

有限呼源 PH/G/1 待ち行列の解析

井出一郎

(日本電信電話公社 武蔵野電気通信研究所)

1. はじめに

計算機システムの性能評価モデルの1つに有限呼源待ち行列モデルがある(図1)。本モデルは、機械を呼源、修理工をサーバとする機械修理(機械干渉)問題として有名であり[1]、無限サーバ(Infinite Server)と単一サーバよりなる循環待ち行列あるいは閉鎖形の待ち行列網モデルの特殊例である。本モデルは、端末を呼源、プロセッサをサーバと見ると、TSSの応答時間のモデル[2][3]となる。また、マルチプロセッサシステムの共通バス(メモリ)競合のモデルとしても用いられる[4][5][6]。

本稿では、呼源の空き間隔が位相型分布[7]、サービス時間が一般分布に従う有限呼源待ち行列モデル(PH/G/1/N)を、補助変数法によって解析する。また数値例によって、空き間隔およびサービス時間の分布形の影響を考察する。

2. 数学モデルと既成理論

2.1 数学モデル

次の仮定を設ける。i) 呼源数 N 、サーバ数 1 。
 ii) 各呼源の空き時間(idle time: サービス終了から新たに発呼するまでの時間)は、呼源間および呼相互に独立で同一(i.i.d.)の位相型分布 $A(x)$ に従う。 $A(x)$ の位相数を m とする。ここで位相型分布とは、唯一の吸収状態($m+1$)を持つ連続時間マルコフ連鎖の、初期確率ベクトル $(\underline{\alpha}, \alpha_{m+1})$ からの吸収時間分布として表しうる $[0, \infty)$ 上の確率分布である。マルコフ連鎖の無限小生成作用素を $\begin{pmatrix} \underline{T} & \underline{t}^T \\ \underline{0}^T & 0 \end{pmatrix}$ とすると、

$$A(x) = 1 - \underline{\alpha} \exp(\underline{T}x) \underline{e}^T, \quad (x \geq 0) \quad (2.1)$$

となる。但し、 $\underline{\alpha} \triangleq (\alpha_1, \dots, \alpha_m)$, $\underline{e} \triangleq (1, \dots, 1)$, $\underline{t}^T \triangleq (t_1^0, \dots, t_m^0)$, $\underline{0} \triangleq (0, \dots, 0)$ 。

$\underline{T} \triangleq (T_{jk})$ は $m \times m$ 行列、 $\underline{\alpha} \underline{e}^T + \alpha_{m+1} = 1$, $\underline{T} \underline{e}^T + \underline{t}^T = \underline{0}^T$, $0 \leq \alpha_j \leq 1$ である。

$A(x)$ の平均 $(-\underline{\alpha} \underline{T}^{-1} \underline{e}^T)$ を $1/\lambda$ とおく。計算を容易にするため、 \underline{T} は(右)上三角行列と仮定する。Cox型分布[8](図2)をはじめとする多くの分布は、上三角行列で表現できるので、これは実用上の制限とはならない。また煩雑さを避けるため $\alpha_{m+1} = 0$ (つまり $A(0) = 0$) の場合に限定する。

iii) サービス時間は i.i.d. で一般分布 $H(x)$ に従う。 $H(x)$ の平均を h 、ラプラス・スチルチェス変換(LST)を $H^*(s) \triangleq \int_0^\infty e^{-sx} dH(x)$ とおく。iv) サービス規律は先着順(FCFS)とする。

2.2 既成理論

空き間隔が指数分布に従う場合(M/G/1/N)を、隠れマルコフ連鎖法により Takács[9] が、補助変数法と離散変換により Jaiswal[10] が解析している。さらに有限呼源の平均応答時間を無限呼源モデル(ボラチェック・ヒンチンの公式)で求めた時の誤差評価[11]や、サービス時間分布による利用率の上下限評価[12]が行われている。サービス時間が指数分布に従う場合(GI/M/1/N)には、定

