# Case Study 3: Queueing Systems

Alain Jean-Marie
INRIA/LIRMM, Université Montpellier, CNRS
Alain.Jean-Marie@inria.fr

RESCOM Summer School 2019
26 June 2019

# Contents

# The simple discrete-time queue

Consider a single-server queue in discrete time.
In each slot:

- arrival of a batch of customers with size distributed according to a general probability distribution:

$$\mathbb{P}\{A = k\} = a_k \ .$$

- service of one customer, if at least one is present.

If the odrer of events is:

1. arrival of batches
2. departure of customers

# Evolution equations

Let $Q_n$ denote the number of customers in queue at time $n$, just after departures, but before new arrivals.
The evolution of this variable is given by:

$$Q_{n+1} \; = \; [Q_n \; + \; A_n \; - \; 1]^+.$$

Let $R_n$ denote the number of customers in queue at time $n$, just before departures and new arrivals.
The evolution of this variable is given by:

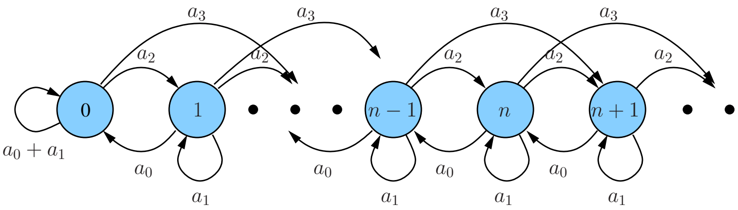
$$R_{n+1} \; = \; [R_n \; - \; 1]^+ \; + \; A_n.$$

Both are connected by:

$$R_n \; = \; Q_n + A_n \; .$$

# Transition diagram, transition Matrix

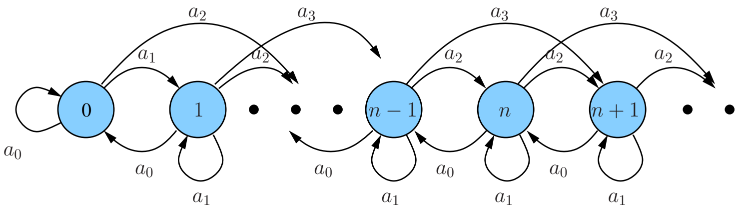
For the Markov chain $\{Q_n; n \in \mathbb{N}\}$:




$$\mathbf{P} = \begin{pmatrix} a_0 + a_1 & a_2 & a_3 & \cdots \\ a_0 & a_1 & a_2 & \ddots \\ 0 & a_0 & a_1 & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{pmatrix}$$

# Transition diagram, transition Matrix

For the Markov chain $\{R_n \; n \in \mathbb{N}\}$:




$$\mathbf{P} = \begin{pmatrix} a_0 & a_1 & a_2 & \dots \\ a_0 & a_1 & a_2 & \ddots \\ 0 & a_0 & a_1 & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{pmatrix}$$

# Solution

Solving for the stationary equations (process $\{R_n\}$):

$$\begin{pmatrix} \pi_0 & \pi_1 & \pi_2 & \dots \end{pmatrix} = \begin{pmatrix} \pi_0 & \pi_1 & \pi_2 & \dots \end{pmatrix} \cdot \begin{pmatrix} a_0 & a_1 & a_2 & \dots \\ a_0 & a_1 & a_2 & \ddots \\ 0 & a_0 & a_1 & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{pmatrix}$$

$$\begin{aligned} \pi_0 &= a_0\pi_0 + a_0\pi_1 \\ \pi_1 &= a_1\pi_0 + a_1\pi_1 + a_0\pi_2 \\ \dots & \quad \dots \\ \pi_n &= a_n\pi_0 + \sum_{k=1}^{n+1} a_{n+1-k}\pi_k \end{aligned}$$

# The generating function approach

Recurrences such as

$$\pi_n = a_n \pi_0 + \sum_{k=0}^{n+1} a_{n+1-k} \pi_k, \qquad n \geq 0$$

can be handled with generating functions.

