USING SIMULATION TO STUDY SERVICE-RATE CONTROLS TO STABILIZE PERFORMANCE IN A SINGLE-SERVER QUEUE WITH TIME-VARYING ARRIVAL RATE

Ni Ma and Ward Whitt

Columbia University

December 5, 2015

Outline

Motivation

Motivation: Stabilizing Performance

- Given the time-varying arrival rates, we are interested in an algorithm that can **stabilize performance** of the queueing system, e.g. expected delay, delay probability, expected queue length.

Motivation: Stabilizing Performance

- Given the time-varying arrival rates, we are interested in an algorithm that can **stabilize performance** of the queueing system, e.g. expected delay, delay probability, expected queue length.
- Earlier papers that study **server-staffing** to stabilize performance in multi-server queues with time-varying arrivals.

Motivation: Stabilizing Performance

- Given the time-varying arrival rates, we are interested in an algorithm that can **stabilize performance** of the queueing system, e.g. expected delay, delay probability, expected queue length.
- Earlier papers that study **server-staffing** to stabilize performance in multi-server queues with time-varying arrivals.

M. Defraeye and I. Van Niewenhuyse (2013) Controlling excessive waiting times in small service systems with time-varying demand: an extension of the ISA algorithm. *Decision Support Systems* 54(4), 1558 – 1567.

Y. Liu and W. Whitt (2012) Stabilizing customer abandonment in many-server queues with time-varying arrivals. *Oper. Res.* 60(6), 1551 – 1564.

O.B. Jennings, A. Mandelbaum, W.A. Massey and W. Whitt (1996) Server staffing to meet time-varying demand. *Manag. Sci.* 42(10), 1383 –1394.

Motivation: Service-Rate Controls

Motivation: Service-Rate Controls

- **Problem**: systems with only a few servers or with inflexible staffing.

Motivation: Service-Rate Controls

- **Problem**: systems with only a few servers or with inflexible staffing.
- In many applications, it is possible to **change the processing rate**.

Motivation: Service-Rate Controls

- **Problem**: systems with only a few servers or with inflexible staffing.
- In many applications, it is possible to **change the processing rate**.

Example (use a service-rate control)

1. Hospital Surgery Rooms
2. Airport Security Inspection Lines

Our Contributions

Our Contributions

We use simulation to study service-rate controls to stabilize performance in a single-server queue with time-varying arrival rates.

Our Contributions

We use simulation to study service-rate controls to stabilize performance in a single-server queue with time-varying arrival rates.

- We conduct simulation experiments to **evaluate the performance** of alternative service-rate controls.

Our Contributions

We use simulation to study service-rate controls to stabilize performance in a single-server queue with time-varying arrival rates.

- We conduct simulation experiments to **evaluate the performance** of alternative service-rate controls.
- We **develop an efficient algorithm** for simulating a time-varying queue with a service-rate control.

The Model

The $G_t/G_t/1$ queue

$G_t/G_t/1$ Single-Server Queueing Model

The $G_t/G_t/1$ queue

$G_t/G_t/1$ Single-Server Queueing Model

- Single server

The $G_t/G_t/1$ queue

$G_t/G_t/1$ Single-Server Queueing Model

- Single server
- Time-varying arrival rate function

The $G_t/G_t/1$ queue

$G_t/G_t/1$ Single-Server Queueing Model

- Single server
- Time-varying arrival rate function
- First-Come First-Served service policy

The $G_t/G_t/1$ queue

$G_t/G_t/1$ Single-Server Queueing Model

- Single server
- Time-varying arrival rate function
- First-Come First-Served service policy
- Unlimited waiting space

The $G_t/G_t/1$ queue

$G_t/G_t/1$ Single-Server Queueing Model

- Single server
- Time-varying arrival rate function
- First-Come First-Served service policy
- Unlimited waiting space
- Service rate is subject to control

The $G_t/G_t/1$ queue

$G_t/G_t/1$ Single-Server Queueing Model

- Single server
- Time-varying arrival rate function
- First-Come First-Served service policy
- Unlimited waiting space
- Service rate is subject to control
- i.i.d. service requirements separate from the service rate

The Arrival Process

The Arrival Process

A **time-transformation** of a stationary counting process:

$$A(t) \equiv N_a(\Lambda(t)) \equiv N_a(\int_0^t \lambda(s)\, ds), \quad t \geq 0, \qquad (1)$$

where

- Λ is the **cumulative** arrival rate function:
  $\Lambda(t) = \int_0^t \lambda(s)\, ds, \quad t \geq 0.$

The Arrival Process

A **time-transformation** of a stationary counting process:

$$A(t) \equiv N_a(\Lambda(t)) \equiv N_a(\int_0^t \lambda(s)\, ds), \quad t \geq 0, \tag{1}$$

where

- Λ is the **cumulative** arrival rate function: $\Lambda(t) = \int_0^t \lambda(s)\, ds, \quad t \geq 0.$
- N_a is a rate-1 counting process with unit jumps.

