待ち行列

1. 待ち行列モデルの分類

待ち行列 · · · 用意された「窓口」でサービスを受けるために, 複数の「客」が並んでいる状態.
コンピュータシステムにおいて

- マルチプログラミングにおいて, プロセッサを巡って実行可能プロセスが競合する.
  $\longrightarrow$ プロセッサが窓口, プロセスが客

- OLTP[*1]におけるトランザクション処理[*2]
  $\longrightarrow$ 処理プログラムが窓口, 各トランザクションが客

といったように, 待ち行列でモデル化できる事象が多く存在する. これらについてはモデルを数学的に解析することで, 所要時間などを予測することができる.

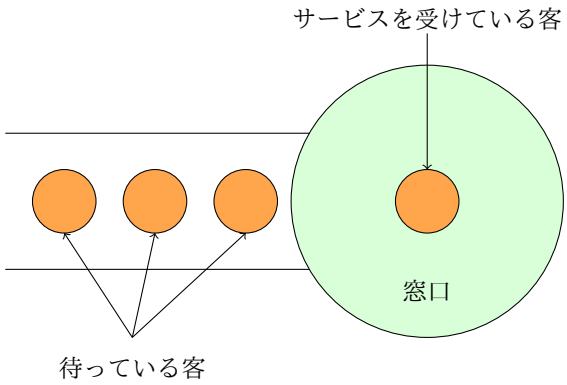

図 1 待ち行列の例

待ち行列モデルは, 窓口に到着する客の頻度分布, 窓口の数などによって, いくつかに分類することができる. ここでよく用いられる表記法に, ケンドール記法がある. これはその待ち行列の性質を

到着の分布 / サービス時間の分布 / 窓口の数

の順で記す方法である.

[*1] コンピュータシステムの処理方式の一種で, 互いに関連する複数の処理を一体化して確実に実行するトランザクション処理を, 端末などからの要求に基づいて即座に実行する方式
[*2] transaction. 預金口座への入出金や電車の座席予約といった一連の不可分な処理単位のことで, データベースの参照や更新を伴うことが一般的

到着およびサービス時間の分布を表す記号としては, 主に以下のものが用いられる.

＜到着＞

M[*3] : ポアソン分布 (到着がランダム)

G : 一般分布 (M と D の中間に該当する. 何らかの係数に従った通常の分布)

D : 一様分布 (到着が一定)

＜サービス時間＞

M : 指数分布 (サービス時間がランダム)

G : 一般分布 (M と D の中間に該当する, 何らかの係数に従った通常の分布)

D : 一様分布 (サービス時間が一定)

たとえば, 到着がランダム (=到着がポアソン分布に従う), サービス時間が一定で, 窓口が 1 つの待ち行列は, ケンドール表記では

M/D/1

と表記される.

◇ 窓口数の考え方

ケンドール記法において「窓口数が複数である」とは, 複数窓口に対して 1 本の待ち行列ができることを表す.

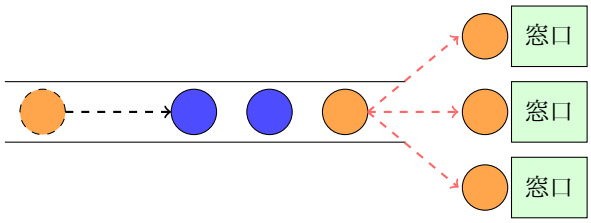

図 2 複数窓口モデル

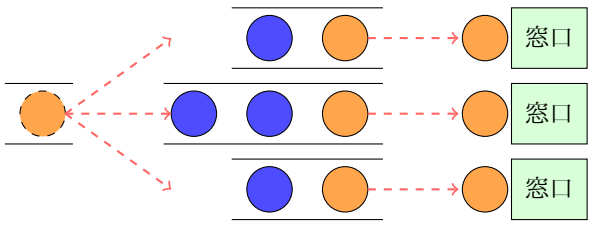

図 3 単一窓口モデル

図 3 においては図 2 と異なり, 単一窓口モデルが複数 (M/M/1 の待ち行列が複数) あると解釈する.

[*3] markovian の頭文字. マルコフ過程に関連した, マルコフ過程によってつくられたという意味.

2. M/M/1 モデル

M/M/1 待ち行列モデルはスーパのレジのようなものである. このモデルは次のことが前提になっている.

- 待ち行列の長さに制限はない
- 一度並んだ客がサービスを受ける前に立ち去ることはない
- 到着した客は必ず待ち行列に並ぶ
- 到着した客の順番が入れ替わることはない

M/M/1 モデルは, 最も基本的な待ち行列モデルである. このモデルでは, 平均待ち時間などを, 比較的単純な計算で得ることができる.

