Information Effects on Performance of Two-tier Service Systems with Strategic Customers

Zhe George Zhang

Department of Decision Sciences, Western Washington University, Bellingham, WA98225, USA

&

Beedie School of Business, Simon Fraser University Vancouver, BC, Canada

****This is a joint work with Dr. Hsing Luh of National Chengchi University***

Outline

- Introduction – system description
- Real-time information scenario
- No real-time information scenario
- Comparison between two scenarios
- Conclusion

I. INTRODUCTION

- A Two-tier Service System: Two service providers (SPs) – free but long wait time v.s. toll but short wait time.
  - Public hospital v.s. private hospital
  - Regular service v.s. expediting service with a fee
- Strategic Customers:
  - Delay/cost sensitive
  - Heterogeneous
  - Selecting an SP to obtain the service based on "congestion information"
- Information Scenarios
  - Real-time queue length at an arrival instant.
  - Long term statistics (or no real-time information)

A Two Queue Model for Two-tier systems

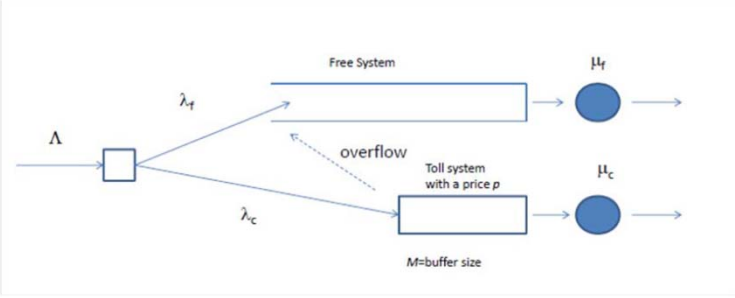


Figure 1. A Two-tier Service System with Strategic Customers.

Research Issues

- How does the real-time queue length information affect the system performance?
- Is there any significant difference in system performance behavior between with real-time information and without real-time information scenarios? (long-asked question by both practitioners and researchers in academia)

Related Literature

- Noar (1969) - a single server queueing system with identical customers influenced by either real-time information or the long-term statistics.
- Schroeter (1982) - heterogeneous customers with uniformly distributed unit-time waiting costs.
- Hassin and Haviv (2003) - a book "To Q or Not to Q" – a survey on more research in this area
- Armony and Maglaras (2004a, b) - service systems where an arriving customer knowing some delay information chooses to balk, wait, or leave a message, in which case the SP calls back within a guaranteed time.
- Guo and Zipkin (2007) - a single server queueing system with heterogeneous customers who either join or balk and three levels of information.
- Guo and Zhang (2012a, b) investigate two-tier service systems motivated by healthcare and border-crossing systems where customers can choose to join a free system or a toll system. Either long-terms statistics with two queues or real-time information with one queue.

Our Approach

- Consider more realistic factors
- Use exact analysis supplemented by simulation studies

II Real-time Information Scenario

- Customers arrive at the system according to Poisson process with arrival rate Λ.
- At a customer's arrival instant, he is provided with the real-time queue length information for both queues, represented by $X_f(t)$, and $X_c(t)$.
- service times are exponentially distributed with rates μ_f and μ_c for the free and toll SPs, respectively.
- The toll lane queue has a finite buffer of size M (for maximum wait time guarantee).
- The customer waiting cost parameter, denoted by θ, is distributed with a cumulative distribution function of F_θ.
- Indifferent SP condition (waiting cost):

$$\theta X_f(t)(1/\mu_f) = p + \theta \min\{X_c(t), M\}(1/\mu_c),$$

Customer's Decision

- An arriving customer makes the following decisions:

if $\theta X_f(t)(1/\mu_f) \leq p + \theta \min\{X_c(t), M\}(1/\mu_c)$,

=> he will join the free queue;

if $\theta X_f(t)(1/\mu_f) > p + \theta \min\{X_c(t), M\}(1/\mu_c)$ and $X_c(t) < M$,

⇒he will join the toll queue;

otherwise, he will be forced to join the free queue when the toll system buffer is full – overflow of the toll lane => no balking assumption.

