論文・総合報告

ポーリングモデル：巡回サービス多重待ち行列

高木　英明

1. はじめに

ポーリングモデル (polling model) とは複数の待ち行列を1つのサーバが巡回してサービスするシステムのことである．類似の巡回的サービスである**同期時分割多重方式** (synchronous time division multiplexing: STDM) では各待ち行列でのサーバの滞在時間が客の有無にかかわらず一定であるのに対して，ポーリングモデルでは各待ち行列にサーバが来るときにそこにいる客数等に応じて動的にサービス期間が決められるのが特徴である．また通常の固定優先順位をもつ待ち行列システムと比べると，ポーリングモデルは優先順位がサーバの位置とともに巡回するシステムであるといえる (図1参照)．

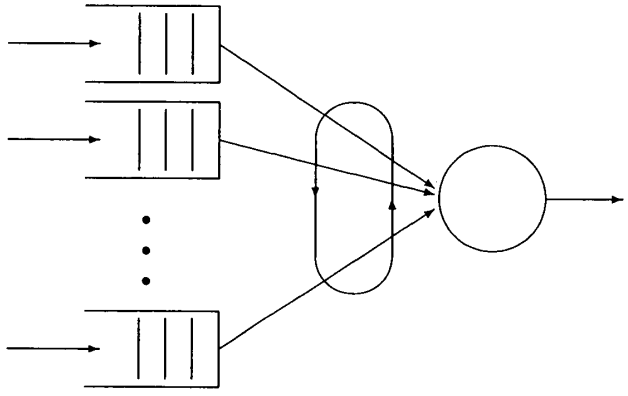

図 1: ポーリングモデル

もともと「ポーリング」とは，中央のコンピュータがそれに接続された複数の端末機に対し順に送りたいデータがあるかどうかを問い合わせ，それに応えてデータのある端末機はデータがあれば送信を行なうというデータ通信の制御方式に付けられた名前である．この場合，中央のコンピュータがサーバの役目を果たし，各端末機の送信メッセージ用バッファが客の待ち行列に対応する．これは1970年代始め頃のポーリングモデルの応用であるが，いろいろ文献を調べると，データ通信以外の分野においても同様の待ち行列モデルが研究されていることがわかった．

現在の観点から整理すると，既に1950年代には単一バッファモデルが綿織物工場の巡回機械修理工の問題に使われた [44]．1960年代には，道路の交差点における交通信号制御のモデルとして [53]，またオペレーションズリサーチの問題として [4]，2つの待ち行列から成る無限バッファモデルが解析された．1960年代の終りには，データ通信の実用化とともに，全処理式とゲート式サービスシステムが研究された [22, 23, 26, 32, 33]．1970年代には上述のようにポーリング方式のデータ通信の理論的モデルとしてよく用いられたが，1980年代になると，ローカルエリアネットワーク (LAN) におけるトークンリング方式のモデルとして再び脚光を浴び [16, 17]，とくにその方式に合った制限式サービスモデルが解析された [29, 56, 66]．さらに，トークンバスやFiber Distributed Data Interface (FDDI) のモデル化では，優先順位付き時間制限式サービスが研究された．最近のLANの標準IEEE802.12 100VG-AnyLAN (Demand Priority) の方式も優先順位付きポーリングモデルである．狭帯域ISDNのDチャネルのアクセス競合はポーリングモデルで解析される [47]．通信分野以外でも，交通や生産システムのモデルとして使われている [5, 14]．

ポーリングモデルがこのように広い応用範囲をもつことは不思議ではない．なぜならば，1つの資源 (サーバ) を複数の利用者 (待ち行列) が協調して使うとき，資源の巡回式割り当て (ポーリングモデル) は自然で公平な方式であると考えられるからである．

ポーリングモデルの解析に関し1980年代半ばまでに種々の分野で多分独立になされていた研究結果を集め，初めて統一的に整理したのが筆者の研究書 [57] と概説論文である [58]．解析のみならず最適化と応用を含むその後の研究成果については筆者や他の研究者による解説論文が多く出ている [8, 18, 21, 30, 41, 48, 59, 60, 61, 67]．

たかぎ　ひであき　筑波大学社会工学系
〒 305 茨城県つくば市天王台 1-1-1

受付：1995.5.9　採択：1995.10.13

本解説文では，第2節で基本的な単一バッファモデルと無限バッファモデルの解析の結果をまとめ，第3節でそれらの拡張モデルを紹介する．なお，文献の引用は代表的なものに限った．

2. 基本モデル

ポーリングモデルはいくつかの**基本モデル** (basic model) とそれらを拡張したモデルに分類することができる．本節では，基本モデルとその解析の結果を紹介する．基本モデルは，それぞれ独立な客の到着過程をもつ複数の待ち行列を1つのサーバが巡回しながらサービスを行なう連続時間システムである．定常状態のみを考える．

すべての基本モデルに共通の条件と記号を示そう．システム内の待ち行列の数を N とする．それらの待ち行列に，サーバが巡回する順序に従って $1,2,\ldots,N$ と番号を付ける．待ち行列 N の次には待ち行列1が訪問される．待ち行列 i には客が率 λ_i の**Poisson過程** (Poisson process) で到着する $(1 \le i \le N)$．待ち行列 i における客の**サービス時間** (service time) の分布関数の Laplace-Stieltjes 変換 (LST)，平均および2次モーメントをそれぞれ $B_i^*(s)$, b_i および $b_i^{(2)}$ で表す．このときシステム全体にかかる**負荷** (load) は

$$\rho = \sum_{i=1}^{N} \rho_i \quad ; \quad \rho_i := \lambda_i b_i \quad (1 \le i \le N) \tag{1}$$

で与えられる．サーバが待ち行列 i でのサービスを終えたあと待ち行列 $i+1$ に移動するのに要する時間をサーバの**歩行時間** (walking time) という．その分布関数の LST，平均および分散をそれぞれ $R_i^*(s)$, r_i および δ_i^2 で表す．各歩行時間は独立であると仮定すれば，負荷が無いときサーバがすべての待ち行列を一巡するのに要する時間の平均と分散はそれぞれ

$$R = \sum_{i=1}^{N} r_i \qquad \Delta = \sum_{i=1}^{N} \delta_i^2 \tag{2}$$

で与えられる．

負荷があるときにサーバがすべての待ち行列を一巡するのに要する時間を**巡回時間** (cycle time) という．待ち行列 i で測った巡回時間の分布関数の LST と平均をそれぞれ $C_i^*(s)$ と $E[C]$ で表す (平均巡回時間はどの待ち行列で測るかに依存しない)．待ち行列 i における任意の客の**待ち時間** (waiting time) W_i の分布関数の LST と平均をそれぞれ $W_i^*(s)$ と $E[W_i]$ で表す．サーバの巡回時間と客の待ち時間はポーリングシステムの主要な**性能指標** (performance measure) である．

