Metody probabilistyczne i statystyka, 2022
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)
- Many strategies, for example: first-in-first-out served
- service time is also random

Examples: Web server

Main parameters

<u>Parameters of a queuing system</u>

λ_A = arrival rate = average number of jobs arriving in one time unit

λ_S = service rate = average numer of jobs served in one unit of time

μ_A = $1/\lambda_A$ = mean interarrival time

μ_S = $1/\lambda_S$ = mean service time

r = $\lambda_A/\lambda_S = \mu_S/\mu_A$ = 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

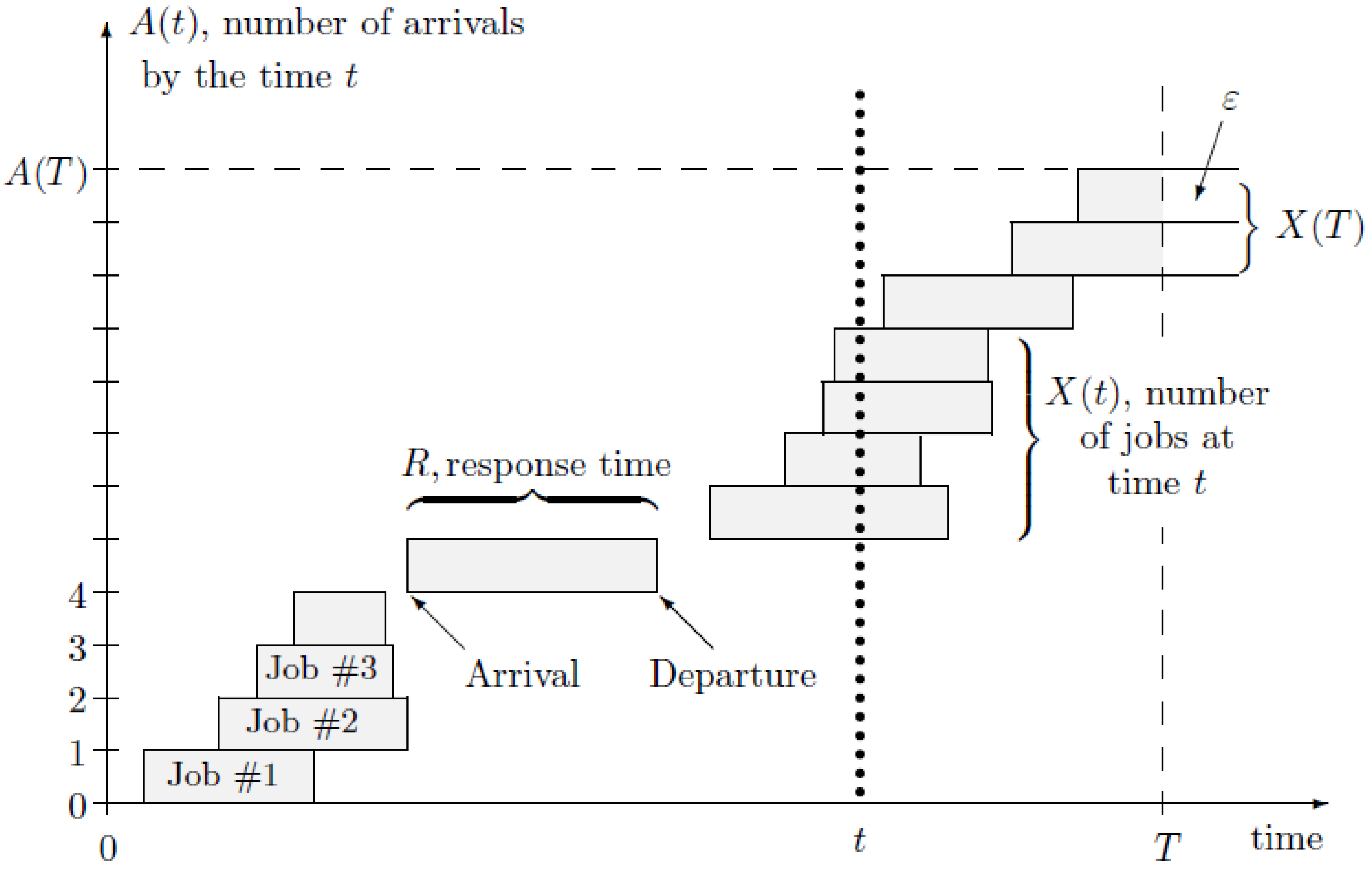
The Little's Law for a stationary system

