数理解析研究所講究録
564巻 1985年 187-196

多様な遅延品質を提供する待ち行列の最適処理規律

NTT武蔵野研究所　横山雅明 (Masaaki Yokoyama)

1　まえがき

パケット交換等，待合せを伴うサービス業務では，客（本文では，客，パケット，ジョブ等を総称して客と呼ぶ）をクラス分けして遅延品質の多様化を図り，サービス性を向上させることが考えられる。

本文では以上の観点から，許容待ち時間の異なる複数の客のクラスを扱う待ち行列を考える。各クラス毎に許容待ち時間に応じた非線形の評価関数（以下では不満度関数と呼ぶ）を導入し，G/G/s/m の一般モデルにおいて評価値（不満度）の総和を最小にする最適処理規律を明らかにする。

2　モデル

2.1　不満度関数 (penalty function)

待ち時間に対して客が感じる不満，あるいはその不満に対

してサービス提供者が支払うペナルティを不満度（penalty）として表わし，本文における処理規律の評価尺度とする。そして許容待ち時間で客をクラス分けし，クラス i の客の不満度関数（待ち時間に対する不満度の関数）を $f_i(x)$ で表わす。各々の客は許容待ち時間より短い時間なら待たされてもあまり不満を感じないが，許容待ち時間以上に待たされると強い不満を感じるようになると考えられる。従って不満度関数 f_i は図1図2に示すような客の許容待ち時間をパラメータとして持つ関数を考えるのが自然である。図1は待ち時間が許容値を越えた時の不満度増加の度合いがどのクラスも等しい様子を表わしている。図2は待ち時間が許容値を越えた時の不満度の増加が，許容待ち時間の大きいクラスほどゆるやかになる様子を表わしている。

これらの不満度関数の関係を次のようにモデル化する。

［定義1］ 第1種不満度関数

不満度関数 $f_i(x)$, $\forall i \in C$（C は自然数の集合の部分集合），に対して $R_+=[0,\infty)$ 上で次式を満たす関数 $f(x)$ が存在する時，$f_i(x)$ は $f(x)$ を基底とする第1種不満度関数であると言う。

$$\begin{cases} f_i(x)=f(x+k_i), & \forall i \in C \\ \min\ k_i=0 \end{cases} \qquad (1)$$

ここで k_i は任意の実定数である。

[定義2]　第2種不満度関数

不満度関数 $f_i(x)$, $\forall_i \in C$, に対して R_+ 上で次式を満たす関数 $f(x)$ が存在する時, $f_i(x)$ は $f(x)$ を基底とする第2種不満度関数であると言う。

$$\begin{cases} f_i(x) = f(h_i x), & \forall_i \in C \\ \min\ h_i = 1 \end{cases} \tag{2}$$

ここで h_i は任意の実定数である。

例えば図1図2において, クラス i の客の許容待ち時間を α_i で表わし, 許容待ち時間の最も大きいクラスを r で表わす。そして $f(x) = f_r(x)$, $k_i = \alpha_r - \alpha_i$ と置くと, 図1の不満度関数は式(1)を満たす。また $f(x) = f_r(x)$, $h_i = \alpha_r / \alpha_i$ と置くと, 図2の不満度関数は式(2)を満たす。