The Arrival Process

A **time-transformation** of a stationary counting process:

$$A(t) \equiv N_a(\Lambda(t)) \equiv N_a(\int_0^t \lambda(s)\, ds), \quad t \geq 0, \tag{1}$$

where

- Λ is the **cumulative** arrival rate function: $\Lambda(t) = \int_0^t \lambda(s)\, ds, \quad t \geq 0.$
- N_a is a rate-1 counting process with unit jumps.
- check: $E[A(t)] = E[N_a(\Lambda(t))] = \Lambda(t) = \int_0^t \lambda(s) ds.$

The Arrival Process

A **time-transformation** of a stationary counting process:

$$A(t) \equiv N_a(\Lambda(t)) \equiv N_a(\int_0^t \lambda(s)\, ds), \quad t \geq 0, \tag{1}$$

where

- Λ is the **cumulative** arrival rate function:
  $\Lambda(t) = \int_0^t \lambda(s)\, ds, \quad t \geq 0.$
- N_a is a rate-1 counting process with unit jumps.
- check: $E[A(t)] = E[N_a(\Lambda(t))] = \Lambda(t) = \int_0^t \lambda(s) ds.$
- All the stochastic variability is separated from the deterministic arrival rate function.

The Service Process

The Service Process

Queue Length and Departure Process

$$Q(t) \equiv A(t) - D(t), \quad t \geq 0, \tag{2}$$

$$D(t) \equiv N_s\left(\int_0^t \mu(s) 1_{\{Q(s)>0\}}\, ds\right), \quad t \geq 0, \tag{3}$$

The Service Process

Queue Length and Departure Process

$$Q(t) \equiv A(t) - D(t), \quad t \geq 0, \tag{2}$$

$$D(t) \equiv N_s\left(\int_0^t \mu(s) 1_{\{Q(s)>0\}}\, ds\right), \quad t \geq 0, \tag{3}$$

where

The Service Process

Queue Length and Departure Process

$$Q(t) \equiv A(t) - D(t), \quad t \geq 0, \tag{2}$$

$$D(t) \equiv N_s(\int_0^t \mu(s) 1_{\{Q(s)>0\}}\, ds), \quad t \geq 0, \tag{3}$$

where

- N_s is a rate-1 counting process with unit jumps, independent of N_a.

The Service Process

Queue Length and Departure Process

$$Q(t) \equiv A(t) - D(t), \quad t \geq 0, \tag{2}$$

$$D(t) \equiv N_s\left(\int_0^t \mu(s) 1_{\{Q(s)>0\}}\, ds\right), \quad t \geq 0, \tag{3}$$

where

- N_s is a rate-1 counting process with unit jumps, independent of N_a.
- $E[D(t)|Q(s), 0 \leq s \leq t] = \int_0^t \mu(s) 1_{\{Q(s)>0\}}\, ds$.

The Service Process

Queue Length and Departure Process

$$Q(t) \equiv A(t) - D(t), \quad t \geq 0, \tag{2}$$

$$D(t) \equiv N_s\left(\int_0^t \mu(s) 1_{\{Q(s)>0\}}\, ds\right), \quad t \geq 0, \tag{3}$$

where

- N_s is a rate-1 counting process with unit jumps, independent of N_a.
- $E[D(t)|Q(s), 0 \leq s \leq t] = \int_0^t \mu(s) 1_{\{Q(s)>0\}}\, ds$.
- The service requirement process N_s is separated from the deterministic service-rate function $\mu(t)$.