One introduces the probability generating functions:

$$R(z) \; := \; \sum_{n=0}^{\infty} \pi_n z^n \; = \; \mathbb{E}(z^R) \qquad\qquad A(z) \; := \; \sum_{n=0}^{\infty} a_n z^n \; = \; \mathbb{E}(z^A) \; .$$

From the recurrence, it is deduced that:

$$R(z) = \pi_0 A(z) \left(1 - \frac{1}{z}\right) \; + \; \frac{A(z)}{z} \; R(z) \; .$$

The function $R(z)$ is therefore solution of the equation:

$$R(z) \; = \; \pi_0 A(z) \, \frac{1 - z}{A(z) - z} \; .$$

# The generating function approach (ctd.)

The quantity $\pi_0$ is unknown in the formula:

$$R(z) \;=\; \pi_0 A(z) \, \frac{1-z}{A(z)-z} \; .$$

One way to determine it is letting $z \to 1$. For probability generating functions: $R(1) = A(1) = 1$. Then, using L'Hôpital's rule,

$$1 \;=\; \pi_0 a_0 \, \frac{-1}{A'(1)-1}$$

and $A'(1) = \mathbb{E}A$. Conclusion:

$$\pi_0 \;=\; 1 - \mathbb{E}A.$$

Or ... directly with Little's formula!
Finally:

$$\boxed{R(z) \;=\; (1-\mathbb{E}A) \, \frac{A(z)(1-z)}{A(z)-z} \; .}$$

# The generating function approach (end)

From the generating function, one recovers the moments of $\boldsymbol{\pi}$, the distributions of $R$ and $Q$:

$$\mathbb{E}Q = \frac{\mathbb{E}A^2 - \mathbb{E}A}{2(1-\mathbb{E}A)}$$

$$\mathbb{E}R = \mathbb{E}Q + \mathbb{E}A$$

$$\mathbb{E}Q^2 = \frac{3(\mathbb{E}A^2)^2 - 9\mathbb{E}A^2\mathbb{E}A + 6(\mathbb{E}A)^2 + 2\mathbb{E}A^3 - 2\mathbb{E}A^3\mathbb{E}A + 3\mathbb{E}A^2 - 3\mathbb{E}A}{6(1-\mathbb{E}A)^2}$$

$$\mathbb{E}R^2 = \mathbb{E}Q^2 + 2\mathbb{E}Q\mathbb{E}A + (\mathbb{E}A)^2 \ .$$

and for specific distributions of $A$, the distribution of $Q$.
For instance, if $A \sim \mathtt{Geom}(\rho)$:

$$\pi_k \ = \ (1-2\rho)\,\frac{\rho^k}{(1-\rho)^{k+1}}\ , \quad k \geq 1, \qquad \pi_0 \ = \ \frac{1-2\rho}{1-\rho}\ .$$

# Model variant #1: finite capacity

New rules:

- Buffer capacity $K$
- Partial batches accepted up to capacity.

New diagram, new $(K+1) \times (K+1)$ matrix:

$$\mathbf{P} = \begin{pmatrix} a_0 & a_1 & a_2 & \dots & a_{K-1} & a_K + a_{K+1} + \dots \\ a_0 & a_1 & a_2 & \dots & a_{K-1} & a_K + a_{K+1} + \dots \\ 0 & a_0 & a_1 & \ddots & & \\ \vdots & \ddots & \ddots & \ddots & & \\ 0 & \dots & & & a_0 & a_1 + a_2 + \dots \end{pmatrix}$$

# Model variant #2: services of geometric duration

The geometric law is memoryless: if $G \sim \mathtt{Geom}(\rho)$ (on the set $\{1, 2, \ldots\}$),

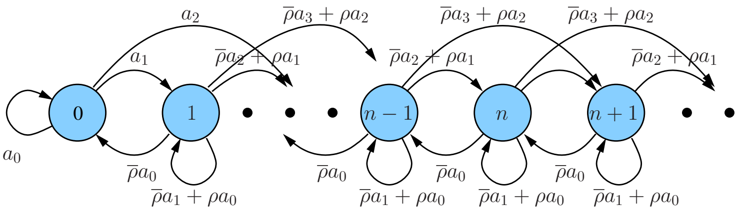
$$\mathbb{P}\{G \geq n + m | G \geq n\} \;=\; \mathbb{P}\{G \geq m\}\ .$$

Another way of looking at it: if $G \sim \mathtt{Geom}(\rho)$,

- with probability $1 - \rho$, $G = 1$
- with probability $\rho$, $G = 1 + G'$ with $G' \sim \mathtt{Geom}(\rho)$.

