Metody probabilistyczne i statystyka, 2021
informatyka algorytmiczna, WIiT PWr

4-Queuing systems

Problem

- jobs arrive as a random process
- server(s) take the jobs from the queue and serve (or drop)
- E.g. first-in-first-out basis
- service time is also random

Examples: Web server

Main parameters

Parameters of a queuing system

$$\lambda_A = \text{arrival rate}$$ = average number of jobs arriving in one time unit

$$\lambda_S = \text{service rate}$$

$$\mu_A = 1/\lambda_A = \text{mean interarrival time}$$

$$\mu_S = 1/\lambda_S = \text{mean service time}$$

$$r = \lambda_A/\lambda_S = \mu_S/\mu_A = \text{utilization, or arrival-to-service ratio}$$

Main parameters

<u>Random variables of a queuing system</u>

$$
\begin{aligned}
X_s(t) &= \text{number of jobs receiving service at time } t \\
X_w(t) &= \text{number of jobs waiting in a queue at time } t \\
X(t) &= X_s(t) + X_w(t), \\
&\quad \text{the total number of jobs in the system at time } t \\
S_k &= \text{service time of the } k\text{-th job} \\
W_k &= \text{waiting time of the } k\text{-th job} \\
R_k &= S_k + W_k, \text{ response time, the total time a job spends in the} \\
&\quad \text{system from its arrival until the departure}
\end{aligned}
$$

A stationary system: S_k, W_k and R_k do not depend on k

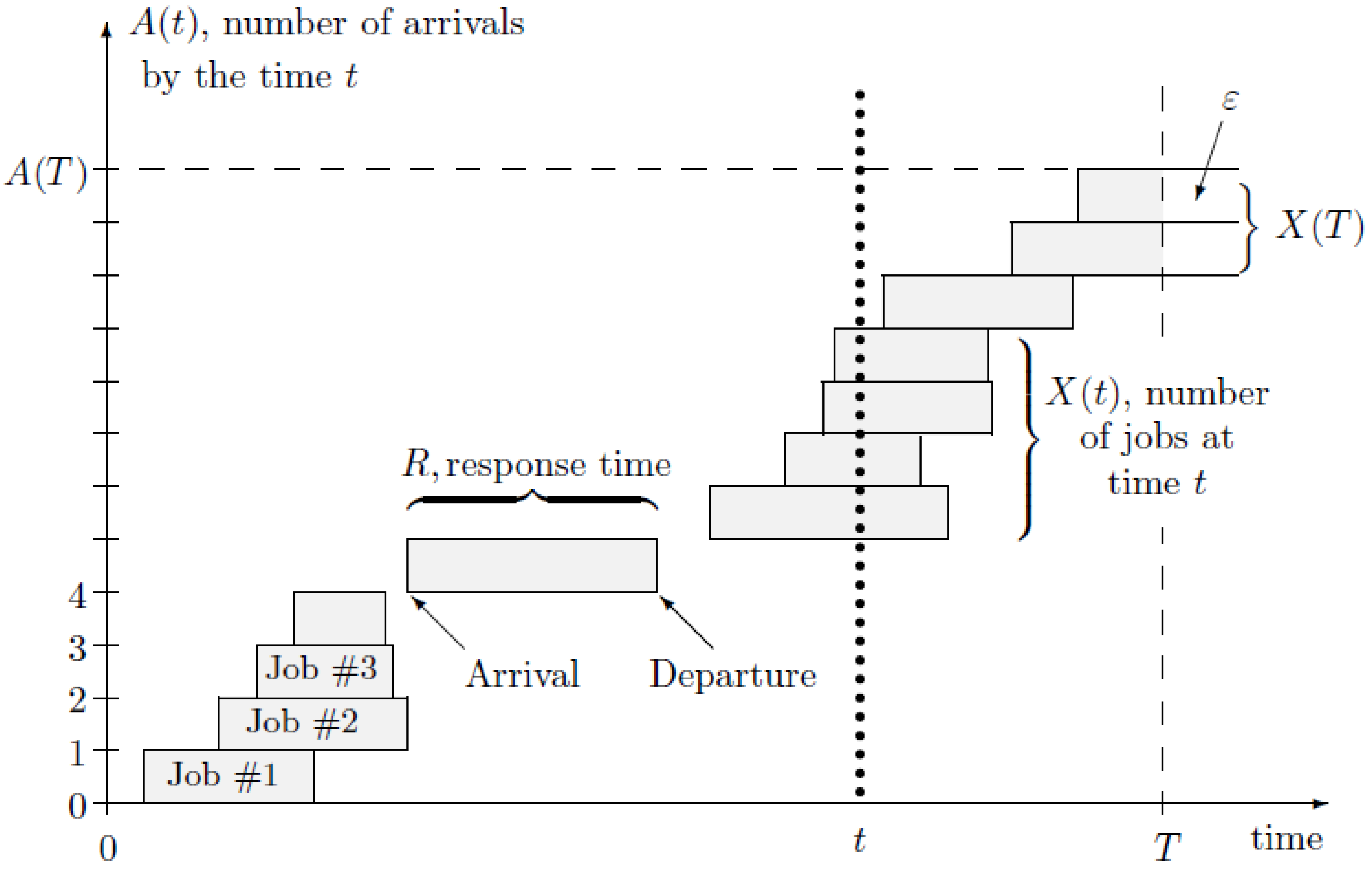
The Little's Law for a stationary system