The Service-Rate Controls

The Service-Rate Controls

- **Rate-matching control**

$$\mu(t) \equiv \frac{\lambda(t)}{\rho}, \quad t \geq 0. \tag{4}$$

The Service-Rate Controls

- **Rate-matching control**

$$\mu(t) \equiv \frac{\lambda(t)}{\rho}, \quad t \geq 0. \tag{4}$$

- **PSA-based square-root control**

$$\mu(t) \equiv \lambda(t) + \frac{\lambda(t)}{2}\left(\sqrt{1 + \frac{\zeta}{\lambda(t)}} - 1\right), \quad t \geq 0, \tag{5}$$

The Service-Rate Controls

- **Rate-matching control**

$$\mu(t) \equiv \frac{\lambda(t)}{\rho}, \quad t \geq 0. \tag{4}$$

- **PSA-based square-root control**

$$\mu(t) \equiv \lambda(t) + \frac{\lambda(t)}{2}\left(\sqrt{1 + \frac{\zeta}{\lambda(t)}} - 1\right), \quad t \geq 0, \tag{5}$$

  - based on the PSA approximation:
    $E[W(t)] \approx \rho(t)V/\mu(t)(1-\rho(t)) = \lambda(t)V/(\mu(t)^2 - \mu(t)\lambda(t))$.

- Supporting treory in Whitt (2015)

The Service-Rate Controls

- **Rate-matching control**

$$\mu(t) \equiv \frac{\lambda(t)}{\rho}, \quad t \geq 0. \tag{4}$$

- **PSA-based square-root control**

$$\mu(t) \equiv \lambda(t) + \frac{\lambda(t)}{2}\left(\sqrt{1 + \frac{\zeta}{\lambda(t)}} - 1\right), \quad t \geq 0, \tag{5}$$

  - based on the PSA approximation: $E[W(t)] \approx \rho(t)V/\mu(t)(1 - \rho(t)) = \lambda(t)V/(\mu(t)^2 - \mu(t)\lambda(t))$.

- Supporting treory in Whitt (2015)

Generating the Arrival Process

Generating the Arrival Process

Let A_k and T_k be arrival times of processes A and N_a

$$A_k = \Lambda^{-1}(T_k). \tag{6}$$

Generating the Arrival Process

Let A_k and T_k be arrival times of processes A and N_a

$$A_k = \Lambda^{-1}(T_k). \tag{6}$$

- **Problem**: need to compute Λ^{-1} for each arrival in each simulation run.

Generating the Arrival Process

Let A_k and T_k be arrival times of processes A and N_a

$$A_k = \Lambda^{-1}(T_k). \quad (6)$$

- **Problem**: need to compute Λ^{-1} for each arrival in each simulation run.
- Compute the inverse function Λ^{-1} for one cycle outside of simulation and do table lookup when simulating.
  - $\Lambda^{-1}(kC + t) = kC + \Lambda^{-1}(t)$ for $0 \le t \le C$, where C is the length of a cycle.
- (See the paper for the details.)

Generating the Service Times

Generating the Service Times

- Let A_k, B_k and D_k be customer's arrival time, begin service time and departure time; V_k and W_k be customer's service time and waiting time in queue.

Generating the Service Times

- Let A_k, B_k and D_k be customer's arrival time, begin service time and departure time; V_k and W_k be customer's service time and waiting time in queue.
- Let sequence of **service requirements** $\{S_k : k \geq 1\}$ be specified as the times between events in the counting process N_s.

Generating the Service Times

- Let A_k, B_k and D_k be customer's arrival time, begin service time and departure time; V_k and W_k be customer's service time and waiting time in queue.
- Let sequence of **service requirements** $\{S_k : k \geq 1\}$ be specified as the times between events in the counting process N_s.
- We have the basic recursions:
  $B_k = max\{D_{k-1}, A_k\}$, $D_k = B_k + V_k$ and $W_k = B_k - A_k$.

Generating the Service Times

- Let A_k, B_k and D_k be customer's arrival time, begin service time and departure time; V_k and W_k be customer's service time and waiting time in queue.
- Let sequence of **service requirements** $\{S_k : k \geq 1\}$ be specified as the times between events in the counting process N_s.
- We have the basic recursions:
  $B_k = max\{D_{k-1}, A_k\}$, $D_k = B_k + V_k$ and $W_k = B_k - A_k$.
- **But V_k is not formulated.**

