Optimal pricing and pre-emptive scheduling in exponential server with two classes of customers

Manu K. Gupta, N. Hemachandra and J. Venkateswaran

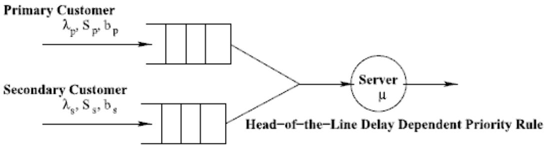
Industrial Engineering and Operations Research
IIT Bombay

ICOCBA - 2012

Outline

- Problem description
- Model by Sinha et. al. (2010) [3]
- Optimization problems
- Search for global optima
- An algorithm for optimal operating parameters
- Conclusions
- Future work

Problem Description



- Primary are the existing customers and their mean waiting time is promised at most S_p.
- Surplus capacity to accommodate new customers.
- The demand of new customers (secondary class) is sensitive to both unit admission price and mean waiting time.
- Objective is to maximize revenue.
- The problem is to quote the unit admission price and service level.

Implementation

- A delay dependent non pre-emptive priority is considered across classes.

$$q_p(t) = delay \times b_p$$

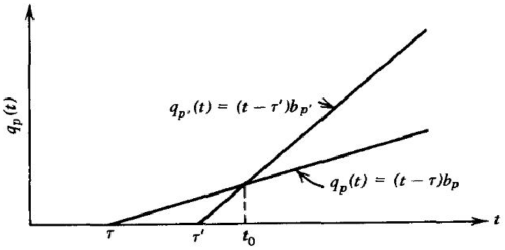


Figure: Illustration of delay dependent priority Kleinrock(1964) [1]

A variation of Sinha et. al.(2010) [3]

Notations:

λ_p	Arrival rate for primary class customers
S_p	Promised mean waiting time of primary class customers
μ	Mean service rate of server
θ	Unit admission price charged to secondary customers
λ_s	Arrival rate of secondary customers
S_s	Promised mean waiting time of secondary class customers

- Delay dependent pre-emptive priority across classes.
- Service times are exponential.
- Remaining settings are similar to [3].

Pre-emptive priority

When a higher priority customer comes, that customer will immediately receive service even if some lower priority customer is currently taking service.

Waiting time expression

Average waiting time in queue for pth class in delay dependent pre-emptive priority (Source Kleinrock (1964)[1])

- There are $1,\ 2\cdots P$ classes.
- Depends on ratios b_i/b_p.

$$W_p = \frac{\frac{W_0}{1-\rho} + \sum_{i=p+1}^{P} \frac{\rho_i}{\mu_p}\left(1 - \frac{b_p}{b_i}\right) - \sum_{i=1}^{p-1} \frac{\rho_i}{\mu_i}\left(1 - \frac{b_i}{b_p}\right) - \sum_{i=1}^{p-1} \rho_i W_i\left(1 - \frac{b_i}{b_p}\right)}{1 - \sum_{i=p+1}^{P} \rho_i\left(1 - \frac{b_p}{b_i}\right)}$$

Waiting time expressions for primary and secondary class

- Waiting time of primary and secondary class customers be W_p and W_s and $\beta = b_s/b_p$.
- expressions for W_p and W_s can be derived from last equation.

$$W_p(\lambda_s,\beta) = \frac{\lambda(\mu-\lambda(1-\beta)) - (\mu-\lambda)\lambda_s(1-\beta)}{\mu(\mu-\lambda)(\mu-\lambda_p(1-\beta))}\mathbf{1}_{\{\beta\leq 1\}} + \frac{\lambda\mu+\lambda_s(\mu-\lambda)(1-\frac{1}{\beta})}{\mu(\mu-\lambda)(\mu-\lambda_s(1-\frac{1}{\beta}))}\mathbf{1}_{\{\beta>1\}} \tag{1}$$

$$W_s(\lambda_s,\beta) = \frac{\lambda\mu+\lambda_p(\mu-\lambda)(1-\beta)}{\mu(\mu-\lambda)(\mu-\lambda_p(1-\beta))}\mathbf{1}_{\{\beta\leq 1\}} + \frac{\lambda(\mu-\lambda(1-\frac{1}{\beta})) - (\mu-\lambda)\lambda_p(1-\frac{1}{\beta})}{\mu(\mu-\lambda)(\mu-\lambda_s(1-\frac{1}{\beta}))}\mathbf{1}_{\{\beta>1\}} \tag{2}$$

where $\lambda = \lambda_p + \lambda_s$.

