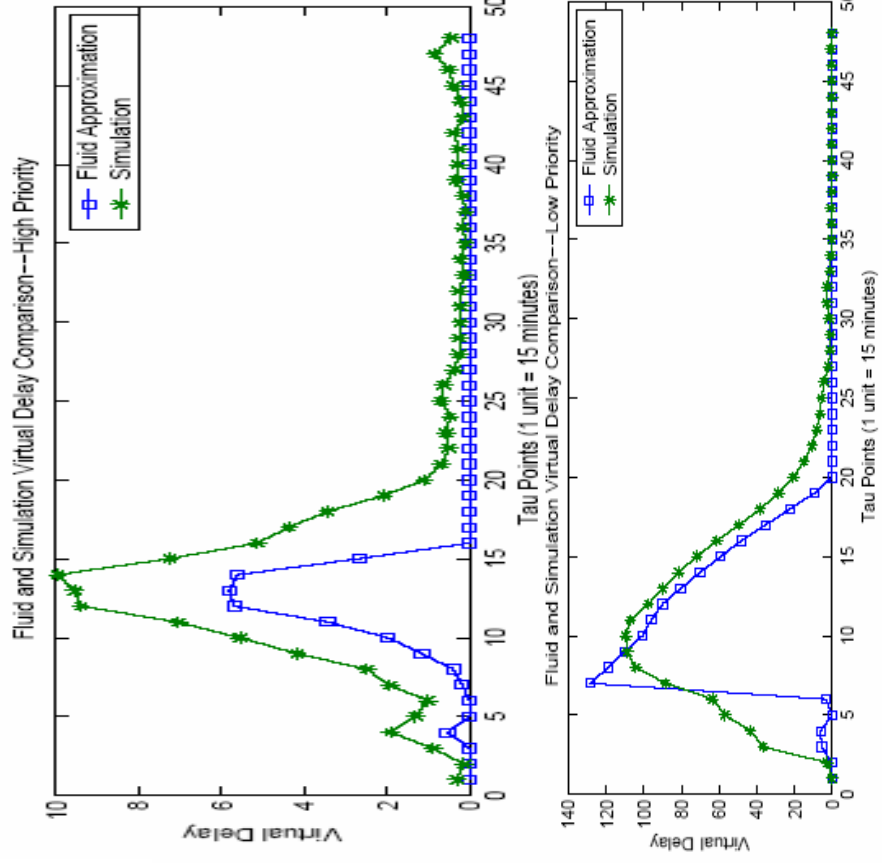


$$L = 50, 1/\mu_1 = 1.0, 1/\mu_2 = 5.0, 1/\beta = 0.5, \psi = 0.5$$

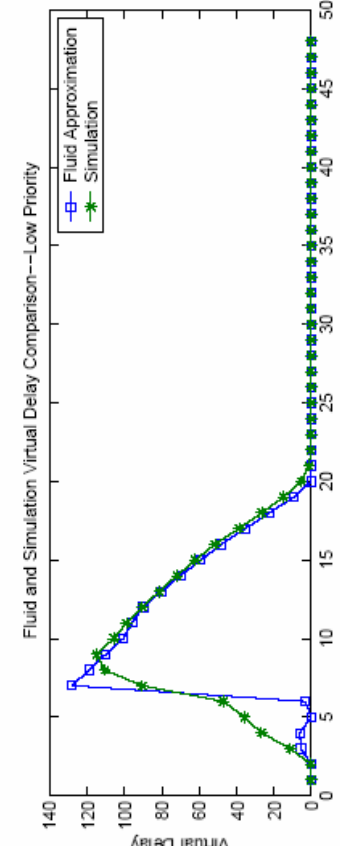
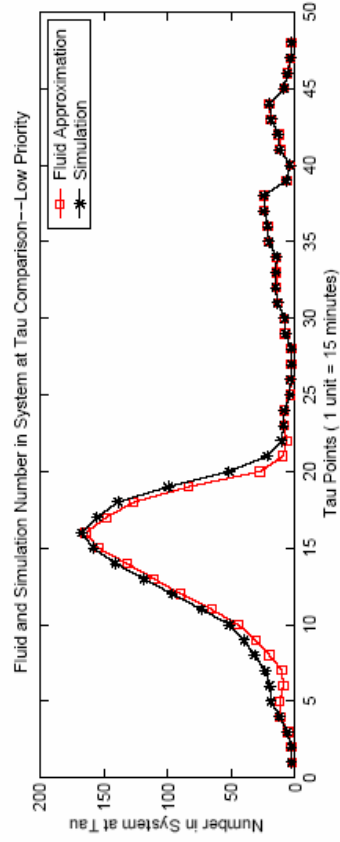
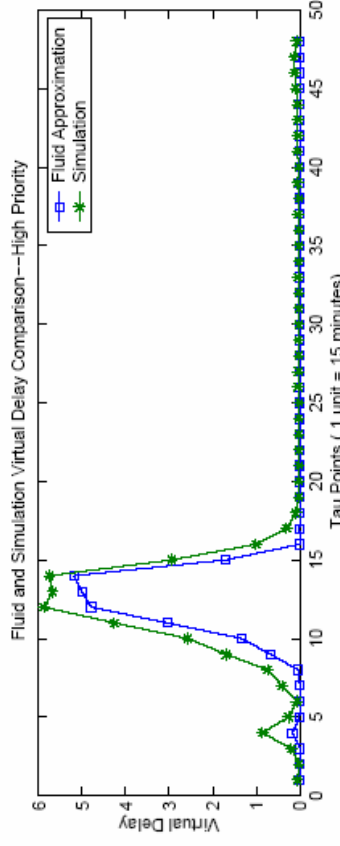
