Pricing shared resources with QoS guarantees in some logistics models

Sudhir Kumar Sinha

Research Scholar
IE&OR, IIT Bombay

Symposium on “Optimization in Supply Chains”

IIT Bombay, Oct 27, 2007

Outline

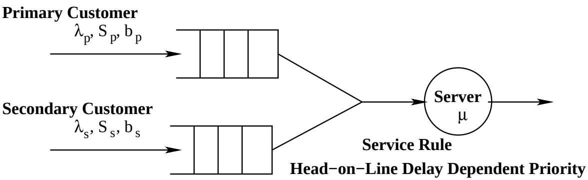
Introduction

- A resource serving its existing customers (Primary customers) under certain service level agreement
- Service level is measured in terms of expected waiting time in the queue
- Agreement results in a steady arrival rate of primary customers
- But, the resource is under-utilized
- Resource owner would like to share the resource to interested new customers (Secondary Customers)

Example:Inland rail container movement in India

- In the recent past, managed by Container Corporation of India Ltd (Concor)
- Now, opened to other private/public sector players
- Interested companies have to
  - arrange rail-linked inland container depot (ICD)
  - procure flat wagons
- Building ICD require high infrastructural set up cost
  - Alternative is to share ICD resources with Concor
- Primary customer: Concor, Secondary customers: New firms

Admission control of secondary customer



- Shared resource resembles a server serving a multi-class queue with heterogeneous service level expectation
- Inclusion of secondary customer
  - Increases traffic intensity
  - Increases resource utilization
  - Adversely affects effective service level offered to the primary customer

Need to control the secondary customer inflow!

Control mechanisms

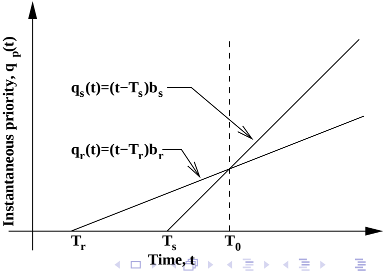
- Assumption: Secondary customer demand depends on price charged and assured service level
- Control mechanisms
  - Pricing policy (Combination of price charged and service level offered)
  - Queue control parameter (Queue discipline) e.g. FCFS, static priority, dynamic priority
- Issues
  1. What should be the suitable pricing policy (a combination of price charged and service level offered) for the secondary customers?
  2. What would be the optimal priority queue management policy?

Delay dependent priority queue [Kleinrock:1964]

- Each class is assigned with a parameter b_p. A high priority class gets high b_p.
- Priority (instantaneous) increases with delay in queue. Specifically, priority at time t of a class p job that arrived at time T_p

$$q_p(t) = (t - T_p)b_p$$

- When server is free, it selects the job with highest instantaneous priority
- Tie is broken by FCFS rule



Key features (M/G/1 head-on-line delay dependent queue)

- Expected waiting time for a class p job

$$W_p = \frac{\frac{W_0}{1-\rho} - \sum_{i=1}^{p-1} \rho_i W_i \left(1 - \frac{b_i}{b_p}\right)}{1 - \sum_{i=p+1}^{P} \rho_i \left(1 - \frac{b_p}{b_i}\right)}, \qquad p \in \{1, 2, \dots, P\}$$

where $\rho_i = \frac{\lambda_i}{\mu_i}$, $\rho = \sum_{p=1}^{P} \rho_i$, $W_0 = \sum_{p=1}^{P} \frac{\lambda_p}{2} \left(\sigma_p^2 + \frac{1}{\mu_p^2}\right)$ and $0 \leq \rho < 1$

- Queue control parameter $\{b_p\}$ appears as ratios (b_p/b_{p+1})
- Waiting time is ordered. If $0 \leq b_1 \leq b_2 \leq \dots \leq b_P$ then $W_1 \geq W_2 \geq \dots \geq W_P$
- Queue control parameter $\{b_p\}$ can be adjusted to meet different performance requirements W_p/W_{p+1} for $p = 1, 2, \dots, P-1$
- An appropriate choice of parameters replicate FCFS and static priority disciplines