Generating the Service Times

Generating the Service Times

Exact service time formula:

$$S_k = \int_{B_k}^{B_k+V_k} \mu(s)\, ds, \quad k \geq 1. \tag{7}$$

If we let

$$M(t) \equiv \int_0^t \mu(s)\, ds, \quad t \geq 0, \tag{8}$$

then

$$V_k = M^{-1}(S_k + M(B_k)) - B_k, \quad k \geq 1. \tag{9}$$

Generating the Service Times

Exact service time formula:

$$S_k = \int_{B_k}^{B_k+V_k} \mu(s)\,ds, \quad k \geq 1. \tag{7}$$

If we let

$$M(t) \equiv \int_0^t \mu(s)\,ds, \quad t \geq 0, \tag{8}$$

then

$$V_k = M^{-1}(S_k + M(B_k)) - B_k, \quad k \geq 1. \tag{9}$$

Simulation Experiments

Simulation Experiments: Arrival Rates

Simulation Experiments: Arrival Rates

The arrival process has the **sinusoidal arrival rate function**

$$\lambda(t) \equiv 1 + \beta \sin(\gamma t) \qquad (10)$$

Simulation Experiments: Arrival Rates

The arrival process has the **sinusoidal arrival rate function**

$$\lambda(t) \equiv 1 + \beta \sin(\gamma t) \tag{10}$$

with

- β=0.2, γ=0.001, 0.01, 0.1, 1, 10.

Simulation Experiments: Arrival Rates

The arrival process has the **sinusoidal arrival rate function**

$$\lambda(t) \equiv 1 + \beta \sin(\gamma t) \tag{10}$$

with

- β=0.2, γ=0.001, 0.01, 0.1, 1, 10.
- To cover a range of difference cycle lengths of $2\pi/\gamma$.

Simulation Experiments: Stochastic Components

Simulation Experiments: Stochastic Components

Use renewal processes with mean 1 for the base process N_a and N_s, and consider three **different i.i.d. interval time distributions**.

- exponential ($c^2 = 1$)

Simulation Experiments: Stochastic Components

Use renewal processes with mean 1 for the base process N_a and N_s, and consider three **different i.i.d. interval time distributions**.

- exponential ($c^2 = 1$)
- hyperexponential (mixture of two exponentials, $c^2 > 1$)

Simulation Experiments: Stochastic Components

Use renewal processes with mean 1 for the base process N_a and N_s, and consider three **different i.i.d. interval time distributions**.

- exponential ($c^2 = 1$)
- hyperexponential (mixture of two exponentials, $c^2 > 1$)
- Erlang (sum of two i.i.d. exponentials, $c^2 = 0.5$)

Simulation and Estimation Methods

Simulation and Estimation Methods

- Consider a **fixed time interval** [0, T] with T$= 2 \times 10^4$ for $\gamma = 0.001$ and T$= 2 \times 10^3$ for the other values of γ.

Simulation and Estimation Methods

- Consider a **fixed time interval** [0, T] with T$= 2 \times 10^4$ for $\gamma = 0.001$ and T$= 2 \times 10^3$ for the other values of γ.
- For each simulation replication, calculate performance measures **at deterministic times** dt, 2dt, 3dt,...T.

Simulation and Estimation Methods

- Consider a **fixed time interval** [0, T] with T$= 2 \times 10^4$ for $\gamma = 0.001$ and T$= 2 \times 10^3$ for the other values of γ.
- For each simulation replication, calculate performance measures **at deterministic times** dt, 2dt, 3dt,...T.
  - Compute virtual waiting time W(t) and number of customers in system Q(t)

Simulation and Estimation Methods

- Consider a **fixed time interval** [0, T] with $T = 2 \times 10^4$ for $\gamma = 0.001$ and $T = 2 \times 10^3$ for the other values of γ.
- For each simulation replication, calculate performance measures **at deterministic times** dt, 2dt, 3dt,...T.
  - Compute virtual waiting time W(t) and number of customers in system Q(t)
- Generate **10,000 independent replications** to estimate mean values and to construct confidence intervals of performance measures.

