Pricing server's surplus capacity

Manu K. Gupta

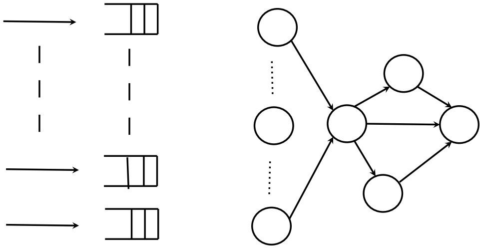
IEOR@IITB

November 2, 2012

Outline

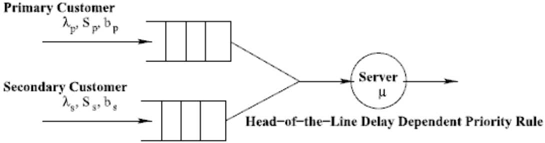
- Introduction
- Model by Sinha et. al. (2010) and some literature
- A proof of conjecture
- Current work
- Future work

Problem definition



- Multi class customers arriving over a network
- Introduce new class without affecting the service level of other classes.
- Scheduling of new class across classes
- Admission control of new class

Single node and two classes (Sinha et. al. (2010))



- There are two classes of customers, primary and secondary.
- Primary are the existing customers and their mean waiting time is promised below S_p.
- There is a surplus capacity to accommodate new customers.
- There is no pre-emption.
- The demand of new customers (secondary class) is sensitive to both unit admission price and mean waiting time.
- The problem is to quote the unit admission price and service level.

Implementation

- A delay dependent non pre-emptive priority is considered across classes (Klienrock (1964)).

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

Notations:

λ_p	Arrival rate for primary class customers
λ_s	Arrival rate of secondary customers
S_p	Promised mean waiting time of primary class customers
S_s	Promised mean waiting time of secondary class customers
μ	Mean service rate of server
σ^2	Variance of service time
θ	Unit admission price charged to secondary customers
ψ	$= \frac{1+\sigma^2\mu^2}{2}$

Original Optimization problem P_0

$$\max_{\lambda_s,\theta,S_s,\beta} \theta\lambda_s \tag{1}$$

Subject to

$$W_p(\lambda_s, \beta) \leq S_p \tag{2}$$

$$S_s \geq W_s(\lambda_s, \beta) \tag{3}$$

$$\lambda_s \leq \mu - \lambda_p \tag{4}$$

$$\lambda_s \leq a - b\theta - cS_s \tag{5}$$

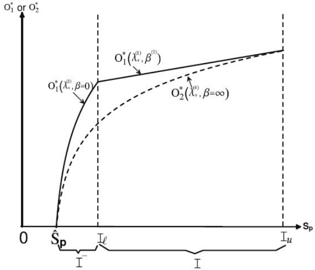
$$\lambda_s, \theta, S_s, \beta \geq 0 \tag{6}$$

- Constraint (2) and (3) are service level constraint for primary and secondary class customers respectively.
- Constraint (4) and (5) are system stability and demand constraint.
- Constraint (3) and (5) will be binding.

Optimization problem P_1 and P_2

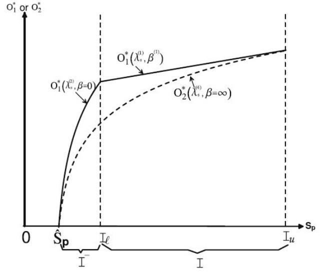
- One can reduce the four dimensional optimization problem (P_0) to two dimensional optimization problem in λ_s and β.
- Range of queue discipline management parameter, $0 \leq \beta \leq \infty$
- $\beta = \infty$ is also a valid decision.
- Optimization problem P_1 reduces to one dimensional convex optimization problem P_2 for $\beta = \infty$.
- To search for global optima, one needs to compare the objective of optimization problems P_1 and P_2

Search for global optima



An algorithm is proposed to find the global optima assuming that conjecture is true.

Search for global optima



An algorithm is proposed to find the global optima assuming that conjecture is true.

Conjecture

For $S_p \in I^-$, the optimal solution of the original problem (P_0) is given by the optimal solution of P_1.

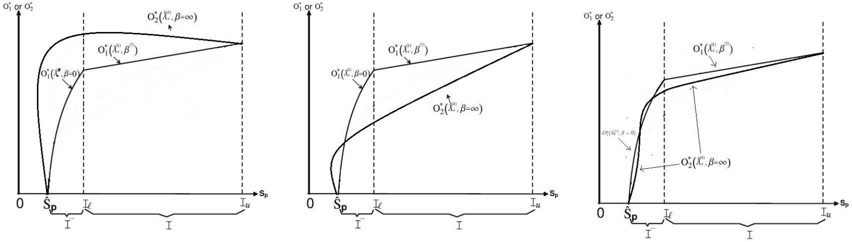
Theorem

Optimal solution for optimization problem P_2, i.e., O_2^* is increasing concave in interval $I^- \cup I$ while O_1^* is increasing concave in I^- and linearly increasing in I

- The fact that O_1^* is increasing concave in I^- and linearly increasing in I follows from Sinha et. al. (2010)
- To prove first part
  - Claim that $I^- \subset J^-$
  - Corollary that $\lambda_s^{(4)}$ is an increasing function of S_p
  - Some optimization and queueing based arguments

Corollary

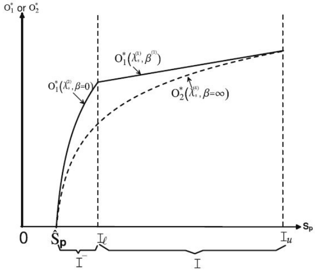
For $S_p \in I^- \cup I$, the optimal solution of P0 is given by optimal solution of P1.



- Contradiction from $O_2^*(\lambda_s^i, \infty) < O_1^*(\lambda_s^f, \beta^f)$ for $S_p \in I$
- Contradiction from infeasibility and $O_2^*(\lambda_s^i, \infty) < O_1^*(\lambda_s^f, \beta^f)$ at $\hat{S}_p + \epsilon$
- Contradiction from concavity of O_2^*

Proof of conjecture

The possible way for O_1^* and O_2^*



Conjecture. *For $S_p \in I^-$, the optimal solution of P0 is given by optimal solution of P1.*
Proof. Follows from above corollary.

A variation of model

- Delay dependent pre-emptive priority across classes
- Service time is exponential.
- Remaining settings are similar to Sinha et. al. (2010)
- Comparison of objectives
- Proposed an algorithm to find the global optimal operating parameters
- A comparative study of two priority policies
- A better algorithm by changing the priority policy
- Sensitivity analysis and numerical examples

Challenges in network

- Departure process of delay dependent priority
  - Departure process of M/M/1 queue (Bruke (1956))
  - Departure process of $\sum M_i/G_i/1$ queue (Stanford (1991))
- Stochastic approximation algorithms for constrained optimization via simulation
  - Scheduling parameter is not compact.
- Relative priority (Haviv et. al. (2007))
  - Relative priority and delay dependent priority are equivalent in two dimension.
  - Relative priority is complete in two dimension.

P. J. Bruke.
The output of a queueing system.
Operations Research, pages 699–704, 1956.

M. Haviv and J. van der Wal.
Waiting times in queues with relative priorities.
Operations Research Letters, pages 591–594, 2007.

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

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.

D. A. Stanford.
Interdeparture-time distribution in the non-preemptive priority $\sum m_i/g_i/1$ queue.
Operations Research, pages 699–704, 1956.

Thank You