Limited Processor Sharing Queue について

山崎源治（工学院大），逆瀬川浩孝（筑波大）

1. はじめに

待ち行列理論において，ジョブの到着過程，サービス時間の分布が一般の場合の複数サーバシステムは，解析が非常に困難な問題として残されてきている。単一サーバの場合と比べて，その困難性の主な理由はシステム内のジョブ数がサーバ数より少ないとき，いわゆる'手空き'のサーバが存在することにある。解析上の詳細な話は別として，システムの能力の面からこれを特徴づけると次のようになる。単一サーバシステムはジョブが1つ以上存在するときは，常にその能力をフルに使う。一方，複数サーバシステムでは，システム内にサーバ数以上のジョブが存在しているときのみその能力をフルに発揮するが，ジョブ数がサーバ数未満のときは，その一部のみを使う。このことから，ジョブの到着過程，サービス時間分布が同一である2つのシステム，一つは単一サーバシステム，他方は2-サーバシステム（個々のサーバによるサービスレイトは，単一サーバシステムのサーバのそれの1/2）を比較したとき，利用率は一致するが，"待ち"からみたとき（ジョブの列待ち時間ではなく，系内滞留時間など），単一サーバシステムの方が有効となることが期待できる。

これは，任意の複数サーバシステムについて，Stidham [3] によってサービス時間がアーランこ分布に従う場合（サービス時間の変動係数が1以下）に証明されている。この結果は，一見より一般的なサービス時間分布のもとでも成立しそうであるが，サービス時間の変動係数が1以上のときには反例すら提出されていて（Brumelle [1]），その成立範囲は未だに明らかではない。

このサービス時間のバラツキが比較的小さい場合の単一サーバの有効性は，上述のように，システムに1つでもジョブが存在するときその能力をフルに発揮できることに起因しているが，複数サーバシステムにおいて，"ジョブが1つでも存在するとき，手空きのサーバはいない"という規律のもとでジョブの処理を行う場合，果してその有効性が保たれるか，否か，という問題が当然生じてくる。より具体的に述べると，単一サーバシステムと，例えば2-サーバシステムを上述の条件のもとで比較する際，その2-サーバシステムが次のように作動するとき，単一サーバの有効性が保たれるか否か，の問題である。システムが空の状態でジョブが到着したときは2人のサーバでそのジョブを処理し（その際のサービスレイトは1人のときのそれの2倍），その処理中に新たに他のジョブが到着したなら，1人のサーバは直ちに新しく到着したジョブの処理を開始する。2つのジョブを処理中のとき新しいジョブが到着したなら，それらのジョブは順に待つ（FCFS）。もちろん，新しいジョブの到着以前に，2つのジョブのどちらかのサービスが完了したなら，直ちに他のジョブの処理が，2人のサーバにより行われる。

ここで，1つのジョブを2人のサーバで処理するとき，そのサービスレイトが1人のサーバのときの2倍であるという条件はかなり理想的で現実的ではないかも知れないが，いわゆる"手が空いたサーバは，他へ手伝いに行く"というシステムは，実際にも数多く見受けられる。2-サーバ以上のシステムで同様の規律の

もとでサービスを行う場合と単一サーバシステムの比較の問題は，計算機のCPUの多重度の面からみると，次のようにより自然に解釈できる。1つのCPUで，ジョブを1つずつ先着順に処理する場合が単一サーバシステムに相当し，あるジョブ数までは先着順にタイムシェアにより処理し，それをこえた場合は待ちが生じる場合が複数サーバシステムに相当するため，多重度を増すことが，“待ち”の面からみて有効であるか，否か？

上述の問題について，1つの答えを与えることが本論の目的である。そのため，§2では，上述の規律の一般的な定式化，§3，4では理論的考察，§5では数値的アプローチ，§6で得られた結果の直観的な解釈を与える。

2. モデル

本稿では，外部から到着したジョブに対して次のようにサービスする待ち行列システムを扱う。

システム内のスペースを適当に分割して前から順に positions 1, 2, 3, … と番号をつける。各 position には1度に1ジョブのみが入ることができる。システム内に n ジョブが存在するとき，(i) 到着したジョブは，position $n+1$ へ入る，(ii) position j ($j=1,2,\cdots,n$) のジョブのサービスが完了したなら，positions $j+1, j+2, \cdots, n$ のジョブは順に positions $j, j+1, \cdots, n-1$ へ移る，(iii) システムには1人だけサーバがいて，position j にその能力を，

$$n \leq m \ (m\text{は整数})\ \text{のとき，}\ \alpha_j\ (>0)\quad \sum_{j=1}^{n}\alpha_j = 1 \quad \text{ふりむけ，}$$

$$n > m\ \text{のとき，}\ \alpha_j\ (j=1,2,\cdots,m)\quad \sum_{j=1}^{m}\alpha_j = 1,\quad \alpha_j = 0\ (j=m+1, m+2, \cdots, n)$$

