ON MEAN WAITING TIME COMPLETENESS AND EQUIVALENCE OF EDD AND HOL-PJ DYNAMIC PRIORITY IN 2-CLASS M/G/1 QUEUE

by

Manu K. Gupta

Along with Prof. N. Hemachandra and Prof. J. Venkateswaran

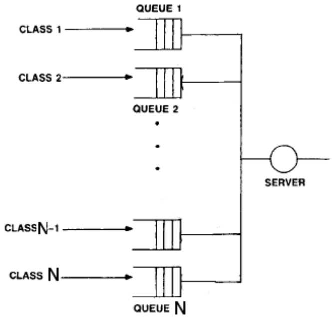
Industrial Engineering and Operations Research
Indian Institute of Technology Bombay

8th International Conference on Performance Evaluation Methodologies and Tools

Outline

Notations

- Single server system with N different classes.
- Independent Poisson arrival rate λ_i and mean service time $1/\mu_i$.
- $\rho_i = \lambda_i/\mu_i$ and $\rho = \sum_{i=1}^{N} \rho_i < 1$
- Performance measure $\mathbf{W} = (w_1, w_2, \ldots, w_N)$.
- All performance vectors are not possible, for example $\mathbf{W} = \mathbf{0}$.



Assumptions

1. Work conserving, non anticipative and non pre-emptive.

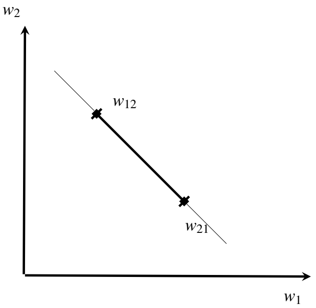
Kleinrock's conservation law (Kleinrock, 1965)

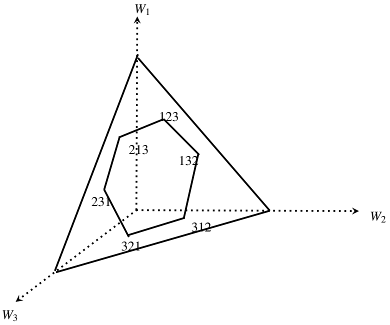
$$\sum_{i=1}^{N} \rho_i w_i = \frac{\rho W_0}{1-\rho} \tag{1}$$

where $W_0 = \sum_{i=1}^{n} \frac{\lambda_i}{2}\left(\sigma_i^2 + \frac{1}{\mu_i^2}\right)$ and σ_i^2 is variance of class i.

Some Properties

- This equation defines a ***hyperplane*** in N-dimensional space of $\mathbf{W}$.
- Dimension of this ***hyperplane*** is $N-1$ for N customer's type.
- In case of two classes, achievable region is a ***straight line segment***.
- In case of three classes, achievable region is a ***polytope***.

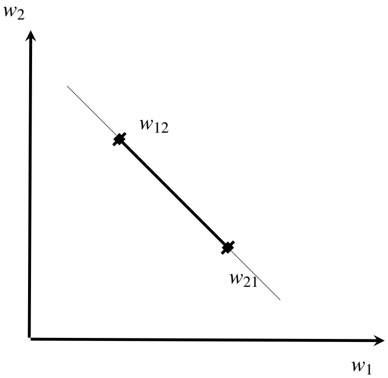




Achievable region in two and three class M/G/1 queue (Mitrani, 2004)

- $(N)!$ extreme points corresponding to non-preemptive strict priority.
- Achievable performance vectors form a *polytope* with these *vertices*.
- A family of scheduling strategy is *complete* if it achieves the *polytope* (Mitrani & Hine, 1977).

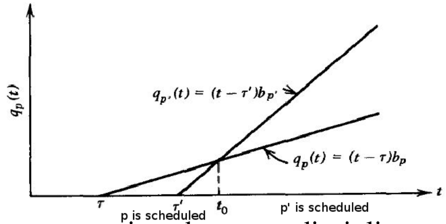
- w_{12} and w_{21} are extreme points on line segment.
- w_{12} is mean waiting time vector when class 1 has strict priority over class 2.
- Every point in the line segment is a convex combination of the extreme points w_{12} and w_{21}.
- $\alpha w_{12} + (1 - \alpha) w_{21}$ achieves all the points in line segment for $\alpha \in [0, 1]$.



