Abstract. In this paper, the duration of perfect simulations for Markovian finite capacity queueing networks is studied. This corresponds to hitting time (or coupling time) problems in a Markov chain over the Cartesian product of the state space of each queue. We establish an analytical formula for the expected simulation time in the one queue case and we provide simple bounds for acyclic networks of queues with losses. These bounds correspond to sums on the coupling time for each queue and are either linear in the queue capacities under light or heavy traffic assumptions or quadratic, when service and arrival rates are similar.

Key words. Perfect simulation, Markov chain, Hitting time

AMS subject classifications. 60J10, 60J22, 65C40, 65C20, 68U20

1. Introduction. Markov chains are an important tool in modelling systems. Amongst others, Markov chains are being used in the theory of queueing systems, which itself is used in a variety of applications as performance evaluation of computer systems and communication networks. In modelling any queueing system, one of the main points of interest is the long run behavior of the system. For an irreducible, ergodic (i.e. aperiodic and positive-recurrent) Markov chain with probability matrix \( P \), this long run behavior is described by the unique vector \( \pi \) which satisfies the linear system

\[
\pi = \pi P.
\]

We shall refer to the vector \( \pi \) as the stationary distribution. In most of the applications, the state space \( S \) of the Markov chain is finite and the chain is irreducible and aperiodic. Because in a finite, irreducible and aperiodic Markov chain all states are positive recurrent, the chain is ergodic. So for a finite Markov chain, irreducibility and aperiodicity are sufficient conditions for the existence of a unique stationary distribution or steady-state. However, it may be hard to compute this stationary distribution, especially when the finite state space is huge which is frequent in queuing models. In that case, steady-state simulation [6] can be be used.

The classical method for simulation has been Monte Carlo simulation for many years. This method is based on the fact that an irreducible aperiodic finite Markov chain with transition matrix \( P \) and initial distribution \( \mu(0) \), the distribution \( \mu(n) \) of the chain at time \( n \) converges to \( \pi \) as \( n \) gets very large. That is:

\[
\lim_{n \to \infty} \mu(n) = \lim_{n \to \infty} \mu(0) P^n = \pi.
\]

So after running the Markov chain long enough, the states of the chain will not depend anymore on the initial state. However, the question is how long is long enough? That is, when is \( n \) sufficiently large so that \( |\mu(n) - \pi| \leq \epsilon \) for a certain \( \epsilon > 0 \)? Moreover, the samples generated by this method will always be biased.

\(^*\)This work was partially supported by the French ACI SurePath project and SMS ANR

\(^1\)Jantien Dopper, Mathematical institute, Leiden University, NL (jgdopper@math.leidenuniv.nl) (this author was partially supported by a grant from INRIA), Bruno Gaujal and Jean-Marc Vincent, Laboratoire ID-IMAG, Mescal Project INRIA-UJF-CNRS-INPG, 51, avenue Jean Kuntzmann, F-38330 Montbonnot, France, e-mail : ({Bruno.Gaujal,Jean-Marc.Vincent}@imag.fr)
There are two approaches to estimate the steady-state from simulations. The method based on one long-run uses the ergodic theorem of Markov chains and estimates the stationary probability of a state \( s \) by the proportion of visits in \( s \) on one trajectory of the process. The drawback of this method is the auto-correlation of the sample. Moreover, a warm-up period is necessary in order to begin estimation when the process is near the steady-state. This initial transient problem remains open in many situations, and computation of confidence intervals needs elaborated techniques depending highly on the model structure and parameter values.

The replication method consists in running independent finite trajectories. The advantage is to obtain independent samples of the steady state and classical convergence theorems could be used to compute confidence intervals. The drawback is the importance of the transient part in each of the replications. The simulation time could then be prohibitive. Discussion on one long-run versus replications could be found in [12]. Extensions to regenerative methods [5] decompose the long-run trajectories into “independent” batches. The simulation strategy replication, one long run with batches is discussed in [1].

In 1996, Propp and Wilson [7] solved these problems of Markov chain simulation by proposing an algorithm which returns exact samples of the stationary distribution. The striking difference between Monte Carlo simulation and this new algorithm is that Propp and Wilson do not simulate into the future, but go backwards in time. The main idea is, while going backwards in time, to run several simulations, starting with all \( s \in S \) until the state at \( t = 0 \) is the same for all of them. If the output is the same for all runs, then the chain has coupled. Because of this coupling element and going backwards, this algorithm has been called Coupling From The Past (from now on: CFTP). A more detailed description of this algorithm will be presented in section 2.