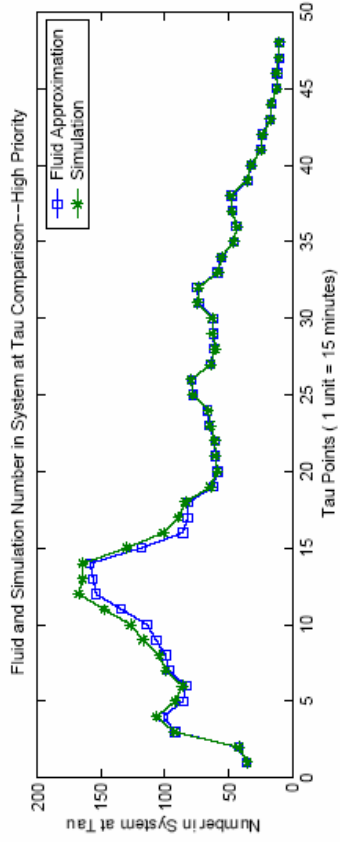
# MULTISERVER QUEUE WITH DYNAMIC, PRE-EMPTIVE PRIORITIES



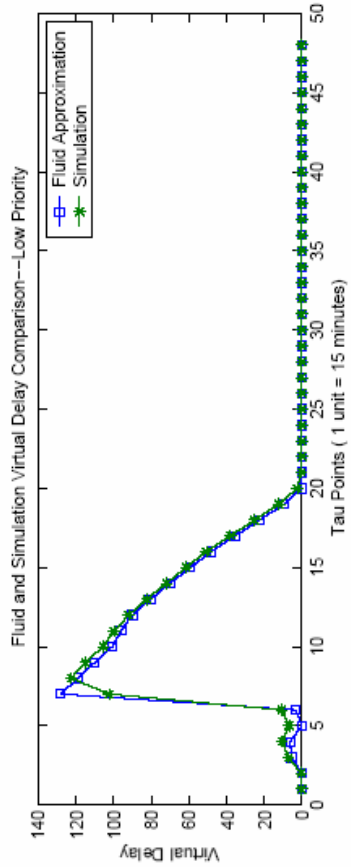
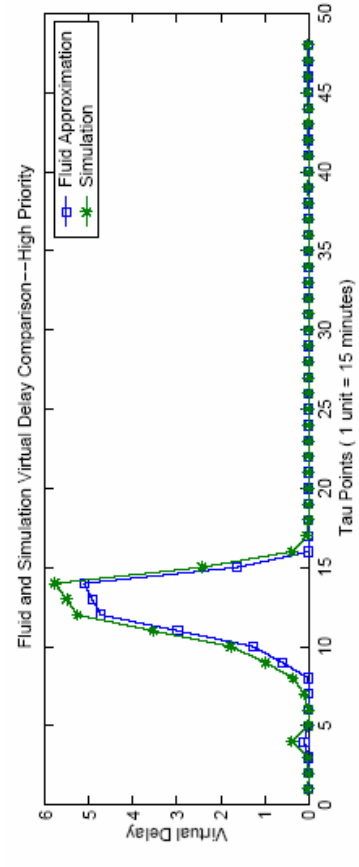
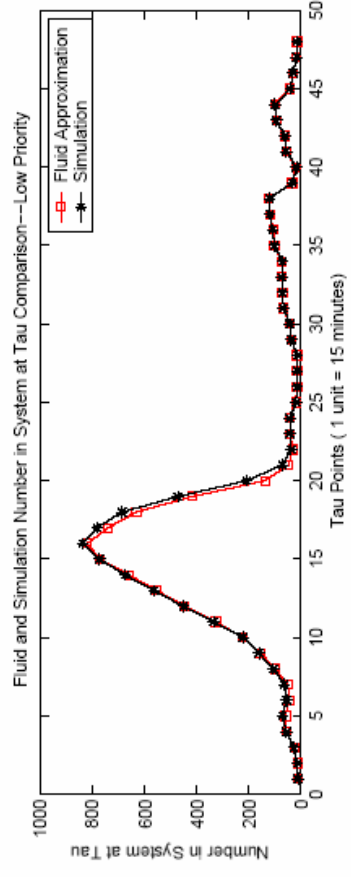
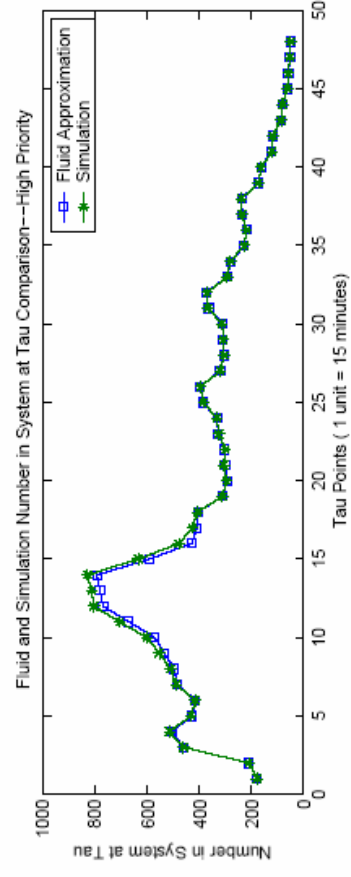