Simulation and Estimation Methods

- Consider a **fixed time interval** [0, T] with T$= 2 \times 10^4$ for $\gamma = 0.001$ and T$= 2 \times 10^3$ for the other values of γ.
- For each simulation replication, calculate performance measures **at deterministic times** dt, 2dt, 3dt,...T.
  - Compute virtual waiting time W(t) and number of customers in system Q(t)
- Generate **10,000 independent replications** to estimate mean values and to construct confidence intervals of performance measures.
  - Take the average over all replications to estimate E(W(t)) and E(Q(t))

Simulation and Estimation Methods

- Consider a **fixed time interval** [0, T] with T$= 2 \times 10^4$ for $\gamma = 0.001$ and T$= 2 \times 10^3$ for the other values of γ.
- For each simulation replication, calculate performance measures **at deterministic times** dt, 2dt, 3dt,...T.
  - Compute virtual waiting time W(t) and number of customers in system Q(t)
- Generate **10,000 independent replications** to estimate mean values and to construct confidence intervals of performance measures.
  - Take the average over all replications to estimate E(W(t)) and E(Q(t))
  - Use t statistics to construct 95% confidence intervals for E(W(t)) and E(Q(t))

Simulation and Estimation Methods

- Consider a **fixed time interval** [0, T] with T$= 2 \times 10^4$ for $\gamma = 0.001$ and T$= 2 \times 10^3$ for the other values of γ.
- For each simulation replication, calculate performance measures **at deterministic times** dt, 2dt, 3dt,...T.
  - Compute virtual waiting time W(t) and number of customers in system Q(t)
- Generate **10,000 independent replications** to estimate mean values and to construct confidence intervals of performance measures.
  - Take the average over all replications to estimate E(W(t)) and E(Q(t))
  - Use t statistics to construct 95% confidence intervals for E(W(t)) and E(Q(t))

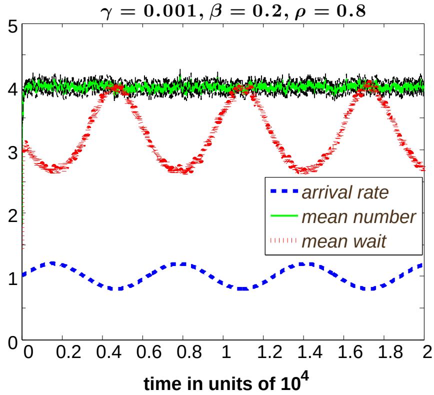
Simulation Results

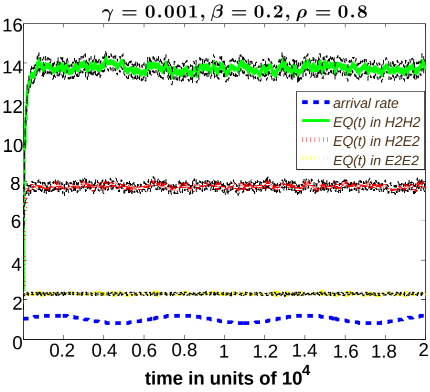
Simulation Results: The Rate-Matching Control

1. γ=0.001

Simulation Results: The Rate-Matching Control

1. γ=0.001
   - Cycle length is 6.28×10^3.
   - The left graph is **Markovian model**; the right graph shows (H_2/H_2), (H_2/E_2) **and** (E_2/E_2).
   - E(Q(t)) stabilized at target, but E(W(t)) is periodic.





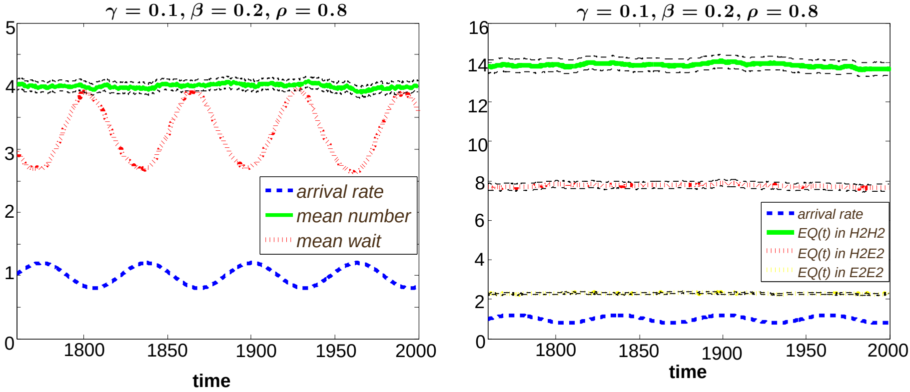
Simulation Results: The Rate-Matching Control

2. γ=0.1

Simulation Results: The Rate-Matching Control

2. γ=0.1

- Cycle length is 62.8, only last 3 to 4 cycles are displayed.
- E(Q(t)) stabilized at target, but E(W(t)) is periodic.