$$\lambda_A \, \mathbf{E}(R) = \mathbf{E}(X)$$

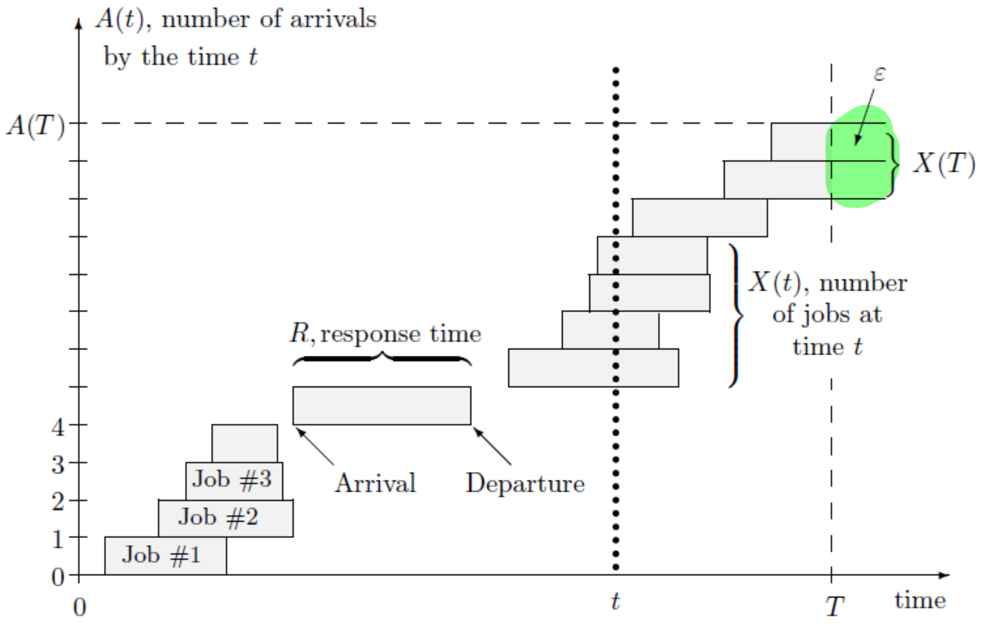
Intuition:

4-queuing systems

Proof of Little's Law



Proof of Little's Law



$$\left(\sum_{k=1}^{A(T)} R_k\right) - \varepsilon = \int_0^T X(t)\,dt.$$

$$\lim_{T\to\infty} \frac{1}{T}\,\mathbf{E}\left(\sum_{k=1}^{A(T)} R_k - \varepsilon\right) = \lim_{T\to\infty} \frac{\mathbf{E}(A(T))\,\mathbf{E}(R)}{T} - 0 = \lambda_A\,\mathbf{E}(R).$$

$$\lim_{T\to\infty} \frac{1}{T}\,\mathbf{E}\int_0^T X(t)\,dt = \mathbf{E}(X).$$

Application

Example 7.1 (Queue in a bank). You walk into a bank at 10:00. Being there, you count a total of 10 customers and assume that this is the typical, average number. You also notice that on the average, customers walk in every 2 minutes. When should you expect to finish services and leave the bank?

