wq-2. 待ち行列

（待ち行列の数理）

URL: https://www.kkaneko.jp/cc/wq/index.html

金子邦彦

CC BY NC SA

アウトライン

2-1 待ち行列

2-2 ケンドール記法

2-3 M/M/1/1 待ち行列

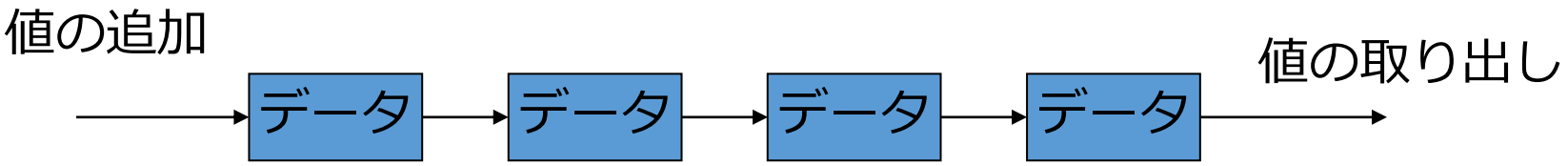
2-4 M/M/1/1 待ち行列の解析

2-5 M/M/1 待ち行列の解析

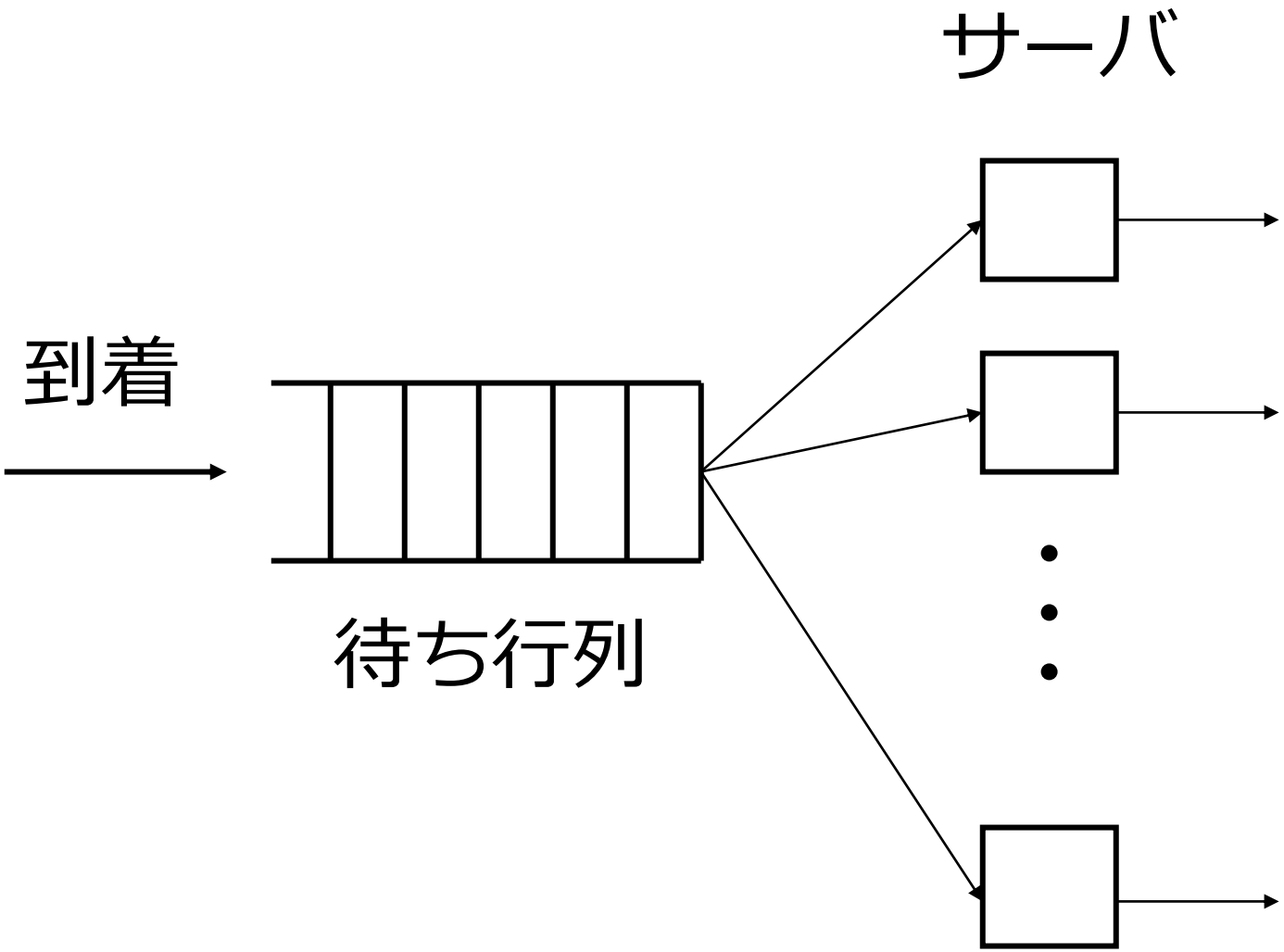
2-1 待ち行列

スタックとキュー

- **スタック**
  - データの挿入と取り出しの両方を列の一方の端から行う
- **キュー**
  - 一方の端から挿入を，もう一方の端から取り出しを行う
  - 取り出されるのは最も古いデータ
  - 最初に入れたデータが最初に取り出される
  - **FIFO**(first-in-first-out，先入れ先出し)と呼ぶ