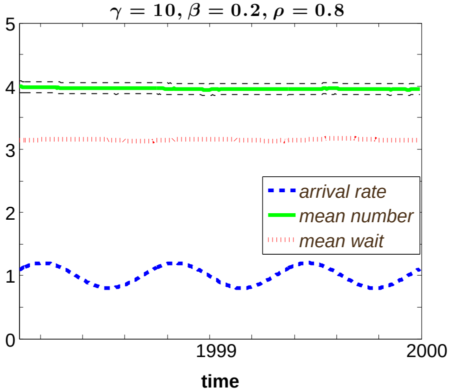
Simulation Results: The Rate-Matching Control

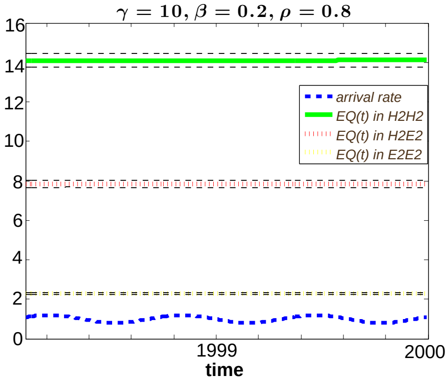
3. $\gamma = 10$

Simulation Results: The Rate-Matching Control

3. γ=10

- Cycle length is 0.63, only last 3 cycles are displayed.
- By Whitt (1984) the system converges to stationary case.



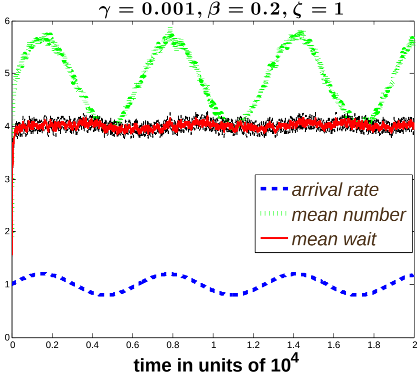


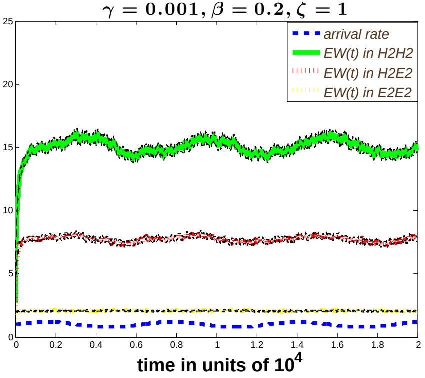
Simulation Results: The Square-Root Control

1. γ=0.001

Simulation Results: The Square-Root Control

1. γ=0.001
   - Cycles are long, and arrival rates change slowly, thus PSA is appropriate. [Whitt, 1991]
   - E(W(t)) is stabilized, while E(Q(t)) is periodic.



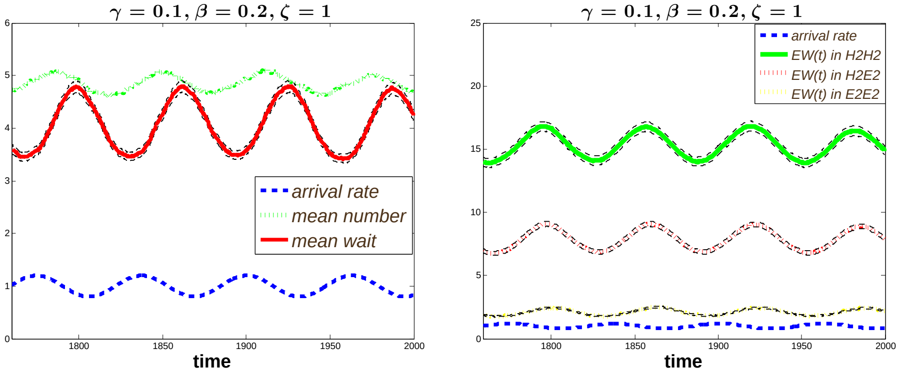


Simulation Results: The Square-Root Control

2. γ=0.1

Simulation Results: The Square-Root Control

2. γ=0.1
   - PSA does not hold as cycles are short.
   - Neither E(W(t)) nor E(Q(t)) is stabilized.



Thank You!

Generating the Arrival Process

Construct the table of Λ^{-1} over an cycle.