Notation

Input parameters

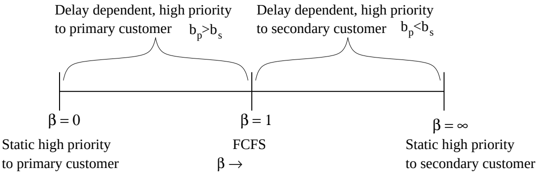
- λ_p: Mean arrival rate of primary customer's jobs
- S_p: Assured service level to primary customers
- $1/\mu$: Mean service time (irrespective of customer class)
- σ^2: Variance of service time (irrespective of customer class)

Decision variables

- θ: Price charged to secondary customer
- S_s: Assured service level to secondary customers
- λ_s: Mean arrival rate of secondary customer's jobs
- $\beta = b_s/b_p$: relative priority of the customers

Key points

- Effect of β on queue discipline



  - $\beta = 0$: Static high priority to primary customer
  - $0 < \beta < 1$: Delay dependent with high priority to primary customer
  - $\beta = 1$: First come first serve
  - $\beta > 1$: Delay dependent with high priority to secondary customer
  - $\beta = \infty$: Static high priority to secondary customer

- Assumption: Potential mean arrival rate of the secondary customer's jobs

$$\Lambda_s(\theta, S_s) = a - b\theta - cS_s$$

where $a, b, c > 0$

Revenue maximization problem

Maximize revenue of the resource owner while maintaining assured service level to existing customers and system capability.

$$P0: \quad \max_{\lambda_s, \theta, S_s, \beta} \theta \lambda_s$$

Subjected to

$$\begin{array}{rcll}
W_p(\lambda_s, \beta) & \leq & S_p & \text{Primary cust. QoS constraint} \\
S_s & \geq & W_s(\lambda_s, \beta) & \text{System capability constraint (QoS)} \\
\lambda_s & \leq & \mu - \lambda_p & \text{System stability constraint} \\
\lambda_s & \leq & a - b\theta - cS_s & \text{Demand constraint} \\
\lambda_s, \theta, S_s, \beta & \geq & 0 & \text{Non-negativity constraint}
\end{array}$$

Reduced revenue maximization problem

- At optimality:

$$\lambda_s^* = a - b\theta^* - cS_s^*$$
$$S_s^* = W_s(\lambda_s^*, \beta^*)$$

- Resulting RM problem

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

  Subjected to

$$W_p(\lambda_s, \beta) \leq S_p$$
$$\lambda_s \leq \mu - \lambda_p$$
$$\lambda_s, \beta \geq 0$$

- $S_s^* = W_s(\lambda_s^*, \beta^*)$ and $\theta^* = [a - \lambda_s^* - cS_s^*]/b$
- Optimal choices of λ_s and β remain insensitive to price sensitivity co-efficient b of the secondary customer.

Expressions for W_p and W_s in current setting

$$W_p(\lambda_s,\beta) = \frac{\lambda\psi\left[\mu-\lambda\left[1-\beta\right]\right]}{\mu\left[\mu-\lambda\right]\left[\mu-\lambda_p\left[1-\beta\right]\right]}\mathbb{1}_{\{\beta\le 1\}} + \frac{\lambda\psi}{\left[\mu-\lambda\right]\left[\mu-\lambda_s\left[1-\frac{1}{\beta}\right]\right]}\mathbb{1}_{\{\beta>1\}}$$