すべての待ち行列の動作が統計的に同じであるとき，システムは**対称** (symmetric) であるという．対称なシステムに対する結果は，以下の方程式等において $\rho_i = \rho/N = \lambda b$ とし他のすべてのパラメタの添字 i を省略することにより得られる．

2.1 単一バッファシステム

それぞれの待ち行列で一時点に高々1人の客しか収容できないシステムを**単一バッファシステム** (single buffer system) という．このシステムでは，客の到着時に待ち行列が占有されていると到着した客は消失すると仮定する．

各サービス時間が定数 b であり，かつ全歩行時間も定数 R であるような対称な単一バッファモデルについては，平均巡回時間 $E[C]$ と平均待ち時間 $E[W]$ に対して次のような閉じた式が得られている [44, 50]．

$$E[C] = R + E[Q]b \tag{3}$$

$$E[W] = (N-1)b - \frac{1}{\lambda} + \frac{NR}{E[Q]} \tag{4}$$

ここで

$$E[Q] = \frac{N\displaystyle\sum_{n=0}^{N-1}\binom{N-1}{n}\prod_{j=0}^{n}\left[e^{\lambda(R+jb)}-1\right]}{1+\displaystyle\sum_{n=1}^{N}\binom{N}{n}\prod_{j=0}^{n-1}\left[e^{\lambda(R+jb)}-1\right]} \tag{5}$$

は1回の巡回時間の間にサービスする客の平均数である．図2に $b = R = 1$ の場合の平均待ち時間 $E[W]$ を $\rho = N\lambda b$ に対して示す ($N = \infty$ は連続ポーリングモデルを表す)．N が有限の場合は，$\rho \to \infty$ のとき $E[W] \approx R + (N-1)b$ となる．

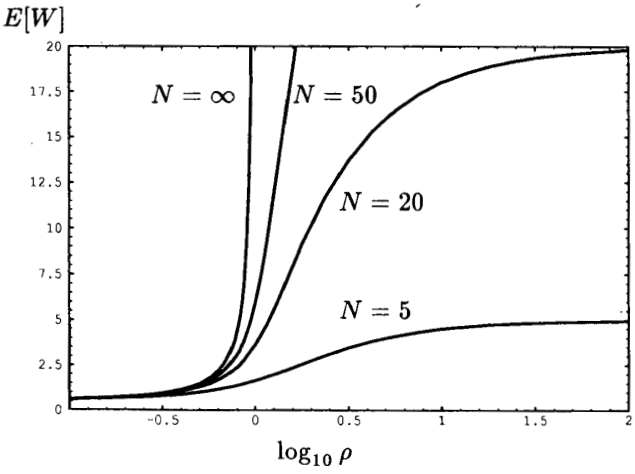

図 2: 対称な単一バッファポーリングモデルにおける平均待ち時間 $E[W]$ $(b = R = 1)$.

非対称な単一バッファシステムについては性能指標が閉じた形に得られていないが，次のような方程式の解として数値的に計算することが可能である．　サーバがk番目に訪れる待ち行列にとっての**駐在時間** (station time) ω_kを，待ち行列$k-1$から待ち行列kへの歩行時間と待ち行列kでのサービス時間 (もし客がいれば) の和と定義する．　ここで番号kはすべての待ち行列についての通し番号で，サーバが1つの待ち行列を訪れる度に1ずつ増えるものとする．　相次ぐN個の駐在時間$\omega_{k-N+1}, \omega_{K-N+2}, \ldots, \omega_k$は互いに独立でないので，それらの結合分布関数のLSTを

$$\Omega_k^*(s_1, \cdots, s_N) := E\left[\exp\left(-\textstyle\sum_{j=1}^{N} \omega_{k-N+j} s_j\right)\right] \quad (6)$$

で定義すると，これは関数方程式

$$\begin{aligned}\Omega_k^*(s_1, \ldots, s_N) &= R_{k-1}^*(s_N + \lambda_k)[1 - B_k^*(s_N)] \\ &\times \Omega_{k-1}^*(0, s_1 + \lambda_k, \ldots, s_{N-1} + \lambda_k) \\ &+ R_{k-1}^*(s_N) B_k^*(s_N) \Omega_{k-1}^*(0, s_1, \ldots, s_{N-1}) \\ &k = 1, 2, \ldots\end{aligned} \quad (7)$$

を満たす [35].

サーバが各待ち行列を去る時刻をその待ち行列にとっての巡回時間の開始点と考えると，k番目に訪問される待ち行列にとっての巡回時間C_kの分布関数のLSTは

$$C_k^*(s) = \Omega_k^*(s, \ldots, s) \quad (8)$$

で表される．各待ち行列をサーバが去ってから次にまた来るまでの時間をサーバの**訪問時間間隔** (intervisit time) という (「巡回時間」との違いは，訪問時間間隔は当該待ち行列でのサービス時間を含まないことである)．待ち行列kの訪問時間間隔I_kの分布関数のLSTは

$$I_k^*(s) = \Omega_{k-1}^*(0, s, \cdots, s) R_{k-1}^*(s) \quad (9)$$

で与えられる．これを使うと，k番目に訪問される待ち行列での客の待ち時間の分布関数のLSTと平均は

$$W_k^*(s) = \frac{\lambda_k [I_k^*(s) - I_k^*(\lambda_k)]}{(\lambda_k - s)[1 - I_k^*(\lambda_k)]} \quad (10a)$$

$$E[W_k] = \frac{E[I_k]}{1 - I_k^*(\lambda_k)} - \frac{1}{\lambda_k} \quad (10b)$$

と書くことができる．

式(7)より$C_k^*(s)$や$W_k^*(s)$を決めるのに必要な$N(2^{N-1} - 1)$個の未知数に対する同数の連立一次方程式が得られるので，原理的には数値計算により$E[C]$や$E[W_k]$を求めることができる．関数$\Omega_k^*(\cdot)$の引数を増やすことにより，これらの方程式の数を$O(2^N)$個にすることもできる (そのかわり係数が複雑になる) [63].