Also assume that the θ follow a uniform distribution over [0,U]

Recall - Two-tier Service Model

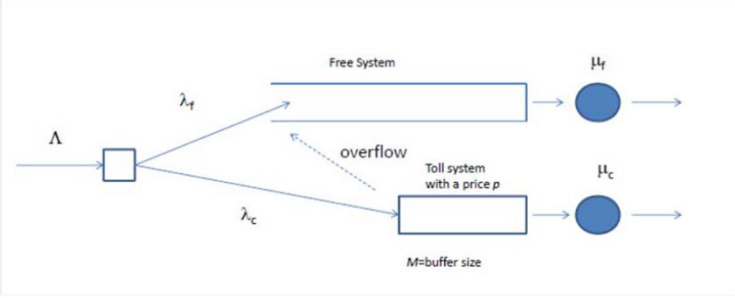


Figure 1. A Two-tier Service System with Strategic Customers.

Quasi-Birth-Death (QBD) process – *A special structure of Q matrix can be obtained based on the customer choice behavior. – the details of this part may be skipped in a time limited presentation.*

The system state is defined as $(X_f(t), X_c(t))$ on the state space

$$\Omega = \{(n, m) : n = 0, 1, ...; m = 0, 1, ..., M\}$$

- Under the steady-state condition, the system reaches steady-state

$$\lim_{t\to\infty} P\{X_f(t) = n, X_c(t) = m) = p_{nm}.$$

- The equilibrium arrival rates to the two systems:

$$\lambda_f(n, m) = \Lambda F_\theta(\frac{p}{n/\mu_f - m/\mu_c}),$$
$$\lambda_c(n, m) = \Lambda\left(1 - F_\theta(\frac{p}{n/\mu_f - m/\mu_c})\right),$$

for states with $m < M$.

Customer self-interest choice behavior => Structure of the QBD generator

- There exists a threshold n_0 for the free queue length

such that $\lambda_f(n, m) < \varepsilon$ whenever $n > n_0$ and $m < M$, where ε is a small positive value. The lower bound n_0 can be determined by

$$n_0 > \frac{p\mu_f}{F_\theta^{-1}(\varepsilon/\Lambda)} + \frac{\mu_f}{\mu_c}M.$$

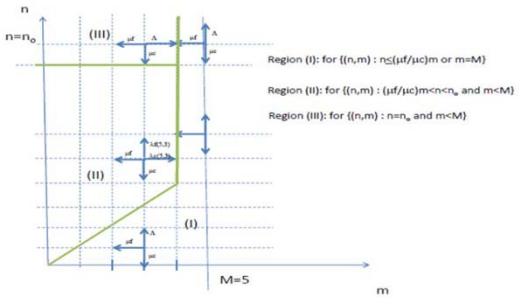
$$n_0 = \text{int}\left(\frac{p}{U}\frac{\Lambda\mu_f}{\varepsilon} + (M-1)\frac{\mu_f}{\mu_c}\right).$$

- Using this property, we can develop a level independent QBD process.
- M determines the number of phases
- n_0 determine the number of boundary states of the QBD process.

State-Dependent Arrival Rates

$$\lambda_f(n,m) = \begin{cases} \Lambda & \text{if } n \le m(\mu_f/\mu_c) \quad \text{or } m = M, \\ \frac{\Lambda}{U}\left(\frac{p\mu_f\mu_c}{n\mu_c - m\mu_f}\right) & \text{if } m(\mu_f/\mu_c) < n < n_0 \text{ and } m < M, \\ 0 & \text{if } n = n_0 \text{ and } m < M. \end{cases}$$

$$\lambda_c(n,m) = \Lambda - \lambda_f(n,m).$$



State Transition Diagram for M=5 case

QBD process (*the next few slides will be skipped in a time limited presentation – the paper is available upon request*)