ふりむける（下では，α. は時刻に依存してもよいが，簡単化のためこれをしない）。

一見，(iii) の規律は抽象的であるが，α_jを適当に設定すると，次のような通常の待ち行列システムとなる。

<u>例・1</u>　$m=1$ のときは，FCFS の単一サーバシステム。

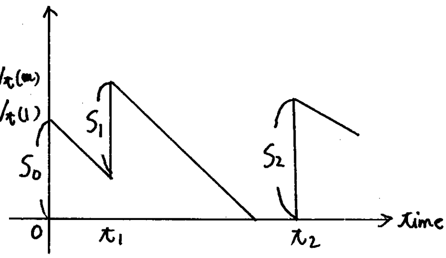
<u>例・2</u>　$n \leq m$ のとき，$\alpha_j = 1/n$ $(j=1,2,\cdots,n)$，$n > m$ のときは，$\alpha_j = 1/m$ $(j=1,2,\cdots,m)$；$\alpha_j = 0$ $(j=m+1,\cdots,n)$ とすると，FCFSの多重度 m のタイム・シェア処理のCPUとなり，特に $m=2$ のときこれは§1で述べた“手が空いたら手伝いに行く”2-サーバシステムとなる。

もちろん，例・2で α. を適当に設定することにより，$m \geq 3$ の場合の，手が空いたら手伝いに行く，という待ち行列システムに対応させることができる。

3. Work Load Process

ジョブのシステムへの相続く到着時点を $t_0\,(=0)$，t_1，t_2，…，これらのジョブのサービス時間を順に S_0，S_1，S_2，… としよう。また前節で述べたシステムで $m=1$ の場合をI-システムと呼び，その時刻 t における残りサービス時間の“総和”

(work load)を，$W_t(1)$で表わす。一方，$m(\geq 2)$の場合を，m-システムと呼び，そのwork loadを$W_t(m)$で表わす。このとき，$W_t(1)$は，単一サーバシステムにおけるvirtual waiting time processと一致することは明らかである。ここで，$t_0, t_1, t_2, \cdots ; S_0, S_1, S_2, \cdots$のある一連の実現値に対して，初期条件$W_{0-}(1)=W_{0-}(m)=0$のもとで$W_t(1)$, $W_t(m)$を図示すると，$W_t(m)$は前節の規律のもとでは$t_0, t_1, t_2, \cdots$で$S_0, S_1, S_2, \cdots$ジャンプし，連続的に45°のスロープで減るため，全く一致する(図・1)。このことと，1-システム，m-システムでの時刻tにおける系内ジョブ数を$L_t(1)$，$L_t(m)$としたとき，$L_t(1)=0$は$W_t(1)=0$に相当し，$L_t(m)=0$は$W_t(m)=0$に相当することから，次の結果を得ることができる。



〔図・1〕

<u>定理1</u>　$W_{0-}(1)=W_{0-}(m)$のとき，

$$W(1) \stackrel{\text{st}}{=} W_t(m),\quad Prob(L_t(1)=0) = Prob(L_t(m)=0),$$

ここで，$\stackrel{\text{st}}{=}$は分布の一致を意味する。

ちなみに，平均到着間隔をλ^{-1}，平均サービス時間をμ^{-1}で表わしたとき，もし$\rho \equiv \lambda/\mu < 1$で$\lim_{t\to\infty} Prob(L_t(1)=0)$が存在するならば，

$$\lim_{t\to\infty} Prob(L_t(m)=0) = 1-\rho$$

となる。本節では，到着間隔，サービス時間に何んら条件をつけていないことを付記する。

4. アーラン サービス

本節では，ジョブのサービス時間がi.i.d. r.v.'sで，それらがフェーズkのアーラン分布(E_k)に従う場合を扱うが，ジョブのシステムへの到着間隔はi.i.d. r.v.'sである必要はなく，次の仮定のみを満たすものとする。

<u>A-1</u>：到着過程は，サービス過程とは独立で，有限時間区間内に到着するジョブ数は，確率1で有限である。

このとき，$L_t(1)$と$L_t(m)$を比較することが本節の目的であるが，便宜上，$W_{0-}(1)=W_{0-}(m)=0$としてそれを行う。

次のような2つの待ち行列システムを想定する。到着の発生源は1つであるが，そこから1-システムへジョブが入る場合と，m-システムへ入る場合である。今，$t_0, t_1, t_2, \cdots, t_{n+1}$が与えられたとし，時刻$t$ ($t_n \leq t < t_{n+1}$)における1-システムとm-システムの存在するジョブの残りサービスフェーズの"総和"を，それぞれ$N_t(1; t_0, t_1, \cdots, t_{n+1})$，$N_t(m; t_0, t_1, \cdots, t_{n+1})$で表わす。同じ到着時点が与えられたとしていることから，tまでに両システムへ到着する仕事の総量(フェーズ数の総和)が等しいことは明らかである。1-システムにおいて，区間$[u, u+\Delta u)$ $(0 \leq u < t)$，