対称モデルにおいてRと$\rho = N\lambda b$を一定値に保ちながら$N \to \infty$としたシステムを**連続ポーリングモデル** (continuous polling model) という．これは，客が円周上の任意の地点に等確率で到着し，サーバが円周に沿って動きながら出会った客をサービスするというモデルである．このモデルでは訪問時間間隔は巡回時間に等しい．歩行速度が一定の場合，平均巡回時間と平均待ち時間は

$$E[C] = \frac{R}{1 - \rho} \quad ; \quad E[W] = \frac{R}{2(1 - \rho)} + \frac{\rho b^{(2)}}{2b(1 - \rho)} \quad (11)$$

で与えられる．さらにサービス時間が一定の場合，巡回時間と待ち時間の分布関数のLSTは

$$C^*(s) = e^{-sR}\left(\frac{1 - \rho}{1 - \rho e^{-sR}}\right)^{R/b} \quad (12a)$$

$$W^*(s) = \frac{1 - C^*(s)}{E[C]s} \quad (12b)$$

で与えられる [20]．サービス時間が確率変数の場合も確率積分と点過程の理論を用いて研究されている [38].

それぞれの待ち行列で一時点にいくらでも多くの客を収容できるシステムを**無限バッファシステム** (infinite buffer system) という．このシステムでは客の損失は無い．無限バッファシステムの基本モデルとして，サーバが各待ち行列で継続してサービスできる最大の客数に応じて，4つのシステムを考える．第1の**全処理式サービス** (exhaustive service) システムでは，サーバは待ち行列が空になるまでサービスを続ける．サービス期間中にその待ち行列に到着する客も同じサービス期間内にサービスされる．第2の**ゲート式サービス** (gated service) システムでは，サーバが待ち行列に来た時点でそこにいる客だけをサービス期間中にサービスし，その期間内に新たに到着する客のサービスは一巡後のサービス期間に行なわれる．第3の**制限式サービス** (limited service) システムでは，サーバが待ち行列に来たとき待ち行列が空であれば素通りし，空でなければ1人の客だけをサービスして次の待ち行列へ移動する．第4の**減少式サービス** (decrementing service) システムでは，サーバが待ち行列に来たとき待ち行列が空であれば素通りし，空でなければ客数が1だけ少なくなるまでサービスを続ける．

これらのシステムが安定であるための必要十分条件は，それぞれ次のように与えられる．

全処理式: $\rho < 1$

ρ	全処理式	ゲート式	制限式	減少式
.00	.5500	.5500	.5500	.5500
.05	.6000	.6053	.6085	.6031
.10	.6556	.6667	.6742	.6628
.15	.7176	.7353	.7485	.7302
.20	.7857	.8125	.8333	.8070
.25	.8667	.9000	.9310	.8953
.30	.9571	1.000	1.045	.9980
.35	1.062	1.115	1.179	1.119
.40	1.183	1.250	1.339	1.263
.45	1.327	1.409	1.535	1.438
.50	1.500	1.600	1.778	1.655
.55	1.711	1.833	2.089	1.931
.60	1.975	2.125	2.500	2.294
.65	2.314	2.500	3.070	2.793
.70	2.767	3.000	3.913	3.523
.75	3.400	3.700	5.286	4.690
.80	4.350	4.750	7.917	6.858
.85	5.933	6.500	15.00	12.27
.90	9.100	10.00	100.0	50.00
.95	18.60	20.50	-	-

表 1: 対称な無限バッファポーリングモデルにおける平均待ち時間 $E[W]$ $(N = 10, r = 0.1, \delta^2 = 0.01, b = 1, b^{(2)} = 1)$.

ゲート式: $\rho < 1$

制限式: すべての i について $\rho + \lambda_i R < 1$

減少式: すべての i について $\rho + \lambda_i(1-\rho_i)R < 1$

これらのの厳密な証明は多次元Markov過程のエルゴード性にかかわり簡単ではない [54].

2.2 無限バッファシステム

4つのサービス方式を含む広範囲の無限バッファシステムにおいて, 平均巡回時間は簡単に

$$E[C] = \frac{R}{1-\rho} \tag{13}$$

で与えられる. 対称な4つの基本モデルにおける平均待ち時間は次のように与えられる.

全処理式: $$E[W]_e = \frac{\delta^2}{2r} + \frac{N\lambda b^{(2)} + r(N-\rho)}{2(1-\rho)} \tag{14a}$$

ゲート式: $$E[W]_g = \frac{\delta^2}{2r} + \frac{N\lambda b^{(2)} + r(N+\rho)}{2(1-\rho)} \tag{14b}$$

制限式:
$$E[W]_l = \frac{\delta^2}{2r} + \frac{N\lambda b^{(2)} + r(N+\rho) + N\lambda\delta^2}{2(1-\rho-N\lambda r)} \tag{14c}$$

減少式:
$$E[W]_d = \frac{\delta^2}{2r} + \frac{N\lambda b^{(2)}(1-\lambda r) + (r+\lambda\delta^2)(N-\rho)}{2[1-\rho-\lambda r(N-\rho)]} \tag{14d}$$

($E[W]_e$ と $E[W]_g$ は橋田温氏と中村義作氏が導出した [32, 33]. 制限式サービスシステムの厳密な解析は野村雅行・塚本克治両氏によりなされたが [46], 上の $E[W]_l$ の式はFuhrmann, Watsonおよび筆者がほぼ同時期に独立に導いた [29, 56, 66]. $E[W]_d$ は筆者による [55].)

表1にこれらの平均待ち時間の数値例を示す. 式(14)の形または数値例を見ると

$$E[W]_e < E[W]_g < E[W]_l$$
$$E[W]_e < E[W]_d < E[W]_l$$

の関係が成り立つことがわかる. また, 負荷が低いとき $E[W]_g > E[W]_d$ であり, 負荷が高いとき $E[W]_g < E[W]_d$ である. なお, 式(14)において $\delta^2 = 0$ とし, $R = Nr$ と $\rho = N\lambda b$ を一定値に保ちながら $N \to \infty$ とすると, 4つの式はいずれも連続ポーリングモデルに対する式(11)に一致する.