So, transitions from a queue size $n > 0$ are now:

- with probability $(1 - \rho)a_i$, to $n - 1 + i$
- with probability $\rho a_i$, to $n + i$.



# Model variant #3: deterministic services

Assume now that services last $s$ slots. The process $\{R_n\}$ is not Markovian anymore.

How to fix this? Two main ideas

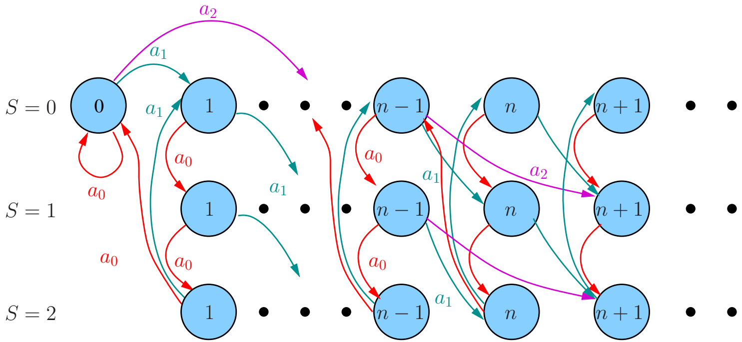
- adding variables to the state space: the "method of phases" of "method of the supplementary variable"
- or embedding the process at specific times

such that the resulting process is now Markovian.

Both methods of phases and embedding can be generalized to general service distributions.

# The method of phases

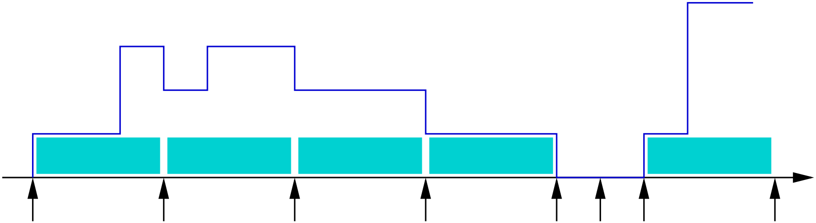
Let $S_n$ be the amount of service given to the customer in service, at the beginning of slot $n$, just after arrivals.
Then the process $(R_n, S_n) \in \mathbb{N} \times \{0, 1, \ldots, s-1\}$ is a Markov Chain.
Its diagram ($s = 3$):



# The method of embedding

Let $\{B_m\}$ be the process of the number of customers in queue when:

- either some service begins
- some slot begins if no customer in service



This is a Markov chain with transitions

- for $n > 0$: $n \to n + a - 1$ with probability $\mathbb{P}\{A_1 + A_2 + A_3 = a\}$
- for $n = 0$: $n \to n + a$ with probability $\mathbb{P}\{A = a\}$.

# Model variant #4: Batch services

Assume that there are $B$ servers that work in parallel.
The evolution equation is now:

$$R_{n+1} \;=\; [R_n - B]^+ + A_n \;.$$

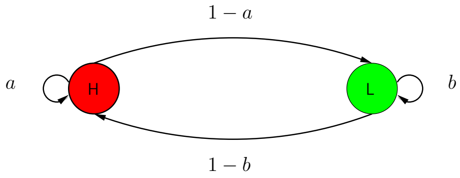
Transition matrix for $B = 2$:

$$\mathbf{P} \;=\; \begin{pmatrix} a_0 & a_1 & a_2 & \cdots \\ a_0 + a_1 & a_2 & a_3 & \ddots \\ a_0 & a_1 & a_2 & \ddots \\ 0 & a_0 & a_1 & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{pmatrix}$$

# Model variant #5: Arrivals with phases

It may happen that the arrival process undergoes phases: e.g. some with high arrival rate, some with low arrival rate.

Assume that the switch from H to L occurs according to a Markov Chain:




When the phase:

- is H, batch sizes have probabilities $h_i = \mathbb{P}\{A = i | H\}$
- is L, batch sizes have probabilities $\ell_i = \mathbb{P}\{A = i | L\}$.

