複数の到着流をもつ単一サーバ待ち行列

滝根　哲哉

1　はじめに

待ち行列モデルはサーバと呼ばれる共有資源に対して多数の利用者（客）が確率的に利用要求を発生するという状況を記述するものです．昨年のＯＲ誌の特集 [2] でも紹介されているように，その応用分野は様々ですが，中でも，通信網の性能評価は最も自然な応用分野といえます．実際，電話網の設計問題を解決するための枠組として，20 世紀初頭に研究が開始されており，通信網への応用を意識したとき，待ち行列理論は通信トラヒック理論とも呼ばれています．

本稿の表題である複数の到着流をもつ単一サーバ待ち行列も，やはり，通信網への応用から自然発生的に現れたモデルです．近年の通信網の高速化に伴い，従来，個別網に収容されていた多様なメディア（音声，動画，データなど）を単一の通信網に収容する統合サービス網の技術開発が進められています．統合サービス網に収容されるトラヒックは，メディアあるいはコンテンツによって，それぞれ固有の特徴を持っているため，通信網内に設置された交換機では，性質の異なるトラヒックが多重化されて伝送されることになります．これを待ち行列モデルで記述すれば，メディアあるいはコンテンツに対応する異なる性質をもつ客（伝送要求）が共有資源であるサーバ（回線あるいは交換機内のバッファ）の利用を競い合うことになります．

このような技術動向に刺激され，複数の到着流をもつ待ち行列の研究がこの十数年来，精力的に行われてきました [13]．一般に，待ち行列モデルは客の到着（発生）パターンと到着した客が要求するサービスに必要な時間の確率的性質を定めることで得られます．本稿で紹介する複数の到着流をもつ待ち行列モデルは (1) 各到着流は異なる確率過程に従い，かつ (2) 到着流毎にサービス要求時間の確率分布が異なるようなものです．後述するように，この種の待ち行列モデル，中でも最も基本的な先着順サービスモデルは，ごく最近まで極めて限られたモデル以外は解析的な扱いが困難なものと思われていました．以下では，筆者がこのようなモデルを考察するに至った経緯から，最終的に解析手法を確立するまでを，着眼点を中心に紹介したいと思います．

2　研究の動機

待ち行列モデルにおける最も基本的な客の到着過程はポワソン過程です．これは客がランダムに到着すると仮定することと等価です[1]．しかし，統合サービス網に対する応用では，必ずしも伝送単位の発生がポワソン過程に従うとは仮定できず，より広い範囲の到着過程を記述できるモデルが必要です [13]．筆者が非ポワソン到着流をもつ待ち行列の研究を始めた 1990 年以前にも，既に幾つかの（解析的取り扱いが可能な）非ポワソン到着過程が提案されていましたが，それらはいずれも記法の面で（筆者にとっては）繁雑に思えました．それゆえ，当時，このような非ポワソン到着過程の枠組で，既知の結果を実質的に発展させることができるとは必ずしも思っていませんでした．今から思えば，それは記法の繁雑さが対象となる問題の構造の見通しを悪くしていたのだと理解できます．

好運にも，筆者が研究を始めたちょうどその頃，マルコフ型到着過程（MAP）[3] と呼ばれる到着過程モデルが提案されました．MAP には，独立な MAP の重畳が MAP で表現できることや，あらゆる定常単純点過程を任意の精度で近似できるなどの利点があります．後に MAP はそれ以前に提案されていた他の非ポワソン到着流と数学的には等価なことが分かるのですが，待ち行列の解析という観点から最も簡潔な表現をもつため，急速に浸透しました．筆者も，この MAP なら何か新しいことができるのではないかと思って勉強を始めました．

到着過程が MAP に従い，サービス時間が独立同一分布に従う単一サーバ待ち行列（MAP/G/1）の客数過程は M/G/1 型と呼ばれる構造を持った離散時間マルコフ連鎖に帰着でき，通常，アルゴリズム的解法の一種である行列解析法 [4] によって解析されます．このアルゴリズム的解法は，昨年，OR 誌で企画された「待ち行列研究の新しい潮流」[6, 7] でも紹介されてい

たきね　てつや　京都大学大学院情報学研究科数理工学専攻
〒 606-8501 京都市左京区吉田本町
E-mail: takine@kuamp.kyoto-u.ac.jp

[1] ポワソン過程の平易な解説が [5] にあります．

るように，系内客数に対応する離散確率変数とシステムの状態変化を記述するために必要な情報を保持する補助的な離散確率変数の組からなる 2 変数離散状態マルコフ連鎖に対して，その構造を利用して定常解を求める数値アルゴリズムと見ることができます．詳細は [19] を参照して下さい．

当時，既に行列解析法の基本的枠組は確立されていましたので，筆者は，統合サービス網への応用という観点から基本的なモデルであり，かつ，従来の行列解析法では解析できないモデルとして，到着流間に非割込み優先権がある待ち行列モデルを研究の対象に据えました．優先権がある場合，系内客数の挙動を表現するためには，各到着流毎の客数を保持する必要があります．その結果，客数に対応する無限個の状態をもつ確率変数が到着流の数だけ必要となり，無限個の状態をもつ確率変数を 1 つしか許さない従来の行列解析法の枠組では解析できないわけです．