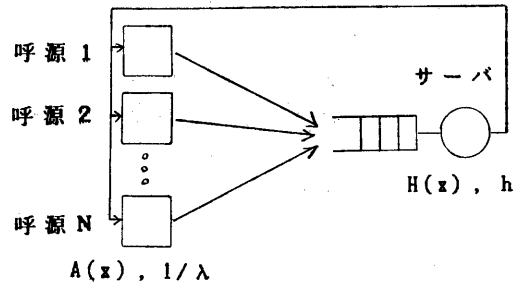


図1 有限呼源待ち行列モデル

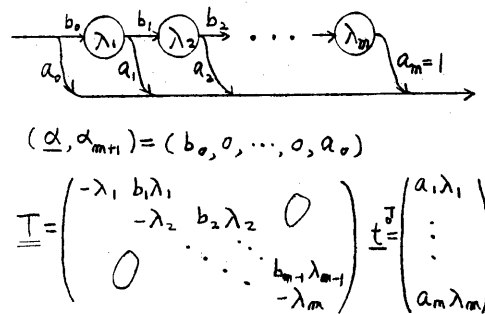


図2 Cox型分布

常状態確率が空き間隔分布形に依存しない（ロバストネスがある）[1] ことが示されている。また、筆者によってPH/PH/1/Nの数値計算が行われている[13]。

FCFS以外の処理規律では、空き間隔が指数分布に従う場合について、割込み優先権[10]、非割込み優先権[14],[15]の解析が行われている。また待ち行列網の理論より、処理規律がプロセッサシェアリング(PS)、後着順割込み継続形(LCFS-PR)[16]、及び一般化PS[17]の時には、空き間隔およびサービス時間分布についてロバストネスがあることが知られている。

評価尺度間の関係

サーバ利用率、平均系内（待ち+サービス中）呼数、平均待ち時間、スループット（単位時間当りにサービスされる呼数）をそれぞれ、 ρ , L , W , λ^* とおく。

平均サイクルタイム $=1/\lambda + W + h$ 、平均空き呼源数 $=N - L$ に注意すると次の関係が成立する。

$$\lambda^* = N / (1/\lambda + W + h) \quad (2.2)$$

$$= (N - L) \lambda \quad (2.3)$$

$$\rho = \lambda^* h = N h / (1/\lambda + W + h) \quad (2.4)$$

従って、これらの評価尺度のいずれかが求まれば、残りを計算できる。例えば ρ が求まったとき、

$$W = h (N/\rho - 1) - 1/\lambda \quad (2.5)$$

$$L = N - \rho / (\lambda h) \quad (2.6)$$

である。

呼源の呼量

呼源の呼量(traffic potential) a_s [18]、つまり待合せが無い場合に生起するはずの呼量を有限呼源モデルの負荷の尺度として用いる。

$$a_s \triangleq N h / (1/\lambda + h) \quad (2.7)$$

1呼源当りの呼源の呼量を $\beta_s \triangleq h / (1/\lambda + h)$ とおく。空き間隔およびサービス時間が単位分布に従う場合(D/D/1/N)の時間平均の意味での利用率は、

$$\rho = \min(a_s, 1) \quad (2.8) \text{ である。}$$

3. 補助変数法による解析

3.1 状態方程式

残余サービス時間を補助変数にとり、系の状態を $\{I(t); J_1(t), \dots, J_m(t); X(t)\}$ で表す。ここで、確率変数 $I(t), \{J_k(t)\} k=1, \dots, m, X(t)$ は、それぞれ時刻 t における系内呼数、呼源（空き間隔）の第 k 位相にいる呼数、及び残余サービス時間である。状態 $(i; j_1, \dots, j_m)$ を表1のように、 i について昇順、 j について降順にならべて番号 $\nu = \nu(i; j_1, \dots, j_m)$ を付ける。但し、 $i + j_1 + \dots + j_m = N$ 。図3にAPLで記述した番号付けの関数を示す。