When the coupling from the past technique is applicable, we get in a finite time one state with steady-state distribution. Then either we use a one long-run simulation from this state avoiding the estimation of the initial transient problem or we replicate independently the CFTP algorithm to get a sample of independent steady-state distributed variables. The analysis of the choice could be done exactly as in [1]. The replication technique has been applied successfully in finite capacity queueing networks with blocking and rejection (very large state-space) [10]. The efficiency of the simulation allows also the estimation of rare events (blocking probability, rejection rate) is done in [9].

The aim of this paper is to study the simulation time needed to generate one state, steady-state distributed, in the context of queueing networks with finite capacities. We will apply CFTP to networks of queues and study the coupling time \( \tau \) of CFTP (i.e. the smallest time \( t \) for which the chain couples. Our main interest is setting bounds on the expected coupling time. We obtain exact analytical values for the expected simulation time for one M/M/1/C queue. As for networks of queues, we show how upper bounds on the mean simulation time can be obtained as sums of coupling times for each queue. This is used to provide explicit bounds which are linear in the queues capacities for acyclic networks with losses, under light or heavy traffic. However, when in input rate and the service rate are close, the bounds become quadratic in the capacities.

The paper is organized as follows. We first introduce the coupling from the past algorithm in Section 2. Then we show general properties of the coupling time for open Markovian queueing networks in Section 3. We will investigate the M/M/1/c
queue in Section 4 providing exact computation for the expected coupling time and the case of acyclic networks in Section 5 where bounds are derived, together with several experimental tests assessing their quality.

2. Coupling from the Past. Let \( \{X_n\}_{n \in \mathbb{N}} \) be an irreducible and aperiodic discrete time Markov chain with a finite state space \( S \) and a transition matrix \( P = (p_{i,j}) \). Let

\[
\phi : S \times E \to S
\]

encoding the chain, so verifying the property that \( \mathbb{P}(\phi(i,e) = j) = p_{i,j} \) for every pair of states \((i,j) \in S\) and \( e \) a random variable distributed on \( E \). The function \( \phi \) could be considered as an algorithm and \( e \) the innovation for the chain. In the context of discrete event systems, \( e \) is an event and \( \phi \) is the transition function. Now, the evolution of the Markov chain is described as a stochastic recursive sequence

\[
X_{n+1} = \phi(X_n, e_{n+1}), \quad (2.1)
\]

with \( X_n \) the state of the chain at time \( n \) and \( \{e_n\}_{n \in \mathbb{N}} \) an independent and identically distributed sequence of random variables.

Let \( \phi^{(n)} : S \times E^n \to S \) denote the function whose output is the state of the chain after \( n \) iterations and starting in state \( s \in S \). That is,

\[
\phi^{(n)}(s, e_{1\to n}) = \phi(\phi(\phi(\phi(s, e_1), e_2), \ldots, e_{n-1}), e_n).
\]  

This notation can be extended to set of states. So for a set of states \( A \subset S \) we note

\[
\phi^{(n)}(A, e_{1\to n}) = \left\{ \phi^{(n)}(s, e_{1\to n}), s \in A \right\}
\]

**Theorem 1.** Let \( \phi \) be a transition function on \( S \times E \). There exists an integer \( l^* \) such that

\[
\lim_{n \to +\infty} \left| \phi^{(n)}(S, e_{1\to n}) \right| = l^* \text{ almost surely.}
\]

This result is based on the following lemma and the fact that \( S \) is finite.

**Lemma 2.1.** The sequence of integers \( \{a_n\}_{n \in \mathbb{N}} \) defined by

\[
a_n = \left| \phi^{(n)}(S, e_{1\to n}) \right|
\]

is non-increasing.

This is clear because

\[
\phi^{(n)}(S, e_{1\to n}) = \phi(\phi^{(n-1)}(S, e_{1\to n-1}), e_n),
\]

and the cardinal \( a_n \) of the image of \( \phi^{(n-1)}(S, e_{1\to n-1}) \) by \( \phi(., e_n) \) is less or equal than the cardinal \( a_{n-1} \) of \( \phi^{(n-1)}(S, e_{1\to n-1}) \).