- Infinitesimal generator of the QBD process

$$Q = \begin{bmatrix} B_{00} & C_{01} & & & & & & & & \\ A_{10} & B_{11} & C_{12} & & & & & & & \\ & A_{21} & B_{22} & C_{23} & & & & & & \\ & & \ddots & \ddots & \ddots & & & & & \\ & & & A_{M,M-1} & B_{M,M} & C_{M,M+1} & & & & \\ & & & & \ddots & \ddots & \ddots & & & \\ & & & & & A_{n_0-1,n_0-2} & B_{n_0-1,n_0-1} & C_{n_0-1,n_0} & & \\ & & & & & & A & B & C & \\ & & & & & & & A & B & C \\ & & & & & & & & \ddots & \ddots & \ddots \end{bmatrix},$$

- where

where all elements are $(M+1)\times(M+1)$ matrices. For $0 \le n \le M$,

$$B_{00} = \begin{bmatrix} -\Lambda & & & \\ \mu_c & -(\Lambda+\mu_c) & & \\ & \ddots & \ddots & \\ & & \mu_c & -(\Lambda+\mu_c) \end{bmatrix},$$

$$C_{01} = \begin{bmatrix} \Lambda & & \\ & \ddots & \\ & & \Lambda \end{bmatrix} = \Lambda I$$

$$A_{10} = \begin{bmatrix} \mu_f & & \\ & \ddots & \\ & & \mu_f \end{bmatrix} = \mu_f I,$$

$$B_{11} = \begin{bmatrix} -(\mu_f+\Lambda) & \lambda_c(1,0) & & \\ \mu_c & -(\Lambda+\mu_c+\mu_f) & & \\ & \ddots & \ddots & \\ & & \mu_c & -(\Lambda+\mu_c+\mu_f) \end{bmatrix},$$

$$C_{12} = \begin{bmatrix} \lambda_f(1,0) & & & \\ & \Lambda & & \\ & & \ddots & \\ & & & \Lambda \end{bmatrix}$$

$$A_{21} = \begin{bmatrix} \mu_f & & \\ & \ddots & \\ & & \mu_f \end{bmatrix} = \mu_f I,$$

$$B_{22} = \begin{bmatrix} -(\mu_f+\Lambda) & \lambda_c(2,0) & & & & \\ \mu_c & -(\Lambda+\mu_c+\mu_f) & \lambda_c(2,1) & & & \\ & \mu_c & -(\Lambda+\mu_c+\mu_f) & & & \\ & & \ddots & \ddots & & \\ & & & \mu_c & -(\Lambda+\mu_c+\mu_f) \end{bmatrix},$$

$$C_{23} = \begin{bmatrix} \lambda_f(2,0) & & & & \\ & \lambda_f(2,1) & & & \\ & & \Lambda & & \\ & & & \ddots & \\ & & & & \Lambda \end{bmatrix}$$

$$A_{M,M-1} = \begin{bmatrix} \mu_f & & \\ & \ddots & \\ & & \mu_f \end{bmatrix} = \mu_f I,$$

$$B_{M,M} = \begin{bmatrix} -(\mu_f+\Lambda) & \lambda_c(5,0) & & & & \\ \mu_c & -(\Lambda+\mu_c+\mu_f) & \lambda_c(5,1) & & & \\ & \ddots & \ddots & \ddots & & \\ & & \mu_c & -(\Lambda+\mu_c+\mu_f) & \lambda_c(M,M-1) & \\ & & & \mu_c & -(\Lambda+\mu_c+\mu_f) \end{bmatrix},$$

$$C_{M,M+1} = \begin{bmatrix} \lambda_f(5,0) & & & \\ & \ddots & & \\ & & \lambda_f(M,M-1) & \\ & & & \Lambda \end{bmatrix}.$$

For $M+1 \leq n \leq n_0 - 1$,

$$A_{n,n-1} = \begin{bmatrix} \mu_f & & \\ & \ddots & \\ & & \mu_f \end{bmatrix} = \mu_f I,$$

$$B_{n,n} = \begin{bmatrix} -(\mu_f+\Lambda) & \lambda_c(n,0) & & & & \\ \mu_c & -(\Lambda+\mu_c+\mu_f) & \lambda_c(n,1) & & & \\ & \ddots & \ddots & \ddots & & \\ & & \mu_c & -(\Lambda+\mu_c+\mu_f) & \lambda_c(n,M-1) & \\ & & & \mu_c & -(\Lambda+\mu_c+\mu_f) & \end{bmatrix},$$

$$C_{n,n+1} = \begin{bmatrix} \lambda_f(n,0) & & & & \\ & \lambda_f(n,1) & & & \\ & & \ddots & & \\ & & & \lambda_f(n,M-1) & \\ & & & & \Lambda \end{bmatrix}.$$