状態確率を

$$P_\nu(x, t) dx \triangleq \Pr\{I(t)=i; J_1(t)=j_1, \dots, J_m(t)=j_m; x \leq X(t) < x+dx, i=1, 2, \dots, N\} \quad (3.1)$$

$$P_\nu(t) \triangleq \Pr\{I(t)=0, J_1(t)=j_1, \dots, J_m(t)=j_m\} \quad (3.2)$$

とおく。記法の簡略化のため、 $i=0$ の ν について $P_\nu(x, t) \triangleq H(x)P_\nu(t)$ とおき、 $P_\nu(x, t), \nu=1, \dots, M$ をならべたベクトルを $\underline{P}(x, t)$ 、その系内呼数が i のサブ

表1 状態番号(PH₃/G/1/3の場合)

ν	i	j_1	j_2	j_3
1	0	3	0	0
2	0	2	1	0
3	0	2	0	1
4	0	1	2	0
5	0	1	1	1
6	0	1	0	2
7	0	0	3	0
8	0	0	2	1
9	0	0	1	2
10	0	0	0	3
11	1	2	0	0
12	1	1	1	0
13	1	1	0	1
14	1	0	2	0
15	1	0	1	1
16	1	0	0	2
17	2	1	0	0
18	2	0	1	0
19	2	0	0	1
20	3	0	0	0

$$\underline{P} = (\underbrace{P_1, \dots, P_{10}}_{P_0}, \underbrace{P_{11}, \dots, P_{16}}_{P_1}, \dots, P_M)$$

$$\underline{P}^{*(n)}(s)(s\underline{I} + \underline{V}) + n\underline{P}^{*(n-1)}(s) + \underline{P}(0)H^{*(n)}(s)\underline{W} = \underline{0} \quad (3.12)$$

従って $\underline{P}_\nu^{*(n)}(s)$ も、 $s \approx \mu_\nu$ の時は $\{P_\ell^{*(n)}(s), \ell \leq \nu-1; P_\nu^{*(n-1)}(s), P(0), H^{*(n)}(s)\}$ によって、 $s = \mu_\nu$ の時は $\{P_\ell^{*(n+1)}(s), \ell \leq \nu-1; P(0), H^{*(n+1)}(s)\}$ によって漸化的に表せる。ここで式(3.9)に注意すると $\underline{P}_\nu^{*(n)}(s)$ は結局、定数ベクトル $\underline{P}(0)$ の線形結合で表現できる。

$$\underline{P}_\nu^{*(n)}(s) = \sum_{\ell=1}^M C_{\nu\ell}(s) P_\ell(0) \quad (3.13)$$

但し、 $C_{\nu\ell}(s)$ は $H^{*(n)}(s)$ 等で表される係数である。 $C_{\nu\ell}(s)$ の値を求めるプログラムは、APL 等の再帰呼出 (recursive call) 機能を用いて容易に実現できる。

P(0) の決定

式(3.8) 及び $i \approx 0$ の ν における式(3.11)より、 $P(0)$ に関する M 個の線形方程式 (但し、ランクは $M-1$) が得られる。また式(3.7) に右から \underline{e}^T をかけ、 $\underline{V} \underline{e}^T = \underline{0}^T$ 、 $\underline{W} \underline{e}^T = \underline{e}^T$ に注意すると

$$s \sum_{i=1}^N q_i^*(s) + [H^*(s) - 1] \sum_{i=1}^N q_i(0) = 0 \quad (3.14)$$

を得る。但し、 $q_i^*(s) \triangleq \underline{P}_i^*(s) \underline{e}^T$ 、 $q_i(0) \triangleq \underline{P}_i(0) \underline{e}^T$ 。

$$\text{式(3.14)を } s \text{ で微分した後 } s=0 \text{ とし、正規化条件 } q_0 + \sum_{i=1}^N q_i(0) = 1 \quad (3.15)$$