◇ 平均到着率 λ

系 (窓口および待ち行列が置かれる領域) に対する, 単位時間当たりの到着客数の平均のこと. 平均到着率 λ が与えられない場合は, 平均到着間隔の逆数で求めることができる.

例えば, 平均して 0.8 秒ごとに 1 件のトランザクションが発生するような場合, 平均到着率は

$$\frac{1}{0.8} = 1.25[\text{件/秒}]$$

となる.

◇ 平均サービス時間 $E(t_s) = 1/\mu$

1 人の客に対するサービスの平均所要時間である.

◇ 窓口利用率 ρ

窓口がサービス中である割合を示す. 窓口利用率 ρ は, 平均到着率と平均サービス時間を用いて,

$$\rho = \lambda \times E(t_s)$$

で計算することができる. 例えば, 平均到着率 $\lambda = 1.25$, 平均サービス時間 $E(t_s) = 0.6$ 秒であれば, 窓口利用率 ρ は

$$\rho = 1.25 \times 0.6 = 0.75$$

となる.

◇ 平均待ち時間 $E(t_w)$

客が系に到着してから, サービスを開始されるまでの時間の平均である.

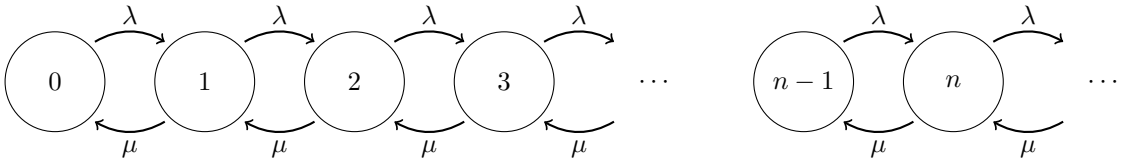

図 4 M/M/1 モデルの待ち行列の状態遷移図

まず, 微分方程式を立てる. 時刻 t の時点で行列に並んでいる人数が n 人である確率を $p_n(t)$ について考える. 微小時間 Δt の間に, 客が 1 人増える確率は $\lambda\Delta t$, 窓口が 1 人処理する確率は $\mu\Delta t$ である. これらの確率は十分小さいので行列に 2 つ以上の変化が起きる確率は無視できる.

$$
\begin{aligned}
p_0(t+\Delta t) &\simeq p_0(t)(1-\lambda\Delta t) + p_1(t)\mu\Delta t \\
p_n(t+\Delta t) &\simeq p_{n-1}(t)\lambda\Delta t + p_n(t)(1-\lambda\Delta t - \mu\Delta t) + p_{n+1}(t)\mu\Delta t
\end{aligned}
$$

これを $\Delta t \to 0$ の極限をとると

$$
\begin{aligned}
\frac{dp_0(t)}{dt} &= -p_0(t)\lambda + p_1(t)\mu \\
\frac{dp_n(t)}{dt} &= p_{n-1}(t)\lambda - p_n(t)(\lambda+\mu) + p_{n+1}(t)\mu
\end{aligned}
$$

となる.
このとき, この $p_n(t)$ の微分方程式を解くことは非常に困難であるため, 定常状態のとき (十分時間が経過したとき), $p_n = \lim_{t\to\infty} p_n(t)$ を求める.
定常状態において $p_n(t)$ は収束することがいえるので微分方程式の左辺 (変化量) は 0 になることがいえる.
よって,

$$
\begin{aligned}
p_0\lambda &= p_1\mu \\
p_{n+1}\mu &= -p_n(t)(\lambda+\mu) + p_{n-1}(t)\lambda
\end{aligned}
$$

$\rho = \lambda/\mu$ であることから ρ でまとめると

$$
\begin{aligned}
p_1 &= \rho p_0 \\
p_{n+1} &= (1+\rho)p_n - \rho p_{n-1}
\end{aligned}
$$

となる. これより三項間漸化式を解けばよい.
下の三項間漸化式は次の二通りに変形できる.

$$
\begin{cases}
p_{n+1} - p_n = \rho(p_n - p_{n-1}) \\
p_{n+1} - \rho p_n = p_n - \rho p_{n-1}
\end{cases} \tag{1}
$$

(1) の上式において,

$$
\begin{aligned}
p_{n+1} - p_n &= \rho(p_n - p_{n-1}) \\
&= \rho^2(p_{n-1} - p_{n-2}) \\
&= \rho^3(p_{n-2} - p_{n-3}) \\
&= \rho^n(p_1 - p_0)
\end{aligned}
$$