$\Delta u > 0$, におけるフェーズの移動（フェーズ数が減る）確率は，もしシステム内に1つでもジョブが存在するときは，

$$k\mu \Delta u + o(\Delta u)$$

で与えられる。一方，m-システムにおけるそれは，

$$\sum_{j=1}^{L_u(m)} \alpha_j k\mu\Delta + o(\Delta u) = k\mu + o(\Delta\mu)$$

となる。従って，次の補助定理から，ジョブが1つでも存在する限り，フェーズ数の減る確率則は両システムで一致する。

補助定理（Cox [2]）：Xを非負のr.v.，$P(x) = \mathrm{Prob}(X > x)$　$(x \geq 0)$　とする。このとき，$P(0)=1$で，かつ　$\mathrm{Prob}(X \leq x+\Delta x \mid X > x) = \eta(x)\Delta x + o(\Delta x)$　$x, \Delta x \geq 0$ なる $\eta(x)$ が存在するなら

$$P(x) = \exp\left\{-\int_0^x \eta(u)du\right\}.$$

一方，§3の議論により，$L_u(1)=0$なる事象と$L_u(m)=0$なる事象は確率的に等価である。従って，次の結果が成立する。

$$N_\tau(1; t_0, t_1, \cdots, t_{n+1}) \overset{st}{=} N_\tau(m; t_0, t_1, \cdots, t_{n+1})$$

ここで，[A-1]のもとで到着に関する条件をはずすことにより，次の結果を得る。

定理2：[A-1]のもとで，$N_\tau(1)$, $N_\tau(m)$をそれぞれ，1-システム，m-システムにおける時刻τにおける残りフェーズ数の総和とする。このとき，

$$N_\tau(1) \overset{st}{=} N_\tau(m).$$

この定理を用いて，$L_\tau(1)$と$L_\tau(m)$の比較を試みる。システム内にnジョブいるときの残りフェーズ数の総和を，1-システムについては$N_n(1)$，m-システムについては$N_n(m)$で表わすと，

$$(n-1)k+1 \leq N_n(1) \leq nk,$$

$$\begin{cases} n \leq N_n(m) \leq nk & \text{for } n \leq m \\ (n-m)k+m \leq N_n(m) \leq nk & \text{for } n > m, \end{cases}$$

となる。上の不等式を直接的に比較することにより，“1-システムとm-システムで残りフェーズ数の総和が等しいとき，システム内ジョブ数はm-システムの方が多くなる可能性がある”，という結果を得ることができる。これと，上の定理から直ちに次の結果を得る。

系： $W_{0-}(1) = W_{0-}(m) = 0$，[A-1] のもとでは，

$$L_t(1) \overset{st}{\leq} L_t(m) \qquad \text{for all } t\ (\geq 0),$$

ここで，$\overset{st}{\leq}$ は確率的大小を意味している。特に，$k=1$ のときは，等号となる。

この系は，本稿で扱っているサービス・システムでも，通常の複数サーバシステムと単一サーバシステムを比較した場合の単一サーバの有効性（[3]）が保たれることを意味している（ただし，アーランこ・サービスのとき）。また，アーランこサービスでは，$L_t(m-1) \overset{st}{\leq} L_t(m)$ も成立しそうであるが，残念ながら，ここで用いた証明法ではこの結果を得ることができなかった。これは，次節で数値的に明らかにされる。さらに，$C_s^2 > 1$ のときは，最早上の系は成立せず，逆の順序関係となることも，次節で明らかになる。

5. 数 値 例

前節では，アーランこサービスの場合の単一サーバの有効性を定性的に明らかにしたが，定量的特性，$L_t(m-1)$ と $L_t(m)$ の比較および $C_s^2 > 1$ のときの特性を明らかにすることはできなかった。本節では，これらを数値的に検討する。以下では，ジョブのシステムへの到着過程は，Poisson process あるとする。

$m=2$，サービス時間の分布が E_2 のとき，§2の例・2の α. のもとでの定常状態における平均系内ジョブ数（サービス中のジョブも含めて）：$L(2)$ は，平衡方程式から母関数を導く通常の方法を経て計算することができ，その結果は次のようになる。

$$L(2) = \frac{\rho(4-\rho^2)}{4(1-\rho)}.$$

一方，$m=1$ のときのそれ（$L(1)$）は，$M/E_2/1$ のよく知られている結果から，

$$L(1) = \frac{\rho(4-\rho)}{4(1-\rho)}.$$

もちろん，$L(2)/L(1) < 1$ であるが，この比は $\rho = 0.535\cdots$ で最小となる。すなわち，“手空きを許さない”という規律の効果は，ρ が 0.5 付近で顕著となる。