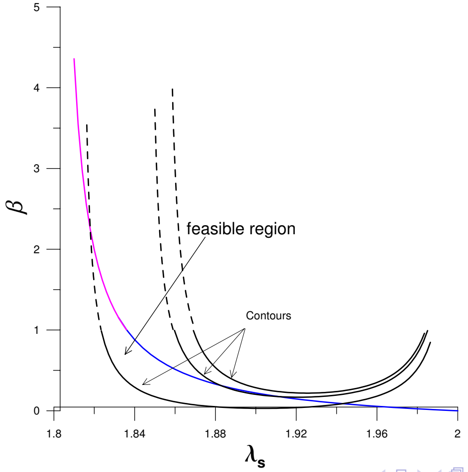
$$W_s(\lambda_s,\beta) = \frac{\lambda\psi}{\left[\mu-\lambda\right]\left[\mu-\lambda_p\left[1-\beta\right]\right]}\mathbb{1}_{\{\beta\le 1\}} + \frac{\lambda\psi\left[\mu-\lambda\left[1-\frac{1}{\beta}\right]\right]}{\mu\left[\mu-\lambda\right]\left[\mu-\lambda_s\left[1-\frac{1}{\beta}\right]\right]}\mathbb{1}_{\{\beta>1\}}$$

where $\lambda = \lambda_p + \lambda_s$, $\psi = \left[1+\sigma^2\mu^2\right]/2$ and $\mathbb{1}_{\{.\}}$ is indicator function.

Remark:

1. $W_p(\lambda_s,\beta)$ and $W_s(\lambda_s,\beta)$ are increasing convex function of λ_s in the interval $[0,\mu-\lambda_p)$.
2. $W_p(\lambda_s,\beta)$ is an increasing concave function of β whereas $W_s(\lambda_s,\beta)$ is a decreasing convex function of β.
3. $W_p(\lambda_s,\beta)$ is a quasi-convex function of β.
4. Hessian matrix $H\left[W_p(\lambda_s,\beta)\right]$ is indefinite.

Illustrative feasible region and objective function contours



Solution algorithm

Based on exhaustive search of KKT point. Inputs: λ_p, μ, σ and S_p

1. Check either $S_p < \hat{S}_p \equiv \frac{\lambda_p \psi}{\mu\left[\mu-\lambda_p\right]}$ or $\frac{a}{c} < \frac{2\mu-\lambda_p}{[\mu-\lambda_p]^2}\psi$, if yes Stop. No feasible solution is possible.

2. Find the only possible real root of the cubic $\tilde{G}(\lambda_s)$ in the interval $[0,\ \mu)$, say $\lambda_s^{(1)}$.

$$\tilde{G}(\lambda_s) = 2\mu\lambda_s^3 - \left[a\mu + c\psi + 4\mu^2\right]\lambda_s^2 + 2\mu\left[a\mu + c\psi + \mu^2\right]\lambda_s - \mu\left[a\mu^2 - c\psi\lambda_p\right]$$

   If $\lambda_s^{(1)} \in [0,\ \mu - \lambda_p)$ and $W_p(\lambda_s^{(1)}, \beta = \infty) < S_p$, then $\lambda_s^* = \lambda_s^{(1)}$ and $\beta^* = \infty$ is optimum. Stop

3. Find the only possible real root of the cubic $G(\lambda_s)$ in the interval $[0,\ \mu - \lambda_p)$, say $\lambda_s^{(2)}$.

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

   where $\phi_0 = \mu - \lambda_p$. Calculate $I_\ell = \frac{\psi\lambda}{\mu[\mu-\lambda_p]}$ and $I_u = \frac{\psi\lambda}{[\mu-\lambda_s^{(2)}][\mu-\lambda]}$

Solution algorithm

4. Define intervals: Low=$[\hat{S}_p, l_\ell)$, Moderate =$[l_\ell, l_u]$ and High $(l_u, a/c]$
   - If $Sp \in Low$ then optimum $\lambda_s^* = \frac{\mu[\mu-\lambda_p]S_p}{\psi} - \lambda_p$ and $\beta^* = 0$
   - If $Sp \in Moderate$, then optimum $\lambda_s^* = \lambda_s^{(2)}$ and

$$\beta^* = \begin{cases} \frac{[\mu-\lambda][\mu S_p[\mu-\lambda_p]-\psi\lambda]}{\psi\lambda^2-\mu S_p\lambda_p[\mu-\lambda]} & \text{for } \frac{\psi\lambda}{\mu[\mu-\lambda_p]} \leq S_p \leq \frac{\psi\lambda}{\mu[\mu-\lambda]} \\ \frac{S_p\lambda_s^*[\mu-\lambda]}{\psi\lambda-S_p[\mu-\lambda_s^*][\mu-\lambda]} & \text{for } \frac{\psi\lambda}{\mu[\mu-\lambda]} < S_p \leq \frac{\psi\lambda}{[\mu-\lambda_s^*][\mu-\lambda]} \end{cases}$$