Some properties of $W_p(\lambda_s, \beta)$ and $W_s(\lambda_s, \beta)$

- $W_p(\lambda_s, \beta)$ and $W_s(\lambda_s, \beta)$ are increasing convex function of λ_s in interval $[0, \mu - \lambda_p)$.
- $W_p(\lambda_s, \beta)$ is an increasing concave function of $\beta \geq 0$ and $W_s(\lambda_s, \beta)$ is a decreasing convex function of $\beta \geq 0$.
- $W_p(\lambda_s, \beta)$ is neither convex nor concave function of (λ_s, β) where $\lambda_s \in [0, \mu - \lambda_p)$ and $\beta \geq 0$. Also, $W_p(\lambda_s, \beta)$ is not a quasi convex function of (λ_s, β).
- $\lambda_s W_s(\lambda_s, \beta)$ is neither convex nor concave function of (λ_s, β) where $\lambda_s \in [0, \mu - \lambda_p)$ and $\beta \geq 0$.

Proof of above statements follows from sign of derivatives of $W_p(\lambda_s, \beta)$ and $W_s(\lambda_s, \beta)$.

The Optimization problem P0

$$Maximize\ \theta\lambda_s \tag{3}$$

Subject to

$$W_p(\lambda_s, \beta) \leq S_p \tag{4}$$
$$S_s \geq W_s(\lambda_s, \beta) \tag{5}$$
$$\lambda_s \leq \mu - \lambda_p \tag{6}$$
$$\lambda_s \leq a - b\theta - cS_s \tag{7}$$
$$\lambda_s, \theta, S_s, \beta \geq 0 \tag{8}$$

- Constraints (4) and (5) are service level constraints.
- Constraints (6) and (7) are system stability and demand constraint respectively.

Optimization problem: P1

$$\textbf{P1:} \quad \max_{\lambda_s, \beta} \frac{1}{b} \left(a\lambda_s - \lambda_s^2 - c\lambda_s W_s(\lambda_s, \beta) \right) \tag{9}$$

Subject to:

$$W_p(\lambda_s, \beta) \leq S_p \tag{10}$$

$$\lambda_s \leq \mu - \lambda_p \tag{11}$$

$$\lambda_s, \beta \geq 0 \tag{12}$$

- $\beta = \infty$ is also a valid decision.
- When $\beta < \infty$, above problem is non convex constraint optimization problem and can be solved using KKT conditions.

Solution of problem P1 ($\beta < \infty$)

Theorem 1: *Suppose* $\frac{a}{c} > \frac{\lambda_p(2\mu-\lambda_p)}{\mu(\mu-\lambda_p)^2}$, *Then there exists* $\lambda_s^{(1)}$ *which is the unique root of cubic* $G(\lambda_s)$ *in the interval* $(0, \mu - \lambda_p)$.

$$G(\lambda_s) \equiv 2\mu\lambda_s^3 - [c + \mu(a + 4\phi_0)]\lambda_s^2 + 2\phi_0[c + \mu(a + \phi_0)]\lambda_s - a\mu\phi_0^2 + c\lambda_p(\mu + \phi_0)$$

where $\phi_0 = \mu - \lambda_p$. *Denote* $\lambda_1 = \lambda_p + \lambda_s^{(1)}$ *and further assume that* S_p *lies in interval* $I \equiv \left(\frac{\lambda_p}{\mu(\mu-\lambda_p)}, \frac{\lambda_1\mu+(\mu-\lambda_1)\lambda_s^{(1)}}{\mu(\mu-\lambda_1)(\mu-\lambda_s^{(1)})}\right)$ *and* $\beta^{(1)}$ *is given by*

$$\beta^{(1)} = \left\{ \begin{array}{ll} \dfrac{(\mu-\lambda_1)\left(\mu S_p(\mu-\lambda_p)-\lambda_p\right)}{\lambda_1^2-(\mu-\lambda_1)(\mu S_p\lambda_p-\lambda_s^{(1)})} & \text{for } \dfrac{\lambda_p}{\mu(\mu-\lambda_p)} < S_p \leq \dfrac{\lambda_1}{\mu(\mu-\lambda_1)} \\ \dfrac{\lambda_s^{(1)}(\mu-\lambda_1)(1+\mu S_p)}{\lambda_1\mu+(\mu-\lambda_1)(\lambda_s^{(1)}+\mu S_p\lambda_s^{(1)}-\mu^2 S_p)} & \text{for } \dfrac{\lambda_1}{\mu(\mu-\lambda_1)} < S_p < \dfrac{\lambda_1\mu+(\mu-\lambda_1)\lambda_s^{(1)}}{\mu(\mu-\lambda_1)(\mu-\lambda_s^{(1)})} \end{array} \right\}$$

then $\lambda_s^{(1)}$ *and* $\beta^{(1)}$ *is strict local maximum of NLP (P1) and constraint* $W_p \leq S_p$ *is binding at this point.*