Generating the Arrival Process

Construct the table of Λ^{-1} over an cycle.

- Calculate Λ value for n_x equally spaced points of $[0, C]$, let spacing be $\eta = C/n_x$.

Generating the Arrival Process

Construct the table of Λ^{-1} over an cycle.

- Calculate Λ value for n_x equally spaced points of $[0, C]$, let spacing be $\eta = C/n_x$.
- Construct approximation J of Λ^{-1} over n_y equally spaced points in $[0, \Lambda(C_] = [0, C]$, let spacing be $\delta = C/n_y$.
  - Using $J(j\delta) = k\eta, \quad 1 \leq j \leq n_y$,
    where $0 \leq k \leq n_x$ and $k\eta$ is closest point greater equal to the true inverse value.

Generating the Arrival Process

Construct the table of Λ^{-1} over an cycle.

- Calculate Λ value for n_x equally spaced points of $[0, C]$, let spacing be $\eta = C/n_x$.
- Construct approximation J of Λ^{-1} over n_y equally spaced points in $[0, \Lambda(C] = [0, C]$, let spacing be $\delta = C/n_y$.
  - Using $J(j\delta) = k\eta, \quad 1 \leq j \leq n_y$, where $0 \leq k \leq n_x$ and $k\eta$ is closest point greater equal to the true inverse value.
- Extend J to $[0, C]$ by letting $J(t) = J(\lfloor t/\delta \rfloor \delta)$.

Results From [Whitt, 2015]: The Rate Matching Control

Results From [Whitt, 2015]: The Rate Matching Control

Theorem 2.1 (*time transformation of a stationary model*)

For (A, D, Q) with the rate-matching service-rate control and the stationary single-server model (A_1, D_1, Q_1),

$$(A(t), D(t), Q(t)) = (A_1(\Lambda(t)), D_1(\Lambda(t)), Q_1(\Lambda(t))), \quad t \geq 0. \tag{11}$$

Results From [Whitt, 2015]: The Rate Matching Control

Theorem 2.1 (*time transformation of a stationary model)*

For (A, D, Q) with the rate-matching service-rate control and the stationary single-server model (A_1, D_1, Q_1),

$$(A(t), D(t), Q(t)) = (A_1(\Lambda(t)), D_1(\Lambda(t)), Q_1(\Lambda(t))), \quad t \geq 0. \tag{11}$$

Theorem 3.2 (*stabilizing the queue-length distribution and the steady-state delay probability)*

Let $Q_1(t)$ be the queue length process when $\lambda(t) = 1$, $t \geq 0$. If $Q_1(t) \Rightarrow Q_1(\infty)$ as $t \to \infty$, where $P(Q_1(\infty) < \infty) = 1$, then also

$$Q(t) \Rightarrow Q_1(\infty) \quad \text{in} \quad \mathbb{R} \quad \text{as} \quad t \to \infty, \tag{12}$$

and

$$P(W(t) > 0) = P(Q(t) \geq 1) \to \rho \quad \text{as} \quad t \to \infty. \tag{13}$$

Results From [Whitt, 2015]: The Square-Root Control

Results From [Whitt, 2015]: The Square-Root Control

Section 6.3 (*stabilizing the expected time-varying virtual waiting time*)

We assume that the Pointwise Stationary Approximation (PSA) is appropriate. Then the square-root control (14) stabilizes $E[W(t)]$ at the target w for all t under heavy traffic.

$$\mu(t) \equiv \lambda(t) + \frac{\lambda(t)}{2}\left(\sqrt{1 + \frac{\zeta}{\lambda(t)}} - 1\right), \quad t \geq 0, \tag{14}$$

where ζ is inversely proportional to w.

Pointwaise Stationary Approximation (PSA)

Performance at different times can be regarded as approximately the same as the performance of the stationary system with the model parameters operating at those separate times.

Theorem From [Whitt, 2015]

Theorem From [Whitt, 2015]

Theorem 6.1 (*impossibility of stabilizing both the waiting time distribution and the mean number in queue*)

Consider a $G_t/G_t/1$ system starting empty in the distant past. Suppose that a service-rate control makes $P(W(t) > x)$ independent of t for all $x \geq 0$. Then the only arrival rate functions for which the mean number waiting in queue $E[(Q(t) - 1)^+]$ is constant, independent of t, are the constant arrival rate functions.