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

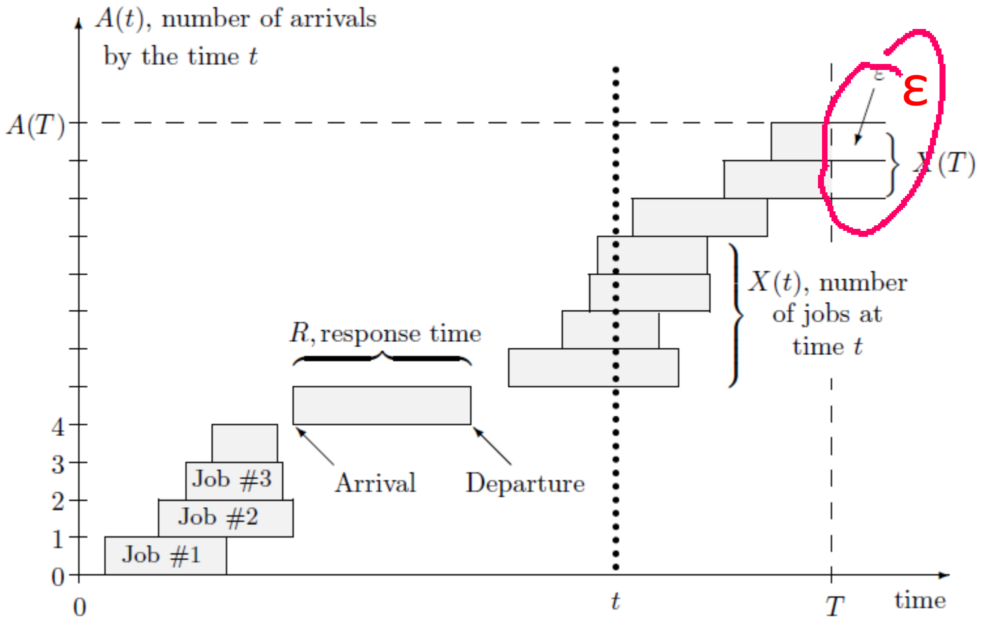
Intuition:

If there are 5 clients coming per minute, each spends 2 minutes, then
This creates 5*2=10 person-minutes per minute. This corresponds to
10 clients
Present.

Proof of Little's Law



Proof of Little’s Law



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

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

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

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,

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

Similar results (and proofs)

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

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

$$\underset{\displaystyle \frac{\lambda_A}{\lambda_S}}{\shortparallel}$$

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

Bernoulli single server system

- discrete time proces
- one server
- unlimited capacity (queue of arbitrary length)
- arrivals in a time unit: 1 new job with pbb p_A
- pbb of completing a job in a time unit: p_S
- arrivals and service completion – independent events

Markov property

changing the queue size does not depend on history

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

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

Applications

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

- Take only a part of the transition matrix

$$P = \begin{pmatrix} 1-p_A & p_A & 0 & \cdots \\ (1-p_A)p_S & \begin{array}{c}(1-p_A)(1-p_S)\\ +p_Ap_S\end{array} & p_A(1-p_S) & \cdots \\ 0 & (1-p_A)p_S & \begin{array}{c}(1-p_A)(1-p_S)\\ +p_Ap_S\end{array} & \cdots \\ 0 & 0 & (1-p_A)p_S & \ddots \\ \vdots & \vdots & \ddots & \ddots \end{pmatrix}$$

Another model: queue maximal size C

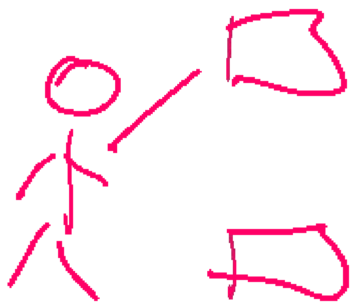
the only changes:

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

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

the transition matrix is of size (C+1) x (C+1)

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

$$
\begin{aligned}
p_{00} &= 1 - p_A &= 1 - \lambda_A \Delta \\
p_{01} &= p_A &= \lambda_A \Delta
\end{aligned}
$$

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

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

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

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

where r is utilization: λ_A / λ_S

service is busy with pbb r
idle with pbb 1-r

Steady distribution

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

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

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

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

What happens if r is close to 1?