Main Results

- Earliest due date based dynamic priority proposed by Goldberg (1977) forms a *complete* class in two class queue.
- Head of Line Priority Jump (HOL-PJ) proposed by Lim & Kobza (1990) forms another *complete* class in two class queue.
- Delay dependent priority (Kleinrock, 1964), earliest due date based dynamic priority and HOL-PJ are mean equivalent.
  - Non linear transformation
- Applications
  - Global FCFS as *minmax* fair policy.
  - A simpler proof of celebrated c/ρ rule for two class M/G/1 queue (Baras et al., 1985), (Yao, 2002).

Delay Dependent Priority (Kleinrock, 1964)



- Class i customers are assigned a queue discipline parameter b_i.
- Instantaneous dynamic priority for customers of class i at time t

$$q_i(t) = (delay) \times b_i, i = 1, 2, \cdots, N.$$

- Customer with highest instantaneous priority receives service.
- Recursion for mean waiting time is derived by Kleinrock (1964) which depends on ratio of b_i.

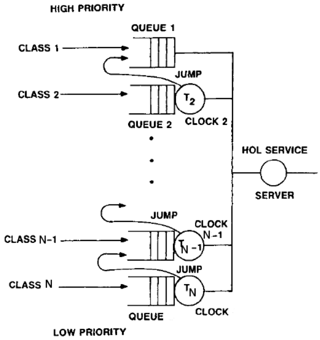
Earliest due date dynamic priority (Goldberg, 1977)

- u_i is the urgency number associated with class i.
- Classes are numbered so that $u_1 \leq u_2 \leq \cdots \leq u_N$ (WLOG).
- A customer from class i is assigned a real number $t_i + u_i$ where t_i is the arrival time of customer.
- Upon service completion, server chooses the customer with minimum value of $\{t_i + u_i\}$.
- Mean waiting time for class r in non preemptive priority is given by:

$$E(W_r) = E(W) + \sum_{i=1}^{r-1} \rho_i \int_0^{u_r - u_i} P(W_r > t)dt - \sum_{i=r+1}^{N} \rho_i \int_0^{u_i - u_r} P(W_i > t)dt$$

Head of Line Priority Jump (Lim & Kobza, 1990)

- Threshold for each class.
- Customers jump to higher class.
- Class 1 has highest priority and class N has lowest.
- Hol-PJ is same as HOL from server's view point.
- Customers are queued according to largeness of excessive delay.



Observations

Mean waiting time of EDD and HOL-PJ are same. Computationally *efficient* and *low* switching frequency.

EDD dynamic priority

In case of two classes, mean waiting time is (Goldberg, 1977, Theorem 2):

$$E(W_h) = E(W) - \rho_l \int_0^u P(T_h[W] > y)dy \tag{2}$$

$$E(W_l) = E(W) + \rho_h \int_0^u P(T_h[W] > y)dy \tag{3}$$

where $u = u_l - u_h \geq 0$. $T_h[W] = \lim_{t \to \infty} T_h[W(t)]$.

$$T_h[W(t)] = \inf\{t' \geq 0;\ \hat{W}_h(t + t' : W(t)) = 0\}$$

where $\hat{W}_h(t + t' : W(t))$ is the workload of the server at time $t + t'$ given an initial workload of $W(t)$ at time t and considering the input workload from class h only after time t.