# Arrivals with phases, ctd.

Transition table:

| origin | destination | probability | restriction |
|---|---|---|---|
| $(H, 0)$ | $(H, n)$ | $ah_n$ | $n \geq 0$ |
| $(H, 0)$ | $(L, n)$ | $(1-a)h_n$ | $n \geq 0$ |
| $(L, 0)$ | $(L, n)$ | $b\ell_n$ | $n \geq 0$ |
| $(L, 0)$ | $(H, n)$ | $(1-b)\ell_n$ | $n \geq 0$ |
| $(H, m)$ | $(H, m+n-1)$ | $ah_n$ | $m > 0, n \geq 0$ |
| $(H, m)$ | $(L, m+n-1)$ | $(1-a)h_n$ | $m > 0, n \geq 0$ |
| $(L, m)$ | $(L, m+n-1)$ | $b\ell_n$ | $m > 0, n \geq 0$ |
| $(L, m)$ | $(H, m+n-1)$ | $(1-b)\ell_n$ | $m > 0, n \geq 0$ |

# Model variant #6: Impatient customers

Assume that customers are impatient: with probability $\alpha$, each of them may leave the queue if not being served.

How to calculate $P_{ij}$?
Introduce $Z_n$, the number of *patient* customers, remaining after

- service has begun
- impatient customers have left

Then

$$R_{n+1} = Z_n + A_n \; .$$

$$\begin{aligned}
P_{ij} &= \mathbb{P}\{R_{n+1} = j | R_n = i\} \\
&= \mathbb{P}\{Z_n + A_n = j | R_n = i\} \\
&= \sum_{z=0}^{i} \mathbb{P}\{Z_n + A_n = j | Z_n = z\} \; \mathbb{P}\{Z_n = z | R_n = i\} \\
&= \sum_{z=0}^{i} \mathbb{P}\{A_n = j - z\} \; \mathbb{P}\{Z_n = z | R_n = i\}
\end{aligned}$$

# Impatient customers, ctd.

There remains to compute $\mathbb{P}\{Z_n = z | R_n = i\}$.
If $i = 0$: $\mathbb{P}\{Z_n = 0 | R_n = 0\} = 1$.
If $i > 0$, one customer enters service. Impatience among the $i - 1$ remaining ones:

$$\begin{aligned}\mathbb{P}\{Z_n = z | R_n = i\} &= \mathbb{P}\{z \text{ stay out of } i - 1\} \\ &= \binom{i-1}{z} \alpha^{z-i+1} (1 - \alpha)^z \ .\end{aligned}$$

Finally, for $i > 0$,

$$\begin{aligned}P_{ij} &= \mathbb{P}\{R_{n+1} = j | R_n = i\} \\ &= \sum_{z=0}^{i-1} a_{j-z} \binom{i-1}{z} \alpha^{z-i+1} (1 - \alpha)^z \ .\end{aligned}$$

# Model variant #7: service with threshold

Assume that the server does not start before there are at least $\nu$ customers in the queue. Add also the buffer capacity $K$.
The evolution equations are now

- if $R_n < \nu$: $R_{n+1} = \min\{R_n + A_n, K\}$
- if $R_n \geq \nu$: $R_{n+1} = \min\{R_n - 1 + A_n, K\}$

The chain now evolves over $\mathcal{E} = \{\nu - 1, \nu, \ldots, K\}$.

Starting the server has cost $C_s$.
Losing a customer because buffer capacity is exceeded has cost $C_L$.
What is the best $\nu$?

# Average cost in the queue

Evaluation of the average cost:

$$J_\nu = \lim_{n\to\infty} \frac{1}{N}\,\mathbb{E}\left(\sum_{n=0}^{N-1}(C_s \mathbf{1}_{\{\text{service starts at } n\}} + C_L \#\{\text{customers lost at } n\}\right) .$$

Considering this is a Markov reward process, we have:

$$\begin{aligned} J_\nu &= C_s \pi_{\nu-1} \mathbb{P}\{A_n > 0\} + C_L \sum_{i=\nu-1}^{K} \pi_i \mathbb{E}(A_n - (K-i)) \\ &= C_s \pi_{\nu-1}(1-a_0) + C_L \sum_{i=\nu-1}^{K} \pi_i \sum_{j=K-i}^{\infty} j a_j \ . \end{aligned}$$

$\rightarrow$ to be evaluated for each possible $\nu$.

# Summary

In this “case study”:

- description of the queue in discrete time; order of the events
- equations of evolution, useful for the
- construction of probability transition matrices
- many variants with service duration, arrival processes, service discipline
- setup of an optimization problem.

More to be done in the Lab!