**$L = 20$** ,  $1/\mu_1 = 1/\mu_2 = 8.69$  minutes,  $1/\beta = 90$  minutes

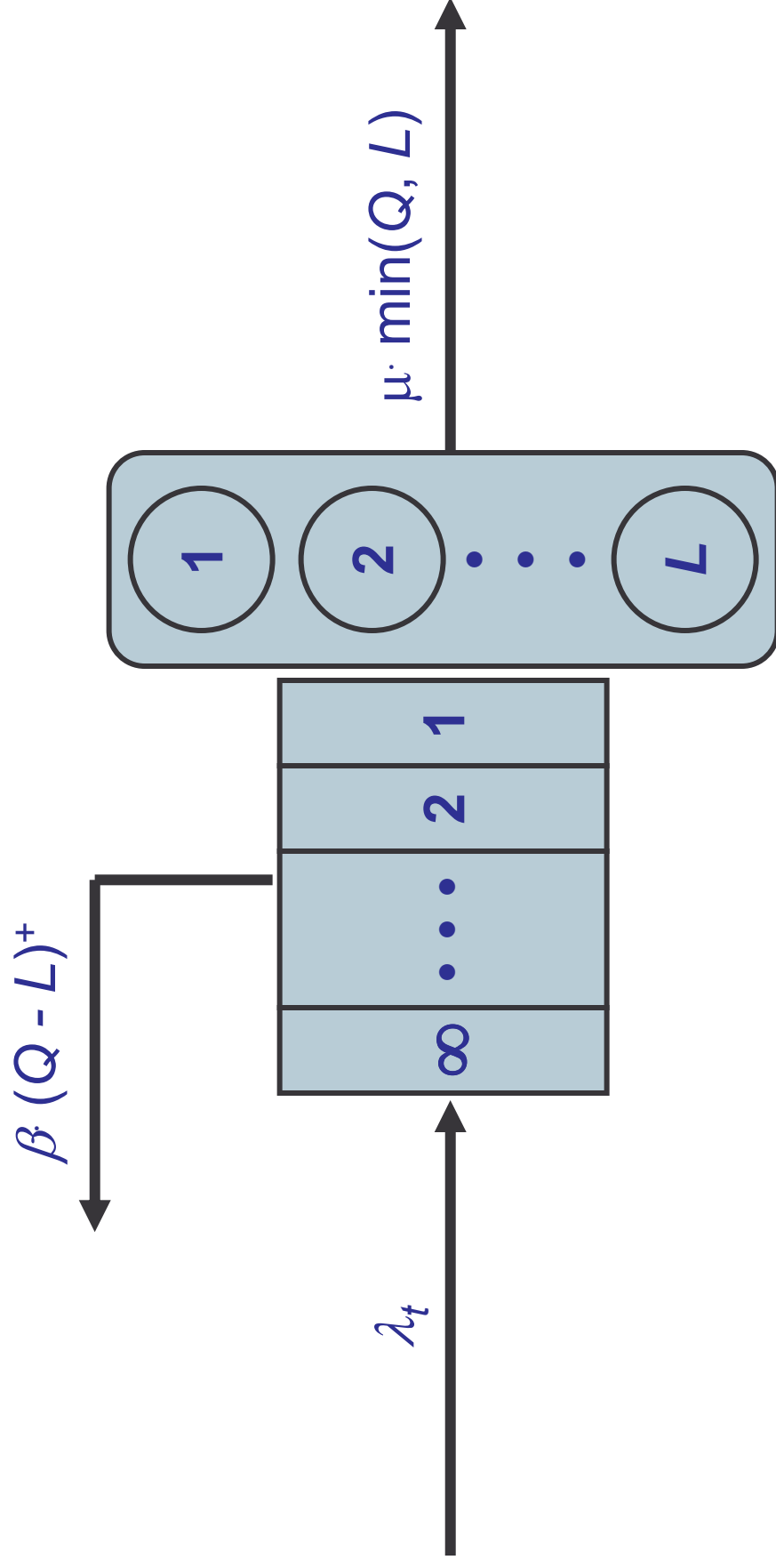


**$L = 100$** ,  $1/\mu_1 = 1/\mu_2 = 8.69$  minutes,  $1/\beta = 90$  minutes



**$L = 500$ ,  $1/\mu_1 = 1/\mu_2 = 8.69$  minutes,  $1/\beta = 90$  minutes**

# CALL CENTERS WITH MUSIC AND IMPATIENCE: THE M/M/L+M QUEUE



We have a Poisson arrival rate  $\lambda_t$ , mean service time  $1/\mu$ , mean time to abandonment  $1/\beta$ , and  $L$  call center agents.

## PRICE AND COST STRUCTURE ASSUMPTIONS

$r$  = revenue rate per customer service completion

$s$  = penalty rate per customer abandonment

and

$c(\cdot)$  = cost rate for the number of agents used.

We assume that this cost function is increasing and concave.