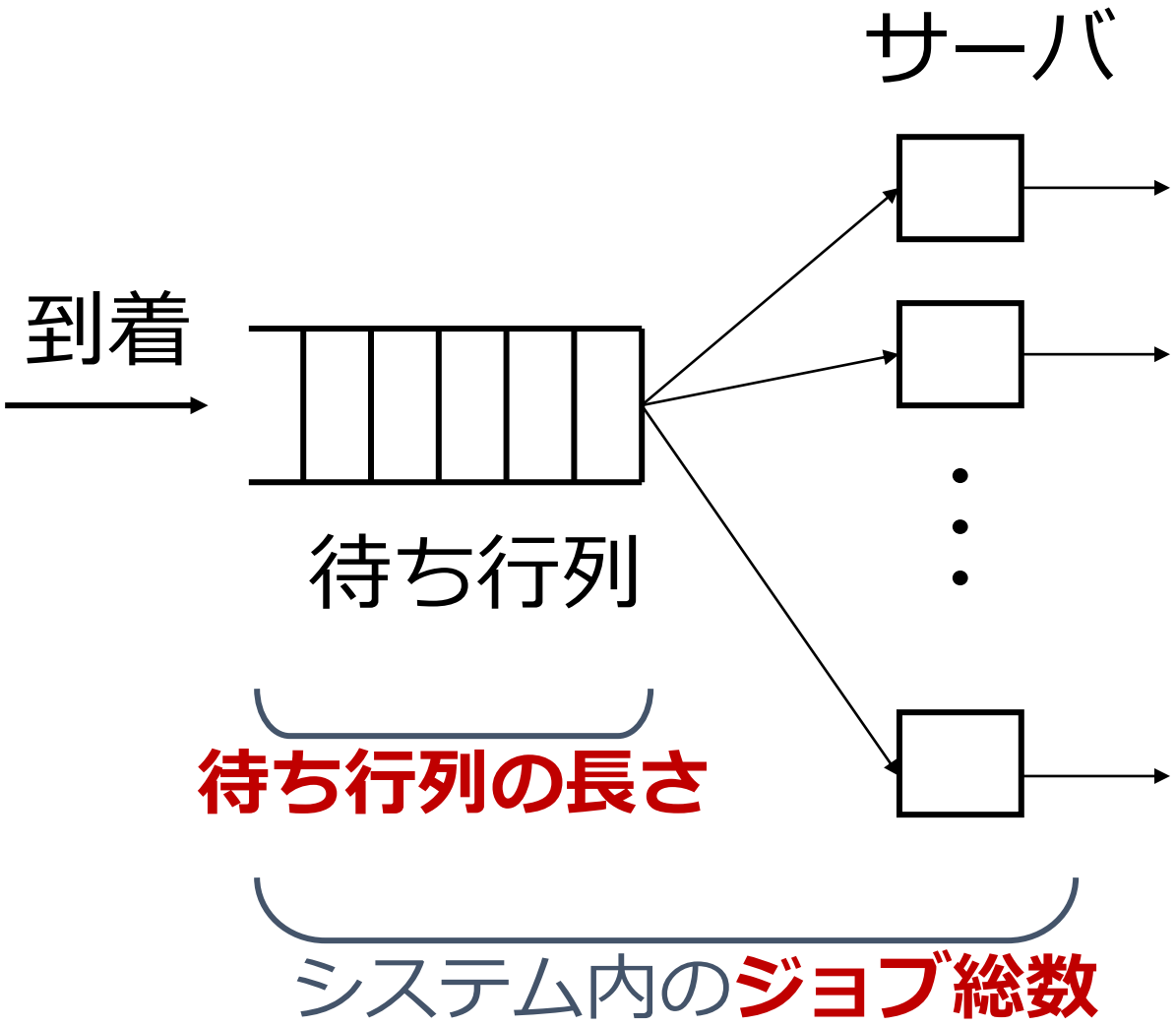
待ち行列



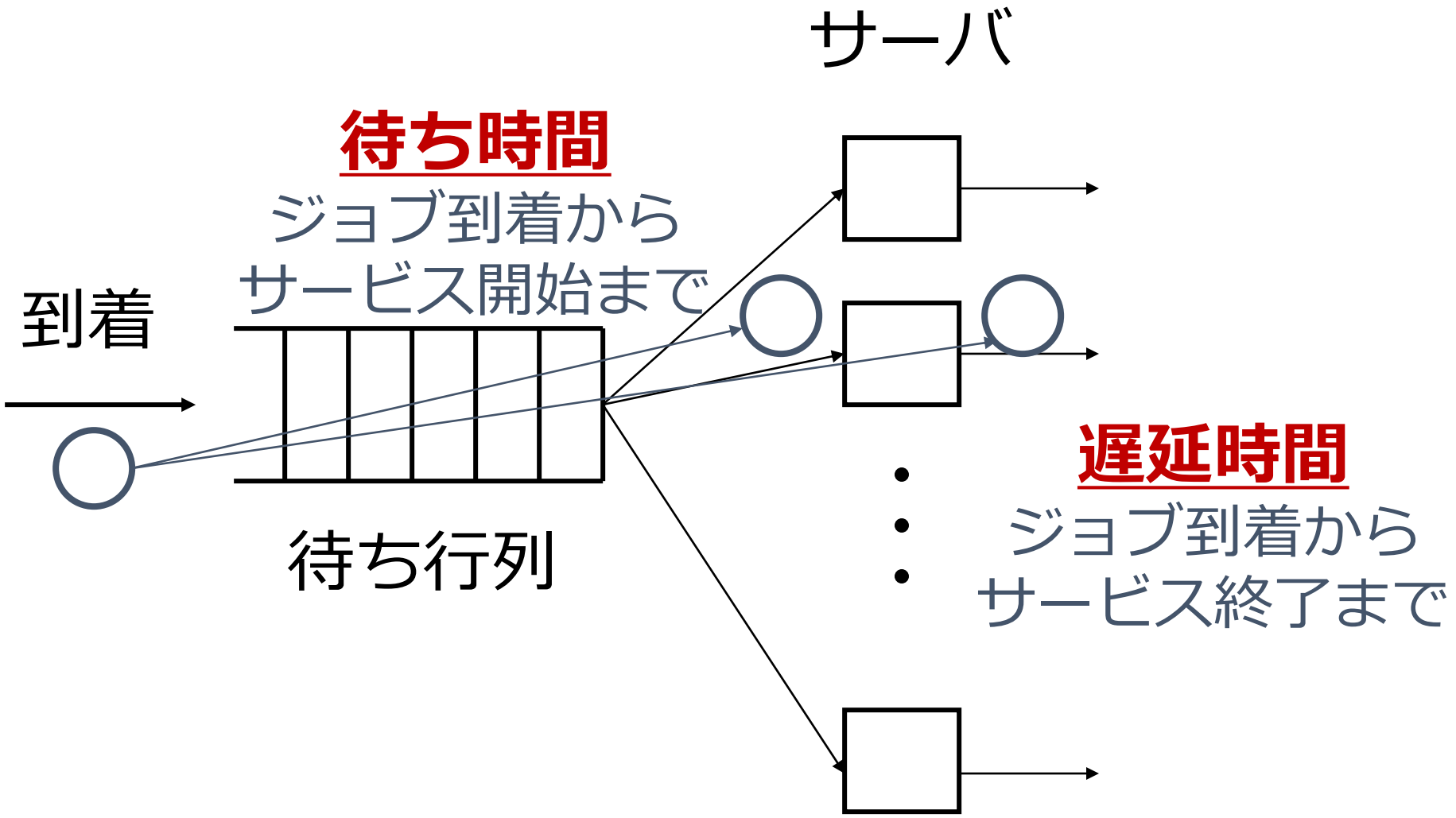
待ち行列

- 処理を受けるために順番待ちをする人がなす列
  - 銀行の窓口や入場券売り場など

待ち行列の長さ，システム内のジョブ総数



遅延時間，待ち時間



遅延時間，待ち時間

$\lambda D = N$

D:　「**遅延時間**」の平均

N:　「システム内の**ジョブ総数**」の平均

$\lambda Dw = Nw$

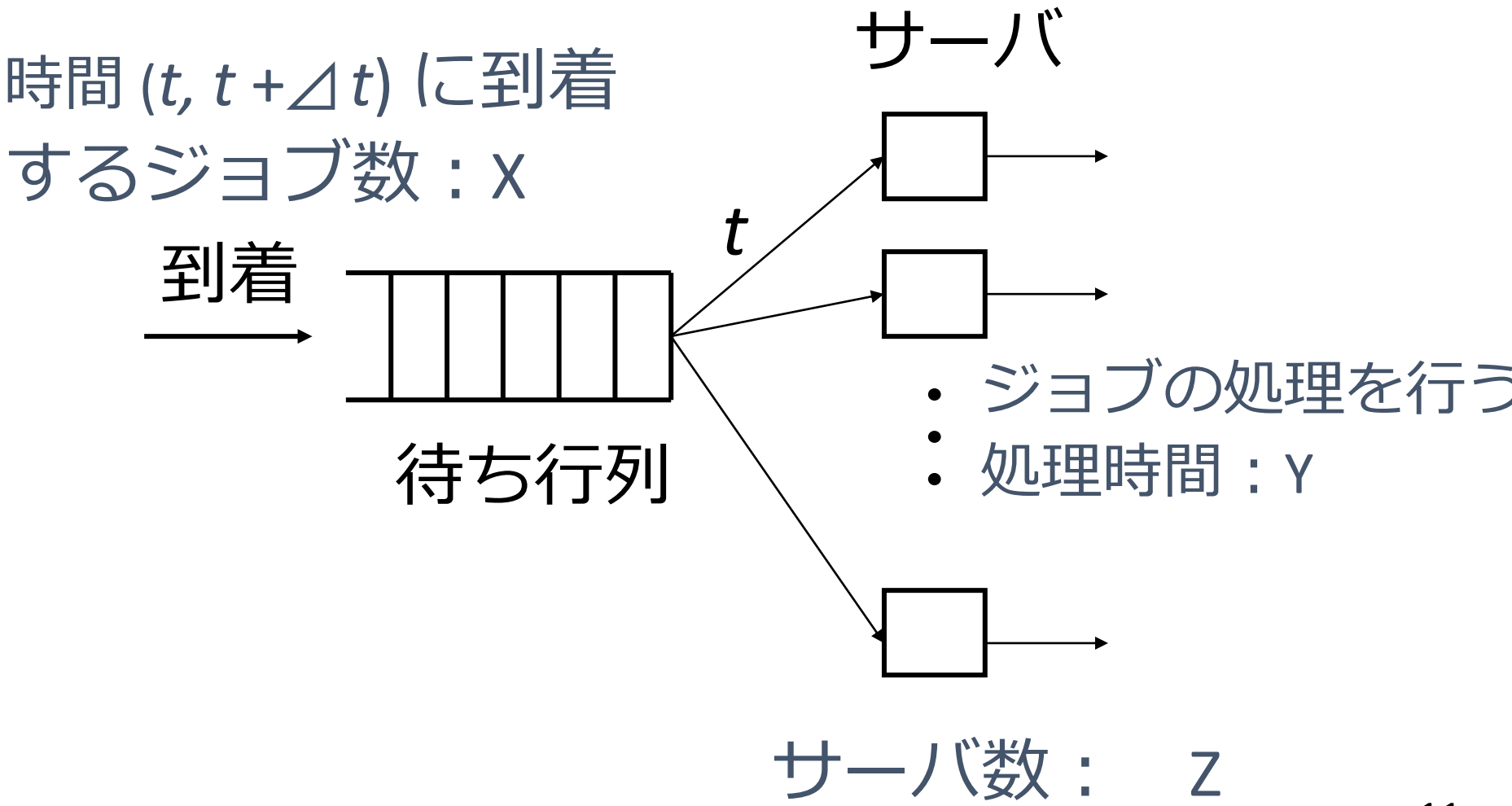