Consider $u_1,\ u_2 \geq 0$ be the weights associated with class 1 and class 2. Let $\bar{u} = u_1 - u_2$. Mean waiting time for this general setting in case of two classes can be written as:

$$
\begin{aligned}
E(W_1) &= E(W) + \rho_2 \left[\int_0^{\bar{u}} P(T_2(W) > y)dy \, \mathbf{1}_{\{\bar{u} \geq 0\}} \right. \\
&\qquad \left. - \int_0^{-\bar{u}} P(T_1(W) > y)dy \, \mathbf{1}_{\{\bar{u} < 0\}} \right] \qquad (4)
\end{aligned}
$$

$$
\begin{aligned}
E(W_2) &= E(W) + \rho_1 \left[\int_0^{\bar{u}} P(T_2(W) > y)dy \, \mathbf{1}_{\{\bar{u} \geq 0\}} \right. \\
&\qquad \left. - \int_0^{-\bar{u}} P(T_1(W) > y)dy \, \mathbf{1}_{\{\bar{u} < 0\}} \right] \qquad (5)
\end{aligned}
$$

$\bar{u} = -\infty$ and $\bar{u} = \infty$ provide corresponding mean waiting times when strict higher priority is given to class 1 and class 2 respectively.

Delay dependent priority

Mean waiting time in two classes can be obtained by recursion in (Kleinrock, 1964):

$$E(W_1) = \frac{\lambda\psi(\mu-\lambda(1-\beta))}{\mu(\mu-\lambda)(\mu-\lambda_1(1-\beta))}\mathbf{1}_{\{\beta\le 1\}} + \frac{\lambda\psi}{(\mu-\lambda)(\mu-\lambda_2(1-\frac{1}{\beta}))}\mathbf{1}_{\{\beta>1\}}$$

$$E(W_2) = \frac{\lambda\psi}{(\mu-\lambda)(\mu-\lambda_1(1-\beta))}\mathbf{1}_{\{\beta\le 1\}} + \frac{\lambda\psi(\mu-\lambda(1-\frac{1}{\beta}))}{\mu(\mu-\lambda)(\mu-\lambda_2(1-\frac{1}{\beta}))}\mathbf{1}_{\{\beta>1\}}$$

$\beta = 0$ and $\beta = \infty$ provide corresponding mean waiting times when strict higher priority is given to class 1 and class 2 respectively.

Mean Equivalence Result

Lemma

Delay dependent priority and earliest due date priority are mean equivalent in two classes and their priority parameters β and $\bar{u}$ are related as:

$$\beta = \frac{\mu-\lambda}{\lambda_2 + \frac{\rho_2}{\mu W_0}(\mu-\lambda)\lambda_1 \tilde{I}(\bar{u})}\left[\frac{\lambda_2}{\mu-\lambda} - \frac{\rho_2(\mu-\lambda_1)\tilde{I}(\bar{u})}{\mu W_0}\right] \times \mathbf{1}_{\{-\infty \le \bar{u} \le 0\}}$$

$$+ \frac{\lambda_2\left(\frac{\mu W_0}{\mu-\lambda} + \rho_2 I(\bar{u})\right)}{\frac{\mu\lambda_2 W_0}{\mu-\lambda} - \rho_2(\mu-\lambda_2)I(\bar{u})}\mathbf{1}_{\{0 \le \bar{u} \le \infty\}}$$

where integrals $\tilde{I}(\bar{u}) = \int_0^{-\bar{u}} P(T_1(W) > y)dy$ *and* $I(\bar{u}) = \int_0^{\bar{u}} P(T_2(W) > y)dy$.

Obtained by equating mean waiting time expressions for two scheduling policies.

Mean Completeness Result

- Delay dependent priority is a mean complete dynamic priority discipline in case of two classes (Federgruen & Groenevelt, 1988).
- An alternate proof for mean completeness of DDP is proposed.
  - One-one correspondence between β of DDP and α, convex combination parameter.
- There is one-one transformation between $\bar{u}$ and β due to monotonicity.
- EDD with two classes of priority is mean complete.
  - A separate proof.
  - One-one correspondence between $\bar{u}$ of EDD and α, convex combination parameter.

HOL-PJ Dynamic Priority

- Mean waiting time expression for HOL-PJ is same as EDD.
    - Urgency number and overdue in EDD correspond to delay requirement and excessive delay in HOL-PJ.
- There is a one-to-one non-linear transformation for mean waiting time between HOL-PJ and DDP discipline.
- Hence, HOL-PJ is mean complete in two class M/G/1 queues.