各到着流が MAP に従う非割込み優先権付き待ち行列の最も単純なモデルは，全ての客のサービス時間が独立同一分布に従う場合です．筆者は，まず，このモデルの研究から始めました．簡単な考察の後，このモデルは，2 つの MAP 到着流をもつ非割込み優先権付き待ち行列の結果を用いれば解析できること，また，この 2 つの到着流をもつモデルは，全ての到着流を 1 つにまとめた到着流をもつ先着順モデルの結果を用いれば解析可能なことが分かりました．全ての客のサービス時間が独立同一分布に従う場合，全ての到着流を 1 つにまとめた到着流をもつ先着順モデルは通常の MAP/G/1 モデルになり，これは従来の行列解析法で解析できます．結果として，サービス時間が独立同一分布に従う場合の非割込み優先権モデルは既知の結果を応用すれば解析できたわけです [10]．

この研究結果は理論的には余り面白いところはありませんが，固定長の伝送単位をもつ ATM 網[2]に直接応用できる利点があります．当時，筆者は米国に滞在しており，帰国前に米国，カナダの大学，研究所を回ってこの結果ならびに離散時間モデルに関する同様の結果の講演をしていました．AT&T Bell 研究所での講演の後，聞きに来てくれた人の中に，もうやることがなくなったと言っていた人がいると聞き，自分が選んだ研究の方向が間違っていなかったと胸を撫で下ろしたことが思い出されます．

さて，サービス時間分布が到着流毎に異なる場合はどうすれば良いのでしょうか．サービス時間が独立同一分布に従う場合の考察から，全ての到着流をまとめて 1 つの到着流とみなし，それを先着順でサービスするモデルの解析結果が必要なことは明らかでした．これが今回紹介する研究の動機となりました．

[2] ATM 網では，伝送される全ての情報はセルと呼ばれる固定長の小さな伝送単位（ヘッダ 5 バイト，情報部 48 バイト）に分割され，通信網へ送り出されます．

3　対象となるモデルと複数の到着流の記述

最初に，以下で対象となるモデルの最も単純な例を示します．

例　2 つの到着流を収容する単一サーバ待ち行列

到着流 1 は到着間隔が λ_1^{-1} のポワソン到着，サービス時間分布関数 $h_1(x)$ をもつ．一方，到着流 2 は到着間隔が平均 λ_2^{-1} の 2 ステージアーラン分布[3]に従い，サービス時間分布関数 $h_2(x)$ をもつ．客は先着順サービス規律に従ってサービスされ，待ち部屋の容量は十分に大きく，到着した客は全てシステム内に入ることができる．

この待ち行列における到着過程を図で表現すると以下のようになります．

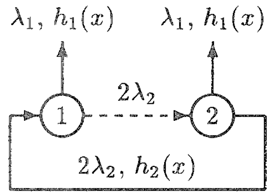

図 1: ポワソン到着と 2 ステージアーラン到着の重畳

一般に，先着順サービス単一サーバ待ち行列において，n 番目の客の離脱直後の客数を Q_n，n 番目の客のサービス時間内に到着する客数を A_n とすると

$$Q_{n+1} = \max(Q_n - 1, 0) + A_{n+1}$$

が成立します．$h_1(x) = h_2(x)$ の場合，Q_n の値とサービスの終了時点でアーラン分布がいずれのステージにあるかを表す確率変数 S_n の値が与えられると A_n の分布が定まります．言い替えますと確率変数の組 (Q_n, S_n) はマルコフ連鎖になります．特にこのマルコフ連鎖は M/G/1 型と呼ばれる構造をもっており，前述の行列解析法で解析できます．

しかし，$h_1(x) \neq h_2(x)$ ならば，客の到着とサービス時間分布の間に相関が生じるため，総客数 Q_n と S_n の組はマルコフ連鎖を構成しません．実際，客数の挙動を表すマルコフ連鎖を構築しようとすれば，サービス中ならびに待っている個々の客がそれぞれ，どちらの到着流から来たかという情報を全て保持する必要があります[4]．

この例からも分かるように，サービス時間分布の異なる複数の非ポワソン到着流を収容する単一サーバ

[3] 平均 $(n\lambda)^{-1}$ の独立な指数分布の n 個の和が従う分布を平均 λ^{-1} の n ステージアーラン分布といいます．

[4] サービス時間が最も扱いやすい（平均の異なる）指数分布に従う場合でも同様です．

待ち行列では，客数の挙動を表現するマルコフ連鎖は極めて複雑な構造をもつため，そのような定式化による解析は極めて困難です．最初に少し触れましたが，待ち行列モデルの解析において，モデルに内在する確率過程（本稿の場合，到着流）をどのような記法で表現するかは，モデルの構造の見通しを良くする上で極めて重要です．簡潔な記法を用いれば，一見，複雑に見えるモデルに対しても何か糸口を得られる可能性があります．

具体的に話を進めるため，ここで前述の MAP の定義を与えます．MAP は有限状態をもつ連続時間マルコフ連鎖を考え，状態遷移が起こったとき，予め決められた確率で客が到着するというものです．具体的には以下のように定義されます [3]．まず，状態集合 $\mathcal{M}=\{1,\ldots,M\}$ をもつ既約な有限状態連続時間マルコフ連鎖 $\{S(t)\}$ を考えます．各状態 $i\in\mathcal{M}$ での滞在時間は平均 η_i^{-1} の指数分布に従います．状態 i での滞在終了後，確率 $p_{i,j}$ $(j\in\mathcal{M}, j\neq i)$ で到着を伴わず状態 j へ遷移します．また，確率 $q_{i,j}$ $(j\in\mathcal{M})$ で 1 人の客の到着を伴って状態 j へ遷移します．ただし

$$\sum_{\substack{j\in\mathcal{M}\\ j\neq i}} p_{i,j}+\sum_{j\in\mathcal{M}} q_{i,j}=1$$