非対称な全処理式とゲート式サービスシステムについては, 各待ち行列での平均待ち時間を数値的に厳密に計算する方法が確立されている. 全処理式サービスシステムの場合, 待ち行列 i での平均待ち時間 $E[W_i]$ は, 訪問時間間隔 I_i の平均と2次モーメントを用いて

$$E[W_i] = \frac{\lambda_i b_i^{(2)}}{2(1-\rho_i)} + \frac{E[(I_i)^2]}{2E[I_i]} \tag{15}$$

と表すことができる. ここで

$$E[I_i] = \frac{R(1-\rho_i)}{1-\rho} \tag{16a}$$

$$E[(I_i)^2] = (E[I_i])^2 + \delta_{i-1}^2 + \frac{1-\rho_i}{\rho_i}\sum_{\substack{j=1\\(j\neq i)}}^{N} r_{ij} \tag{16b}$$

であり, r_{ij} は待ち行列 i と j の駐在時間 ω_i と ω_j の共分散である. 全処理サービスシステムにおける待ち行列 i の駐在時間は, 待ち行列 $i-1$ のサービス期間の終了から待ち行列 i のサービス期間の終了までの時間であると定義される. N^2 個の未知数 $\{r_{ij}; i, j = 1, 2, \cdots, N\}$ は次の N^2 個の連立1次方程式の解として数値計算で求められる [28].

$$r_{ij} = \frac{\rho_i}{1-\rho_i}\left(\sum_{m=i+1}^{N} r_{jm} + \sum_{m=1}^{j-1} r_{jm} + \sum_{m=j}^{i-1} r_{mj}\right)$$
$$j < i \tag{17a}$$

$$r_{ij}=\frac{\rho_i}{1-\rho_i}\left(\sum_{m=i+1}^{j-1}r_{jm}+\sum_{m=j}^{N}r_{mj}+\sum_{m=1}^{i-1}r_{mj}\right)\qquad j>i \tag{17b}$$

$$r_{ii}=\frac{\delta_{i-1}^2}{(1-\rho_i)^2}+\frac{\lambda_i b_i^{(2)}E[I_i]}{(1-\rho_i)^3}+\frac{\rho_i}{1-\rho_i}\sum_{\substack{j=1\\(j\neq i)}}^{N}r_{ij} \tag{17c}$$

さらに，係数ははるかに複雑であるが，同じものを計算するための N 個の連立1次方程式も導かれている [49]．ゲート式サービスシステムについても同様の解法が得られている．

非対称システムにおいて，高負荷の待ち行列と低負荷の待ち行列での任意の客の平均待ち時間を比べた興味深い結果が報告されている．全処理式サービスシステムでは，高負荷の待ち行列のほうが低負荷の待ち行列よりも平均待ち時間が短い．それは，サーバが高負荷の待ち行列に長く滞在しがちであるので，そこに到着する多くの客はそのサービス期間のうちにサービスされるからであると説明されている [28]．ゲート式および制限式サービスシステムにおいては，逆の傾向が観測される．すなわち，高負荷の待ち行列に到着する多くの客は，少なくともサーバが一巡するのを待たなければならないので，平均待ち時間が長くなる [34]．

制限式と減少式サービスシステムに対する同様の厳密な解析は今までのところ成功していない．最も簡単な2つの待ち行列から成る制限式または減少式サービスシステムに対して，待ち行列長の結合確率分布の母関数の方程式を複素平面上の境界値問題に帰着させる厳密解法が知られているだけである [10]．制限式サービスシステムは応用上よく現れるので，そこでの待ち時間を評価するために多くの近似解法が提案されている．代表的なものは，条件付き巡回時間の考えを用いる数値計算法 [39] と，下記の擬保存則を満たすように調整して作られる陽公式である．後者の1つの例は

$$E[W_i]\approx\frac{1-\rho+\rho_i}{1-\rho-\lambda_i R}\cdot\frac{1-\rho}{(1-\rho)\rho+\sum_{j=1}^{N}\rho_j^2}\cdot\left[\frac{\rho}{2(1-\rho)}\sum_{j=1}^{N}\lambda_j b_j^{(2)}+\frac{\rho\Delta^2}{2R}+\frac{R}{2(1-\rho)}\sum_{j=1}^{N}\rho_j(1+\rho_j)\right] \tag{18}$$

である [13]．この式はLANの階層モデルにおけるトークンリング部分のサブモデルに使われた [45]．

非対称な基本モデルにおける各待ち行列の厳密な平均待ち時間は，全処理式とゲート式サービスシステムでは数値的に求めることができても陽な表現式は得られていないし，制限式および減少式サービスシステムについては数値計算法さえ無い．しかしながら，驚くべきことに，各待ち行列の平均待ち時間の適当な重み付け平均が簡単な式で表されることが知られている．待ち行列毎にサービス方式が異なってもよいシステムについて，その式は

$$\begin{aligned}&\sum_{i\in E,G}\rho_i E[W_i]+\sum_{i\in L}\rho_i\left(1-\frac{\lambda_i R}{1-\rho}\right)E[W_i]\\&+\sum_{i\in D}\rho_i\left[1-\frac{\lambda_i(1-\rho_i)R}{1-\rho}\right]E[W_i]\\&=\frac{\rho\sum_{i=1}^{N}\lambda_i b_i^{(2)}}{2(1-\rho)}+\rho\frac{\Delta^2}{2R}+\frac{R\left(\rho-\sum_{i=1}^{N}\rho_i^2\right)}{2(1-\rho)}\\&+\frac{R\sum_{i\in G,L}\rho_i^2}{1-\rho}-\frac{R\sum_{i\in D}\rho_i\lambda_i^2 b_i^{(2)}}{2(1-\rho)}\end{aligned} \tag{19}$$

で与えられる．ここで，E,G,L および D はそれぞれ全処理式，ゲート式，制限式および減少式サービス方式をもつ待ち行列の番号の集合を表す [9]．この式は**擬保存則** (pseudoconservation law) とよばれる．その理由は，もし歩行時間がすべて0なら，客がいる限りサーバは必ずサービスを行なうという**仕事量保存** (work conserving) システムに対する**Kleinrockの保存則** (Kleinrock's conservation law) に帰着するからである．擬保存則から，対称システムにおける平均待ち時間 (14) が直ちに得られる．

3. 拡張モデル

前節で述べた基本モデルの条件を様々に変形・拡張したモデルの解析が多くの研究者によってなされている．それらの変形・拡張モデルには，現実のシステムの特徴をより忠実に採り入れようとする試みがある一方で，単に数学的練習問題のようなものも多い．

以下では，いくつかの本質的な拡張モデルについて最近の研究成果をまとめる．

3.1 客の動き方に関する拡張

基本モデルでは，客はそれぞれの待ち行列に独立なPoisson過程で到着し，サービス終了後直ちにシステムから退去すると仮定された．到着過程を集団Poisson到着に拡張するとき，各待ち行列において独立な集団Poisson到着がある場合と，システム全体に対して集団Poisson到着があり客は確率的に各待ち行列に振り分けられる場合とが考えられるが，どちらの場合も基本モデルと同様の解

析が可能である [7, 42]. マルチメディアトラヒックをモデル化するといわれるMarkov変調Poisson到着過程をもつポーリングモデルの厳密解は得られていない. 再生過程や流体型の到着過程をもつポーリングモデルはいくつかの近似解が提案されている.

