

制限型プロセッサシェアリングモデルについて

01009830 駒澤大学 小沢利久 OZAWA Toshihisa

1. はじめに

TCP/IP に代表されるパケット通信では、リンクなどの共通リソースを複数のコネクション（ユーザー）でシェアしながら利用しているが、そのような仕組みはひとつのリソースに注目し、適当な時間スケールでの平均的な挙動を考えることでプロセッサシェアリングモデル（PS モデル）のバリエーションとして近似的に表すことができる [2, 3, 4]. その中のひとつに制限型 PS モデル（LPS モデル） [2] がある。これは、客にクラスを設け、クラス毎にサーバー能力割り当て（配分率）の上限値を設定する PS モデルである。パケット通信ではその上限値が、例えばアクセスラインの帯域に相当する。よって、クラス（上限値）を変更したときに平均ファイル転送時間などの通信性能がどのように改善（悪化）するのかが、クラス所属の選択権を持つ客にとって重要な関心事となる。そこで本報告では、LPS モデルについて、配分率の上限値の変更が性能に及ぼす影響を考察する¹。

2. LPS モデル

2.1. 複数クラス PS モデル

ここでは複数クラス PS モデルを広く次のように定義する。まず、客のクラス数を K とし、クラス k の時点 t での系内容数を $X_k(t)$ 、そのベクトルを $X(t)$ とする。サーバーが単位時間に提供可能なサービス時間を 1 とし、クラス k の客一人当たりへのサーバー能力の瞬間的な配分率を $\gamma_k(X(t))$ 、そのベクトルを $\gamma(X(t))$ で与える。ここで、 $\gamma_k(X(t)) \geq 0$, $\gamma_1(X(t)) + \dots + \gamma_K(X(t)) \leq 1$ とする。このとき、クラス k のある客が時間区間 $[0, t]$ で受けたサービスの量は $\int_0^t \gamma_k(X(s)) ds$ で与えられる。

2.2. DPS モデルと LPS モデル

ここでは代表的な PS モデルである Discriminatory PS (DPS) モデルと LPS モデルについて示す。

DPS モデル [5]: 配分にクラス依存の重みを考慮した PS モデルのことで、

$$\gamma_k(\mathbf{x}) = g_k / (\mathbf{g}, \mathbf{x}), k = 1, \dots, K.$$

ここで、 $\mathbf{x} = (x_1, \dots, x_K)$, g_k はクラス k の重みで \mathbf{g} はそのベクトル、 (\mathbf{g}, \mathbf{x}) は内積である。 $g_k = \text{const.}$ の場合が標準的な PS モデルであり、Egalitarian PS (EPS) モデルと呼ばれ、 $\gamma_k(\mathbf{x}) = 1/\|\mathbf{x}\|$ となる。ここで、 $\|\mathbf{x}\| = x_1 + \dots + x_K$ である。

¹この報告の内容は NTT 研究所の川原亮一氏と石橋圭介氏とのコミュニケーションが切っ掛けとなっており、ベースとなっている。

LPS モデル [2]: サーバー能力の配分率にクラス依存の制限（上限）があるモデルのことで、 $k = 1, \dots, K$ に対して、

$$\gamma_k(\mathbf{x}) = \min \left\{ r_k, \max_{1 \leq j \leq k} \frac{1 - \sum_{i=1}^{j-1} r_i x_i}{\|\mathbf{x}\| - \sum_{i=1}^{j-1} x_i} \right\}.$$

ここで、 $0 < r_k \leq 1$ はクラス k の配分率の上限（上限配分率）であり、 $r_1 \leq r_2 \leq \dots \leq r_K$ とする。LPS モデルは max-min 公平性（最小の配分率を最大にする）に従う。制限値がクラスに依存しない LPS モデル ($r_k = \text{const.}$) を hLPS モデルと呼ぶことにする。

3. LPS モデルの解析（数値計算）

クラス k の客は強度 λ_k のポアソン過程に従って到着し、サービス時間は平均 $1/\mu_k$ の一般分布に従うとする。EPS モデルと hLPS モデルは対称待ち行列であり、系内容数の定常分布は積形式解によって表すことができる [3]. しかし、DPS モデルと LPS モデルは対称待ち行列の性質を満たさないため、系内容数の定常分布を簡単な式で表すことはできない。DPS モデルについては、同時にサービスを受けているクラス k とクラス l の客の配分率の比が状態に依存せず常に g_k/g_l であることを利用して平均応答時間を求めることができるが [1], LPS ではそのような単純な関係は成り立たない。

そこで、サービス時間を指数分布として、マルコフ連鎖 $\{X(t)\}$ の定常分布を数値計算によって求めることにする。数値計算の方法は様々考えられるが、ここでは以下の方法を用いる。

○まず、レベル n を $\mathcal{L}_n = \{\mathbf{x} \in Z_+^K : \|\mathbf{x}\| = n\}$ で与える。このレベル分けを基にして推移速度行列 Q をブロック行列化する。すると Q は非均質な (nonhomogeneous) ブロック 3 重対角行列となる ($\{X(t)\}$ は nonhomogeneous QBD になる)。この Q に対して縮約/非縮約法を適用して定常分布を数値的に求める。

○ただし、 Q は無限次元なので、適当なレベル以上ではレベル内での分布（条件付き分布）が多項分布に従うとして有限次元に落とす。これは、 $n^* = \lceil 1/r_1 \rceil$ として、LPS モデルではレベル n^* 以上での状態推移率が EPS モデルでのそれと同じになることから、適当なレベル以上ではレベル内での分布が EPS モデルでのそれと十分近いという仮定に基づく。