を用いると

$$q_0 + h \sum_{i=1}^N q_i(0) = 1 \quad (3.16)$$

を得る。上記の M 個の線形方程式と式(3.16)を連立させて解くことにより $\underline{P}(0)$ が決定され、定常状態確率 $\underline{P}^*(0)$ より各種評価尺度を得る。例えば、利用率は、

$$\rho = 1 - q_0 \quad (3.17)$$

で与えられる。

4. 数値例と考察

4.1 分布形の影響

表 2, 3 に空き間隔分布とサービス時間分布に対する平均待ち時間、利用率を示す。ここで $E_{k,\lambda}$ は k -アーラン分布、 H_2 は対称条件を付与した 2 次の超指数分布、つまり確率密度関数が

$$f(x) \triangleq \alpha_1 \lambda_1 e^{-\lambda_1 x} + \alpha_2 \lambda_2 e^{-\lambda_2 x} \quad (4.1)$$

で、平均の小さい方の位相の比率

$$r \triangleq (\alpha_1 / \lambda_1) / ((\alpha_1 / \lambda_1) + (\alpha_2 / \lambda_2))$$

が 1/2 の分布である。但し、 $\alpha_1 + \alpha_2 = 1, \lambda_1 \geq \lambda_2$ 。分布のばらつきの尺度として

変動係数 \triangleq 標準偏差 / 平均

を示す。 C_a, C_s はそれぞれ、空き間隔およびサービス時間の変動係数である。呼源の呼量と利用率の関係を図 4 に、変動係数に対する平均待ち時間特性を図 5 に示す。これらより次のことが推測できる。

i) サービス時間分布のばらつきが小さい程、特性が良い (利用率が高く、平均待ち時間が短い)。

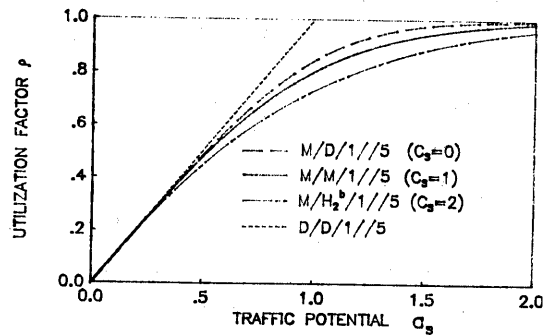


図4 呼源の呼量と利用率

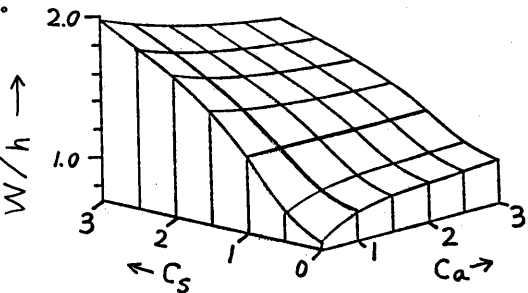


図5 変動係数と平均待ち時間 (PH/G/1//5, $a_s=1$)

ii) サービス時間分布のばらつきの大小によって空き間隔分布形の影響の傾向が変わる。つまりサービス時間のばらつきが指数分布より小さい時には、空き間隔のばらつきが小さい方が特性が良いが、サービス時間のばらつきが指数分布よりも大きいときには逆になる。

iii) 空き間隔分布形の影響はサービス時間分布形の影響に比べて小さい。

表2 正規化平均待ち時間 W/h ($N=5, a_s=1$)

サービス時間 C _s C _a	空き間隔			H ₂ ^b			
	E ₄ 1/2	E ₂ 1/√2	M 1	√2	2	3	10
D 0	0.7508	0.8222	0.9053	0.9481	0.9777	1.0008	1.0212
E ₉ 1/3	0.8259	0.8832	0.9518	0.9863	1.0132	1.0330	1.0505
E ₄ 1/2	0.9090	0.9525	1.0058	1.0334	1.0550	1.0710	1.0852
E ₂ 1/√2	1.0374	1.0622	1.0936	1.1105	1.1239	1.1338	1.1428
M 1	1.2427	1.2427	1.2427	1.2427	1.2427	1.2427	1.2427
√2	1.4657	1.4436	1.4148	1.3928	1.3736	1.3582	1.3435
H ₂ ^b 2	1.7123	1.6755	1.6244	1.5777	1.5337	1.4958	1.4569
3	1.9622	1.9258	1.8699	1.8042	1.7329	1.6619	1.5773
10	2.2603	2.2536	2.2413	2.2173	2.1753	2.0960	1.8058