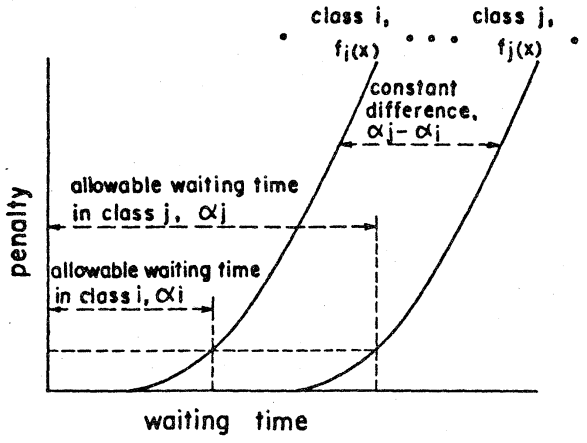

図1　不満度関数の例
（第1種不満度関数）

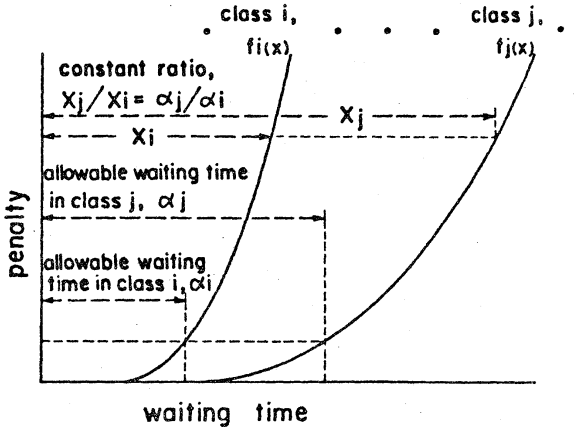

図2　不満度関数の例
（第2種不満度関数）

不満度関数はサービスの種類に応じて適当に選ぶ必要があるが，その多くは第1種または第2種の不満度関数でモデル化できると考えられる。本文では非線形評価関数の下での処理規律最適化の第1歩として，第1種不満度関数を検討の対象とする。また、ゆるやかな制約条件として、その基底が凸関数または凹関数であることを仮定する。

2.2 待ち行列

G/G/s/m の一般モデルを考える。すなわち，サーバの数は s（$1 \leqq s \leqq \infty$）で各々同じ処理能力を持ち，待ち室の数は m（$0 \leqq m \leqq \infty$）で，客の到着は任意の定常過程に従うものとする。さらに以下の条件を仮定する。

[仮定]

(1) サービス時間は互に独立で同一の分布に従う確率変数（i.i.d.r.v.）で，客の到着過程と独立である。

(2) 処理規律は非割込みで，サービス時間とは独立である。すなわち，サービス中の客に割込むことはしない。また個々の客のサービス時間を調べてサービスの順番を変えることはしない。

(3) 系内において仕事は保存される。すなわち，系内でサービス要求が新たに生じたり無くなったりしない。

(4) 待ち室に客がいる限りサーバは空き状態にならない。

(5) 全稼動期間（全てのサーバが連続的に塞っている期間）は確率1で有限である。

3 最適処理規律

[記号の説明]

$f_i(x)$： クラス i の不満度関数，$i \in C$

$c(i)$： 全稼動期間において i 番目に到着した客のクラス番号，$c(i) \in C$

$\pi(i)$： 全稼動期間において i 番目に到着した客が，処理規律 π の下でサービスされる順番

$w_\pi(i)$： 全稼動期間において i 番目に到着した客の処理規律 π の下での待ち時間

$t_a(i)$： 全稼動期間において i 番目に到着した客の到着時刻

$t_s(i)$： 全稼動期間における i 番目のサービス開始時刻

[定義3] 最適

全ての全稼動期間において次式を満たす処理規律 π_0 は最適である。

$$\sum_{i=1}^{n} E\left[f_{c(i)}\left(w_{\pi_0}(i)\right)\right] = \min_{\pi} \sum_{i=1}^{n} E\left[f_{c(i)}\left(w_{\pi}(i)\right)\right] \qquad (3)$$