です．このとき，(i,j) 要素 $C_{i,j}, D_{i,j}$ がそれぞれ

$$C_{i,j}=\eta_i p_{i,j}\quad (j\neq i),\quad C_{i,i}=-\eta_i,\quad D_{i,j}=\eta_i q_{i,j}$$

で与えられる $M\times M$ 行列 $\boldsymbol{C},\boldsymbol{D}$ を用いて，この到着過程を表現 $(\boldsymbol{C},\boldsymbol{D})$ をもつ MAP といいます．定義より $\boldsymbol{D}$ の (i,j) 要素は到着を伴う状態 i から状態 j への遷移率を表し，$\boldsymbol{C}$ の非対角 (i,j) 要素は到着を伴わない状態 i から状態 j への遷移率を表します．

例えば率 λ_1 のポワソン過程は到着間隔が平均 λ_1^{-1} の指数分布に従うので $\boldsymbol{C}=-\lambda_1$，$\boldsymbol{D}=\lambda_1$ となります．また，平均 λ_2^{-1} の 2 ステージアーラン分布は

$$\boldsymbol{C}=\begin{pmatrix}-2\lambda_2 & 2\lambda_2\\ 0 & -2\lambda_2\end{pmatrix},\quad \boldsymbol{D}=\begin{pmatrix}0 & 0\\ 2\lambda_2 & 0\end{pmatrix}$$

となります．以下では MAP を支配しているマルコフ連鎖の状態を相と呼びます．

さて，複数の独立な MAP を重畳したとき，どのように記述できるでしょうか．例として，表現 $(\widetilde{\boldsymbol{C}}_1,\widetilde{\boldsymbol{D}}_1)$ ならびに $(\widetilde{\boldsymbol{C}}_2,\widetilde{\boldsymbol{D}}_2)$ をもつ独立な 2 つの MAP の重畳を考えます．それぞれの MAP の相がとる状態集合を $\mathcal{M}_1$，$\mathcal{M}_2$ とすると，重畳後の到着過程において相がとる状態集合 $\mathcal{M}$ はこれらの直積集合 $\mathcal{M}_1\times\mathcal{M}_2$ となります．すなわち

$$\boldsymbol{C}=\widetilde{\boldsymbol{C}}_1\otimes\boldsymbol{I}_2+\boldsymbol{I}_1\otimes\widetilde{\boldsymbol{C}}_2,\quad \boldsymbol{D}=\widetilde{\boldsymbol{D}}_1\otimes\boldsymbol{I}_2+\boldsymbol{I}_1\otimes\widetilde{\boldsymbol{D}}_2$$

としたとき，重畳された到着過程は表現 $(\boldsymbol{C},\boldsymbol{D})$ をもつ MAP となります．ここで $\otimes$ はクロネッカ積，$\boldsymbol{I}_i$ は $\mathcal{M}_i$ の要素数と同じ次元の単位行列です．特に $\boldsymbol{D}$ を構成している 2 つの要素はそれぞれの到着流からの到着を表現していることに注意して下さい．例えば前述のポワソン到着と 2 ステージアーラン到着の重畳は

$$\boldsymbol{C}=\begin{pmatrix}-\lambda_1-2\lambda_2 & 2\lambda_2\\ 0 & -\lambda_1-2\lambda_2\end{pmatrix},\ \boldsymbol{D}=\begin{pmatrix}\lambda_1 & 0\\ 2\lambda_2 & \lambda_1\end{pmatrix}$$

で記述されます．

上記の例では，重畳された到着流を 1 つの到着流として見ていますが，複数の到着流を明示的に記述するためにマーク付き MAP が提案されています [1]．これは，到着を伴う遷移率を表現する行列 $\boldsymbol{D}$ を各到着流毎に個別に用意するというものです．一般に，マーク付き MAP を用いて P 個の到着流を記述する際，MAP にならって $(\boldsymbol{C},\boldsymbol{D}_1,\ldots,\boldsymbol{D}_P)$ で表現されます．ここで $\boldsymbol{D}_k$ $(k=1,\ldots,P)$ の (i,j) 要素は，k 番目の到着流からの客が MAP の相の i から j への遷移に伴って到着する率を与えています．行列 $\boldsymbol{C}$ の定義は MAP の場合と同じです．例えば 2 つの MAP の重畳の場合，$\boldsymbol{D}_1=\widetilde{\boldsymbol{D}}_1\otimes\boldsymbol{I}_2$，$\boldsymbol{D}_2=\boldsymbol{I}_1\otimes\widetilde{\boldsymbol{D}}_2$ となります．また，上述の例において，ポワソン流を到着流 1，2 ステージアーラン流を到着流 2 としたとき，以下のように記述されます．

$$\boldsymbol{D}_1=\begin{pmatrix}\lambda_1 & 0\\ 0 & \lambda_1\end{pmatrix},\quad \boldsymbol{D}_2=\begin{pmatrix}0 & 0\\ 2\lambda_2 & 0\end{pmatrix}$$

マーク付き MAP の表現はサービス時間に関する情報を含みません．前述の例で示したような，各到着流がそれぞれ異なるサービス時間分布をもつ場合，それらを明示的に表し，かつ，簡潔な（例えば図 1 と等価なものを表現する）記法が必要となります．そこでマーク付き MAP を拡張し，$(\boldsymbol{C}, \boldsymbol{D}_1(x), \ldots, \boldsymbol{D}_P(x))$ という形の表現を導入します．ここで $\boldsymbol{D}_k(x)$ $(k=1,\ldots,P)$ の (i,j) 要素は，サービス時間が x 以下であるような k 番目の到着流からの客が，MAP の相の i から j への遷移に伴って到着する率を与えています．行列 $\boldsymbol{C}$ の定義はマーク付き MAP の場合と同じです．例えば図 1 の到着流は以下のように表現されます．

$$\boldsymbol{D}_1(x)=\begin{pmatrix}\lambda_1 h_1(x) & 0\\ 0 & \lambda_1 h_1(x)\end{pmatrix},\ \boldsymbol{D}_2(x)=\begin{pmatrix}0 & 0\\ 2\lambda_2 h_2(x) & 0\end{pmatrix}$$