バッファの大きさが1と無限大の中間の有限値である(到着時にバッファが一杯であれば受け入れられない)システムや, 客の母集団が有限である(サービスが終わると母集団に帰る)システムは, 状態空間が有限であるので, 複雑ではあるが厳密解が可能である. 後者は, 通信ネットワークにおける**流量制御** (flow control) のための**ウィンドウ制御** (window control) および**等数制御** (isarithmic control) 方式のモデルとして使うことができる.

1つの待ち行列でのサービスを終了した客が確率σでもとの待ち行列にもどり, 確率$1-\sigma$でシステムから退去する($0 \leq \sigma < 1$)という**Bernoulli フィードバック** (Bernoulli feedback) は, サービスのやり直し(たとえば伝送エラーによる再送信)や分割サービス(たとえば大きなファイルをブロック化して送る)のモデルとなる [62, 63]. さらに, サービスを終えた客が他の待ち行列に移動する可能性も考慮に入れると, **待ち行列網** (network of queues) を1つのサーバが巡回的にサービスするというモデルが得られる. 全処理式およびゲート式サービスをもつ待ち行列網は厳密解が可能である [52].

3.2 サービス方式の拡張

基本モデルでは, システムのすべての待ち行列において, 全処理式, ゲート式, 制限式またはゲート式のうちどれか1つのサービス方式が採用されると仮定した. それぞれの方式の拡張がいくつも提案されているが, 応用上重要な拡張は, トークンリングLANの動作で各ノードからの継続送信に最大値が設定されていることに対応する, 個数制限式と時間制限式サービスである. **個数制限式サービス** (number-limited service) では, 各待ち行列で継続してサービスされる客数の最大値がkで与えられる. (理論上, kの値は待ち行列毎に異なってもよい. このk個にサービス期間中に新たに到着する客を含めるか否かによって, 「全処理型k-制限式」と「ゲート型k-制限式」が区別される. これらは, $k=1$の場合が基本モデルの制限式になり, $k=\infty$の場合がそれぞれ全処理式とゲート式になる.) **時間制限式サービス** (time-limited service) では, 各待ち行列においてサービス期間の長さが制限される. 客のサービスの途中で制限時間に達したとき, サービスを打ち切る場合と, FDDIの**非同期オーバーラン** (asynchronous overrun) のように, そのサービスだけは最後まで続ける場合が考えられる. このような拡張された制限式サービスシステムは近似的に解析されている [25].

数学的な拡張の1つは, **Bernoulli スケジューリング** (Bernoulli scheduling) とよばるもので, サーバがある待ち行列において1人の客のサービスを終えたときにまだ他の客が同じ待ち行列にいる場合, 確率pで再びその待ち行列でサービスを行ない, 確率$1-p$で次の待ち行列に移動する($0 \leq p \leq 1$. $p=0$が基本モデルの制限式に, $p=1$が全処理式に対応する). 非対称なBernoulliスケジューリングシステムについて, 各待ち行列での平均待ち時間はもちろん得られていないが, 擬保存則は得られており, したがって対称システムにおける平均待ち時間も得られる [64].

客にサービスの優先順位のクラスがあるシステムについては, 優先順位に基づく処理が各待ち行列内部で行なわれる場合は簡単に厳密解が得られる(バケーションを取るサーバをもつ優先処理待ち行列システムの解を使う) [51] が, トークンリングやFDDIの仕様に近いシステム全体に渡っての優先順位に基づく処理をするモデルは近似解のみが提案されている.

ゲート式サービスの変形として, 大局的ゲート式とCRMAを紹介する. **大域的ゲート式サービス** (globally gated service) では, システム内に「主待ち行列」とよばれる特定の待ち行列が1つあり, 待ち行列iでのサービスはサーバが最後に主待ち行列を訪問したときに既に待ち行列iにいた客についてのみ行なわれ, その後に到着した客は次回まで待たされる. このシステムに対しては平均待ち時間が陽に得られている [12]. この方式を拡張して, システム内の待ち行列がn個の主待ち行列とそれぞれの主待ち行列に従属する待ち行列に分割されている方式を**同期ゲート式サービス** (synchronous gated service) という($n=N$の場合が基本モデルのゲート式であり, $n=1$の場合が大局的ゲート式である) [36]. また, **巡回式予約多重アクセス** (cyclic reservation multiple access: CRMA) とよばれる方式では, ちょうど郵便集配車が各地のポストをまわって郵便物を集め郵便局で処理をするように, サーバがシステム内を一巡して客を集めてから1人ずつサービスするもので, 一方向回線で接続された高速通信ネットワークのプロトコルのモデルとして提案された [12].

3.3 サーバの動き方

基本モデルでは, サーバはすべての待ち行列を番号順に巡回すると仮定された. 解析されている非巡回サービス

システムを次のように分類することができる．(1) **固定的** (deterministic) な動き．たとえば，エレベータや磁気ディスクのアームのようにサーバが往復する方式をとくに **SCAN** という．(2) **確率的** (probabilistic) な動き．待ち行列 i の次に確率 p_{ij} で待ち行列 j が訪問される方式を **Markov 的経路選択** (Markovian routing) という．および (3) **状態依存的** (state-dependent) な動き．たとえば，最も長い待ち行列に移動するサーバや，最も近くの客のサービスに向かう**貪欲** (greedy) なサーバ [31]，あるいはシステム内に客がいなくなると基地に帰るとか最後にサービスした待ち行列のところに止まるサーバ [27] 等が考えられる．SCAN システムで大局的ゲート式サービスを用いると，すべての待ち行列の平均待ち時間が同じになるという特異な結果が報告されている [3]．