ここで n は全稼動期間中に到着する客の数である。

定義3は，不満度の期待値が存在する場合には，不満度の

期待値を最小にすることと同等である。

次に第1種不満度関数を前提とした2つの処理規律π_1, π_2を導入する。

［定義4］　処理規律π_1, π_2

処理規律π_1, π_2は以下の(1)～(3)によって定義される。

(1)　π_1, π_2はクラスiの客が到着した時，その客が実際の到着時刻よりk_i時間前に到着したものと見なす到着時刻の変換を行う。

(2)　π_1は(1)により変換された到着時刻に基づいて客を到着順にサービスする。

(3)　π_2は(1)により変換された到着時刻に基づいて客を到着順と逆の順番でサービスする。

ここで，ある全稼動期間に到着した客の任意の組(U, V)を考え，その全稼動期間における客U, Vの到着順序を各々u, vで表わすと，定義4より直ちに次の命題が得られる。

［命題1］

$t_a(u)-k_{c(u)} < t_a(v)-k_{c(v)}$とする。この時，処理規律$\pi_1$の下では，$t_a(u) > t_s(\pi_1(v))$ ならば $t_s(\pi_1(u)) > t_s(\pi_1(v))$，$t_a(u) \leqq t_s(\pi_1(v))$ ならば $t_s(\pi_1(u)) \leqq t_s(\pi_1(v))$ である。また処理規律π_2の下では，$t_a(v) > t_s(\pi_2(u))$ ならば $t_s(\pi_2(v)) > t_s(\pi_2(u))$，$t_a(v) \leqq t_s(\pi_2(u))$ ならば $t_s(\pi_2(v)) \leqq t_s(\pi_2(u))$である。

以上の準備に基づいて，本文の主題である最適処理規律に関する定理を次に示す。

［定理1］

$f(x)$を基底とする第1種不満度関数において，$f(x)$がR_+上の凸関数であれば処理規律π_1は最適である。また$f(x)$がR_+上の凹関数であれば処理規律π_2は最適である。

（証明）

$f(x)$を凸関数とする。ある全稼動期間における客の到着時刻の見本過程とサービス開始時刻の見本過程を考える。その全稼動期間に到着する客の数をnとする。次の条件を満たす客の組(U, V)が存在する処理規律πを考える。

$$\begin{cases} t_a(u) - k_{c(u)} < t_a(v) - k_{c(v)} \\ t_a(u) \leqq t_s(\pi(v)) \\ t_s(\pi(u)) > t_s(\pi(v)) \end{cases} \tag{4}$$

命題1より処理規律π_1の下では式(4)の条件を満たす客の組は存在しない。すなわち，$t_a(u) - k_{c(u)} < t_a(v) - k_{c(v)}$，$t_a(u) > t_s(\pi(v))$のとき$t_s(\pi(u)) > t_s(\pi(v))$であるのは自明であることから，式(4)を満たす客の組(U, V)が存在しない処理規律は，全ての客の待ち時間に関して処理規律π_1と等しい。そこで客Uを$\pi(v)$番目にサービスし，客Vを$\pi(u)$番目にサービスし，その他の客は処理規律πと同じ順番でサ

ービスする処理規律 π^* を考える。ここで着目した全稼動期間における不満度の総和を

$$M_\pi = \sum_{i=1}^{n} f_{c(i)}\left(t_s(\pi(i)) - t_a(i)\right) \tag{5}$$

とおくと，サービス時間が i.i.d.r.v. で到着過程と独立であることから次式が成り立つ。

$$\begin{aligned} M_\pi - M_{\pi^*} = {} & f_{c(u)}\left(t_s(\pi(u)) - t_a(u)\right) + f_{c(v)}\left(t_s(\pi(v)) - t_a(v)\right) \\ & - f_{c(v)}\left(t_s(\pi(u)) - t_a(v)\right) - f_{c(u)}\left(t_s(\pi(v)) - t_a(u)\right) \end{aligned} \tag{6}$$

ここで，

$$\lambda = \frac{t_a(v) - k_{c(v)} - t_a(u) + k_{c(u)}}{t_s(\pi(u)) - t_s(\pi(v)) + t_a(v) - k_{c(v)} - t_a(u) + k_{c(u)}} \tag{7}$$