Theorem From [Whitt, 2015]

Theorem 6.1 (*impossibility of stabilizing both the waiting time distribution and the mean number in queue*)

Consider a $G_t/G_t/1$ system starting empty in the distant past. Suppose that a service-rate control makes $P(W(t) > x)$ independent of t for all $x \geq 0$. Then the only arrival rate functions for which the mean number waiting in queue $E[(Q(t) - 1)^+]$ is constant, independent of t, are the constant arrival rate functions.

Theorem From [Whitt, 1984]

Theorem From [Whitt, 1984]

Theorem 1 (*Convergence of point processes*)

The point process $D(t)$ has predictable stochastic intensity $\Lambda(t)$, then it can be represented as the random-time transformation

$$D(t) = \Pi(C(t)), \quad t \geq 0, \tag{15}$$

where $\Pi(t)$ is a Poisson process with unit intensity and $C(t) = \int_0^t \Lambda(u)du$. If $C_n(t) \Rightarrow ct$ in $\mathbb{R}$ as $n \to \infty$ for each t, then $D_n \Rightarrow \Pi_c$ in $D[0, \infty)$ as $n \to \infty$, where Π_c is a Poisson process with intensity c.

Theorem From [Whitt, 1984]

Theorem 1 (*Convergence of point processes*)

The point process $D(t)$ has predictable stochastic intensity $\Lambda(t)$, then it can be represented as the random-time transformation

$$D(t) = \Pi(C(t)), \quad t \geq 0, \tag{15}$$

where $\Pi(t)$ is a Poisson process with unit intensity and $C(t) = \int_0^t \Lambda(u)du$. If $C_n(t) \Rightarrow ct$ in $\mathbb{R}$ as $n \to \infty$ for each t, then $D_n \Rightarrow \Pi_c$ in $D[0, \infty)$ as $n \to \infty$, where Π_c is a Poisson process with intensity c.

Section 3.2 (*Applications to queues*)

Apply rescaling

$$D_n(t) = \hat{D}_n(t)(t/n) \quad \textit{and} \quad C_n(t) = \hat{C}_n(t)(t/n) \tag{16}$$

for $t \geq 0$.

Theorem From [Whitt, 1984]

Theorem 1 (*Convergence of point processes*)

The point process $D(t)$ has predictable stochastic intensity $\Lambda(t)$, then it can be represented as the random-time transformation

$$D(t) = \Pi(C(t)), \quad t \geq 0, \tag{15}$$

where $\Pi(t)$ is a Poisson process with unit intensity and $C(t) = \int_0^t \Lambda(u)du$. If $C_n(t) \Rightarrow ct$ in $\mathbb{R}$ as $n \to \infty$ for each t, then $D_n \Rightarrow \Pi_c$ in $D[0, \infty)$ as $n \to \infty$, where Π_c is a Poisson process with intensity c.

Section 3.2 (*Applications to queues*)

Apply rescaling

$$D_n(t) = \hat{D}_n(t)(t/n) \quad \textit{and} \quad C_n(t) = \hat{C}_n(t)(t/n) \tag{16}$$

for $t \geq 0$.

References

- M. Defraeye and I. Van Niewenhuyse (2013) Controlling excessive waiting times in small service systems with time-varying demand: an extension of the ISA algorithm. *Decision Support Systems* 54(4), 1558 – 1567.
- B. He, Y. Liu and W. Whitt (2015) Staffing a service system with non-Poisson nonstationary arrivals. *Working Paper.*
- O.B. Jennings, A. Mandelbaum, W.A. Massey and W. Whitt (1996) Server staffing to meet time-varying demand. *Manag. Sci.* 42, 1383 –1394.
- Y. Liu and W. Whitt (2012) Stabilizing customer abandonment in many-server queues with time-varying arrivals. *Oper. Res.* 60(6), 1551 – 1564.
- W. Whitt (1984) Departures from a queue with many busy servers. *Math. Oper. Res.* 9(1984), 534–544.
- W. Whitt (1991) The pointwise stationary approximation for $M_t/M_t/s$ queues is asymptotically correct as the rates increase. *Management Science* 37(3), 307–314.
- W. Whitt (2015) Stabilizing performance in a single-server queue with time-varying arrival rate. *Queueing Systems* 81, 341–378.
- G. Yom-Tov and A. Mandelbaum (2014) Erlang-R: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing and Service Oper. Management* 16(2), 283 – 299.