もしサーバの動きを制御することができれば，適当な目的関数を設定してそれを最適化するようにサーバを動かすことが考えられる．このような最適化には，静的，半動的および動的の 3 通りがある．たとえば，平均巡回時間や任意時刻におけるシステム内の全客数あるいは全負荷が目的関数とされる．

静的最適化 (static optimization) では，システムの稼働前にサーバの最適な動き方を決め，そのとおりにサーバを動かす．サーバが訪れる待ち行列の番号順は**ポーリング順序表** (polling order table) で表される．全処理式またはゲート式サービスシステムに対し，$\sum_{i=1}^{N} \rho_i E[W_i]$ を最小にするように，ポーリング順序表の大きさと待ち行列番号の頻度と順序を最適化する方法が提案されている [11]．

半動的最適化 (semidynamic optimization) では，各巡回前にその時点でのシステムの状態に応じて巡回順序を決める．たとえば，歩行時間の無い全処理式およびゲート式システムにおいては，各巡回直前の待ち行列 i の客数を L_i とすると，L_i/λ_i の昇順にまわるとき平均巡回時間が最小になることが知られている [15]．

動的最適化 (dynamic optimization) では，1 人の客のサービスを終了する度にサーバの最適な動きを決める．たとえば，対称なシステム内の全客数を任意の時刻において最小にするためには，(i) 1 つずつの待ち行列で空になるまでサービスを続ける，(ii) その待ち行列が空になれば，システム内で最も客数の多い待ち行列へ移動する，そして (iii) システム全体が空になれば，最後にサービスした待ち行列のところに止まっている，というのが最適である [43]．半動的および動的最適化問題は **Markov 決定過程** (Markov decision process) として解かれる．

また，閉じた待ち行列網を 1 つのサーバがサービスするときには，サービス中の待ち行列においてのみシステムの状態が変化するので，どのようにサーバを動かすかという問題は**複数の腕をもつスロットマシンの問題** (multi-armed bandit problem) として定式化される [65]．

3.4 その他のシステム

最後にその他の代表的なモデルの研究状況を紹介する．

歴史的には，歩行時間の無い全処理式およびゲート式サービスモデルが既に 1960 年代に解析されている [4, 22, 23]．第 2 節で見たように，歩行時間が有るモデルの解析では，サーバの 1 巡回時間に渡ってシステムの変数を評価する方法が取られる．ところが，歩行時間が零のモデルでは，システム内に客がいなくなる (安定なシステムではそのようなことが確率 1 で無限回起こる) と，サーバは有限時間に無限回の巡回を行なうので平均巡回時間が零となり，巡回時間に基づく解析ができない．このような理由により，歩行時間が無いモデルと有るモデルは別々に取り扱われてきたが，最近になり 2 つのモデルにおける平均待ち時間が関連付けられるようになった [19, 24]．

システムの時間軸が**スロット** (slot) とよばれる一定の間隔で区切られ，システムの動作はスロットの境界のみで起こると仮定する**離散時間** (discrete time) モデルも早くから解析された [37]．離散時間ポーリングモデルは本質的に客の集団到着がある連続時間モデルと同様に扱うことができるので，対応する結果が得られている．

通信システムやその他の現実のシステムにおいては効率向上のため複数のサーバをよく用いるが，複数のサーバをもつポーリングモデルの解析や最適化は数値計算によってのみ可能である．**一般化確率 Petri ネット** (generalized stochastic Petri nets) を用いた研究 [1, 2] によれば，サービス率の総和が与えられているとき，(高速都市型通信ネットワークに典型的である) 長い歩行時間と高負荷をもつシステムは，複数のサーバを使うことによって平均待ち時間を大きく減らすことができる．また，1 つの待ち行列だけが高負荷であるシステムを 2 つのサーバが巡回サービスを行うと，高負荷の待ち行列が 1 つのサーバを独占し，他のすべての待ち行列が他方のサーバを共同使用するという傾向が発見されている．さらに，多くの待ち行列をもつシステムを少数のサーバでサービスするとき，サーバの最適な動き方は巡回サービスであることが確認されている．複数サーバのポーリングモデルに適用できる別の数値計算法として，システムの状態変数と方程式をすべて ρ のべき級数に展開し，その係数に対する簡単化され

た方程式を数値的に解くというべき級数法 (power-series algorithm) が提案されている [6].

4. おわりに

1968 年の *Scientific American* 誌に掲載された非専門家向けの記事 [40] で「興味深い重要な待ち行列モデル」として紹介されたポーリングモデルは, 30 年間に渡り, 解析・最適化・応用について多くの人々により研究されてきた. 筆者が個人的に集めた約 760 篇 (平成 7 年 10 月現在) の参考文献のリストはインターネットのワールドワイドウェブ (WWW) で http://www.sk.tsukuba.ac.jp/~takagi/polling.html を開くと見ることができる. ポーリングモデルの解析において, 日本での研究が早い時期からインパクトのある貢献をしたことは特筆に値する. 今後の課題は, マルチメディアトラヒックの到着過程をもつシステムや複数のサーバをもつシステムの研究であると思われる.

謝辞

本論文の研究は (財) 電気通信普及財団の平成 7 年度助成を受けています. また, 査読者の方々の有益なコメントに感謝します.

参考文献

[1] Ajmone Marsan, M., Donatelli, S., and Neri, F., GSPN models of Markovian multiserver multiqueue systems. *Performance Evaluation*, Vol.11, No.4, pp.227–240, November 1990.

[2] Ajmone Marsan, M., Donatelli, S., Neri, F., and Rubino, U., Good and bad dynamic polling orders in symmetric single buffer Markovian multiserver multiqueue systems. *IEEE INFOCOM '93*, pp.176–185, San Francisco, California, March 30–April 1, 1993.

[3] Altman, E., Khamisy, A., and Yechiali, U., On elevator polling with globally gated regime. *Queueing Systems*, Vol.11, No.1–2, pp.85–90, July 1992.

[4] Avi-Itzhak, B., Maxwell, W. L., and Miller, L.W., Queueing with alternating priorities. *Operations Research*, Vol.13, No.2, pp.306–318, March–April, 1965.

[5] Bertsimas, D. J., and Van Ryzin, G., Stochastic and dynamic vehicle routing in the Euclidean plane with multiple capacitated vehicles. *Operations Research*, Vol.41, No.1, pp.60–76, January–February 1993.

[6] Blanc, J. P. C., Performance evaluation of polling systems by means of the power-series algorithm. *Annals of Operations Research*, Vol.35, No.1–4, pp.155–186, April 1992.