Solution of problem P1 ($\beta < \infty$)

Theorem 2: *Suppose* $\frac{a}{c} > \frac{\lambda_p(2\mu-\lambda_p)}{\mu(\mu-\lambda_p)^2}$ *and* $S_p = \frac{\lambda_p}{\mu(\mu-\lambda_p)}$, *Then there exists* $\lambda_s^{(1)}$ *which is the unique root of cubic* $G(\lambda_s)$ *in the interval* $(0, \mu - \lambda_p)$. *Then* $\lambda_s^{(1)}$ *and* $\beta^{(2)} = 0$ *is the strict local maximum of NLP(P1) and constraint* $W_p \leq S_p$ *is binding.*

- Used KKT necessary and sufficient condition for problem P1.
- Used a claim that $G(\lambda_s)$ has a unique root in $(0, \mu - \lambda_p)$ if $\frac{a}{c} > \frac{\lambda_p(2\mu-\lambda_p)}{\mu(\mu-\lambda_p)^2}$
- Interval I is obtained by using waiting time expression.
- Intuition behind results.

Optimization problem P2 ($\beta = \infty$)

$$\textbf{P2} \quad max_{\lambda_s} \frac{1}{b}[a\lambda_s - \lambda_s^2 - c\lambda_s \tilde{W}_s(\lambda_s)] \tag{13}$$

subject to:

$$\tilde{W}_p(\lambda_s) \leq S_p \tag{14}$$

$$\lambda_s \leq \mu - \lambda_p \tag{15}$$

$$\lambda_s \geq 0 \tag{16}$$

- Here notation $\tilde{W}_p(\lambda_s) = W_p(\lambda_s, \beta = \infty)$ and $\tilde{W}_s(\lambda_s) = W_s(\lambda_s, \beta = \infty)$.
- Above one dimensional optimization problem turns out to be convex optimization problem.
- KKT necessary conditions will be enough to find optimal solution.

Solution of problem P2 ($\beta = \infty$)

<u>**Theorem 3:**</u> *Suppose* $(\mu - \lambda_p)(2\mu\lambda_p^2 + c(\mu + \lambda_p)) > a\mu\lambda_p^2$ *holds then there exist* $\lambda_s^{(3)}$ *which is the unique root of cubic* $\tilde{G}(\lambda_s)$ *in the interval* $(0, \mu - \lambda_p)$

$$\tilde{G}(\lambda_s) \equiv 2\mu\lambda_s^3 - (c + \mu(a + 4\mu))\lambda_s^2 + 2\mu(c + a\mu + \mu^2)\lambda_s - a\mu^3 = 0 \qquad (17)$$

Denote $\lambda_3 = \lambda_p + \lambda_s^{(3)}$ *and further assume that* S_p *lies in the interval* $J \equiv \left(\frac{\lambda_3\mu + \lambda_s^{(3)}(\mu - \lambda_3)}{\mu(\mu - \lambda_3)(\mu - \lambda_s^{(3)})}, \infty \right)$. *Then* $\lambda_s^{(3)}$ *is the global maxima of NLP (P2) and constraint* $\tilde{W}_p \leq S_p$ *is non binding at this point.*

- Used KKT necessary condition for problem P2.
- $\tilde{G}(\lambda_s)$ has a unique root in $(0, \mu - \lambda_p)$ under given condition.
- Interval J is obtained by using waiting time expression.
- For $S_p \notin J$, waiting time constraint will be binding.

Solution of problem P2 ($\beta = \infty$)

Theorem 4: *Given that S_p lies in the interval J^-. Defined by*

$$J^- = \left\{ \begin{array}{ll} \left(\frac{\lambda_p}{\mu(\mu-\lambda_p)}, \frac{\lambda_3\mu+\lambda_s^{(3)}(\mu-\lambda_3)}{\mu(\mu-\lambda_3)(\mu-\lambda_s^{(3)})}\right) & \text{if } (\mu-\lambda_p)(2\mu\lambda_p^2 + c(\mu+\lambda_p)) > a\mu\lambda_p^2 \\ \left(\frac{\lambda_p}{\mu(\mu-\lambda_p)}, \infty\right) & \text{otherwise} \end{array} \right\}$$

where $\lambda_3 = \lambda_p + \lambda_s^{(3)}$ and $\lambda_s^{(3)}$ is the unique root of cubic $\tilde{G}(\lambda_s)$ in the interval $(0, \mu - \lambda_p)$ whenever $(\mu-\lambda_p)(2\mu\lambda_p^2 + c(\mu+\lambda_p)) > a\mu\lambda_p^2$, then $\lambda_s^{(4)}$ is the global maximum of NLP (P2) and constraint 14 is binding.