Dw:　「**待ち時間**」の平均

Nw:　「**待ち行列の長さ**」の平均

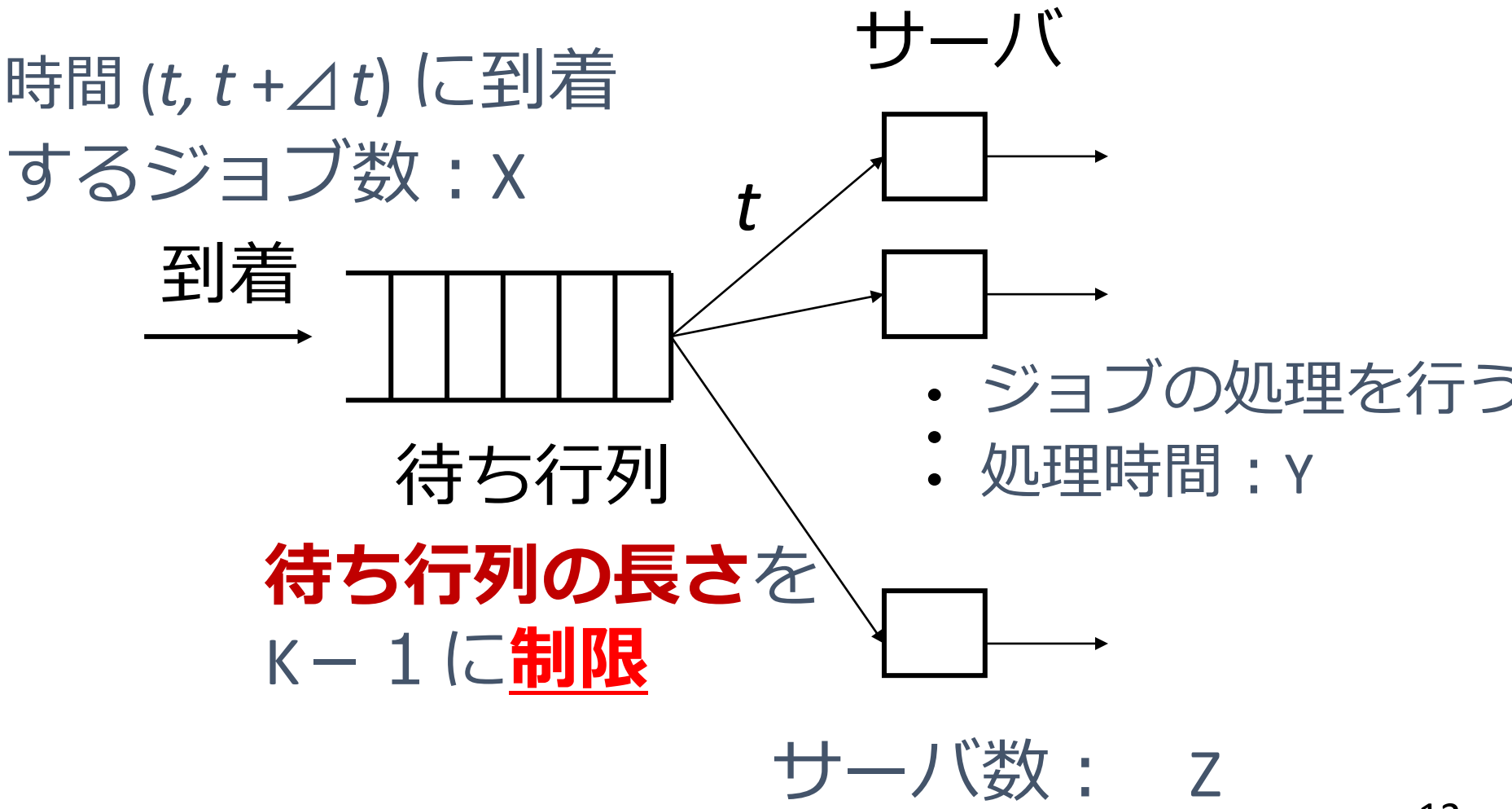
以下，システム内の**ジョブ総数**，**待ち行列の長さ**を考える

2-2 ケンドール記法

ケンドール記法 X/Y/Z



ケンドール記法 X/Y/Z/K



待ち行列の長さの制限

- **待ち行列の長さ**に限りがある
  - **待ち行列の長さ**が「**最大で Ｋ－１**」に制限されるとき，
    システム内の**ジョブ総数**は　Ｋ　に制限される