To complete the proof of the theorem, consider an arbitrary sequence of events \( \{e_n\}_{n \in \mathbb{N}} \). Lemma 2.1 implies that the sequence \( \{a_n\}_{n \in \mathbb{N}} \) converges to a limit \( l \). Because the sizes of these sets belong to the finite set \( \{1, \ldots, |S|\} \), there exists \( n_0 \in \mathbb{N} \) such that

\[
a_{n_0} = \left| \phi^{(n_0)}(S, e_{1\to n_0}) \right| = l.
\]
Consider now $l^*$ the minimal value of $l$ among all possible sequences of events. Then there exists a sequence of events $\{e^*_n\}_{n \in \mathbb{N}}$ and an integer $n_0^*$ such that

$$\left| \phi^{(n_0^*)} (S, e^*_{1-n_0^*}) \right| = l^*. $$

As a consequence of the Borel-Cantelli lemma, almost all sequences of events $\{e_n\}_{n \in \mathbb{N}}$ include the pattern $e^*_{1-n_0^*}$. Consequently, the limit of the cardinality of $\phi^{(n)} (S, e_{1-n})$ is less than $l^*$. The minimality of $l^*$ finishes the proof.

**Definition 1.** The system couples if

$$\lim_{n \to +\infty} \left| \phi^{(n)} (S, e_{1-n}) \right| = 1 \text{ with probability } 1,$$

Then the forward coupling time $\tau^f$ defined by

$$\tau^f = \min\{n \in \mathbb{N}; \text{ such that } \left| \phi^{(n)} (S, e_{1-n}) \right| = 1\},$$

is almost surely finite. The coupling property of a system $\phi$ depends only on the structure of $\phi$. The probability measure on $\mathcal{E}$ does not affect the coupling property, provided that all events in $\mathcal{E}$ have a positive probability. Moreover, the existence of some pattern $e^*_{1-n_0^*}$ that ensures coupling, guarantees that $\tau^f$ is stochastically upper bounded by a geometric distribution

$$\mathbb{P}(\tau^f \geq k.n_0^*) \leq (1 - p(e_1^*)p(e_2^*)\ldots p(e_{n_0^*}))^k,$$

where $p(e) > 0$ is the probability of event $e$.

At time $\tau^f$, all trajectories issued from all initial states at time 0 have collapsed in only one trajectory. Unfortunately, the distribution of $X_{\tau^f}$ is not stationary. In [4] is given an example that illustrates why it is not possible to consider that the process reaches the stationary regime.

In fact, this iteration scheme could be reversed in time as it is usually done in analysis of stochastic point processes. Extend the sequence of events to the negative part and build the reversed scheme on sets by

$$A_n = \phi^{(n)} (S, e_{-n+1-n}).$$

It is clear that the sequence of sets $A_n$ is non-decreasing ($A_{n+1} \subseteq A_n$). Consequently, if the system couples, the sequence $A_n$ converges almost surely to a set with only one element. There exists almost surely a finite time $\tau^b$, the backward coupling time, defined by

$$\tau^b = \min\{n \in \mathbb{N}; \text{ such that } \left| \phi^{(n)} (S, e_{1-n}) \right| = 1\}.$$

**Proposition 1.** The coupling time $\tau^b$ backward and $\tau^f$ forward are governed by the same probability distribution.

For a detailed proof of this proposition, we refer to [11]. Here is the main idea of the proof. Compute the probability

$$\mathbb{P}(\tau^f > n) = \mathbb{P}(\phi^{(n)} (S, e_{1-n}) > 1).$$
But the process \( \{e_n\}_{n \in \mathbb{Z}} \) is stationary, then shifting the process to the left leads to

\[
P(\|\phi^{(n)}(S, e_{1-n})\| > 1) = P(\|\phi^{(n)}(S, e_{-n+1-0})\| > 1) = P(\tau^b > n).
\]

Hence, if we want to make any statement about the probability distribution of the coupling time \( \tau^b \) of CFTP, we can use the conceptually easier coupling time \( \tau^f \).

The main result of the backward scheme is the following theorem [7].

**Theorem 2.** Provided that the system couples, the limit element of the backward scheme is steady state distributed.

From this fact, a general algorithm (1) sampling the steady state has been established.