   - If $Sp \in High$ then optimum $\lambda_s^* = \frac{1}{2S_p}\left[S_p\left[2\mu-\lambda_p\right]+\psi-\sqrt{\left[S_p\lambda_p+\psi\right]^2+4\mu\psi S_p}\right]$ and $\beta^* = \infty$

Numerical example

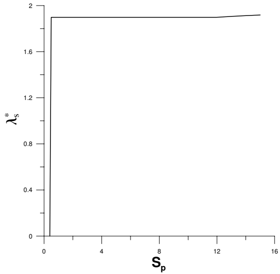
$$\lambda_p = 8,\ \mu = 10,\ \sigma = 0.1,\ a = 100,\ b = 0.2, c = 0.1$$
$$\hat{S}_p = 0.4,\ \lambda_s^{(2)} = 1.898,\ I_\ell = 0.4949,\ I_u = 11.97$$

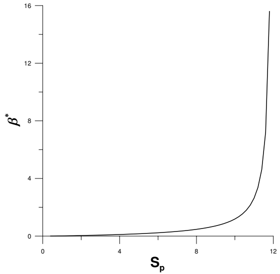
S_p	β^*	λ_s^*	θ^*	S_s^*	Revenue*
0.41	0	0.2	497.86	2.28	99.57
0.42	0	0.4	496.69	2.62	198.68
0.45	0	1	492.75	4.5	492.75
0.4949	0	1.898	466.25	48.52	884.94
1	0.01	1.898	467.32	46.39	886.96
3	0.07	1.898	471.53	37.95	894.97
6	0.23	1.898	477.85	25.32	906.96
9.703	1	1.898	485.66	9.70	921.78
10	1.18	1.898	486.27	8.48	922.96
10	2.64	1.898	488.39	4.24	926.96
11.97	∞	1.898	490.46	0.122	930.87
13	∞	1.905	490.41	0.122	934.65
14	∞	1.912	490.37	0.123	937.82
15	∞	1.918	490.35	0.123	940.58

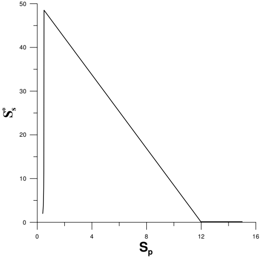
- S_p region is subdivided into intervals, viz: Low, Moderate, High, Very High.
- Either λ_s^* or β^* remains constant in within an interval.

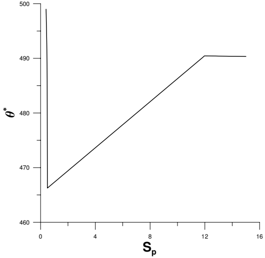
Sensitivity w.r.t. S_p

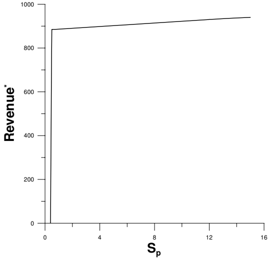
S_p- Interval	S_p	β^*	λ_s^*	θ^*	S_s^*	Revenue*
Low	Increase	Constant (0)	Increase (linear)	Decrease (Concave)	Increase (Convex)	Increase (Concave)
Moderate	Increase	Increase (Convex)	Constant	Increase (Convex)	Decrease (Concave)	Increase (Linear)
High	Increase	Constant (∞)	Increase (Concave)	Decrease	Increase	Increase
Very High	Unconstraint problem, Decision is independent of S_p					











Limitation: Model does not consider variance of service levels.

Thank You.