- Ｋ＝０ の場合は
  - すでにサーバが他のジョブを処理中のとき
    - 到着したジョブは棄却される（待ち行列に入らない）
  - サーバがジョブを処理していないとき
    - 到着したジョブは直ちに処理される

ケンドール記法

X/Y/Z/K

X:　到着過程

Y：　処理時間分布

Z:　サーバ数

K：　**待ち行列の長さ**制限

（**待ち行列の長さ**の最大長：K－１）

X/Y/Z

待ち行列の長さに制限**無し**

2-3 M/M/1/1 待ち行列

ケンドール記法

X:　到着過程

→**ポアソン分布**のとき「M」と書く

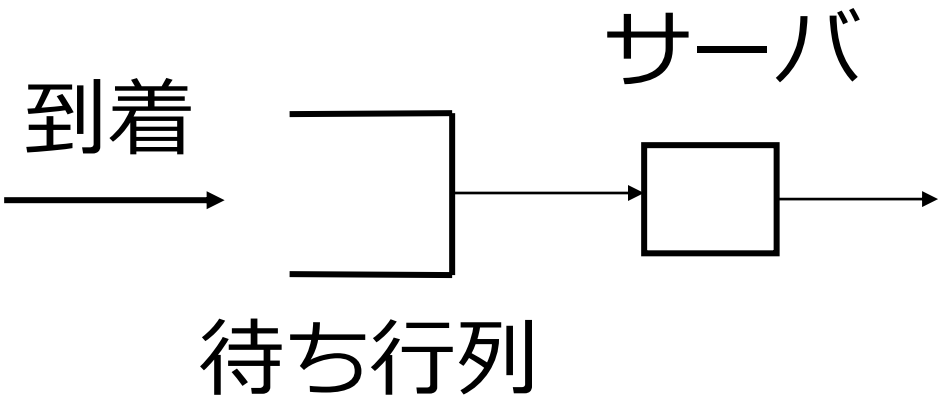
Y：　処理時間分布

→**指数分布**のとき「M」と書く

Z:　サーバ数

K：　**待ち行列の長さ**の制限

M/M/1/1 待ち行列



- 到着過程：　**ポアソン分布**
- 処理時間分布：　**指数分布**
- サーバの個数：　１個
- **待ち行列の長さ**の制限：　K＝1

「分布」の種類

M:　**ポアソン分布／指数分布**

Ek:　k相のアーラン分布

D:　一定分布

G:　一般分布

GI:　独立性を有する一般分布

平均到着率

- 単位時間に到着するジョブの平均値
- 待ち行列に加わろうとするジョブのやってくる頻度

到着率λのポアソン分布

- ジョブの到着がランダム

- 「時間 $(t, t+\Delta t)$ に到着するジョブ数」に注目
  - Δt　に比例して増加
  - 平均値： $\lambda \Delta t$

- λは単位時間あたりの平均ジョブ数

ポアソン分布の特徴

- 同じ幅をもった時間区間あたりの到着の仕方は，時刻に依存しない
- 共通部分のない時間区間たちのそれぞれの到着の仕方は独立である
- 同時刻に2人のジョブがやってくることはない
- ごく短い時間 ⊿tの間にジョブが1人来る確率はλ⊿t

平均処理率

- 単位時間に処理を受けるジョブの平均値
- 処理がどの程度で行われているかの尺度

処理率μの指数分布

- ジョブの完了時刻がランダム
- 「あるジョブの処理の完了から次のジョブの完了までの時間」に着目
  - 平均値： $1/\mu$
- μは単位時間あたりの平均ジョブ完了数
  - サーバがジョブを処理中の間，Δt 内に完了する処理数： $\mu\Delta t$

指数分布の特徴

- 進行中の処理が終了する確率は，それまでに処理に要した時間に依存しない
- ある時刻に開始される処理は，それ以前に行われた処理や到着に依存しない
- ごく短い時間 ⊿tの間に処理が1つ終了する確率は μ⊿t

微小時間の意味

- 微小時間 ⊿tの間に到着するジョブ：
  - たかだか1人
- 時間 ⊿tの間に終了する処理：
  - たかだか1つ
- 時間 ⊿tの間に「ジョブの到着」「処理の終了」が同時になされることはない

2-4 待ち行列の解析

待ち行列の解析での解析の対象

- **待ち行列の長さ**
- システム内の**ジョブ総数**
- ジョブの**遅延時間**
- ジョブの**待ち時間**

解析手順

- **確率的に解析**
- 待ち行列の**システムの状態**と**状態遷移**による解析

  システムの状態：$P_0, P_1, P_2, \cdots$

  （添字は，システム内のジョブ総数）
- **定常状態**での**待ち行列の長さ**，**ジョブ総数**を算出する
  - $t \rightarrow \infty$では，**システムの状態**は**定常確率**に**漸近**する（初期状態を無視できる）