$\boldsymbol{D}_k=\boldsymbol{D}_k(\infty)$ $(k=1,\ldots,P)$ に注意して下さい．

4 系内仕事量分布と待ち時間

本章以降は，表現 $(\boldsymbol{C},\boldsymbol{D}_1(x),\ldots,\boldsymbol{D}_P(x))$ をもつ到着流を収容する先着順サービス単一サーバ待ち行列の解析を考えます．まず，この待ち行列における MAP の相の定常確率分布 $\boldsymbol{\pi}$ は次式で決定されます．

$$\boldsymbol{\pi}\left(\boldsymbol{C}+\sum_{k=1}^{P}\boldsymbol{D}_k\right)=\boldsymbol{0},\qquad \boldsymbol{\pi}\boldsymbol{e}=1$$

ここで e は全ての要素が1の列ベクトルです．さらに π を用いて，この待ち行列の利用率 ρ は

$$\rho = \pi \int_0^\infty x d\boldsymbol{D}(x) e$$

で与えられます．ただし

$$\boldsymbol{D}(x) = \boldsymbol{D}_1(x) + \cdots + \boldsymbol{D}_P(x) \quad (1)$$

です．以下では $\rho < 1$ を仮定します[5]．

前述したように，客数の解析は簡単そうではありません．では，待ち時間はどうでしょうか．先着順サービスの場合の待ち時間は，客が到着した時点における系内仕事量で与えられます．この系内仕事量とは，その客の到着時点以降，到着がないと仮定したとき，システムが空になるまでの時間で与えられる量です．

系内仕事量は以下のようなものを想像すれば理解しやすいかもしれません．待ち行列では通常，客が行列を作ってサービスを待つと捉えられますが，客が待つ代わりに，客はシステムに仕事を運んで来ると考えてみます．例えば，客が到着した時，その客は書類の束をサーバの前に積み上げるとします．サーバは書類がある限り，一定のペースで書類を処理していくとします．すなわち，客が持ち込んだ書類を処理するのに必要な時間がサービス時間に対応します．また，先着順でサービスされる場合，ある到着客が持ち込んだ書類の処理が始まるまでの時間（待ち時間に対応）は，到着時点において，サーバの前に積まれている未処理の書類を全て処理し終えるまでの時間で与えられます．この未処理の書類を全て処理し終えるまでの時間を系内仕事量といいます．

書類の例からも明らかなように，もし，到着した客がサービス終了まで退去することがなく，かつ仕事がある限りサーバは稼働し続けるとすると，系内仕事量はサービスの順序とは無関係な量となります[6]．このような待ち行列を仕事量保存型といいます．さらに仕事量保存型の場合，時刻 t における系内仕事量 V_t は $V_t > 0$ のとき率1で減少し，客の到着があると，その客のサービス時間分だけ上にジャンプします．

さて，この系内仕事量に関しては，率 λ のポワソン到着，分布関数 $h(x)$ に従う独立同一なサービス時間をもつ通常の M/G/1 の場合，時間依存分布関数 $v(t,x) = \Pr(V_t \le x)$ に対して，次のような結果が古くから知られていました [8]．まず，微小時間 δt での遷移を考えることにより

$$v(t+\delta t, x) = (1-\lambda\delta t) v(t, x+\delta t) + \lambda\delta t \int_{0-}^x v(t, x-y) dh(y) + o(\delta t)$$

が成立し，これより

$$\frac{\partial}{\partial t} v(t,x) = \frac{\partial}{\partial x} v(t,x) - \lambda v(t,x) + \lambda \int_{0-}^x v(t,x-y) dh(y)$$

を得ます．これは Takács の積分微分方程式と呼ばれています．特に，定常状態における系内仕事量 V の分布関数 $v(x) = \Pr(V \le x)$ は $\partial v(t,x)/\partial t = 0$ より

$$\frac{d}{dx} v(x) = \lambda v(x) - \lambda \int_{0-}^x v(x-y) dh(y)$$

となり，両辺のラプラス・スティルチェス変換（LST）を取ることにより，系内仕事量分布の LST $v^*(s)$ は

$$v^*(s)(s - \lambda + \lambda h^*(s)) = v(0)s \quad (2)$$

を満たすことがわかります（$h^*(s)$ は $h(x)$ の LST）．未知数 $v(0)$ は系内仕事量が0，すなわちシステムが空である確率であり，これは平均サービス時間を h とすると $v(0) = 1 - \lambda h$ で与えられます．

そこで，この議論を表現 $(\boldsymbol{C}, \boldsymbol{D}_1(x), \ldots, \boldsymbol{D}_P(x))$ をもつ到着流[7]を収容する先着順サービス単一サーバ待ち行列の場合に拡張します [9]．時刻 t における MAP の相を S_t で表し，j 番目の要素が $Pr(V_t \le x, S_t = j)$ で与えられる行ベクトル $\boldsymbol{v}(t,x)$ を考えると，M/G/1 と同様の議論により次式が成立します．

$$\boldsymbol{v}(t+\delta t, x) = (\boldsymbol{I} + \boldsymbol{C}\delta t)\boldsymbol{v}(t, x+\delta t) + \int_{0-}^x \boldsymbol{v}(t, x-y) d\boldsymbol{D}(y)\delta t + o(\delta t)$$