Solution. We have $\mathbf{E}(X) = 10$ and $\mu_A = 2$ min. By the Little's Law,

$$\mathbf{E}(R) = \frac{\mathbf{E}(X)}{\lambda_A} = \mathbf{E}(X)\mu_A = (10)(2) = \underline{20 \text{ min}}.$$

Since $\lambda_A \cdot E(R) = E(X)$

Other corollaries

$$\mathbf{E}(X_w) = \lambda_A \, \mathbf{E}(W),$$

$$\mathbf{E}(X_s) = \lambda_A \, \mathbf{E}(S) = \lambda_A \mu_S = r.$$

r = utilization $= \frac{\lambda_A}{\lambda_S}$

Bernoulli single server system

Bernoulli single-server queuing process is a discrete-time queuing process with the following characteristics:

- one server
- unlimited capacity -- the waiting queue can by arbitrarily long
- arrivals occur according to a Binomial process, and the probability of a new arrival during each frame is p_A
- the probability of a service completion (and a departure) during each frame is p_S provided that there is at least one job in the system at the beginning of the frame
- service times and interarrival times are independent

Markov property

changing the queue size does not depend on history

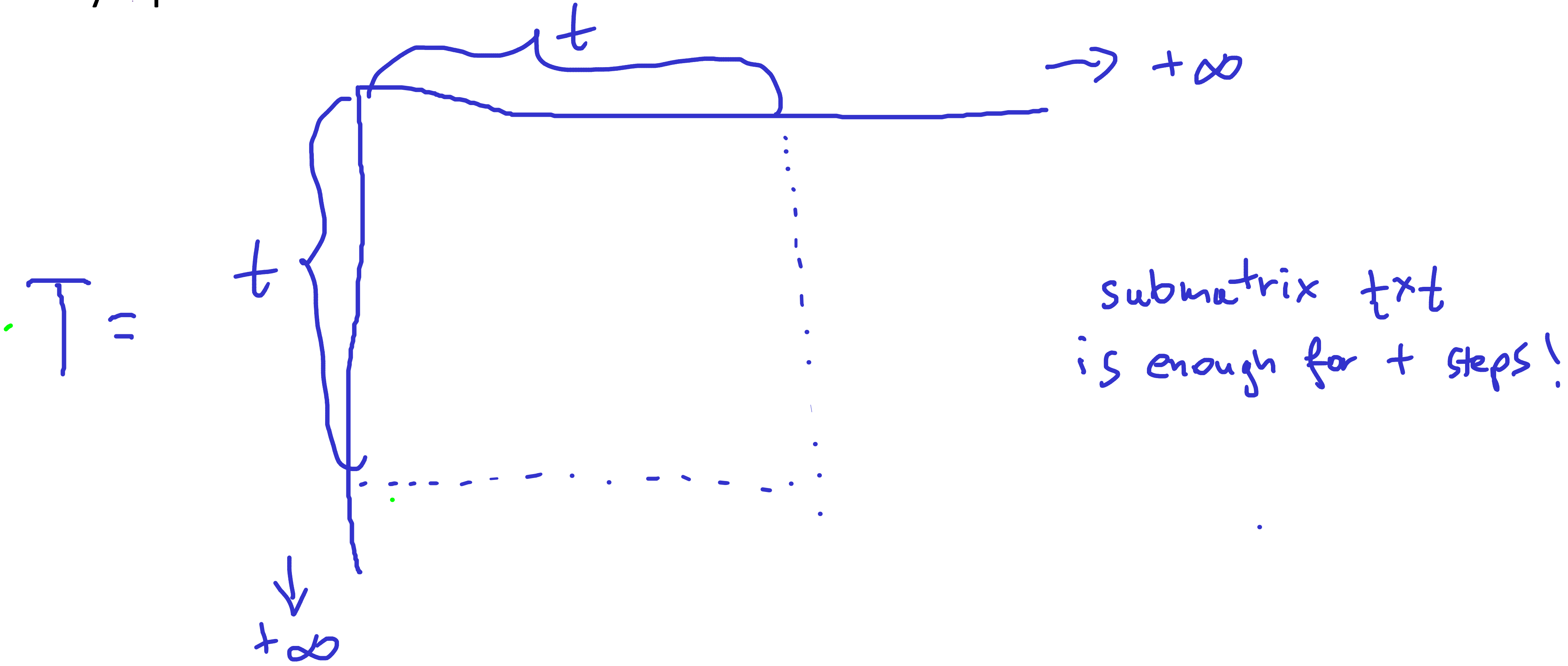
$$
\begin{aligned}
p_{00} &= P\{\text{ no arrivals }\} &= 1 - p_A \\
p_{01} &= P\{\text{ new arrival }\} &= p_A
\end{aligned}
$$