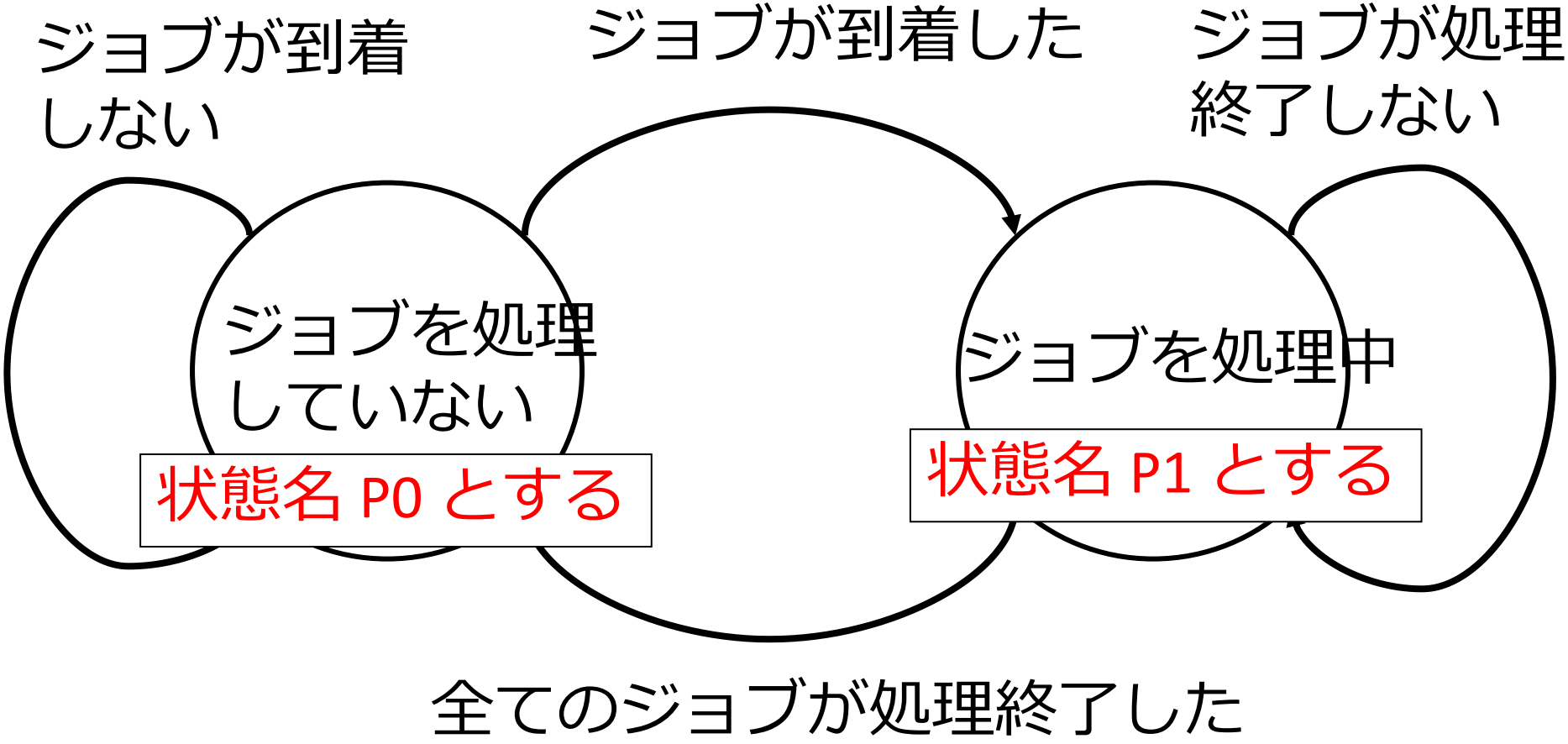
システム処理能力ρ

ρ= λ/μ

- λ⊿ｔ：　「時間 $(t, t+\Delta t)$ に到着するジョブ数」の平均
- μ⊿ｔ：　「サーバがジョブを処理中の間，⊿ｔ 内に完了する処理数」の平均

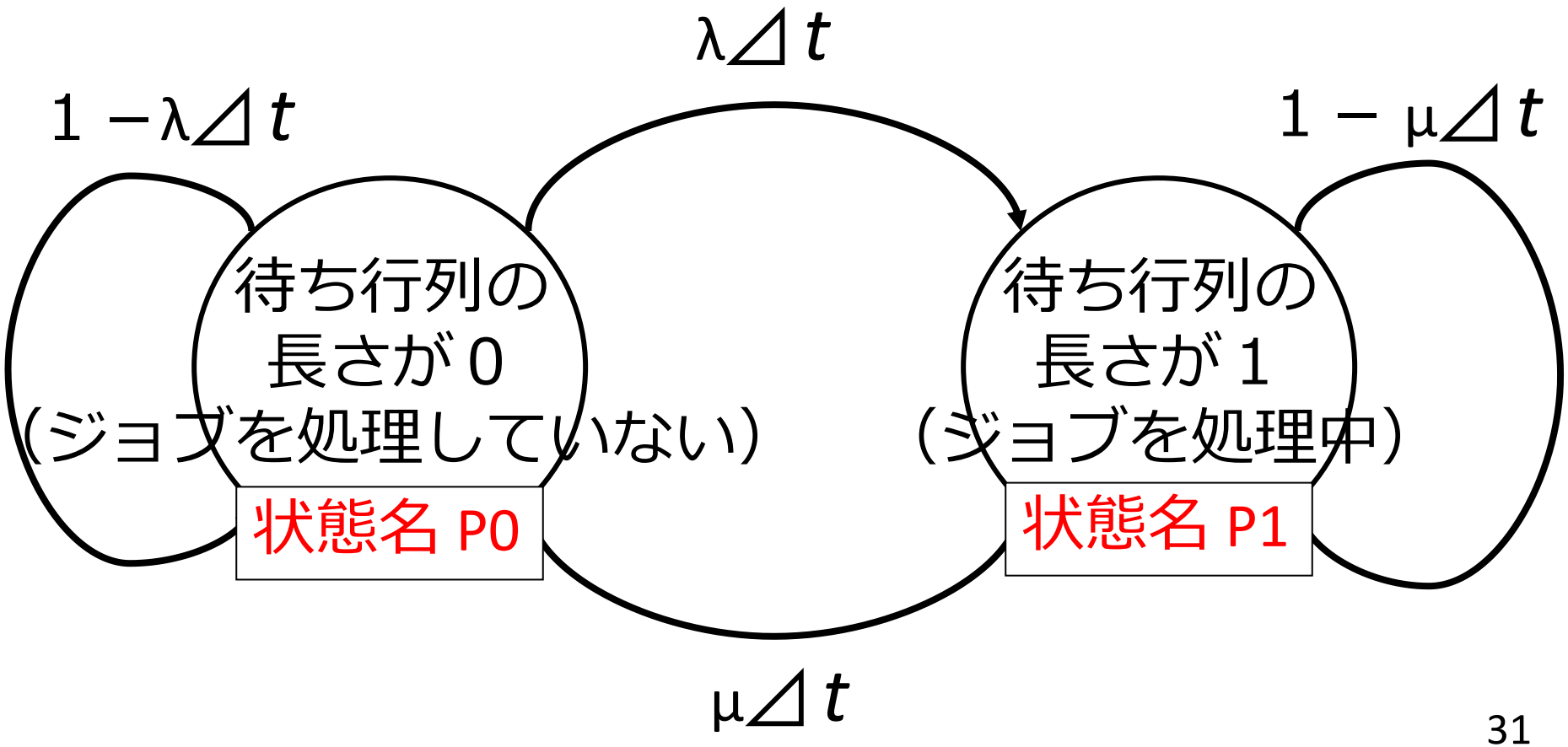
※ **待ち行列の長さ**に限りがないとすると：
λ＜μ（つまりρ＜１）である必要がある
（さもないと待ち行列が**あふれる**）

個々のサーバの状態と状態遷移



M/M/1/1待ち行列でのサーバの状態遷移

・K=1 なので，**待ち行列の長さ**は 0 か 1
・状態遷移の確率は，下図の通り



M/M/1/1待ち行列でのサーバの状態遷移

$$P_0(t+\Delta t) = (1-\lambda\Delta t)P_0(t) + \mu\Delta t P_1(t)$$

$$P_1(t+\Delta t) = \lambda\Delta t P_0(t) + (1-\mu\Delta t)P_1(t)$$

M/M/1/1待ち行列での定常確率

$\lim_{t\to\infty} P_0(t)$, $\lim_{t\to\infty} P_1(t)$ を求めよう

$t\to\infty$ のとき $P_0(t)\to P_0$, $P_1(t)\to P_1$ （収束する）と仮定する

$$\frac{dP_0(t)}{dt} = -\lambda P_0(t) + \mu P_1(t)$$ だが（理由は後述）

仮定より，$t\to\infty$ のとき $\frac{dP_0(t)}{dt} = 0$ なので

$$-\lambda P_0 + \mu P_1 = 0$$

これと $P_0 + P_1 = 1$ から，$P_0 = \frac{\mu}{\lambda+\mu}$, $P_1 = \frac{\lambda}{\lambda+\mu}$