Global FCFS

- Fairness is in terms of minimizing the maximum dissatisfaction.
- Dissatisfaction of a customer is quantified in terms of mean waiting time of that customer's class.

$$\min_{\alpha \in \mathcal{F}} \max_{i \in \mathcal{I}} \; E(W_\alpha^{(i)}) \tag{6}$$

$\mathcal{I}$: Set of classes
$\mathcal{F}$: Work conserving, non pre-emptive and non anticipative scheduling.
$E(W_\alpha^{(i)})$: Mean waiting time for class i customers when scheduling policy $\alpha \in \mathcal{F}$ is employed.

$$\min_{\alpha} \epsilon_{\alpha}$$

$$\begin{aligned} E(W_{\alpha}^{(i)}) &\leq \epsilon_{\alpha} \quad \alpha \in \mathcal{F}, i \in \mathcal{I} &(7)\\ \epsilon_{\alpha} &\geq 0, &(8)\end{aligned}$$

Since EDD is a *complete* parametrized dynamic priority discipline in case of two classes, it can be re-written as

$$\min_{\bar{u}} \epsilon_{\bar{u}}$$

$$\begin{aligned} E(W_{\bar{u}}^{(i)}) &\leq \epsilon_{\bar{u}} \quad \bar{u} \in [-\infty, \infty], i \in \mathcal{I} &(9)\\ \epsilon_{\bar{u}} &\geq 0, &(10)\end{aligned}$$

Solution

$\bar{u} = 0$ is optimal solution and this corresponds to global FCFS scheduling.

Optimal Scheduling Policy

$$\textbf{P1} \quad \min_{\alpha \in \mathcal{F}} \; c_1 E(W_\alpha^{(1)}) + c_2 E(W_\alpha^{(2)})$$

where $\mathcal{F}$ is set of all work conserving, non pre-emptive and non anticipative scheduling policies. Problem **P1** is equivalent to **P2** defined below:

$$\textbf{P2} \quad \min_{\bar{u} \in [-\infty, \infty]} \; c_1 E(W_{\bar{u}}^{(1)}) + c_2 E(W_{\bar{u}}^{(2)})$$

Solution

Optimization problem **P2** can be easily solved to yield the optimal c/ρ rule

Conclusions and Future Work

- The notion of completeness is discussed for work conserving queueing systems.
- Certain parametrized dynamic priorities (EDD and HOL-PJ) are shown to be mean complete in two class M/G/1 queue.
- Mean waiting time equivalence between EDD, DDP and HOL-PJ is established.
- An explicit one-to-one nonlinear transformation is given between EDD and DDP.
- Significance of these results is discussed.
- It will be interesting to extend these ideas in higher dimensions.

References I

Baras, J., Dorsey, A., & Makowski, A. (1985). Two competing queues with linear costs and geometric service requirements: The μ c-rule is often optimal. *Advances in Applied Probability*, (pp. 186–209).

Federgruen, A., & Groenevelt, H. (1988). M/G/c queueing systems with multiple customer classes: Characterization and control of achievable performance under nonpreemptive priority rules. *Management Science*, *9*, 1121– 1138.

Goldberg, H. M. (1977). Analysis of the earliest due date scheduling rule in queueing systems. *Mathematics of Operations Research*, *2(2)*, 145–154.

Kleinrock, L. (1964). A delay dependent queue discipline. *Naval Research Logistics Quarterly*, *11*, 329–341.

Kleinrock, L. (1965). A conservation law for wide class of queue disciplines. *Naval Research Logistics Quarterly*, *12*, 118–192.

Lim, Y., & Kobza, J. E. (1990). Analysis of delay dependent priority discipline in an integrated multiclass traffic fast packet switch. *IEEE Transactions on Communications*, *38 (5)*, 351–358.

Mitrani, I. (2004). *Probabilistic Modelling*. Cambridge University Press.

Mitrani, I., & Hine, J. (1977). Complete parametrized families of job scheduling strategies. *Acta Informatica*, *8*, 61– 73.

Yao, D. D. (2002). Dynamic scheduling via polymatroid optimization. In *Performance Evaluation of Complex Systems: Techniques and Tools* (pp. 89–113). Springer.

Thank you!!!