For $n \geq n_0$,

$$A = \begin{bmatrix} \mu_f & & \\ & \ddots & \\ & & \mu_f \end{bmatrix} = \mu_f I = A_{i,i-1}, \text{ for } i \geq 1.$$

$$B = \begin{bmatrix} -(\mu_f+\Lambda) & \Lambda & & \\ \mu_c & -(\Lambda+\mu_c+\mu_f) & \Lambda & \\ & \ddots & \ddots & \ddots \\ & & \mu_c & -(\Lambda+\mu_c+\mu_f) \end{bmatrix},$$

$$C = \begin{bmatrix} \ddots & \\ & \Lambda \end{bmatrix}.$$

Stability Condition

Proposition 1 *With real-time delay information, the two-tier service system reaches the steady state if*

$$\mu_f > \frac{\left(1-\frac{\Lambda}{\mu_c}\right)\left(\frac{\Lambda}{\mu_c}\right)^M \Lambda}{1-\left(\frac{\Lambda}{\mu_c}\right)^{M+1}}. \quad (3)$$

Corollary 2 *If $\Lambda/\mu_c \leq 1$, as $M \longrightarrow \infty$, (3) becomes $\mu_f > 0$ or there is no requirement for a positive μ_f . If $\Lambda/\mu_c > 1$, as $M \longrightarrow \infty$, (3) becomes $\mu_f + \mu_c > \Lambda$.*

Under the stability condition, the stationary probability vector is defined as

$$\mathbf{p}_n = [p_{n0}, p_{n1}, ..., p_{nM}],$$

where $p_{nm} = \lim_{t\to\infty} P\{X_f(t) = n, X_c(t) = m\}$.We know that when $n \geq n_0$, the matrix geometric solution is given by

$$\mathbf{p}_{n+1} = \mathbf{p}_n \mathbf{R}.$$

Like any regular QBD process, the rate matrix R should satisfy $R^2A + RB + C = 0$

A major contribution for this work

- We have developed a new computational algorithm to calculate the stationary distribution of the QBD process, called K-Matrix approach.
- Details are presented in the paper.

Performance measures

The boundary state vector $\mathbf{p_0}$ has n_0 components (or $M+1$ dimensional vectors) and is the unique solution of the equation system of $\mathbf{p_0}(B_0 + RA) = 0$ and $\mathbf{p_0}(I - R)^{-1}1 = 0$.

- Marginal Probabilities:

$$p_{\bullet j} = \sum_{n=0}^{\infty} p_{nj} \qquad p_{n\bullet} = \sum_{j=0}^{M} p_{nj}$$

- Expected Queue Lengths:

$$E(L^c) = \sum_{j=0}^{M} j p_{\bullet j} \text{ and } E(L^f) = \sum_{n=0}^{\infty} n p_{n\bullet}.$$

III No Real-time Information Scenario

- Customers make the lane selection based on the long-term statistics (e.g. expected waiting time).
- Planned and effective arrival rates for a stable system:

Let $\pi_i, (0 \le i \le M)$ be the probability that there are i customers in the toll system.

- We have $\lambda_f^{\text{eff}} = \lambda_f + \pi_M \lambda_c$ and $\lambda_c^{\text{eff}} = (1 - \pi_M)\lambda_c$.
- Stability condition:

Proposition 3 *With no real-time delay information, the two-tier service system reaches steady state if* $\mu_f > \lambda_f + \pi_M \lambda_c$.

Equilibrium Analysis

Let $W(\lambda_f^{\text{eff}}, \mu_f)$ and $W(\lambda_c, \mu_c)$ denote the expected waiting times of the steady-state free and toll systems.