M/M/1/1待ち行列での定常確率

$$P_0(t+\Delta t) = (1-\lambda\Delta t)P_0(t) + \mu\Delta t P_1(t)$$

$$P_0(t+\Delta t) - P_0(t) = -\lambda\Delta t\, P_0(t) + \mu\Delta t P_1(t)$$

$$\frac{P_0(t+\Delta t) - P_0(t)}{\Delta t} = -\lambda P_0(t) + \mu P_1(t)$$

$$\frac{P_0(t+\Delta t) - P_0(t)}{\Delta t} = -(\lambda+\mu)P_0(t) + \mu$$

（$P_0(t) + P_1(t) = 1$から）

$P_0(t)$の方程式が求まった

M/M/1/1待ち行列での定常確率

$$\lim_{\Delta t \to 0} \frac{P_0(t+\Delta t) - P_0(t)}{\Delta t} = -(\lambda+\mu)P_0(t)+\mu$$

$$\frac{d\,P_0(t)}{dt} = -(\lambda+\mu)P_0(t)+\mu$$

これは $P_0(t)$ の微分方程式

$$P_0(t) = \frac{\mu}{\lambda+\mu} + \left(P_0(0) - \frac{\mu}{\lambda+\mu}\right)e^{-(\lambda+\mu)t}$$

$$\lim_{\Delta t \to 0} P_0(t) = \frac{\mu}{\lambda+\mu}$$

定常状態における性質

$$-\lambda P_0 + \mu P_1 = 0$$

つまり　$\lambda \Delta t \lim_{t\to\infty} P_0(t) = \mu \Delta t \lim_{t\to\infty} P_1(t)$

$\lambda \Delta t \lim_{t\to\infty} P_0(t)$：新たなジョブが$\Delta t$以内に到着する確率

$\mu \Delta t \lim_{t\to\infty} P_1(t)$：処理中のジョブが$\Delta t$以内に完了する確率

まとめ

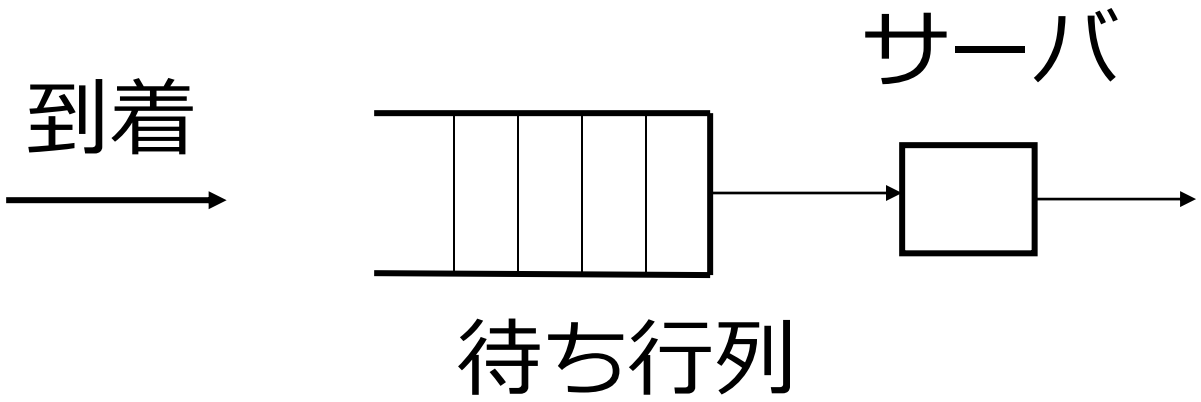
- 定常状態
  - $\lim_{t\to\infty} P_0(t) = \frac{1}{1+\rho}$
  - $\lim_{t\to\infty} P_1(t) = \frac{\rho}{1+\rho}$

（但し　$\rho = \lambda/\mu$）

- 定常状態でのサーバ内のジョブ総数
  - 0である確率：$\lim_{t\to\infty} P_0(t)$, 1である確率：$\lim_{t\to\infty} P_1(t)$

M/M/1 待ち行列

M/M/1 待ち行列



- 到着過程：　**ポアソン分布**
- 処理時間分布：　**指数分布**
- サーバの個数：　１個
- **待ち行列の長さ**の制限：　**制限なし**

M/M/1 待ち行列

- 処理の窓口は1つ
- 処理を受けるための列は1つ
- いったん行列に加わったら，処理を受けるまで待ち続ける
- ジョブの到着の仕方は**ポアソン分布**に従う
- 処理時間の分布は**指数分布**に従う

時刻 t+⊿t の時点で，ジョブが1個もない確率

- 時刻tにジョブが n 個ある確率：　$P_n(t)$とおく
  (n=0,1,2,・・・)
- 時刻 t+⊿t にジョブが1個もいないのは，次のいずれかの場合
  1. ジョブが1個もいなくて，⊿tに新たなジョブが来なかった

     $P(A)=P_0(t)*(1- \lambda \Delta t)$
  2. 1 個のジョブが処理を受けていて，⊿tの間に処理が終了した

     $P(B)=P_1(t)*\mu \Delta t$
- これらは独立な事象なので，

$$P_0(t+\Delta t) = P_0(t)*(1-\lambda \Delta t) + P_1(t)* \mu \Delta t$$

続き

- $P_n(t)$ は，時刻に依存しない（定常状態）と考えて

$$P_n(t + \Delta t) = P_n(t)$$

- 前ページの式に代入すると

$$P_0(t)\ \lambda \Delta t = P_1(t) \mu \Delta t$$

- つまり

$$P_0(t)\ \lambda = P_1(t) \mu$$

$P_n(t)$

- 時刻が t から t + ⊿t になった時点で，ジョブの総数が n である場合は以下の3通り
  - 時刻tに n 個で，新たなジョブが到着せず，処理も終了しなかった

    $P(A) = P_n(t)*(1-\lambda⊿ t)*(1-\mu⊿ t)$

    $= P_n(t)*(1-\lambda⊿t – \mu⊿t)$
  - 時刻 t に n-1個で，新たなジョブが1つ到着した

    $P(B) = P_{n-1}(t)*\lambda⊿t$
  - 時刻 t に n+1 個で，ジョブの処理が1つ終了した

    $P(C) = P_{n+1}(t)*\mu⊿t$

続き

- $P_n(t+\Delta t)=P(A)+P(B)+P(C)$から

$$P_n(t)(\lambda + \mu) = P_{n-1}(t)\lambda + P_{n+1}(t)\mu$$

まとめ

以上から，定常状態における状態遷移については，次が成り立つ

$$\lambda P_0 = \mu P_1$$

$$(\lambda+\mu)P_1 = \lambda P_0 + \mu P_2$$

$$\vdots$$

$$(\lambda+\mu)P_i = \lambda P_{i-1} + \mu P_{i+1}$$

すべての状態の確率を P0 の式で書く

$$P_1 = \rho P_0$$

$$P_2 = (1+\rho)P_1 - \rho P_0$$

$$= (1+\rho)\,\rho\, P_0 - \rho P_0$$

$$= \rho_2 P_0$$

$$P_i = \rho_i P_0$$

一方，$\Sigma P_i = 1$ なので　$P_i = (1-\rho)\,\rho_i$

続き

$$\Sigma Pi = 1$$

つまり，$P_0(t)*(1+\rho_1+\rho_2+\cdots)=1$

- $\rho \geqq 1$
  - 処理できるジョブ数より，やってくる方が多い
  - $1+\rho_1+\rho_2+\cdots$は発散する
- $\rho < 1$
  - $1+\rho_1+\rho_2+\cdots=1/(1-\rho)$　　（発散しない）
  - $P_0(t)=1-\rho$