$$
\begin{aligned}
p_{i,i-1} &= P\{\text{ no arrivals } \bigcap \text{ one departure }\} &= (1-p_A)p_S \\
p_{i,i} &= P\{\text{ no arrivals } \bigcap \text{ no departures }\} \\
&\quad + P\{\text{ one arrival } \bigcap \text{ one departure }\} &= (1-p_A)(1-p_S) + p_A p_S \\
p_{i,i+1} &= P\{\text{ one arrival } \bigcap \text{ no departures }\} &= p_A(1-p_S)
\end{aligned}
$$

applications

Distribution of the number of jobs in a queue after t steps

- Take only a part of the transition matrix



Another model: Queue maximal size C

$$p_{C,C-1} = (1 - p_A)p_S.$$

$$p_{C,C} = (1 - p_A)(1 - p_S) + p_A p_S + p_A(1 - p_S) = 1 - (1 - p_A)p_S.$$

Example 7.3 (TELEPHONE WITH TWO LINES). Having a telephone with 2 lines, a customer service representative can talk to a customer and have another one "on hold." This is a system with limited capacity $C = 2$. When the capacity is reached and someone tries to call, (s)he will get a busy signal or voice mail.

Steady distribution?

Average 10 calls per hour, average duration 4 minutes

$$
\begin{aligned}
p_A &= \lambda_A \Delta = 1/6, \\
p_S &= \lambda_S \Delta = 1/4.
\end{aligned}
$$

$$
P = \begin{pmatrix} 1-p_A & p_A & 0 \\ (1-p_A)p_S & (1-p_A)(1-p_S)+p_A p_S & p_A(1-p_S) \\ 0 & (1-p_A)p_S & 1-(1-p_A)p_S \end{pmatrix}
$$

$$
= \begin{pmatrix} 5/6 & 1/6 & 0 \\ 5/24 & 2/3 & 1/8 \\ 0 & 5/24 & 19/24 \end{pmatrix}.
$$

Steady distribution

$$\pi P = \pi \;\Rightarrow\; \begin{cases} \frac{5}{6}\pi_0 + \frac{5}{24}\pi_1 = \pi_0 \\ \frac{1}{6}\pi_0 + \frac{2}{3}\pi_1 + \frac{5}{24}\pi_2 = \pi_1 \\ \frac{1}{8}\pi_1 + \frac{19}{24}\pi_2 = \pi_2 \end{cases}$$

$$\pi_0 = 25/57 = \underline{0.439}, \quad \pi_1 = 20/57 = \underline{0.351}, \quad \pi_2 = 12/57 = \underline{0.210}.$$

Continuous time queuing system

An **M/M/1 queuing process** is a continuous-time Markov queuing process with the following characteristics,

- one server;
- unlimited capacity;
- Exponential interarrival times with the arrival rate λ_A;
- Exponential service times with the service rate λ_S;
- service times and interarrival times are independent.