サービス時間が E_2 でも，$m \geq 3$ となると母関数の方法は使えず，m が増えるとかなり面倒な数値計算を必要とし，E_3 になるとさらにやっかいになるが，数値解法，および詳細な数値例は，ここでは省略する（当日，示す）。その結果のみを要約すると，サービス時間が E_2，E_3 に従う場合は，$\rho=0.5$ 付近の上述の特性は $m \geq 3$ でも同様であり，

$$\{L(2)/L(1)\}_{E_2} < \{L(2)/L(1)\}_{E_3} \quad \text{および} \quad L(2)/L(1) < L(3)/L(1) < \cdots,$$

となる。

$C_s^2>1$ の場合の特性をみるため，サービス時間の密度関数：$f(t)$ が次の場合（超指数分布）を考える。

$$f(t) = 2\alpha^2\mu e^{-2\alpha\mu t} + 2(1-\alpha)^2\mu e^{-2(1-\alpha)\mu t} \qquad 0<\alpha<1/2 .$$

$m=2$ で，§2の例2の規律のもとでの $L(2)$ は母関数から導くことができ，その結果の1例を［表・1］に示す。

この表から明らかなように，$C_s^2>1$ のときは，$L(2)/L(1)<1$（1-システムより m-システムの方が効果的）で，C_s^2 が大きい程この比は小さくなる。また，やはり $\rho=0.5$ 付近で，"手空きを許さない"という規律の効果は著しくなる。$m\geq3$ の場合は母関数による方法は難しく，直接的な数値解法に頼らざるを得ない。その詳細は省略し，結果のみを要約すると，次のようになる。$\rho=0.5$ 付近の特性は $m\geq3$ でも同じで，

	$L(2)/L(1)$	
	$\rho=0.5$	0.9
$C_s^2=1.5$ ($\alpha=0.28$)	0.940	0.981
$C_s^2=3$ ($\alpha=0.15$)	0.816	0.951

［表・1］

$$L(2)/L(1) > L(3)/L(1) > \cdots,$$

となる。すなわち，サービス時間のバラツキが大きければ大きい程，m（多重度）を増すことは，"待ち"の面からは有効となる。

6. おわりに

§4で用いた証明法は，アーラン分布の特性に強く依存していて，より一般的なサービス分布に対しては適用することは難しいように思われる。また，§5の結果は，数値計算上のため特定の分布に限定せざるを得なかった。しかしながら，§4,5で得られた結果は，次のように解釈すると合理的であるように思われる。

Fを確率分布関数として，次の分布関数のクラスを導入する。

$$NBUE = \left\{ F \,;\, F(0)=0,\ F(t)<1 \text{ なるすべての } t \text{ について } \int_0^\infty x\,dF(x) \geq \int_0^\infty \frac{1-F(x+t)}{1-F(t)}dx \right\}$$

いま，ジョブのサービス時間の分布がNBUEであるとする（$C_s^2\leq1$）。このとき，ある時点で新しいジョブのサービスを開始したときの平均サービス時間は，その時点までサービスを受けていたジョブの残り平均サービス時間よりも大きい ⇒ ジョブを1つずつ処理していく方が能率的である，ということになる。従って，$L_t(1) \overset{st}{\leq} L_t(m)$ が成立する。もちろん，$E_k \subset NBUE$ である。

一方，サービス時間の分布が，NBUEの条件の不等号を逆にした分布のクラス（NWUE）に属するときは（$C_s^2\geq1$），逆の議論が成立し，古いジョブはできるだけ放っておいて，新しいジョブが到着したとき直ちにそのサービスに着手した方が能率的 ⇒ 結果として $L_t(1) \overset{st}{\geq} L_t(m)$ が成立する。もちろん，前節で扱った超指数分布はNWUEに属する。

以上のことから，§4,5で得られた特性は，サービス時間の分布がNBUE（NWUE）に属する場合，一般的に成立することが期待される。さらに，§5では，ジョブのシステムへの到着は Poisson process としているが，§4（あるいは[3]）

の議論のように，1-システムとm-システムの能率の比較では，同じ到着過程のもとで考えられるため，得られる特性はサービス分布のみに依存し，到着過程の特性には影響されないことを付記する。

＜参　考　文　献＞

1. Brumelle, S. L., "Some inequalities for parallel service queues", Opns. Res., 19, 402-413 (1971).
2. Cox, D. R., "Renewal theory", Wiley, New York, (1962).
3. Stidham, S. Jr., "On the optimality of single-server queueing systems", Opns. Res., 18, 708-732(1970).