$$\lambda_s^{(4)} = \mu - \frac{\lambda_p}{2} - \frac{1}{2}\sqrt{\lambda_p^2 + \frac{4\mu^2}{\mu S_p + 1}} \tag{18}$$

- Used KKT necessary condition to problem P2.
- exploited the fact that waiting time constraint is binding.
- To search for global optima, one needs to compare the objectives of optimization problems P1 and P2.

Search for global optima

- Solution is given by both optimization problems P1 and P2 in interval I.
- Objectives of two optimization problems are compared using arguments similar to non pre-emptive case.

Theorem 5:

1. *Suppose* $0 < \frac{a}{c} \leq \frac{\lambda_p(2\mu-\lambda_p)}{\mu(\mu-\lambda_p)^2}$, *then we can write* $(\hat{S}_p, \infty) = J^- \cup J$ *with* J *being possibly empty. Then optimization problem P2 has a solution but P1 is infeasible. For* $S_p \in (\hat{S}_p, \infty)$, *the optimal solution to P0 is given by optimal solution to P2 with* $\beta^* = \infty$ *and* λ_s^* *is either* $\lambda_s^{(3)}$ *or* $\lambda_s^{(4)}$.
2. *Suppose* $\frac{a}{c} > \frac{\lambda_p(2\mu-\lambda_p)}{\mu(\mu-\lambda_p)^2}$ *holds then*
   - *For* $S_p = \hat{S}_p$, *optimal solution of P0 is given by P1 with* $\lambda_s^* = \lambda_s^{(1)}$ *and* $\beta^* = 0$ *as optimal solution.*
   - *We can write* $(\hat{S}_p, \infty) = I \cup I^+ \cup J$, *with* J *being possibly empty. then optimization problem P1 and P2 have optimal solution. Optimal solution to P0 is given by P1 with* $\lambda_s^* = \lambda_s^{(1)}$ *and* $\beta^* = 0$ *in interval* I *and for* $S_p \in I^+ \cup J$ *optimal solution to P0 is given by P2 with* $\beta^* = \infty$ *and* $\lambda_s^* = \lambda_s^{(3)}$ *or* $\lambda_s^{(4)}$.

Algorithm to find the global optima

Inputs: λ_p, μ, a, b, c and S_p
Steps:

1. if $S_p < \hat{S}_p = \frac{\lambda_p}{\mu(\mu-\lambda_p)}$ or $\frac{a}{c} \leq 0$, then there does not exist a feasible solution. assign $\lambda_s^* = 0$ and stop else go to step 2.
2. if $\frac{a}{c} \leq \frac{\lambda_p(2\mu-\lambda_p)}{\mu(\mu-\lambda_p)^2}$ then go to step 3 else go to step 7.
3. if $S_p = \hat{S}_p$, there does not exist a feasible solution, assign $\lambda_s^* = 0$ and stop else go to step 4.
4. if $\frac{\mu-\lambda_p}{\mu\lambda_p} \leq \frac{a\lambda_p}{2\mu\lambda_p^2+c(\mu+\lambda_p)}$ then $J_l = \infty$ and go to step 6, else define $J_l = \frac{\lambda_3\mu+\lambda_s^{(3)}(\mu-\lambda_3)}{\mu(\mu-\lambda_3)(\mu-\lambda_s^{(3)})}$, $J = (J_l, \infty)$ and find $\lambda_s^{(3)}$ which is the unique root of cubic $\tilde{G}(\lambda_s)$ in the interval $(0, \mu - \lambda_p)$ where
$$\tilde{G}(\lambda_s) \equiv 2\mu\lambda_s^3 - (c + \mu(a + 4\mu))\lambda_s^2 + 2\mu(c + a\mu + \mu^2)\lambda_s - a\mu^3.$$
5. if $S_p \in J$ then $\lambda_s^* = \lambda_s^{(3)}$, $\beta^* = \infty$ go to step 10, else go to step 6.
6. define $J^- = (\hat{S}_p, J_l)$ if J_l is finite and $J^- = (\hat{S}_p, \infty)$ if $J_l = \infty$. Assign $\lambda_s^* = \lambda_s^{(4)} = \mu - \frac{\lambda_p}{2} - \frac{1}{2}\sqrt{\lambda_p^2 + \frac{4\mu^2}{\mu S_p+1}}$, $\beta^* = \infty$ go to step 10.
7. if $S_p = \hat{S}_p$ then find $\lambda_s^{(1)}$, unique root of cubic $G(\lambda_s)$ in the interval $(0, \mu - \lambda_p)$ with $\phi_0 = \mu - \lambda_p$ where
$$G(\lambda_s) \equiv 2\mu\lambda_s^3 - [c + \mu(a + 4\phi_0)]\lambda_s^2 + 2\phi_0[c + \mu(a + \phi_0)]\lambda_s - a\mu\phi_0^2 + c\lambda_p(\mu + \phi_0)$$
and assign $\lambda_s^* = \lambda_s^{(1)}$, $\beta^* = 0$ go to step 10, else go to step 8.