Limit of Bernoulli queueing system

$$
\begin{aligned}
p_{00} &= 1 - p_A &= 1 - \lambda_A \Delta \\
p_{10} &= p_A &= \lambda_A \Delta
\end{aligned}
$$

$$
\begin{aligned}
p_{i,i-1} &= (1 - p_A) p_S = (1 - \lambda_A \Delta) \lambda_S \Delta \approx \lambda_S \Delta \\
p_{i,i+1} &= p_A (1 - p_S) = \lambda_A \Delta (1 - \lambda_S \Delta) \approx \lambda_A \Delta \\
p_{i,i} &= (1 - p_A)(1 - p_S) + p_A p_S \approx 1 - \lambda_A \Delta - \lambda_S \Delta
\end{aligned}
$$

$$
P \approx \begin{pmatrix}
1 - \lambda_A \Delta & \lambda_A \Delta & 0 & 0 & \cdots \\
\lambda_S \Delta & 1 - \lambda_A \Delta - \lambda_S \Delta & \lambda_A \Delta & 0 & \cdots \\
0 & \lambda_S \Delta & 1 - \lambda_A \Delta - \lambda_S \Delta & \lambda_A \Delta & \cdots \\
0 & 0 & \lambda_S \Delta & 1 - \lambda_A \Delta - \lambda_S \Delta & \ddots \\
\vdots & \vdots & \vdots & \ddots & \ddots
\end{pmatrix}
$$

Limit of Bernoulli queueing system – steady distribution

Looking for π such that
$$\begin{cases} \pi P = \pi \\ \sum \pi_i = 1 \end{cases}$$

$$\pi_0(1-\lambda_A\Delta)+\pi_1\lambda_S\Delta=\pi_0 \quad\Rightarrow\quad \lambda_A\Delta\pi_0=\lambda_S\Delta\pi_1 \quad\Rightarrow\quad \boxed{\lambda_A\pi_0=\lambda_S\pi_1}.$$

$$\pi_0\lambda_A\Delta+\pi_1(1-\lambda_A\Delta-\lambda_S\Delta)+\pi_2\lambda_S\Delta=\pi_1 \quad\Rightarrow\quad (\lambda_A+\lambda_S)\pi_1=\lambda_A\pi_0+\lambda_S\pi_2.$$

$$\boxed{\lambda_A\pi_1=\lambda_S\pi_2}.$$

And so on …

$$\boxed{\lambda_A\pi_{i-1}=\lambda_S\pi_i} \quad \text{or} \quad \boxed{\pi_i=r\,\pi_{i-1}}$$

Steady distribution

$$\sum_{i=0}^{\infty} \pi_i = \sum_{i=0}^{\infty} r^i \pi_0 = \frac{\pi_0}{1-r} = 1 \quad \Rightarrow \quad \begin{cases} \pi_0 & = \quad 1 - r \\ \pi_1 & = \quad r\pi_0 = r(1-r) \\ \pi_2 & = \quad r^2\pi_0 = r^2(1-r) \\ & \qquad \text{etc.} \end{cases}$$

where $r = \frac{\lambda_A}{\lambda_S}$ (utilization)

Steady distribution

This distribution of $X(t)$ is *Shifted Geometric*, because $Y = X + 1$ has the standard Geometric distribution with parameter $p = 1 - r$,

$$P\{Y = y\} = P\{X = y - 1\} = \pi_{y-1} = r^{y-1}(1-r) = (1-p)^{y-1}p \text{ for } y \geq 1,$$

$$\mathbf{E}(X) = \mathbf{E}(Y-1) = \mathbf{E}(Y) - 1 = \frac{1}{1-r} - 1 = \frac{r}{1-r}$$

$$\mathrm{Var}(X) = \mathrm{Var}(Y-1) = \mathrm{Var}(Y) = \frac{r}{(1-r)^2}$$

Waiting time for X jobs

$$W = S_1 + S_2 + S_3 + \ldots + S_X$$

$$\mathbf{E}(W) = \mathbf{E}(S_1 + \ldots + S_X) = \mathbf{E}(S)\,\mathbf{E}(X) = \frac{\mu_S\, r}{1-r} \quad \text{or} \quad \frac{r}{\lambda_S(1-r)}$$

Response time

$$\mathbf{E}(R) = \mathbf{E}(W) + \mathbf{E}(S) = \frac{\mu_S\, r}{1-r} + \mu_S = \frac{\mu_S}{1-r} \quad \text{or} \quad \frac{1}{\lambda_S(1-r)}.$$

Queue size

$$X_w = X - X_s.$$

$$\mathbf{E}(X_w) = \mathbf{E}(X) - \mathbf{E}(X_s) = \frac{r}{1-r} - r = \frac{r^2}{1-r}.$$