ここで $\boldsymbol{I}$ は単位行列です．これより

$$\frac{\partial}{\partial t} \boldsymbol{v}(t,x) = \frac{\partial}{\partial x} \boldsymbol{v}(t,x) + \boldsymbol{v}(t,x)\boldsymbol{C} + \int_{0-}^x \boldsymbol{v}(t,x-y) d\boldsymbol{D}(y)$$

を得ます．特に，定常状態における系内仕事量を V，MAP の相を S としたとき，j 番目の要素が $\Pr(V \le x, S = j)$ で与えられる行ベクトル $\boldsymbol{v}(x)$ は

$$\frac{d}{dx} \boldsymbol{v}(x) = \boldsymbol{v}(x)\boldsymbol{C} - \int_{0-}^x \boldsymbol{v}(x-y) d\boldsymbol{D}(y)$$

を満たし，両辺の LST を取ることにより，系内仕事量分布のベクトル LST $\boldsymbol{v}^*(s)$ は

$$\boldsymbol{v}^*(s)(s\boldsymbol{I} + \boldsymbol{C} + \boldsymbol{D}^*(s)) = \boldsymbol{v}(0)s$$

を満たします（$\boldsymbol{D}^*(s)$ は $\boldsymbol{D}(x)$ の LST）．よって j 番目の要素が $\Pr(V = 0, S = j)$ で与えられる未知ベクトル $\boldsymbol{v}(0)$ を求めれば良いことが分かります．これは以下のように導出されます．

時刻0で系内仕事量が x であり，かつ，MAP の相が i であるという条件の下で，これ以降初めてシステムが空になった時点での MAP の相が j である確率

[5] これにより定常状態の存在が保証されます．

[6] どの仕事から手をつけても，全て処理するまでに必要な時間は同じなので，優先権がある場合でも先着順でサービスされる場合でも系内仕事量の分布は同じです．

[7] 系内仕事量を考える際には，表現 $(\boldsymbol{C}, \boldsymbol{D}(x))$ をもつ到着流と等価です（(1) 参照）．

を (i,j) 要素にもつ行列 $\boldsymbol{P}(x)$ を考えます．定義より $o(\delta x)$ の項を無視すると

$$\boldsymbol{P}(x+\delta x)=\boldsymbol{P}(x)\left[\boldsymbol{I}+\boldsymbol{C}\delta x+\int_0^\infty d\boldsymbol{D}(y)\delta x\boldsymbol{P}(y)\right]$$

が成立し，これより

$$\frac{d}{dx}\boldsymbol{P}(x)=\boldsymbol{P}(x)\left[\boldsymbol{C}+\int_0^\infty d\boldsymbol{D}(y)\boldsymbol{P}(y)\right]$$

を得ます．ここで

$$\boldsymbol{Q}=\boldsymbol{C}+\int_0^\infty d\boldsymbol{D}(x)\exp(\boldsymbol{Q}x)$$

とおくと $d\boldsymbol{P}(x)/dx=\boldsymbol{P}(x)\boldsymbol{Q}$，すなわち $\boldsymbol{P}(x)=\exp(\boldsymbol{Q}x)$ を得ます．行列 $\boldsymbol{Q}$ はシステムが空である時間区間のみに注目した場合のMAPの相の遷移率行列を表しています．

この $\boldsymbol{Q}$ を用いて $\boldsymbol{v}(0)$ を求めることができます．システムが空である確率は $1-\rho$ ですので $\boldsymbol{v}(0)\boldsymbol{e}=1-\rho$ となります．よって $\boldsymbol{\kappa Q}=0$ かつ $\boldsymbol{\kappa e}=1$ を満たす行ベクトル $\boldsymbol{\kappa}$ を用いて $\boldsymbol{v}(0)=(1-\rho)\boldsymbol{\kappa}$ となります．以上の考察より，最終的な結果

$$\boldsymbol{v}^*(s)[s\boldsymbol{I}+\boldsymbol{C}+\boldsymbol{D}^*(s)]=(1-\rho)s\boldsymbol{\kappa} \tag{3}$$

を得ました．この両辺を微分することにより系内仕事量の積率を計算することができます．特にサービス時間分布が独立同一な分布関数 $h(x)$（LST $h^*(s)$）をもつ MAP/G/1 の場合，

$$\boldsymbol{v}^*(s)[s\boldsymbol{I}+\boldsymbol{C}+h^*(s)\boldsymbol{D}]=(1-\rho)s\boldsymbol{\kappa} \tag{4}$$

となります．(2), (3), (4) が極めて似た形になっていることに注意して下さい．この意味で，表現 $(\boldsymbol{C}, \boldsymbol{D}_1(x),\ldots,\boldsymbol{D}_P(x))$ をもつ到着流を収容する単一サーバ待ち行列はポワソン到着をもつ M/G/1 の自然な拡張になっていることが分かります．

この結果は，元々，非割込み優先権付き MAP/G/1 待ち行列モデルの解析のために考察したものです．しかし，思いもかけず，非常にきれいな結果が出てきたことで著者自身，非常に驚いた覚えがあります．

次に，この系内仕事量分布を用いて各到着流から来る客の実待ち時間を考えます．最初に述べたように，客の待ち時間は到着時点における系内仕事量に等しくなります．しかし，客は MAP に従って到着するため，MAP の相によって到着する頻度が異なります．よって定常状態において客が到着時に見る系内仕事量は，MAP の各相における到着率と相の出現頻度の積の比で重み付けられたものとなります．以下では k 番目の到着流から来た客をクラス k の客と呼びます．