Algorithm to find the global optima

Inputs: λ_p, μ, a, b, c and S_p
Steps:

8. if $\frac{\mu-\lambda_p}{\mu\lambda_p} \leq \frac{a\lambda_p}{2\mu\lambda_p^2+c(\mu+\lambda_p)}$ then $J_l = \infty$ else define $J_l = \frac{\lambda_3\mu+\lambda_s^{(3)}(\mu-\lambda_3)}{\mu(\mu-\lambda_3)(\mu-\lambda_s^{(3)})}$ and find $\lambda_s^{(3)}$, root of cubic $\tilde{G}(\lambda_s)$.
9. find $\lambda_s^{(1)}$, the root of cubic $G(\lambda_s)$, define $I_u = \frac{\lambda_1\mu+\lambda_s^{(1)}(\mu-\lambda_1)}{\mu(\mu-\lambda_3)(\mu-\lambda_s^{(3)})}$. Also define $I = (\hat{S}_p, I_u)$, $I^+ = [I_u, J_l)$ if J_l is finite otherwise take I^+ as $I^+ = [I_u, \infty)$ take $J = (J_l, \infty)$ if J_l is finite otherwise $J = \phi$.
   1. if $S_p \in I$ then $\lambda_s^* = \lambda_s^{(1)}$ and ,

      $$\beta^* = \begin{cases} \dfrac{(\mu-\lambda_1)\left(\mu S_p(\mu-\lambda_p)-\lambda_p\right)}{\lambda_1^2-(\mu-\lambda_1)(\mu S_p\lambda_p-\lambda_s^{(1)})} & \text{for } \dfrac{\lambda_p}{\mu(\mu-\lambda_p)} < S_p \leq \dfrac{\lambda_1}{\mu(\mu-\lambda_1)} \\ \dfrac{\lambda_s^{(1)}(\mu-\lambda_1)(1+\mu S_p)}{\lambda_1\mu+(\mu-\lambda_1)(\lambda_s^{(1)}+\mu S_p\lambda_s^{(1)}-\mu^2 S_p)} & \text{for } \dfrac{\lambda_1}{\mu(\mu-\lambda_1)} < S_p < \dfrac{\lambda_1\mu+(\mu-\lambda_1)\lambda_s^{(1}}{\mu(\mu-\lambda_1)(\mu-\lambda_s^{(1)}} \end{cases}$$

   2. if $S_p \in I^+$ then $\lambda_s^* = \lambda_s^{(3)}$, $\beta^* = \infty$,
   3. if $S_p \in J$ then $\lambda_s^* = \lambda_s^{(4)}$, $\beta^* = \infty$
10. if given problem is feasible then optimum assured service level to the secondary class customers is $S_s^* = W_s(\lambda_s^*, \beta^*)$ and optimal unit admission price charged to secondary class customers is $\theta^* = (a - cS_s^* - \lambda_s^*)/b$.

Conclusions and future work

- Solved non convex and convex optimization problems.
- Comparison of two optimization problems.
- An algorithm to find optimal parameters.
- Comparative study of two queueing systems.
- Numerical study and sensitivity analysis
- Network variation of the model.

L. Kleinrock.
A delay dependent queue discipline.
Naval Research Logistics Quarterly, 11:329–341, September-December 1964.

B. S. Raghav.
Performance analysis of delay dependent priority queue.
Master's thesis, IIT Bombay, 2011.

S. K. Sinha, N. Rangaraj, and N. Hemachandra.
Pricing surplus server capacity for mean waiting time sensitive customers.
European Journal of Operational Research, 205(1):159 – 171, 2010.

Thank you!!!