[7] Boxma, O. J., Workloads and waiting times in single-server systems with multiple customer classes. *Queueing Systems*, Vol.5, No.1–3, pp.185–214, November 1989.

[8] Boxma, O. J., Analysis and optimization of polling systems. In: *Queueing, Performance and Control in ATM (ITC-13)*, J. W. Cohen and C. D. Pack (editors), pp.173–183, Elsevier Science Publishers B.V. (North-Holland), Amsterdam, 1991.

[9] Boxma, O. J., and Groenendijk, W. P., Pseudo-conservation laws in cyclic-service systems. *Journal of Applied Probability*, Vol.24, No.4, pp.949–964, December 1987.

[10] Boxma, O. J., and Groenendijk, W. P., Two queues with alternating service and switching times. In: *Queueing Theory and Its Applications - Liber Amicorum for J. W. Cohen*, O. J. Boxma and R. Syski (editors), pp.261–282, Elsevier Science Publishers B.V. (North-Holland), Amsterdam, 1988.

[11] Boxma, O. J., Levy, H., and Weststrate, J. A., Efficient visit orders for polling systems. *Performance Evaluation*, Vol.18, No.2, pp.103–123, September 1993.

[12] Boxma, O. J., Levy, H., and Yechiali, U., Cyclic reservation schemes for efficient operation of multiple-queue single-server systems. *Annals of Operations Research*, Vol.35, No.1–4, pp.187–208, April 1992.

[13] Boxma, O. J., and Meister, B., Waiting-time approximations for cyclic-service systems with

switchover times. *Performance Evaluation*, Vol.7, No.4, pp.299–308, November 1987.

[14] Bozer, Y. A., and Srinivasan, M. M., Tandem configurations for automated guided vehicle systems and the analysis of single vehicle loops. *IIE Transactions*, Vol.23, No.1, pp.72-82, March 1991.

[15] Browne, S., and Yechiali, U., Dynamic priority rules for cyclic-type queues. *Advances in Applied Probability*, Vol.21, No.2, pp.432–450, June 1989.

[16] Bux, W., Local-area subnetworks: A performance comparison. *IEEE Transactions on Communications*, Vol.COM-29, No.10, pp.1465–1473, October 1981.

[17] Bux, W., Token-ring local-area networks and their performance. *Proceedings of the IEEE*, Vol.77, No.2, pp.238–256, February 1989.

[18] Campbell, G. M., Cyclic queueing systems. *European Journal of Operational Research*, Vol.51, No.2, pp.155–167, March 1991.

[19] Choudhury, G. L., Polling with a general service order table: Gated service. *IEEE INFOCOM '90*, pp.268–276, San Francisco, California, June 3-7, 1990.

[20] Coffman, E. G. Jr., and Gilbert, E. N., A continuous polling system with constant service times. *IEEE Transactions on Information Theory*, Vol.IT-33, No.4, pp.584–591, July 1986.

[21] Conti, M., Gregori, E., and Lenzini, L., Metropolitan area networks (MANs): protocols, modeling and performance evaluation. In: *Performance Evaluation of Computer and Communication Systems, Joint tutorial papers of Performance '93 and Sigmetrics '93*, L. Donatiello and R. Nelson (editors), pp.81–120, Lecture Notes in Computer Science 729, Springer-Verlag, Berlin, 1993.

[22] Cooper, R. B., Queues served in cyclic order: Waiting times. *The Bell System Technical Journal*, Vol.49, No.3, pp.399–413, March 1970.

[23] Cooper, R. B., and Murray, G., Queues served in cyclic order. *The Bell System Technical Journal*, Vol.48, No.3, pp.675–689, March 1969.

[24] Cooper, R. B., Niu, S.-C., and Srinivasan, M. M., A decomposition theorem for polling models: the switchover times are effectively additive. *Operations Research* に掲載予定.

[25] de Souza e Silva, E., Gail, H. R., and Muntz, R. R., Polling systems with server timeout. *IEEE/ACM Transactions on Networking* に掲載予定.

[26] Eisenberg, M., Queues with periodic service and changeover times. *Operations Research*, Vol.20, No.2, pp.440–451, March–April 1972.

[27] Eisenberg, M., The polling system with a stopping server. *Queueing Systems*, Vol.18, Nos.3,4, pp.387–431, November 1994.

[28] Ferguson, M. J., and Aminetzah, Y. J., Exact results for nonsymmetric token ring systems. *IEEE Transactions on Communications*, Vol.COM-33, No.3, pp.223–231, March 1985. 訂正: Choudhury, G. L., and Takagi, H., Comments on "Exact results for nonsymmetric token ring systems." *IEEE Transactions on Communications*, Vol.38, No.8, pp.1125–1127, August 1990.

[29] Fuhrmann, S. W., Symmetric queues served in cyclic order. *Operations Research Letters*, Vol.4, No.3, pp.139–144, October 1985.

[30] Grillo, D., Polling mechanism models in communication systems - Some application examples. In: *Stochastic Analysis of Computer and Communication Systems*, H. Takagi (editor), pp.659–698, Elsevier Science Publishers B. V. (North-Holland), Amsterdam, 1990.

[31] Harel, A., and Stulman, A., Polling, greedy and horizon servers on a circle. *Operations Research*, Vol.43, No.1, pp.177–186, January–February 1995.

[32] Hashida, O., Analysis of multiqueue. *Review of the Electrical Communication Laboratories*, Vol.20, No.3–4, pp.189–199, March–April, 1972.

[33] 橋田温, 中村義作: 複数待ち行列の解析. 経営科学, 第13巻第1号 pp.30–47, 1969年10月.

[34] Ibe, O. C., and Cheng, X., Approximate analysis of asymmetric single-service token-passing systems. *IEEE Transactions on Communications*, Vol.37, No.6, pp.572–577, June 1989.

[35] Ibe, O. C., and Cheng, X., Performance analysis of asymmetric single-buffer polling systems. *Performance Evaluation*, Vol.10, No.1, pp.1–14, October 1989.

[36] Khamisy, A., Altman, E., and Sidi, M., Polling systems with synchronization constraints. *Annals of Operations Research*, Vol.35, No.1–4, pp.231–267, April 1992.

[37] Konheim, A. G., and Meister, B., Waiting lines and times in a system with polling. *Journal of the Association for Computing Machinery*, Vol.21, No.3, pp.470–490, July 1974.