平均ジョブ数，平均待ちジョブ数

$$N = \Sigma n P_n$$

$$= \Sigma n (1-\rho) \rho^{n}$$

$$= \frac{\rho}{1-\rho}$$

$$N_w = \Sigma (n-1) P_n$$

$$= \Sigma n P_n - (1 - P_0)$$

$$= N - (1 - P_0)$$

$$= \frac{\rho^2}{1-\rho}$$

平均ジョブ数

- 平均ジョブ数をLとおくと

- Lを計算すると

$$L = \sum_{n=0}^{\infty} n P_n$$

$$L = \frac{\rho}{1-\rho}$$

- 処理を受けているジョブを含まない待ち行列内の平均ジョブ数: L_q

$$L_q = \sum_{n=1}^{\infty} (n-1) P_n$$

$$= \sum_{n=1}^{\infty} (n-1)(1-\rho)\rho^n$$

$$= \rho \sum_{n=1}^{\infty} (n-1)(1-\rho)\rho^{n-1}$$

$$= \rho \sum_{n=1}^{\infty} n(1-\rho)\rho^n$$

$$= \rho \sum_{n=1}^{\infty} n P_n$$

$$= \rho L$$

$$L_q = \frac{\rho^2}{1-\rho}$$

平均待ち時間

- ジョブが並びはじめて処理を受け始めるまでの時間の平均： W_q

$$L_q=\lambda W_q$$

- このことから

$$W_q = \frac{L_q}{\lambda} = \frac{\rho}{\lambda(1-\rho)} = \frac{\rho^2}{\mu(1-\rho)}$$

おわりに

- 待ち行列の数理
  - システム内のジョブ総数
  - 待ち行列の長さ

- 待ち行列の定常状態
  - 状態遷移
  - 定常確率

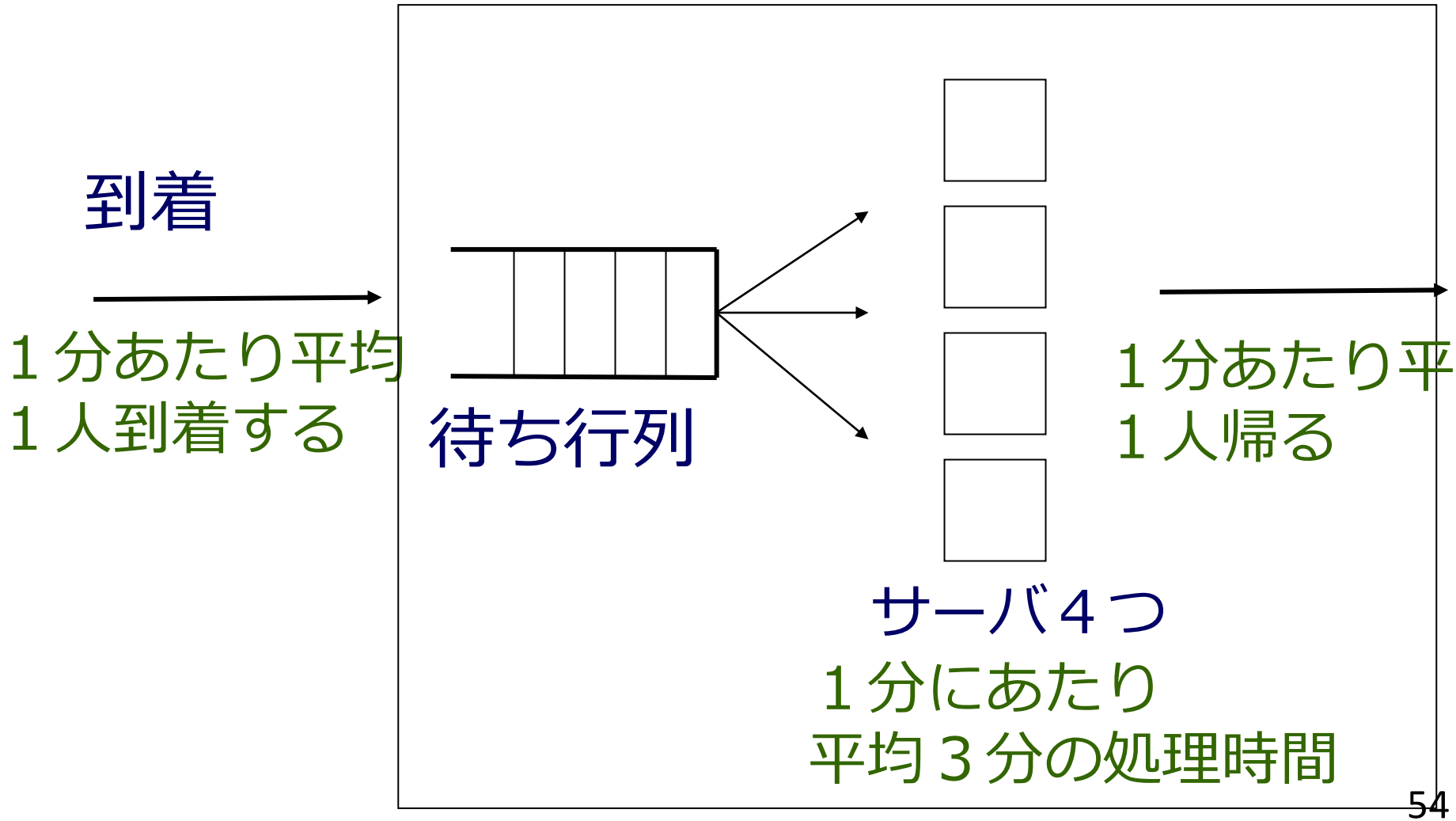
「確率」を使って，待ち行列の振る舞いをとらえる

練習１

- ある銀行には４つの窓口がある．この銀行に，１分に1人の割合で客がポアソン到着するとしよう．また，１人の客のサービス時間は平均3分の指数分布に従うものとする．単純のために，1度銀行に入った客は，サービスを受けるまで必ず待つということを仮定しよう
  （１）この場合のケンドール表記を書け