Waiting time after arrival

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

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

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

$$X_w = X - X_s.$$

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

- ❑X_s is either 0 or 1
- ❑X_s is 1 iff the system is busy
- ❑System is busy with pbb r

Multiserver systems, k servers

$$
\begin{aligned}
p_{i,i+1} &= \lambda_A\Delta \cdot (1-\lambda_S\Delta)^n \approx \lambda_A\Delta = p_A \\
p_{i,i} &= \lambda_A\Delta \cdot n\lambda_S\Delta(1-\lambda_S\Delta)^{n-1} + (1-\lambda_A\Delta)\cdot(1-\lambda_S\Delta)^n \\
&\approx 1-\lambda_A\Delta - n\lambda_S\Delta = 1 - p_A - np_S \qquad (7.12) \\
p_{i,i-1} &\approx n\lambda_S\Delta = np_S \\
p_{i,j} &= 0 \text{ for all other } j.
\end{aligned}
$$

Again, $n = \min\{i,k\}$ is the number of jobs receiving service among the total of i jobs in the system.

Multiserver systems, matrix for k=3

$$P \approx \begin{pmatrix} 1-p_A & p_A & 0 & 0 & 0 & \ddots \\ p_S & 1-p_A-p_S & p_A & 0 & 0 & \ddots \\ 0 & \mathbf{2}p_S & 1-p_A-2p_S & p_A & 0 & \ddots \\ 0 & 0 & \mathbf{3}p_S & 1-p_A-3p_S & p_A & \ddots \\ 0 & 0 & 0 & \mathbf{3}p_S & 1-p_A-3p_S & \ddots \\ 0 & 0 & 0 & 0 & \mathbf{3}p_S & \ddots \\ \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}$$

Multiserver systems –steady distribution

$$\begin{cases} \pi_0(1-p_A)+\pi_1 p_S=\pi_0 \quad \Rightarrow \quad \pi_0 p_A=\pi_1 p_S \quad \Rightarrow \quad \pi_0 p_A=\pi_1 p_S \quad \Rightarrow \quad \boxed{\pi_1=r\pi_0}, \\ \pi_0 p_A+\pi_1(1-p_A-p_S)+2\pi_2 p_S=\pi_1 \quad \Rightarrow \quad \pi_1 p_A=2\pi_2 p_S \quad \Rightarrow \quad \boxed{\pi_2=2r\pi_1}. \end{cases}$$

$$\left\{ \begin{array}{lllll} \pi_1 & = & r\pi_0 & & \\ \pi_2 & = & r\pi_1/2 & = & r^2\pi_0/2! \\ \pi_3 & = & r\pi_2/3 & = & r^3\pi_0/3! \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \pi_k & = & r\pi_{k-1}/k & = & r^k\pi_0/k! \end{array} \right. \qquad \left\{ \begin{array}{lllll} \pi_{k+1} & = & (r/k)\pi_k & = & (r/k)\, r^k\pi_0/k! \\ \pi_{k+2} & = & (r/k)\pi_{k+1} & = & (r/k)^2\, r^k\pi_0/k! \\ \text{etc.} & & & & \end{array} \right.$$

Multiserver systems –steady distribution

$$\begin{aligned} 1 &= \pi_0 + \pi_1 + \dots \\ &= \pi_0 \left(1 + r + \frac{r^2}{2!} + \frac{r^3}{3!} + \dots + \frac{r^k}{k!} + \frac{r^k}{k!}(r/k) + \frac{r^k}{k!}(r/k)^2 + \dots \right) \\ &= \pi_0 \left(\sum_{i=0}^{k-1} \frac{r^i}{i!} + \frac{r^k}{(1 - r/k)k!} \right), \end{aligned}$$