表3 利用率 ρ ($N=5, a_s=1$)

サービス時間 C _s C _a	空き間隔			H ₂ ^b			
	E ₄ 1/2	E ₂ 1/√2	M 1	√2	2	3	10
D 0	0.8694	0.8588	0.8467	0.8409	0.8364	0.8332	0.8304
E ₉ 1/3	0.8582	0.8499	0.8401	0.8352	0.8315	0.8288	0.8264
E ₄ 1/2	0.8462	0.8400	0.8325	0.8287	0.8258	0.8236	0.8217
E ₂ 1/√2	0.8282	0.8248	0.8205	0.8183	0.8165	0.8152	0.8140
M 1	0.8009	0.8009	0.8009	0.8009	0.8009	0.8009	0.8009
√2	0.7733	0.7760	0.7794	0.7821	0.7845	0.7864	0.7882
H ₂ ^b 2	0.7449	0.7490	0.7548	0.7601	0.7653	0.7697	0.7744
3	0.7182	0.7219	0.7278	0.7348	0.7426	0.7505	0.7602
10	0.6887	0.6893	0.6905	0.6928	0.6968	0.7046	0.7347

4.2 その他の分布形状パラメタ

一般に到着間隔やサービス時間の分布のLSTが関係する待ち行列の特性は、分布の低次モーメントの情報のみでは定まらない。分布形を示すパラメタとして変動係数では不十分な場合があることが指摘されている[12][19]。分布の形状パラメタとしては変動係数の他に分散係数 (peakedness factor) [20]や

パーセントイル[19]が提案されているが、分布形の影響は対象とする待ち行列モデルによって異なるため決定的なものはない。ここではWhitt [21]と同様に、変動係数の等しい2次の超指数分布H₂(式(4.1))における位相の比率パラメタrの影響を表4に示す。有限呼源モデルにはどんな形状パラメタがふさわしいかは、今後の課題である。

4.3 空き間隔を指数分布で代用した時の精度

空き間隔分布が指数分布に従う場合には、簡単なM/G/1//Nの公式がある。空き間隔が指数分布でない場合にも、この公式で代用できれば応用上便利である。図6にPH/D/1//NをM/D/1//Nで代用した時の利用率の精度つまり、

$$\text{相対誤差} = (P_M - P_{PH}) / P_{PH} \times 100 \%$$

を示す。利用率の誤差は、実用上無視できる程度であることがわかる。また、呼源数が多くなるにしたがい誤差は小さくなる。なお、平均待ち時間の誤差は、10%程度である。

表4 比率パラメタrの影響 (平均待ち時間W/h) (H₂^b/H₂^b/1//5, a_s=1, C_a=C_s=2)

サービス時間 C _s C _a	r	空き間隔 r				
		0.01	0.1	0.5	0.9	0.99
D	0.01	1.5622	1.6402	1.7752	1.8556	1.8773
	0.1	1.5384	1.5865	1.7134	1.7966	1.8195
	0.5	1.4719	1.4766	1.5337	1.5998	1.6218
	0.9	1.3427	1.3413	1.3434	1.3563	1.3646
	0.99	1.2573	1.2573	1.2572	1.2573	1.2579