**Algorithm 1** Backward-coupling simulation (general version)

```plaintext
for all \( s \in S \) do
    \( y(s) \leftarrow s \) \{choice of the initial value of the vector \( y \), \( n = 0 \}\}
end for
repeat
    \( e \leftarrow \) Random event; \{generation of \( e_{-n+1} \)\}
    for all \( s \in S \) do
        \( y(s) \leftarrow y(\phi(s, e)) \); \{\( y(s) \) state at time 0 of the trajectory issued from \( s \) at time \(-n+1\)\}
    end for
until All \( y(x) \) are equal
return \( y(x) \)
```

The complexity \( c_{\phi} \) of this algorithm for a given sequence of events equals

\[
c_{\phi} = \tau^b |S|.
\]

Coupling time is of fundamental importance for the efficiency of the sampler. To improve the sampler complexity, we could reduce the factor \( |S| \) and reduce the coupling time. When the state space is partially ordered by a partial order \( \preceq \) and the transition function is monotone for each event \( e \), it is sufficient to simulate trajectories issued from maximal and minimal states [7]. Denote by \( M \) and \( m \) the set of maximal, respectively minimal elements of \( S \) for the partial order \( \preceq \). The monotone version of algorithm (1) is given by algorithm (2). In this case, we need to store the sequence of events in order to preserve the coherence between the trajectories driven from \( M \cup m \).

The doubling scheme (first step in the loop) leads to a complexity

\[
c_{\phi} \leq 2.\tau^b(|M| + |m|), \tag{2.3}
\]

which could be small as it will be shown in examples.

3. **Open Markovian queueing networks.** Consider an open network \( Q \) consisting of \( K \) queues \( Q_1, \ldots, Q_K \). Each queue \( Q_i \) has a finite capacity, denoted by \( C_i \), \( i = 1, \ldots, K \). Thus the state space of a single queue \( Q_i \) is \( S_i = \{0, \ldots, C_i\} \). Hence, the state space \( S \) of the network is \( S = S_1 \times \cdots \times S_K \). The state of the system is described by a vector \( s = (s_1, \ldots, s_K) \) with \( s_i \) the number of customers in queue \( Q_i \).
Algorithm 2 Backward-coupling simulation (monotone version)

\[ n=1; \]
\[ R[n]=\text{Random\_event;\{array will R stores the sequence of events \}} \]
\[ \text{repeat} \]
\[ n=2.n; \]
\[ \text{for all } s \in M \cup m \text{ do} \]
\[ y(s) \leftarrow s \text{ \{Initialize all trajectories at time } -n\text{\}} \]
\[ \text{end for} \]
\[ \text{for } i=n \text{ downto } n/2+1 \text{ do} \]
\[ R[i]=\text{Random\_event;\{generates all events from time } -n+1 \text{ to } \frac{n}{2}+1\text{\}} \]
\[ \text{end for} \]
\[ \text{for } i=n \text{ downto } 1 \text{ do} \]
\[ \text{for all } s \in M \cup m \text{ do} \]
\[ y(s) \leftarrow \Phi(y(s),R[i]) \]
\[ \text{end for} \]
\[ \text{end for} \]
\[ \text{until All } y(s) \text{ are equal} \]
\[ \text{return } y(s) \]

The state space is partially ordered by the component-wise ordering and there are a maximal state \( M \) when all queues are full and a minimal state when all queues are empty.

The network evolves in time due to exogenous customer arrivals from outside of the network and to service completions of customers. After finishing his service at a server, a customer is either directed to another queue by a certain routing policy or he leaves the network. A routing policy determines to which queue a customer will go, taking into account the global state of the system. Moreover, the routing policy also decides what happens with a customer if he is directed to a queue which buffer is filled with \( C_i \) customers.

An event in this network is characterized by the movements of some clients between queues modeling the routing strategy and the Poisson process defining the occurrence rate of the event. For example consider the acyclic queueing network (figure 3.1) is characterized by 4 queues and 6 events.

\[
\begin{array}{c}
\lambda_1 \\
\downarrow \\
C_4 \\
\downarrow \\
\lambda_2 \\
\downarrow \\
C_2 \\
\downarrow \\
\lambda_3 \\
\downarrow \\
C_1 \\
\downarrow \\
\lambda_4 \\
\downarrow \\
C_3 \\
\downarrow \\
\lambda_5
\end{array}
\]