この方法では、 Q のブロック対角成分が対角行列とな

り、非対角成分も階層的な構造を持つ。また、縮約されたマルコフ連鎖は出生死滅過程になる。しかし、やはり次元との関係で大きなモデルに適用するのは困難である。

4. 数値例

クラス数が $K = 2$ の場合について、一方が上限配分率を増加させた場合の平均応答時間 W_k , $k = 1, 2$, の変化について調べる。どの数値例も $\mu_1 = \mu_2 = 1$ とした。

表 1 は両クラスの到着率が等しい場合に、 r_2 または r_1 を増加させていったときの平均応答時間である。これらの表から、

- r_2 を増加させても W_1 はあまり変化しない、
 - r_1 を増加させても W_2 はあまり変化しない、
- ことが分かる。このことから、LPS モデルではあるクラスの上限配分率の変化は他のクラスにはあまり影響しないことが予想される。また、
- r_2 を増加させると W_1, W_2 とも単調に減少している、
 - r_1 を増加させると W_1 は単調に減少するが、 W_2 は増加の後、減少する、

ことより、あるクラスが上限配分率を増加させると他のクラスの平均応答時間も減る可能性があることを示している。これについては次節で取り上げる。

表 2 は両クラスの到着率が異なる場合に、 r_2 または r_1 を増加させていったときの平均応答時間である。比較のために到着率が等しい場合の数値例も示す。表より、両クラスの到着率が異なる場合もそれらが等しい場合と同様の数値的傾向を示していることが分かる。ところで、表 2 は、 $\lambda_1 = \lambda_2 = 0.4$ であったものを $\lambda_1 = 0.1, \lambda_2 = 0.7$ に変えた場合の比較も示している。これは、クラス 1 のトラヒックを一部分、クラス 2 へ移したことを意味するが、そのような移行をしても平均応答時間はあまり変化しないことが分かる。パケット通信でいえば、より広いアクセス帯域のサービスに多くの人が移行したときの状況に対応する。

以上に示したことは文献 [2] でもシミュレーションによって確認されている。

5. 考察

数値例で示したことに関連し、サンプルパスの議論を用いて次が成立することが分かる。

- 一般の到着過程、一般のサービス時間分布、一般のクラス数 K であるエルゴード的な LPS モデルにおいて、最上位クラスの上限配分率 r_K を増加させると系内客数の定常分布は確率順序の意味で単調非増加となる。

このことより、平均系内客数及びリトルの式から平均応答時間についても同様の単調性が成り立つことが分かる。

ところで、LPS モデルについて数値例で示したように、あるクラスの上限配分率の変化が他のクラスにあまり影響を与えないことが確かであれば、平均応答時間を近似するひとつの方法が導かれる。すなわち、クラス k を注目するクラスとし、そのクラス以外のクラスの上限配分率を全て r_k にした hLPS モデルを構成し、そのモデルにおけるクラス k の平均応答時間を W_k の近似として用いるのである。文献 [2] で提案された近似式はこのようにして構成された近似式と解釈することもできる。

6. おわりに

数値例から予想される性質がどこまで正確に成り立つかの確認をこれからの課題としたい。

参考文献

- [1] G.Fayolle and I.Mitrani, Sharing a processor among many job classes, J. of ACM 27(3) (1980).
- [2] R.Kawahara, et al., A method of bandwidth dimensioning and management for aggregated TCP flows with heterogeneous access, submitted (2004).
- [3] 小沢利久, プロセッサシェアリングモデルを用いた通信応答時間の近似解析, 信学技報 NS2002-12 (2002).
- [4] J.W.Roberts and L.Massoulié, Bandwidth sharing and admission control for elastic traffic, 11th ITC Specialist Seminar (1998).
- [5] S.F.Yashkov, Processor-sharing queues: some progress in analysis, Queueing systems 2 (1987).

表 1: 平均応答時間 : 到着率が等しい場合

r_1	r_2	$\lambda_1 = \lambda_2 = 0.3$		$\lambda_1 = \lambda_2 = 0.45$	
		W_1	W_2	W_1	W_2
0.05	0.05	20.0	20.0	25.1	25.1
0.05	0.1	19.9	10.1	22.5	17.3
0.05	0.15	19.9	6.91	21.1	13.4
0.05	0.2	19.9	5.30	20.4	11.3

r_1	r_2	$\lambda_1 = \lambda_2 = 0.3$		$\lambda_1 = \lambda_2 = 0.45$	
		W_1	W_2	W_1	W_2
0.05	0.2	19.9	5.30	20.4	11.3
0.1	0.2	10.0	5.60	14.3	12.4
0.15	0.2	7.01	5.67	13.3	12.9
0.2	0.2	5.59	5.59	12.5	12.5

表 2: 平均応答時間 : 到着率が異なる場合

r_1	r_2	$\lambda_1 = \lambda_2 = 0.4$		$\lambda_1 = 0.1, \lambda_2 = 0.7$	
		W_1	W_2	W_1	W_2
0.05	0.05	21.2	21.2	21.2	21.2
0.05	0.1	20.2	12.2	20.3	12.1
0.05	0.15	20.0	8.96	20.2	9.31
0.05	0.2	19.9	7.15	20.1	7.95

r_1	r_2	$\lambda_1 = \lambda_2 = 0.4$		$\lambda_1 = 0.1, \lambda_2 = 0.7$	
		W_1	W_2	W_1	W_2
0.05	0.2	19.9	7.15	20.1	7.95
0.1	0.2	11.0	7.92	11.0	7.92
0.15	0.2	8.81	7.97	8.70	7.85
0.2	0.2	7.76	7.76	7.77	7.77