- Indifferent SP condition: A customer with θ satisfying:

$$p + \overline{\theta} W(\lambda_c, \mu_c) = \overline{\theta} W(\lambda_f^{\text{eff}}, \mu_f), \quad (4)$$

$$\lambda_f + \lambda_c = \Lambda.$$

is indifferent between the two systems.

Proposition 4 *For a stable two-tier service system without real-time queue length information, there exist a unique $\overline{\theta}$ satisfying (4) and the equilibrium planned arrival rates λ_f and λ_c.*

Toll system

- Toll system is treated as an *M/M/1/*M queue with the planned equilibrium arrival rate:

$$W(\lambda_c, \mu_c) = \begin{cases} \frac{1}{\lambda_c^{\text{eff}}} \left\{ \frac{1}{1-\rho_c} - \frac{1+M\rho_c^{M+1}}{1-\rho_c^{M+1}} \right\} & \text{if } \rho_c \neq 1, \\ \frac{M}{2} & \rho_c = 1, \end{cases}$$

$$\pi_i = \begin{cases} \frac{(1-\rho_c)\rho_c^i}{1-\rho_c^{M+1}} & \text{if } \rho_c \neq 1, \\ \frac{1}{M+1} & \rho_c = 1, \end{cases} \quad i = 0, 1, \ldots, M,$$

where $\rho_c = \lambda_c/\mu_c$ and $\lambda_c^{\text{eff}} = \lambda_c(1-\pi_M)$.

Free System

- Treated as an MMPP/M/1 queue to compute $W(\lambda_f^{\text{eff}}, \mu_f)$
- The Markov Process modulating the arrival process with

$$Q' = \begin{bmatrix} -\lambda_c & \lambda_c & & & & \\ \mu_c & -(\mu_c+\lambda_c) & \lambda_c & & & \\ & \mu_c & -(\mu_c+\lambda_c) & \lambda_c & & \\ & & \ddots & \ddots & \ddots & \\ & & & & \mu_c & -\mu_c \end{bmatrix}_{(M+1)\times(M+1)}$$

- The arrival rate matrix

$$L = \begin{bmatrix} \lambda_f & 0 & \cdots & & \\ 0 & \lambda_f & \cdots & & \\ & \ddots & \ddots & \ddots & \\ & & & \lambda_f & \\ & & & & \Lambda \end{bmatrix}_{(M+1)\times(M+1)}.$$

Introducing matrix $B = (-Q') + L$, we can utilize the matrix analytic method (MAM) approach to evaluate the performance measures of the MMPP/M/1 queue (see the Appendix) for the free system.

However..

- The challenge is that we do not have any explicit expressions for the equilibrium arrival rates.
- A search algorithm has to be used to numerically compute these arrival rates and the resulting expected waiting times.

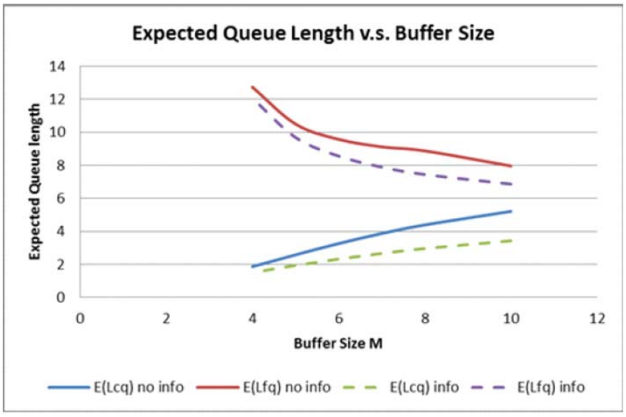
Computing the Nash equilibrium performance measures (Detailed explanations are in the paper)

A Search Algorithm of Computing $W(\lambda_c, \mu_c)$ and $W(\lambda_f^{*\theta}, \mu_f)$:

- Step 1: Initialization - Select an initial small (or large) θ_0 for the right (or left) search, compute $\lambda_f = \Lambda F_\theta(\theta_0)$, $\lambda_c = \Lambda(1-F_\theta(\theta_0))$, π_M from (6), $W(\lambda_c, \mu_c)$ from (5), and $W(\lambda_f^{*\theta}, \mu_f)$ using the MMPP/M/1 queue algorithm such that $\lambda_f + \lambda_c \pi_M < \mu_f$ and $f(\theta_0) > 0$. Use a positive $\Delta\theta$ to search to the right (or a negative $\Delta\theta$ to search to the left). Set $n = 0$.
- Step 2: Let $\theta_{n+1} = \theta_n + \Delta\theta$. Use $\lambda_f = \Lambda F_\theta(\theta_{n+1})$ and $\lambda_c = \Lambda(1 - F_\theta(\theta_{n+1}))$ to compute $W(\lambda_c, \mu_c)$ from (6) and check the stability condition for the free system, $\lambda_f + \lambda_c \pi_M < \mu_f$. If the free system is not stable, increase M until the stability condition is satisfied. Then compute $W(\lambda_f^{*\theta}, \mu_f)$ based on the MMPP/M/1 algorithm. Calculate $f(\theta_{n+1})$ and compared with $f(\theta_n)$.
- Step 3: If the sign of f changes between θ_n and θ_{n+1}, the solution to (9), $\overline{\theta}$, exists on the interval (θ_n, θ_{n+1}), go to next step. If the sign of f does not change between θ_n and θ_{n+1}, let $n = n + 1$, go to Step 2.
- Step 4: If $\Delta\theta > \epsilon$, let $\theta_0 = \theta_n$, $\Delta\theta = \Delta\theta/10$, $n = 0$ go to Step 2. Otherwise, $\overline{\theta} = (\theta_n + \theta_{n+1})/2$. Then use $\lambda_f = \Lambda F_\theta(\overline{\theta})$, $\lambda_c = \Lambda(1 - F_\theta(\overline{\theta}))$ to compute the equilibrium $W(\lambda_c, \mu_c)$ using (5) and $W(\lambda_f^{*\theta}, \mu_f)$ using the MMPP/M/1 queue algorithm (see Appendix) and Little's Law. □

IV. Comparison of Two Scenarios

- How much benefit does real-time queue length information bring to the system? - For a given p=10, vary M.



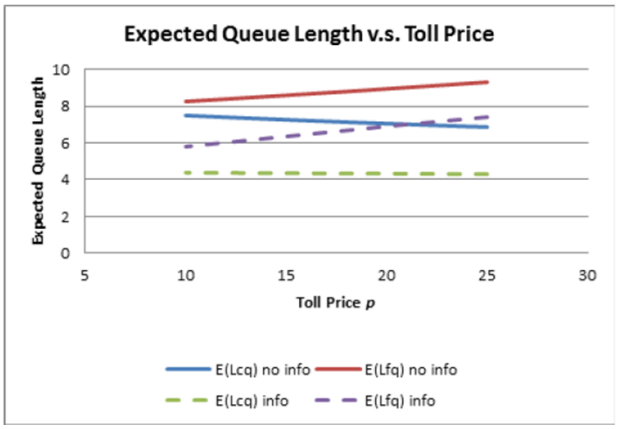
A Two-tier Service System with $\mu_f = 0.5, \mu_c = 0.6, \Lambda = 1, p = 10, U = 10$.

Observations

- At M=10, if real-time info is provided, for the toll system, the expected queue length can be reduced by almost 45%; for the free system, the expected length can be reduced by almost 19%.
- The performance behaviors for the two systems remain similar in the two information scenarios:
  - When M is getting smaller, the difference between the toll and free systems is getting bigger (common sense).
  - When M is getting smaller, the information effect is getting smaller.