<table>
<thead>
<tr>
<th>rate</th>
<th>origin</th>
<th>destination</th>
<th>enabling condition</th>
<th>routing policy</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \lambda_0 )</td>
<td>( Q_{-1} )</td>
<td>( Q_0 )</td>
<td>none</td>
<td>rejection if ( Q_0 ) is full</td>
</tr>
<tr>
<td>( \lambda_1 )</td>
<td>( Q_0 )</td>
<td>( Q_1 )</td>
<td>( s_0 &gt; 0 )</td>
<td>rejection if ( Q_1 ) is full</td>
</tr>
<tr>
<td>( \lambda_2 )</td>
<td>( Q_0 )</td>
<td>( Q_2 )</td>
<td>( s_0 &gt; 0 )</td>
<td>rejection if ( Q_2 ) is full</td>
</tr>
<tr>
<td>( \lambda_3 )</td>
<td>( Q_1 )</td>
<td>( Q_3 )</td>
<td>( s_1 &gt; 0 )</td>
<td>rejection if ( Q_3 ) is full</td>
</tr>
<tr>
<td>( \lambda_4 )</td>
<td>( Q_2 )</td>
<td>( Q_3 )</td>
<td>( s_2 &gt; 0 )</td>
<td>rejection if ( Q_3 ) is full</td>
</tr>
<tr>
<td>( \lambda_5 )</td>
<td>( Q_3 )</td>
<td>( Q_{-1} )</td>
<td>( s_3 &gt; 0 )</td>
<td>none</td>
</tr>
</tbody>
</table>

Fig. 3.1. Network with rejection
Since the network is open, clients are able to enter and leave the network. We assume that customers who enter from outside the network to a given queue arrive according to a Poisson process. Furthermore, suppose that the service times at server $i$ are independent and exponentially distributed with parameter $\mu_i$.

**Definition 3.1.** An event $e$ is an application defined on $S$ that associates to each state $x \in S$ a new state denoted by $\phi(x, e)$. The function $\phi$ is the transition function of the network.

For example, to event $e_1$ (fig 3.1) we get

$$\phi(\cdot, e_1) : (s_0, s_1, s_2, s_3) \mapsto \begin{cases} (s_0 - 1, s_1 + 1, s_2, s_3) & \text{if } (s_0 \geq 1) \text{ and } (s_1 < C_1); \\ (s_0 - 1, s_1 + 1, s_2, s_3) & \text{if } (s_0 \geq 1) \text{ and } (s_1 = C_1)(Q_1 \text{ full}); \\ (s_0, s_1, s_2, s_3) & \text{if } (s_0 \geq 0)(Q_0 \text{ empty}). \end{cases}$$

**Definition 3.2.** An event $e$ is monotone if $\phi(x, e) \leq \phi(y, e)$ for every $x, y \in S$ with $x \leq y$.

It is clear that the previous event $e_1$ is monotone. Moreover usual events such as routing with overflow and rejection, routing with blocking and restart, routing with a index policy rule (eg Join the shortest queue) are monotone events [2, 10].

Denote by $E = \{e_1, \ldots, e_M\}$ the finite collection of events of the network. With each event $e_i$ is associated a Poisson process with parameter $\lambda_i$. If an event occurs which does not satisfy the enabling condition the state of the system is unchanged.

To achieve the construction of the discrete-time Markov chain, the system is uniformized by a Poisson process with rate $\Lambda = \sum_{i=1}^{M} \lambda_i$. Hence, we can regard this Poisson process as a clock which determines when an event transition takes place. To determine which specific transition, we use the collection $E$ of events of the network.

Let

$$p_i = \mathbb{P} \text{ (event } e_i \text{ occurs)} = \frac{\lambda_i}{\Lambda}.$$

By construction, the following theorem is clear.

**Theorem 3.3.** The uniformized Markov chain has the same stationary distribution as the queueing network, and so the embedded discrete time Markov chain.

Provided that events are monotone, the CFTP algorithm applies on queueing networks to build steady-state sampling of the network.

In our example of Figure 3.1 we ran the CFTP algorithm and produced samples of coupling time. The parameters used for the simulation are the following:

- Queues capacity : 10
- $\lambda_1 = 1.4$, $\lambda_2 = 0.6$, $\lambda_3 = 0.8$, $\lambda_4 = 0.5$ and $\lambda_5 = 0.4$
- $\lambda_0$ the global input rate is varying.
- Number of samples used to estimate the mean coupling time: 10000

The result is displayed in Figure 3.2.

This type of curve is of fundamental importance because the coupling time corresponds to the simulation duration and is involved in the simulation strategy (long run versus replication). These first results can be surprising because they exhibit a strong dependence on parameters values. The aim of this paper is now to understand more deeply what are the critical values for the network and to build bounds on the coupling time that are non-trivial.
Fig. 3.2. Mean coupling time for the acyclic network of Figure 3.1 when the input rate varies from 0 to 4, with 95% confidence intervals.

Definition 3.4. Let \( N_i \) be the function from \( S \) to \( S_i \) with
\[
N_i(s_1, \ldots, s_K) = s_i.
\]
So \( N_i \) returns the number of customers in queue \( Q_i \).