定常状態におけるクラス k の客の待ち時間を W_k とし，定常状態においてクラス k の客の到着直後の MAP の相を S_k^* とします．このとき j 番目の要素が $\Pr(W_k\le x, S_k^*=j)$ であるベクトル $\boldsymbol{w}_k(x)$ は

$$\boldsymbol{w}_k(x)=\frac{\boldsymbol{v}(x)\boldsymbol{D}_k}{\boldsymbol{\pi D}_k\boldsymbol{e}} \tag{5}$$

で与えられます．また，定常状態におけるクラス k の客の系内滞在時間を R_k としたとき，j 番目の要素が $\Pr(R_k\le x, S_k^*=j)$ であるベクトル $\boldsymbol{r}_k(x)$ は

$$\boldsymbol{r}_k(x)=\frac{\int_{0-}^x d\boldsymbol{v}(y)\boldsymbol{D}_k(x-y)}{\boldsymbol{\pi D}_k\boldsymbol{e}}$$

で与えられます．

5　待ち行列長分布と不変式

系内仕事量の解析が終った後，暫く筆者は非割込み優先権付き待ち行列に関する研究 [11,15] や系内仕事量と同様の性質をもつ連続時間マルコフ連鎖に関する研究 [12] などを行っていました．待ち行列長分布をどう攻めるか，ずっと心に残っていたものの，手つかずの状態でした．しかし，昨年，後で述べるある思いつきがきっかけとなり攻略できたのです．

まず，従来とは視点を変え，客が先着順でサービスされるという仮定の下で，待ち行列長分布を系内仕事量分布，あるいは待ち時間分布と関連つけることを考えました．そのために，系内経過時間（attained waiting time）を定義します．これはもしシステムが空であるなら 0 であり，サービス中の客がいる場合は，この客の到着時点から現在までの時間で定義されます（図 2 参照）．

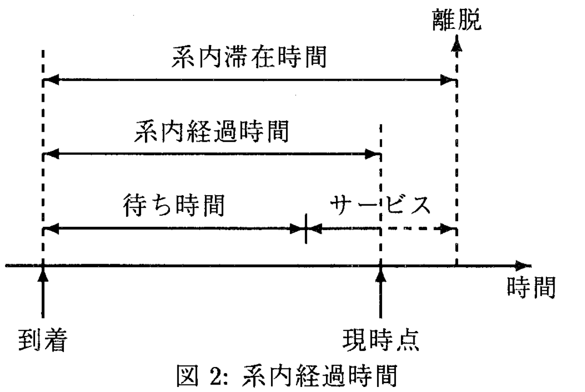

図 2: 系内経過時間

もし系内仕事量が 0 であるならば，客はシステム内にいません．一方，もし客がサービス中であるならば，サービスは先着順で行われるため，サービスを待っている客はサービス中の客が到着してから現時点までの間に到着したことになります．すなわち，ある時点で待っている客は，その時点での系内経過時間の間に到着した客になります．このような考え方に基づいて待ち行列長分布が満たす式を導出したのですが，その結果は，より一般的な状況でも成立する不変式の特

別な場合であることが分かりました．ここでは，まずこの不変式を紹介したいと思います．

$L_k(t)$ を時刻 t におけるクラス k の系内客数とし，$N_k(t), J_k(t)$ をそれぞれ，時間間隔 $(0,t]$ の間のクラス k の到着客数ならびに離脱客数とします．このとき

$$(L_1(t),\dots L_P(t)) = (L_1(0),\dots L_P(0)) + (N_1(t),\dots N_P(t)) - (J_1(t),\dots J_P(t))$$

が成立し，$(N_1(t),\dots N_P(t))$ が表現 $(\boldsymbol{C},\boldsymbol{D}_1,\dots,\boldsymbol{D}_P)$ をもつ定常なマーク付き MAP に支配されていると仮定します．ここで，クラス k の系内客数 $L_k(t)$ の定常状態における値を L_k，$z=(z_1,\dots,z_P)$ とし，

$$l_j(z) = E\left[z_1^{L_1}\cdots z_P^{L_P}1_{\{S=j\}}\right]$$

を j 番目の要素にもつ結合母関数ベクトルを $\boldsymbol{l}(z)$ とします．ここで $1_{\{\chi\}}$ は事象 χ の指示関数です．また，定常状態において，クラス k の客の離脱直後におけるクラス n の系内客数 $L_n(t)$ の値を $Q_n(k)$ とし，

$$q_{k,j}(z) = E\left[z_1^{Q_1(k)}\cdots z_P^{Q_P(k)}1_{\{S=j\}}\right]$$

を j 番目の要素にもつ結合母関数ベクトルを $\boldsymbol{q}_k(z)$ とします．このとき $\boldsymbol{l}(z)$ と $\boldsymbol{q}_k(z)$ の間には次式が成立します [18].

$$\boldsymbol{l}(z)\left(\boldsymbol{C}+\sum_{k=1}^{P} z_k\boldsymbol{D}_k\right) = \sum_{k=1}^{P}\lambda_k(z_k-1)\boldsymbol{q}_k(z) \qquad (6)$$

ここで λ_k はクラス k の到着流の到着率です．

$$\lambda_k = \boldsymbol{\pi}\boldsymbol{D}_k\boldsymbol{e} \qquad (7)$$

(6) はサービス時間分布の構造，サービス規律あるいはサーバ数などに依存せず，非常に広い範囲のモデルで成立します．また，単一の到着流の場合の結果 [14] の拡張になっています．