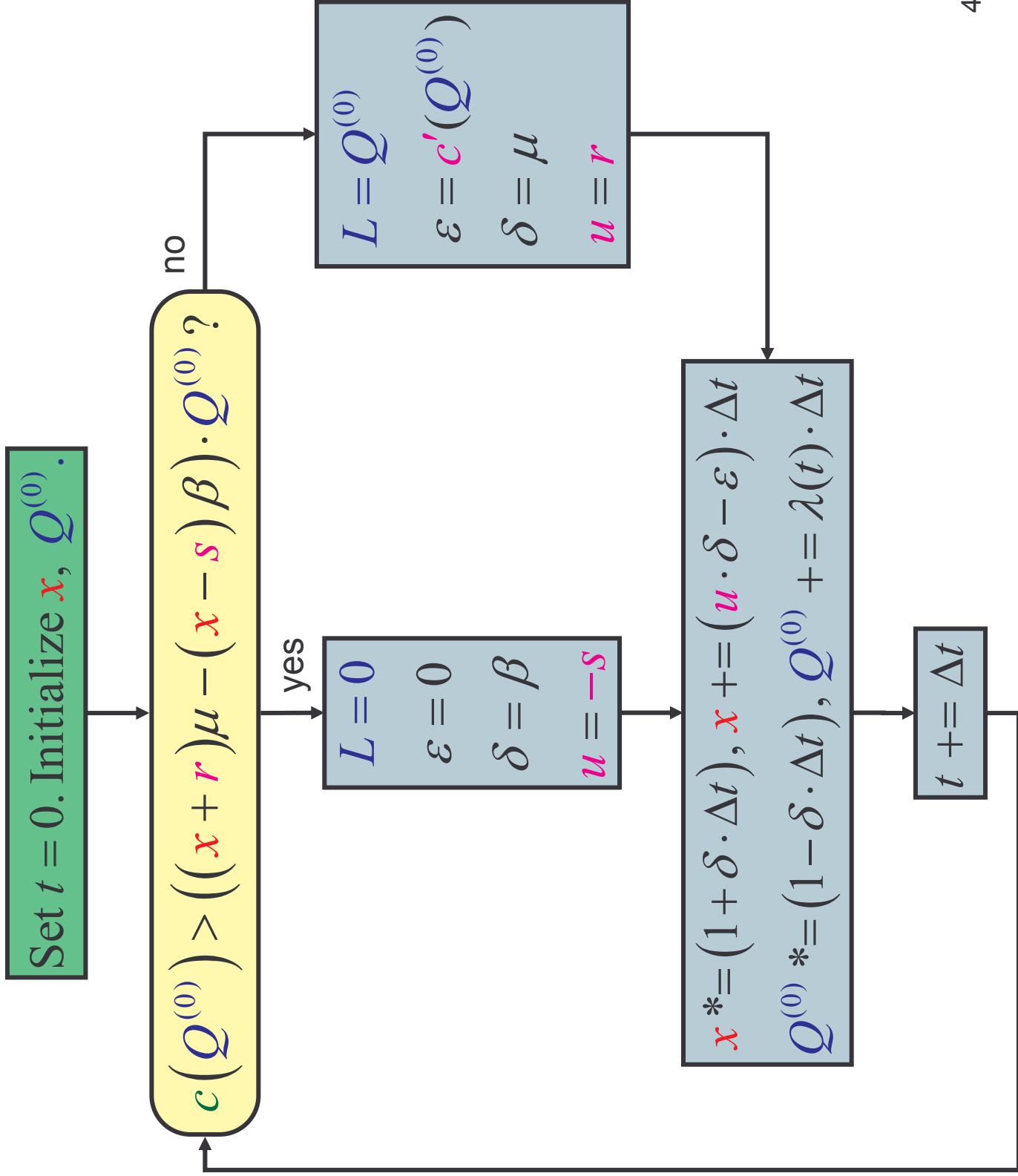
# FLUID APPROXIMATION OF CALL CENTER STAFFING FOR PROFIT OPTIMALITY

Find  $L_t$  so that we can maximize:

$$\int_0^T r\mu \cdot (Q^{(0)}(t) \wedge L_t) - s\beta \cdot (Q^{(0)}(t) - L_t)^+ - c(L_t) dt$$

with the constraint:

$$\frac{d}{dt} Q^{(0)}(t) = \lambda_t - \beta \cdot (Q^{(0)}(t) - L_t)^+ - \mu \cdot (Q^{(0)}(t) \wedge L_t).$$



# CALL CENTER MODELING FEATURES OF MARKOVIAN SERVICE NETWORKS

- Queues with Time-Varying Rates
- Multi-server Queues
- Customer Impatience
- Approximate Finite Buffers Through Customer Impatience
- Dynamic and Pre-emptive Priorities
- Network Routing Due to Service Completion
- Network Routing Due to Impatience

## USEFULNESS OF MARKOVIAN SERVICE NETWORKS

- Expands the family of tractable queueing models relevant to call centers.
- Captures transient effects such as sudden surges in demand.
- Approximates mean of call center metrics by fluid models.
- Approximates variance and distribution of call center metrics by diffusion models that are usually Gaussian.
- Fluid and diffusion models are both characterized by low dimensional dynamical systems (set of differential equations).
- The dimension of these equations is only a function of the number of network nodes and is independent of the number of call center agents.