(1) の下式において,

$$
\begin{aligned}
p_{n+1} - \rho p_n &= p_n - \rho p_{n-1} \\
&= p_{n-1} - \rho p_{n-2} \\
&= p_1 - \rho p_0
\end{aligned}
$$

となる. $p_1 = \rho p_0$ であることを用いれば,

$$
\begin{aligned}
p_{n+1} - \rho p_n &= p_1 - \rho p_0 \\
&= 0
\end{aligned}
$$

したがって,

$$\begin{aligned} p_{n+1} &= \rho p_n \\ &= \rho^{n+1} p_0 \end{aligned}$$

また, $\sum_{n=0}^{\infty} p_n = 1$ であることから

$$\begin{aligned} \sum_{n=0}^{\infty} p_n &= \sum_{n=0}^{\infty} \rho^n p_0 \\ &= (1 + \rho + \rho^2 + \rho^3 + \cdots) p_0 \\ &= \frac{1}{1-\rho} p_0 \end{aligned}$$

より,

$$\begin{aligned} \sum_{n=0}^{\infty} p_n = 1 &\iff \frac{1}{1-\rho} p_0 = 1 \\ &\iff p_0 = 1 - \rho \end{aligned}$$

よって,

$$\begin{aligned} p_{n+1} - p_n &= \rho^n (p_1 - p_0) \\ (\rho - 1) p_n &= \rho^n (\rho - 1) p_0 \\ p_n &= \rho (1 - \rho) \end{aligned}$$

行列の長さが n 人のとき, 平均待ち時間は $nE(t_s)$ であるから, 平均待ち時間は

$$\begin{aligned} E(t_w) &= \sum_{n=1}^{\infty} p_n \cdot nE(t_s) \\ &= (1-\rho) \sum_{n=1}^{\infty} np_n \cdot E(t_s) \\ &= (1-\rho) \cdot \frac{\rho}{(1-\rho)^2} E(t_s) \\ &= \frac{\rho}{1-\rho} E(t_s) \end{aligned}$$

例えば, $\lambda = 1.25,\ E(t_s) = 0.6,\ \rho = 0.75$ の場合,

$$\begin{aligned} E(t_w) &= \frac{0.75}{1 - 0.75} \times 0.6 \\ &= 3 \times 0.6 \\ &= 1.8[\text{秒}] \end{aligned}$$

となり, 客は平均して 1.8 秒待つことになる.

♢ 平均応答時間 $E(t_q)$

客は待ち行列に並んで自分の順番が来るのを待ち, その後サービスを受けることになる. すなわち平均応答時間は, 平均待ち時間 + サービス時間に等しいことになる. したがって,

$$\begin{aligned} E(t_q) &= E(t_w) + E(t_s) \\ &= \frac{\rho}{1-\rho} \times E(t_s) + E(t_s) \\ &= \frac{\rho + 1 - \rho}{1-\rho} \times E(t_s) \\ &= \frac{1}{1-\rho} \times E(t_s) \end{aligned}$$

と計算することができる.

♢ 平均待ち客数 (平均待ち行列長) $E(L_w)$

待っている客の数 (= サービス中の客を除いた, 待ち行列の長さ) の平均である.

$$\begin{aligned} E(L_w) &= \lambda \times E(t_w) \\ &= \lambda \times \frac{\rho}{1-\rho} \times E(t_s) \\ &= \frac{\rho^2}{1-\rho} \end{aligned}$$

となる.

♢ 平均系内滞留客数 (平均系長) $E(L_q)$

サービス中の客も含めて, 系内にいる客数の平均である.

$$\begin{aligned} E(L_q) &= \lambda \times E(t_q) \\ &= \lambda \times \frac{1}{1-\rho} \times E(t_s) \\ &= \frac{\rho}{1-\rho} \end{aligned}$$

常に窓口がサービス中とは限らないので, 単純に $E(L_q) = E(L_w) + 1$ とはならない.

3. 複数窓口モデル

複数モデルは,

- 到着時間間隔やサービス時間の分布は M/M/1 と同じ
- 窓口が複数存在する
- 窓口ごとに並列に客を処理する

であるモデルで, M/M/S と表す.

窓口数が増えると解析が複雑になり, M/M/1 のような単純な式で待ち時間を得ることはできない. 一般には, あらかじめ用意した表やグラフを用いて, 待ち時間を求める. その中でも, 平均サービス時間を 1 に正規化し, ρ と窓口数が変化した場合の待ち時間をまとめた表がよく用いられる. この表から正規化された平均待ち時間を得た後, 平均サービス時間を乗じることで, 実際の平均待ち時間を求めることができる.