(6) を用いれば待ち行列長分布が満たす式を簡単に得ることが出来ます．以下では、待ち部屋をシステムと捉え，待ち客数を考えます．定常状態において，クラス k の待ち客数を X_k とし，定常状態における MAP の相を S としたとき，j 番目の要素 $x_j^*(z)$ が

$$x_j^*(z) = E\left[z_1^{X_1}\cdots z_P^{X_P}1_{\{S=j\}}\right]$$

で与えられるような結合母関数ベクトル $\boldsymbol{x}^*(z)$ を定義します．これは (6) の $\boldsymbol{l}(z)$ に対応します．一方 $\boldsymbol{q}_k(z)$ に対応するのは，クラス k の客が待ち部屋から出たとき（サービスの開始時点）に待ち部屋に残っている客数です．客は先着順でサービスされるので，これらの客は待ち部屋から出た客の待ち時間の間に到着した客です．よって (6) より次式を得ます[8].

$$\boldsymbol{x}^*(z)\left(\boldsymbol{C}+\sum_{k=1}^{P} z_k\boldsymbol{D}_k\right) = \sum_{k=1}^{P}\lambda_k(z_k-1) \cdot\int_{0-}^{\infty} d\boldsymbol{w}_k(x)\exp\left[\left(\boldsymbol{C}+\sum_{i=1}^{P} z_i\boldsymbol{D}_i\right)x\right] \qquad (8)$$

さて，(8) からどのようにして系内客数分布を計算すれば良いのでしょうか．上記の系内経過時間に着目するアイデアはかなり以前からあったのですが，数値計算ができる形に処理できるとは当時思えませんでした．話は逸れますが，昨年の春，集団到着を許す MAP/G/1 の客数分布の新しい計算アルゴリズムを思いつきました [16]．これは，時間平均の客数分布が，一見，直接は関係なさそうな M/G/1 型マルコフ連鎖の解と同じ分布を持つという点に注目したものです．

これを思いついた直後，ひょっとすると同じような構造が隠れているのではないかと思い，一気に計算を進めたところ，$x_j(n_1,\dots,n_P) = \Pr(X_1=n_1,\dots, X_P=n_P, S=j)$ を j 番目の要素にもつ定常状態における待ち客数の結合確率ベクトル $\boldsymbol{x}(n_1,\dots,n_P)$ を計算する数値アルゴリズムを得ることができました．以下にその概略を紹介します．まず定義より，

$$\boldsymbol{x}^*(z) = \sum_{n_1=0}^{\infty}\cdots\sum_{n_P=0}^{\infty} z_1^{n_1}\cdots z_P^{n_P}\boldsymbol{x}(n_1,\dots,n_P)$$

です．また (5), (7) より $\boldsymbol{w}_k(x) = \boldsymbol{v}(x)\boldsymbol{D}_k/\lambda_k$ です．よって $\theta = \max_i|C_{i,i}|$ で与えられる θ を用いて (8) を変形すると[9]

$$\boldsymbol{x}^*(z)\left(\boldsymbol{C}+\sum_{k=1}^{P} z_k\boldsymbol{D}_k\right) = \sum_{k=1}^{P}(z_k-1)\sum_{m=0}^{\infty}\boldsymbol{v}^{(m)}(\theta)\boldsymbol{D}_k \cdot\sum_{\substack{n_1\ge 0\\ n_1+\cdots+n_P\le m}}\cdots\sum_{n_P\ge 0} z_1^{n_1}\cdots z_P^{n_P}\boldsymbol{F}_m(n_1,\dots,n_P) \qquad (9)$$

と書くことができます．ただし

$$\boldsymbol{v}^{(m)}(\theta) = \int_{0-}^{\infty} e^{-\theta x}\frac{(\theta x)^m}{m!}d\boldsymbol{v}(x) \qquad (10)$$

$$\sum_{\substack{n_1\ge 0\\ n_1+\cdots+n_P\le m}}\cdots\sum_{n_P\ge 0} z_1^{n_1}\cdots z_P^{n_P}\boldsymbol{F}_m(n_1,\dots,n_P) = \left[\boldsymbol{I}+\theta^{-1}\left(\boldsymbol{C}+\sum_{i=1}^{P} z_i\boldsymbol{D}_i\right)\right]^m$$

です．(9) の両辺の $z_1^{n_1}\cdots z_P^{n_P}$ の係数を比較することにより $\boldsymbol{x}(n_1,\dots,n_P)$ に対する再帰式を $\boldsymbol{v}^{(m)}(x)$ と $\boldsymbol{F}_m(n_1,\dots,n_P)$ を用いて構築することができます．

---

[8] (8) の右辺にある積分はクラス k の待ち時間の間に到着する客数を表現しています．

[9] この操作は連続時間マルコフ連鎖を離散時間マルコフ連鎖に変換する一様化手法として知られています．

すなわち，$\boldsymbol{F}_m(n_1,\ldots,n_P)$ と $\boldsymbol{v}^{(m)}(\theta)$ が計算できれば $\boldsymbol{x}(n_1,\ldots,n_P)$ が計算できるわけです．

$\boldsymbol{F}_m(n_1,\ldots,n_P)$ は，$P=1$ のときの数値計算アルゴリズムが既に知られており，$P\geq 2$ への拡張は容易です．問題は $\boldsymbol{v}^{(m)}(\theta)$ の数値計算アルゴリズムでしたが，この $\boldsymbol{v}^{(m)}(\theta)$ は必ずしも自明でない構造を含んでおり，これを利用することで解決できました．