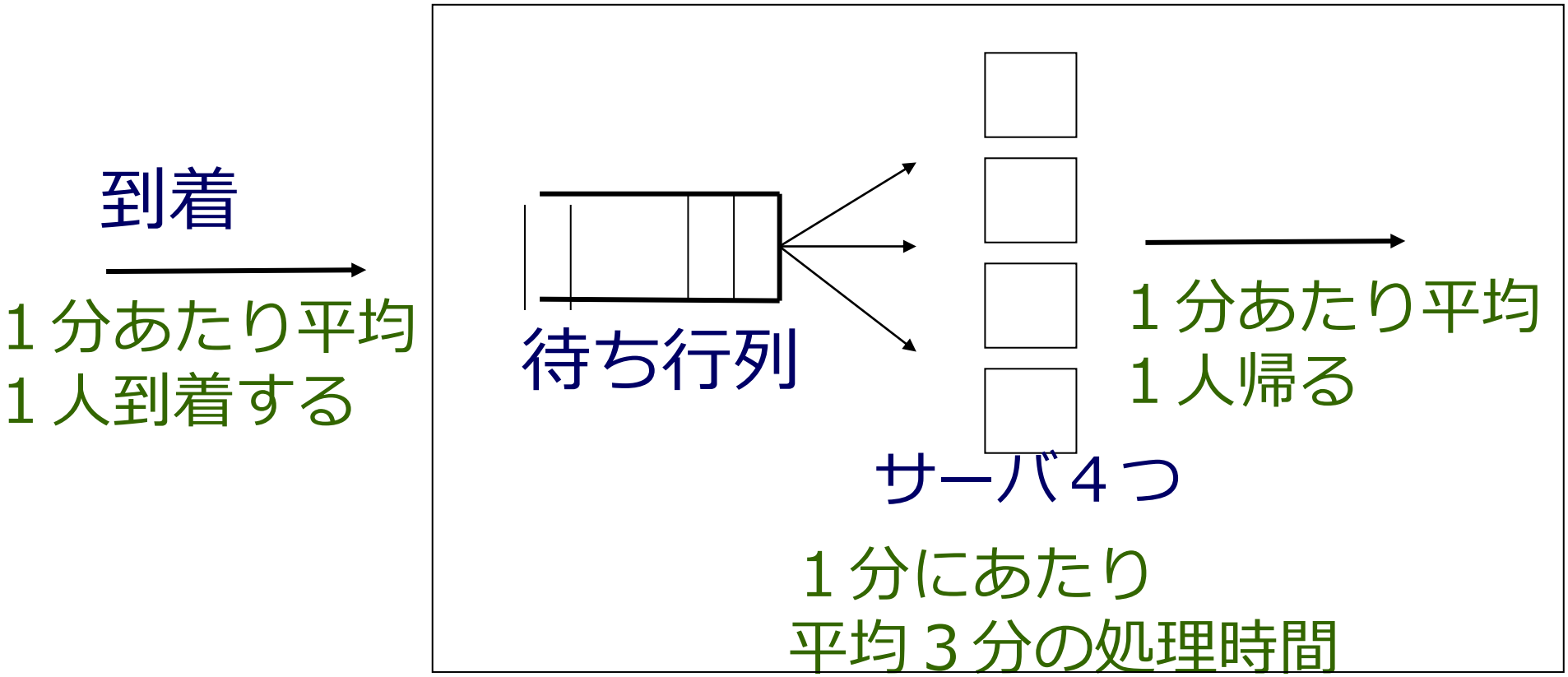
Ｍ／Ｍ／４

（２）「定常状態」において，１分あたりに平均で，客は何人帰るか



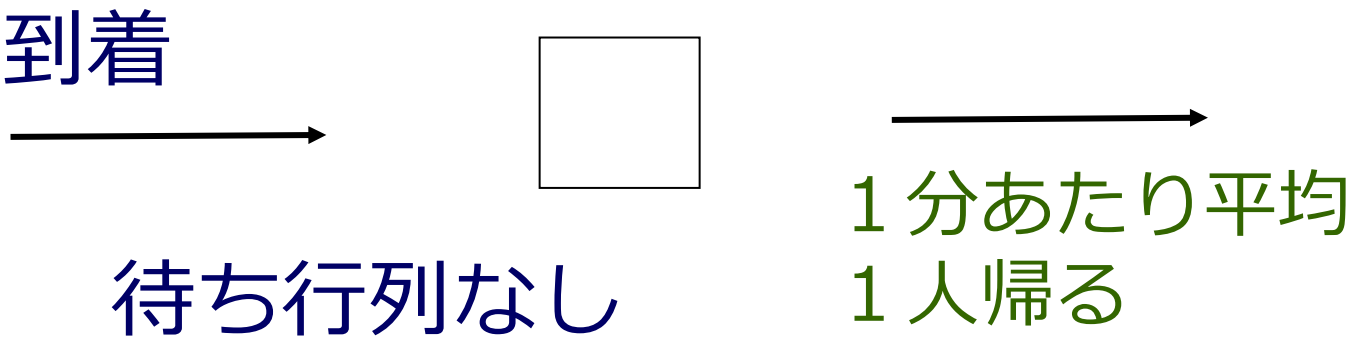
（３）システム処理能力（ρ）を答えよ.

λ=1, S=4, μ=1/3 より,
ρ= λ／ Sμ　= ３／４



練習２

- １つの窓口しかない銀行を考える
  - 客は，銀行にやってきたとき窓口が空いているかどうか確認する
  - 空いていれば窓口で用事を済ませる
  - 空いていなければあきらめて家に帰り，銀行の用事を忘れるものとする

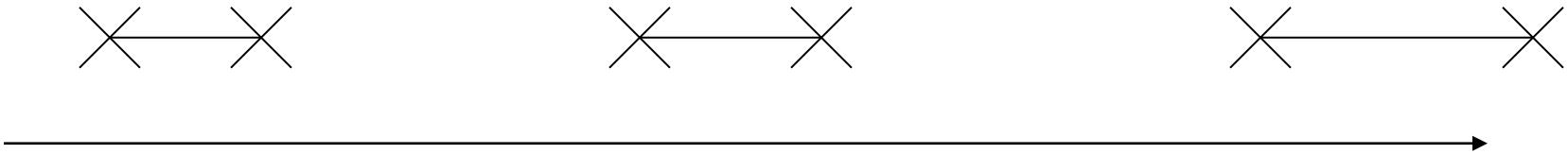


- 客は，平均２０分あたり１人のポアソン分布で銀行に到着し，すべての客は窓口に１０分間（等しい時間）滞在するとしよう.
  - 客があきらめて帰る確率はいくらか

  - 客の到着はλ = 1/20 のポアソン分布
  - t = 10 を $\left.\frac{(\lambda t)^k e^{-\lambda t}}{k!}\right|_{k=0}$ に代入
  - 客が来てから次の客が来るまでの時間は，λ = 1/20 の指数分布
  - t = 10 を $1- e^{-\lambda t}$ に代入

練習３

- Ａさんの電話の話し時間は，平均５分）の指数分布に従うものと仮定する.
  - Ａさんに電話をかけたらたまたま話し中であった.では，３分後に電話をかけたら再び話し中である確率はいくらか
  - 話し時間は，λ = 1/5 の指数分布
  - t = 3 を $1- e^{-\lambda t}$ に代入



話し時間がランダム