Comparison between two scenarios

- For a given M=20, vary p:



A Two-tier Service System with $\mu_f = 0.5, \mu_c = 0.6, \Lambda = 1, M = 20, U = 10.$

Observations

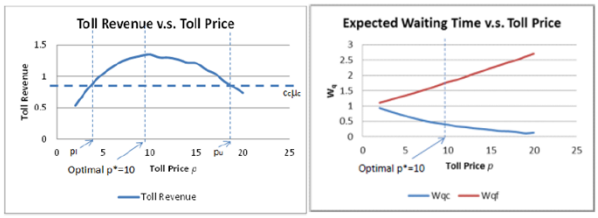
- At M=20, if real-time info is provided, for the toll system, the expected queue length can be reduced by average of 40%; for the free system, the expected length can be reduced by average of 24% over the entire price range tested.
- The performance behaviors remain similar for the two information scenarios.
- **Implication**: *Using the "no real-time information" model to study the "real-time information" model provides some bounds on "first-moment" based performance measures. . . . but not tight bounds!*

The Effects of Toll System's Pricing and Service Guarantee

- Necessary condition for the toll system to exist:

$$\lambda_c p(1-\pi_M) \geq c_c \mu_c$$

- Inconsistency between the toll system's goal (max profit/revenue) and customer service goal (min the overall waiting time/reduce the performance difference)



(a) (b)

Figure 5. A Two-tier System with No Real-time Information,

$\Lambda = 1, \mu_f = 1.25, \mu_c = 0.67, M = 5, \theta \sim U(0,10)$

Ethical Issue

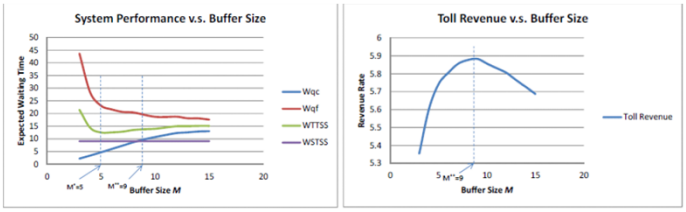
- For the heavy traffic load case (high demand), the toll system may try to charge very high price and serve only very rich customers.
- This problem occurred in some practical two-tier service systems (VIP queue in some Walk-in clinics in Taiwan).

Possible solutions:

- Toll price should be regulated.
- Government may subsidize the customers seeking toll system service -> reduce the burden of the public free service system.

The Effect of Buffer Size M

- Provide a service guarantee for the maximum expected delay time in the toll system.
- Again we see the difference between minimizing the overall expected waiting time and maximizing the toll system's revenue: implying some kind of "regulation on service guarantee" may be needed if social welfare goal must be met.

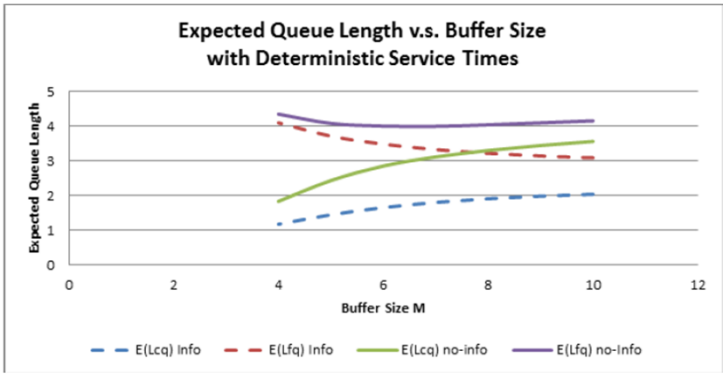


(a) (b)