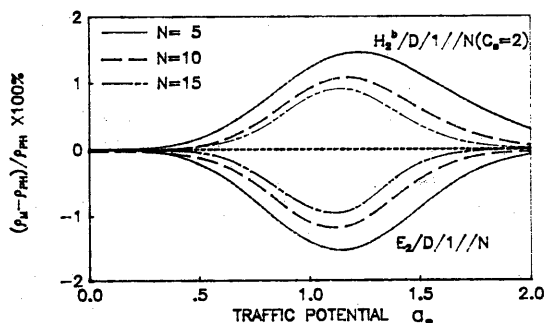


図6 空き間隔を指数分布で代用した時の利用率の誤差

5. おわりに

空き間隔が位相型分布、サービス時間が一般分布に従う有限呼源待ち行列モデルを、残余サービス時間を用いる補助変数法によって解析した。また、数値例により、i) サービス時間のばらつきが指数分布より大きいかわいかにによって空き間隔分布形の影響が逆転すること、ii) 利用率に対する空き間隔分布形の影響は小さいこと、等を示した。

[参考文献]

- [1] Barlow, R. E. : "Repairman Problems", in K. J. Arrow (eds), *Studies in Applied Probability and Management Science*, Stanford Univ. Press, pp.18-33(1962)
- [2] Scherr, A. L. : *An Analysis of Time-shared Computer Systems*, MIT Press (1967)
- [3] 益田、亀田 : オペレーティングシステムの性能解析, 情報処理学会(1982)
- [4] 池原 悟 : "マルチプロセッサ方式における共用メモリアクセス競合の解析", 信学論, J63-D, pp.334-341(1980)
- [5] Marsan, M.A. , Gregoretti, F. : "Memory interference models for a multi-microprocessor system with a shared bus and a single external common memory", *Multiprocessing and Microprogramming*, 7, pp.124-133(1981)
- [6] 住田 修一 : "非割込み優先権処理のある共通メモリアクセス競合の解析", 信学論, J68-B, pp.1-8 (1985)
- [7] Neuts, M. : *Matrix-Geometric Solutions in Stochastic Models*, The Johns Hopkins Univ. Press(1981)
- [8] Cox, D. R. : "A use of complex probabilities in the theory of stochastic processes", *Proc. Camb. Phil. Soc.*, 51, pp.313-319(1955)
- [9] Takács, L. : "On a stochastic process concerning some waiting time problems", *Theory of Probability and its Applications*, 2, pp.90-103(1957)
- [10] Jaiswal, N. K. : *Priority Queues*, Academic Press (1968)
- [11] Buzen, J. P. , Goldberg, P. S. : "Guidelines for the use of infinite source queueing models in the analysis of computer system performance", *Proc. NCC*, pp.371-374(1974)
- [12] Price, T. G. : "A note on the effect of the central processor service time distribution on processor utilization in multiprogrammed computer systems", *J. ACM*, 23, pp.342-346 (1976)
- [13] 井出 : "共通バス要求間隔分布の競合に対する影響", 信学会通信部門全国大会, 274(1984)
- [14] Hodgson, V. , Hebble, T. L. : "Nonpreemptive priorities in machine interference", *Operations Research*, 15, pp.245-253(1967)
- [15] Chandra, M.J. , Sargent, R. G. : "A numerical method to obtain the equilibrium results for the multiple finite source priority queueing model", *Management Science*, 29, pp.1298-1308(1983)
- [16] Baskett, F. , Chandy, K. M. , Muntz, R. R. , Palacios, F. G. : "Open, closed, and mixed networks of queues with different class of customers", *J. ACM*, 22, pp.248-260 (1975)
- [17] Cohen, J.W. : "The multiple phase service network with generalized processor sharing", *Acta Informatica*, 12, pp.245-284(1979)
- [18] 藤木、雁部 : 通信トラヒック理論, 丸善(1980)
- [19] Lazowska, E. D. : "The use of percentiles in modeling CPU service time distributions", K.M. Chandy and M. Reiser(eds.) *Computer Performance*, North Holland, pp.53-66(1977)
- [20] Holtzman, J. M. : "The accuracy of the equivalent random method with renewal inputs", *Bell Syst. Tech. J.*, 52, pp.1673-1679(1973)
- [21] Whitt, W. : "On approximations for queues, III: mixture of exponential distributions", *Bell Lab. Tech. J.*, 63, pp.163-175(1984)