まず $\boldsymbol{v}^*(s)$ の定義と (10) より

$$\sum_{m=0}^{\infty}\boldsymbol{v}^{(m)}(\theta)z^m=\boldsymbol{v}^*(\theta-\theta z)$$

であることがわかります．よって (3) より

$$\sum_{m=0}^{\infty}\boldsymbol{v}^{(m)}(\theta)z^m[\theta(1-z)\boldsymbol{I}+\boldsymbol{C}+\boldsymbol{D}^*(\theta-\theta z)] =\theta(1-z)(1-\rho)\boldsymbol{\kappa}$$

を得ます．この両辺の z^n の係数を比較することにより $\boldsymbol{v}^{(m)}(\theta)$ の満たす無限個の連立方程式が得られます．この連立方程式を良く調べると，$\boldsymbol{v}^{(m)}(\theta)$ は，M/G/1 型とよばれる構造をもった，あるマルコフ連鎖の定常解と同一であることが分かりました．すなわち，通常の M/G/1 型マルコフ連鎖に対する行列解析法を適用することで $\boldsymbol{v}^{(m)}(\theta)$ が計算できることが明らかになったわけです[10]．

6 おわりに

サービス時間分布が到着流毎に異なる複数の到着流をもつ単一サーバ待ち行列に関する筆者の研究成果を紹介しました．待ち行列理論は，その名が示すとおり待ち行列長を中心とした議論が多く，通常，待ち行列長を元に待ち時間分布を導出するという手順を経ます．しかし，サービス時間分布が到着流毎に異なる場合，客数の挙動を直接解析することは極めて困難です．一方，系内仕事量，あるいは待ち時間分布を最初に解析し，それを利用して客数分布の解析へと向かうアプローチを取れば，従来，解析が困難と見られていたクラスの問題も解決できる可能性があることが分かります．

本稿では触れることができませんでしたが，この種の研究では，実際に数値計算するための数値的に安定したアルゴリズムを示すことが極めて重要です．今まで各種の問題に対して様々なアルゴリズムが提案されていますが，大規模な問題に対しては，まだ，十分とは言えません．今後，解析が困難と思われているモデルに対して挑戦していくとともに．より効率の良いアルゴリズムの開発を進め，将来，それらを一つのソフトウェアパッケージにまとめたいと思っています．

[10] サービス中の客も含めた系内客数の結合分布も同様に計算できます．

最後になりましたが，本稿の草稿に対して貴重な御意見を頂いた東北大学の山下英明氏に深謝致します．

参考文献

[1] Q.-M. He: "Queues with marked customers," *Adv. Appl. Prob.*, 28, 567–587, 1996.

[2] 紀一誠（編）: 特集「待ち行列理論の最近の応用」オペレーションズ・リサーチ，43 (5), 258–288, 1998.

[3] D. M. Lucantoni K. S. Meier-Hellstern and M. F. Neuts: "A single-server queue with server vacations and a class of non-renewal arrival processes," *Adv. Appl. Prob.* 22, 676–705, 1990.

[4] M. F. Neuts: *Structured Stochastic Matrices of M/G/1 Type and Their Applications*, Marcel Dekker, New York, 1989.

[5] 高橋幸雄: "やさしい待ち行列 (2) — 等間隔運転は待ちを減らす," オペレーションズ・リサーチ，40 (12), 716–721, 1995.

[6] 高橋幸雄：“待ち行列研究の新しい潮流 (1) — 待ち行列研究の変遷,” オペレーションズ・リサーチ，43 (9), 495–499, 1998.

[7] 高橋幸雄，牧本直樹：“待ち行列研究の新しい潮流 (3) — 相型分布と行列解析法,” オペレーションズ・リサーチ，43 (11), 618–623, 1998.

[8] L. Takács: "Investigation of waiting time problems by reduction to Markov processes," *Acta Math. Acad. Sci. Hung.*, 6, 101–129, 1955.

[9] T. Takine and T. Hasegawa: "The workload in the MAP/G/1 queue with state-dependent services: Its application to a queue with preemptive resume priority," *Stoch. Mod.*, 10, 183–204, 1994.

[10] T. Takine et al.,: "Mean waiting times in nonpreemptive priority queues with Markovian arrival and i.i.d. service processes," *Perfor. Eval.*, 20, 131–149, 1994.

[11] T. Takine: "A nonpreemptive priority MAP/G/1 queue with two classes of customers," *JORSJ*, 39, 266–290, 1996.

[12] T. Takine: "A continuous version of matrix-analytic methods with the skip-free to the left property," *Stoch. Mod.*, 12, 673–682, 1996.

[13] 滝根哲哉，村田正幸：“通信網における待ち行列 –理論の応用と課題–,” オペレーションズ・リサーチ，43 (5), 264–271, 1998.

[14] T. Takine and Y. Takahashi: "On the relationship between queue lengths at a random instant and at a departure in the stationary queue with BMAP arrivals," *Stoch. Mod.*, 14, 601–610, 1998.

[15] T. Takine: "The nonpreemptive priority MAP/G/1 queue," to appear in *Opns. Res.*, 1999.

[16] T. Takine: "A new recursion to compute the queue length in the stationary BMAP/GI/1 queue," submitted for publication.

[17] T. Takine: "Queue length distribution in a FIFO single-sever queue with multiple arrival streams having different service time distributions," submitted for publication.

[18] T. Takine: "Joint queue length distribution in a stationary queue with multiple arrival streams governed by a Markov chain," submitted for publication.

[19] 滝根哲哉: "構造化されたマルコフ連鎖と待ち行列," システム／制御／情報，43 (3), 1999.