Figure 7. A Two-tier Service System with No Real-time Information and

Robustness of Results

- Non exponential service time case – deterministic service time system



A Two-tier Service System with the deterministic service times and

$\mu_f = 0.5, \mu_c = 0.6, \Lambda = 1, p = 10, U = 10.$

Service times with a range of COVs

Dist.	Erlang	k=infty	Erlang	k=9	Erlang	k=8	Erlang	k=7	Erlang	k=6
COV	0		0.333333		0.353553		0.377964		0.408248	
M	E(Lc) Imp	E(Lf) Imp	E(Lc) Imp	E(Lf) Imp	E(Lc) Imp	E(Lf) Imp	E(Lc) Imp	E(Lf) Imp	E(Lc) Imp	E(Lf) Imp
10	-42.85%	-25.70%	-42.73%	-24.80%	-42.38%	-24.33%	-42.73%	-23.95%	-43.37%	-23.41%
9	-42.45%	-23.34%	-42.35%	-22.44%	-42.06%	-21.43%	-43.63%	-23.11%	-43.47%	-21.18%
8	-42.19%	-20.40%	-42.64%	-19.37%	-42.78%	-18.63%	-42.56%	-18.77%	-42.03%	-16.46%
7	-42.21%	-16.77%	-43.13%	-17.00%	-43.86%	-16.33%	-42.30%	-14.50%	-43.43%	-15.17%
6	-42.13%	-13.07%	-41.36%	-10.11%	-41.39%	-9.95%	-42.76%	-13.61%	-42.81%	-13.26%
5	-40.63%	-9.04%	-40.11%	-11.04%	-39.19%	-7.63%	-40.45%	-12.24%	-38.86%	-7.64%
4	-36.05%	-5.75%	-34.42%	-6.99%	-33.78%	-3.40%	-33.76%	-6.50%	-32.87%	-4.93%
Average Imp	-41.21%	-16.30%	-40.96%	-15.97%	-40.78%	-14.53%	-41.17%	-16.10%	-40.98%	-14.58%

Dist.	Erlang	k=5	Erlang	k=4	Erlang	k=3	Erlang	k=2	Erlang	k=1
COV	0.447214		0.5		0.57735		0.707107		1	
M	E(Lc) Imp	E(Lf) Imp	E(Lc) Imp	E(Lf) Imp	E(Lc) Imp	E(Lf) Imp	E(Lc) Imp	E(Lf) Imp	E(Lc) Imp	E(Lf) Imp
10	-42.74%	-22.50%	-42.28%	-21.09%	-43.89%	-22.29%	-44.86%	-18.93%	-46.26%	-19.03%
9	-41.95%	-19.02%	-43.36%	-19.96%	-44.21%	-19.86%	-44.48%	-15.45%	-44.00%	-16.92%
8	-42.03%	-15.89%	-43.41%	-16.57%	-44.09%	-16.76%	-44.65%	-13.70%	-41.59%	-13.77%
7	-42.57%	-13.25%	-42.87%	-12.63%	-44.57%	-16.33%	-43.57%	-13.10%	-38.45%	-13.34%
6	-41.64%	-10.08%	-41.46%	-9.74%	-41.28%	-11.17%	-39.49%	-10.30%	-33.76%	-9.94%
5	-38.46%	-7.57%	-38.19%	-7.76%	-36.89%	-6.85%	-34.92%	-8.70%	-27.67%	-7.22%
4	-32.35%	-5.71%	-32.23%	-5.69%	-30.62%	-4.20%	-28.01%	-4.19%	-21.35%	-7.02%
Average Imp	-40.25%	-13.43%	-40.54%	-13.35%	-40.79%	-13.92%	-40.00%	-12.05%	-36.16%	-12.46%

Dist.	Hyperexponential	
COV	1.732050808	1.457737974
M	E(Lc) Imp	E(Lf) Imp
10	-35.56%	-18.53%
9	-32.76%	-19.39%
8	-28.98%	-10.11%
7	-25.61%	-11.59%
6	-21.74%	-11.95%
5	-16.28%	-7.57%
4	-10.52%	-0.81%
Average Imp	-24.49%	-11.42%

V Conclusion

- For a real-time information scenario, we develop a QBD process model where the transition matrix structure is determined by the customer's choice behavior.
- Under a stability condition, we compute the stationary performance measures by utilizing the special structure of QBD process.
- For a non real-time information scenario, we treat the toll and free systems as an M/M/1/K and MMPP/M/1 models, respectively.
- Under the equilibrium condition, we develop a search algorithm for computing the performance measures.

Conclusion…

- With the two models developed, we discover:
  - Real-time queue length information reduces expected waiting times of both free and toll systems – about 40-45% reduction for the toll system and 15-20% reduction for the free system.
  - The performance characteristics remain similar for both information scenarios.
- We also examine:
  - The effect of decision variables of the toll system – toll price and buffer size => inconsistency between the social welfare goal and the private firm' profit goal.
  - The effect of the government subsidy on the performance of the two-tier service systems.

The End

Q & A