とおいて，式(1)式(7)を用いて式(6)を変形する。式(4)の条件より $0 < \lambda < 1$ となることと，$f(x)$ が凸関数であることから，式(6)は次のように変形できる。

$$\begin{aligned} M_\pi - M_{\pi^*} \geqq {} & f\left((1-\lambda)\left(t_s(\pi(u)) - t_a(u) + k_{c(u)}\right)\right. \\ & \left. + \lambda\left(t_s(\pi(v)) - t_a(v) + k_{c(v)}\right)\right) \\ & + f\left((1-\lambda)\left(t_s(\pi(v)) - t_a(v) + k_{c(v)}\right)\right. \\ & \left. + \lambda\left(t_s(\pi(u)) - t_a(u) + k_{c(u)}\right)\right) \\ & - f\left(t_s(\pi(u)) - t_a(v) + k_{c(v)}\right) \\ & - f\left(t_s(\pi(v)) - t_a(u) + k_{c(u)}\right) \\ = {} & 0 \end{aligned} \tag{8}$$

すなわち $M_{\pi} \geqq M_{\pi^*}$ である。従って任意の置換が互換の積で表わせることから $M_{\pi} \geqq M_{\pi_1}$ となり，次式が得られる。

$$M_{\pi_1} = \min_{\pi} M_{\pi} \tag{9}$$

ところで，処理規律はサービス時間とは独立であるから，n人の客のサービス開始時刻の結合分布は処理規律と独立である。従って式(9)より次式が成り立つ。

$$\sum_{i=1}^{n} E\left[f_{c(i)}(w_{\pi_1}(i))\right] = \min_{\pi} \sum_{i=1}^{n} E\left[f_{c(i)}(w_{\pi}(i))\right] \tag{10}$$

すなわち定義より処理規律 π_1 は最適である。また $f(x)$ が凹関数の場合も同様の手法により証明される（省略）。

（証明終り）

4 むすび

遅延品質多様化の観点から，許容待ち時間の異なる複数の客のクラスを扱う待ち行列の最適処理規律について検討した。許容待ち時間を考慮した評価関数として第1種不満度関数と第2種不満度関数を定義し，第1種不満度関数が適用される G/G/s/m の一般モデルにおいて不満度の総和を最小にする最適処理規律を明らかにした。

謝辞　確率見本過程の扱いおよび客の到着過程が任意の場合の待ち時間の期待値の存在性等について有益な御討論をして頂いた、NTT武蔵野電気通信研究所、高橋敬隆研究主任に深く感謝致します。

文　献

(1) N.K.Jaiswal, "Priority queues," Academic Press, New York (1968).
(2) D.R.Cox and W.L.Smith, "Queues," Methuen, London (1961).
(3) I.Brosh and P.Naor, "On optimal disciplines in priority queuing," Bull. Inst. Internat. Statist. 40, pp.593-609 (1963).
(4) K.R.Balachandran, "Parametric priority rules: an approach to optimization in priority queues," Operations Res. 18, 3, pp.526-540 (1970).
(5) E.B.Veklerov, "Optimal priorities in queuing systems," Automation and Remote Control 32, 6, pp.986-989 (1971).
(6) A.Beja and E.Sid, "Optimal priority assignment with heterogeneous waiting costs," Operations Res. 23, 1, pp.107-117 (1975).
(7) C.E.bell, "Efficient operation of optional-priority queuing systems," Operations Res. 21, 3, pp.777-786 (1973).
(8) E.Kofman and S.A.Lippman, "An M/M/1 dynamic priority queue with optional promotion," Operations Res. 29, 1, pp.174-188 (1981).
(9) O.A.Vasicek, "An inequality for the variance of waiting time under a general queuing discipline," Operations Res. 25, 5, pp.879-884 (1977).
(10) B.T.Doshi and E.H.Lipper, "Comparisons of service disciplines in a queueing system with delay dependent customer behaviour," Proc. Applied Probability-Computer Science: The Interface, II, Birkhauser, pp.269-301 (1982).