- The approximations become more accurate as one scales up customer demand and the number of agents.

## RELATED PAPERS

- Fluid Approximation of a Priority Call Center With Time-Varying Arrivals, M. Fu; W. A. Massey; & A. Ridley. *The Telecommunications Review*; 2004, pp. 69-77.
- The Analysis of Queues with Time-Varying Rates for Telecommunication Models, W. A. Massey. *Telecommunications Systems*; 2002, pp. 173-204.
- Queue Lengths and Waiting Times for Multiserver Queues With Abandonment and Retrials, A. Mandelbaum; W. A. Massey; M.I. Reiman; B. Rider; & A. Stolyar. *Telecommunications Systems*; 2002, pp. 149-171.

Time Varying Multiserver Queues with Abandonment and Retrials, A. Mandelbaum; W. A. Massey; M. I. Reiman; & B. Rider; *Proceedings of ITC 16*; 1999.

Strong Approximations for Markovian Service Networks, A. Mandelbaum; W. A. Massey; & M. I. Reiman; *QUESTA* 30; 1998, pp. 149-201.

Strong Approximations for Time Dependent Queues, W. A. Massey & A. Mandelbaum; *MOR*; 20:1; February 1995, pp. 33-64.

Asymptotic Analysis of the Time Dependent  $M/M/1$  Queue, W. A. Massey; *MOR*; 10 (May 1985); pp. 305-327.

Nonstationary Queues, W. A. Massey; Stanford University; 1981; PhD Thesis.