[38] Kroese, D. P., and Schmidt, V., A continuous polling system with general service times. *The Annals of Applied Probability*, Vol.2, No.4, pp.906–927, 1992.

[39] Kuehn, P. J., Multiqueue systems with nonexhaustive cyclic service. *The Bell System Technical Journal*, Vol.58, No.3, pp.671–698, March 1979.

[40] Leibowitz, M. A., Queues. *Scientific American*, Vol.219, No.2, pp.96–103, August 1968.

[41] Levy, H., and Sidi, M., Polling systems: applications, modeling, and optimization. *IEEE Transactions on Communications*, Vol.38, No.10, pp.1750–1760, October 1990.

[42] Levy, H., and Sidi, M., Polling systems with simultaneous arrivals. *IEEE Transactions on Communications*, Vol.39, No.6, pp.823–827, June 1991.

[43] Liu, Z., Nain, P., and Towsley, D., On optimal polling policies. *Queueing Systems*, Vol.11, No.1–2, pp.59–83, July 1992.

[44] Mack, C., Murphy, T., and Webb, N. L., The efficiency of N machines uni-directionally patrolled by one operative when walking time and repair times are constants. *Journal of the Royal Statistical Society*, Series B, Vol.19, No.1, pp.166–172, 1957.

[45] Murata, M., and Takagi, H., Two-layer modeling for local area networks. *IEEE Transactions on Communications*, Vol.COM-36, No.9, pp.1022–1034, September 1988.

[46] 野村雅行, 塚本克治: ポーリングシステムのトラヒック解析. 電子通信学会論文誌, Vol.J61-B, No.7, pp.600–607, 1978年7月.

[47] Ozawa, T. 1990., Analysis of a multiqueue model for an ISDN access interface. *Performance Evaluation*, Vol.15, No.2, pp.65–76, June 1992.

[48] Rubin, I., and Baker, J. E., Media access control for high-speed local area and metropolitan area communication networks. *Proceedings of the IEEE*, Vol.78, No.1, pp.168–203, January 1990.

[49] Sarkar, D., and Zangwill, W. I., Expected waiting time for nonsymmetric cyclic queueing systems - Exact results and applications. *Management Science*, Vol.35, No.12, pp.1463–1474, December 1989.

[50] Scholl, M., and Potier, D., Finite and infinite source models for communication systems under polling. IRIA Rapport de Recherche, No.308, Institut de Rescherche en Informatique et en Automatique, Le Chesnay, France, 1978.

[51] Shimogawa, S., and Takahashi, Y., A note on the pseudo-conservation law for a multi-queue with local priority. *Queueing Systems*, Vol.11, No.1–2, pp.145–151, July 1992.

[52] Sidi, M., Levy, H., and Fuhrmann, S. W., A queueing network with a single cyclically roving server. *Queueing Systems*, Vol.11, No.1–2, pp.121–144, July 1992. 訂正: Correction to equation (5.6) in the paper: A queueing network with a single cyclically roving server. *Queueing Systems*, Vol.16, No.1–2, p.193, April 1994.

[53] Stidham, S., Jr., Optimal control of a signalized intersection, Part I: Introduction: Structure of intersection models, Part II: Determining the optimal switching policies, and Part III: Descriptive stochastic models. Technical Reports No. 94, 95 and 96, Department of Operations Research, Cornell University, Ithaca, New York, 1969.

[54] Szpankowski, W., Towards computable stability criteria for some multidimensional stochastic processes. In: *Stochastic Analysis of Computer and Communication Systems*, H. Takagi (editor), pp.131–172, Elsevier Science Publishers B. V. (North-Holland), Amsterdam, 1990.

[55] Takagi, H., Mean message waiting time in a symmetric polling system. In: *Performance '84*, E. Gelenbe (editor), pp.293–302, Elsevier Science Publishers B.V. (North-Holland), Amsterdam, 1985.

[56] Takagi, H., Mean message waiting times in symmetric multi-queue systems with cyclic service. *Performance Evaluation*, Vol.5, No.4, pp.271–277, November 1985.

[57] Takagi, H., *Analysis of Polling Systems.* The MIT Press, Cambridge, Massachusetts, 1986.

[58] Takagi, H., Queuing analysis of polling models. *ACM Computing Surveys*, Vol.20, No.1, pp.5–28, March 1988.

[59] Takagi, H., Queueing analysis of polling models: An update. In: *Stochastic Analysis of Computer and Communication Systems*, H. Takagi (editor), pp.267–318, Elsevier Science Publishers B. V. (North-Holland), Amsterdam, 1990.

[60] Takagi, H., Application of polling models to computer networks. *Computer Networks and ISDN Systems*, Vol.22, No.3, pp.193–211, October 1991.

[61] Takagi, H., Queueing analysis of polling models: progress in 1990–1994. In: *Frontiers in Queueing: Models, Methods and Problems*, J. H. Dshalalow (editor), CRC Press, 1995 (出版予定).

[62] Takine, T., Takagi, H., and Hasegawa, T., Sojourn times in vacation and polling systems with Bernoulli feedback. *Journal of Applied Probability*, Vol.28, No.2, pp.422–432, June 1991.

[63] Takine, T., Takahashi, Y., and Hasegawa, T. Average message delay of an asymmetric single-buffer polling system with round-robin scheduling of services. In: *Modelling Techniques and Tools for Computer Performance Evaluation*, D. Potier and R. Puigjaner (editors), pp.179–187, Plenum Publishing Corporation, New York, 1989.

[64] Tedijanto, Exact results for the cyclic-service queue with a Bernoulli schedule. *Performance Evaluation*, Vol.11, No.2, pp.107–115, July 1990.

[65] Walrand, J., Queueing networks. In: *Handbooks in Operations Research and Management Science, Volume 2: Stochastic Models*, D. P. Heyman and M. J. Sobel (editors), pp.519–603, Elsevier Science Publishers B.V. (North-Holland), Amsterdam, 1990.

[66] Watson, K. S., Performance evaluation of cyclic service strategies – A survey. In: *Performance '84*, E. Gelenbe (editor), pp.521–533, Elsevier Science Publishers B.V. (North-Holland), Amsterdam, 1985.

[67] Yechiali, U., Analysis and control of polling systems. In: *Performance Evaluation of Computer and Communication Systems, Joint tutorial papers of Performance '93 and Sigmetrics '93*, L. Donatiello and R. Nelson (editors), pp.630–650, Lecture Notes in Computer Science 729, Springer-Verlag, Berlin, 1993.