As in section 2, \( \tau^b \) refers to the backward coupling time of the chain, which is in case the coupling time from the past of the queueing network.

Definition 3.5. Let \( \tau^b_i \) denote the backward coupling time on coordinate \( i \) of the state space. Thus \( \tau^b_i \) is the smallest \( n \) for which
\[
\| \{ N_i(\phi^n(s, e_{-n+1}, \ldots, e_0)) \mid s \in S \} \| = 1.
\]
Because coordinate \( s_i \) refers to queue \( Q_i \), the random variable \( \tau^b_i \) represents the coupling time from the past of queue \( Q_i \). Once all queues in the network have coupled, the CFTP algorithm returns one value and hence the chain has coupled. Thus
\[
\tau^b = \max_{1 \leq i \leq K} \{ \tau^b_i \} \leq \sum_{i=1}^{K} \tau^b_i.
\]
By taking expectation and interchanging sum and expectation we get:
\[
E[\tau^b] = E\left[ \max_{1 \leq i \leq K} \{ \tau^b_i \} \right] \leq E\left[ \sum_{i=1}^{K} \tau^b_i \right] = \sum_{i=1}^{K} E[\tau^b_i].
\]
It follows from Proposition 1 that \( \tau^b \) and \( \tau^f \) have the same distribution. The same holds for \( \tau^b_i \) and \( \tau^f_i \). Hence \( E[\tau^b_i] = E[\tau^f_i] \) and
\[
E\tau^b \leq \sum_{i=1}^{K} E[\tau^f_i].
\]
The bound given in Equation 3.3 is interesting because $E[\tau_{f}^i]$ is sometimes amenable to explicit computations, as shown in following sections. In order to derive those bounds, one may provide yet other bounds, by making the coupling state explicit.

**Definition 3.6.** The first hitting time $h_{j\rightarrow k}$ of a Markov chain $X_n$ is defined as

$$h_{j \rightarrow k} = \inf \{ n \in \mathbb{N} : X_n = k | X_0 = j \} \text{ with } j, k \in S.$$ 

The first hitting time $h_{j \rightarrow k}$ with $j, k \in S_i$ is the first hitting of a single queue $Q_i$ of the network. Now $h_{0 \rightarrow C_i}$ represents the number of steps it takes the queue $Q_i$ to go from state 0 to state $C_i$. Now we take queue $Q_i$ out of the network and examine it independently. Suppose that $h_{0 \rightarrow C_i} = n$ for the sequence of events $e_1, \ldots, e_n$. Because of monotonicity of $\phi$ we have

$$\phi^{(n)} (0, e_1, \ldots, e_n) \leq \phi^{(n)} (s, e_1, \ldots, e_n) \leq \phi^{(n)} (C_i, e_1, \ldots, e_n) = 0$$

with $s \in S_i$. Hence, coupling has occurred. So $h_{0 \rightarrow C_i}$ is an upper bound on the forward coupling of queue $Q_i$. The same argumentation holds for $h_{C_i \rightarrow 0}$. Thus

$$E[\tau_{f}^i] \leq E[\min \{ h_{0 \rightarrow C_i}, h_{C_i \rightarrow 0} \}] \quad (3.4)$$

by Jensen inequality.

**4. Coupling time in a M/M/1/C queue.** In a M/M/1/C model, we have a single queue with one server. Customers arrive at the queue according to a Poisson process with rate $\lambda$ and the service time is distributed according to an exponential distribution with parameter $\mu$. In the queue there is only place for $C$ customers. So the state space $S = \{0, \ldots, C\}$. If a customer arrives when there are already $C$ customers in the queue, he immediately leaves without entering the queue. After uniformization, we get a discrete time Markov chain which is governed by the events $e_a$ with probability $p = \frac{\lambda}{\lambda + \mu}$ and $e_d$ with probability $q = 1 - p$. Event $e_a$ represents an arrival and event $e_d$ represents an end of service with departure of the customer.

In order to estimate the expectation of the coupling time from the past $E[\tau_{f}^b]$ we use inequality 3.5. Since there is only one queue, the first two inequalities in 3.5 become equalities. Indeed, when applying forward simulation, the chain only can couple in state 0 or state $C$. This follows since for $r, s \in S$ with $0 < r < s < C$ we have

$$\phi (r, e_a) = r + 1 < s + 1 = \phi (s, e_a),$$

$$\phi (r, e_d) = r - 1 < s - 1 = \phi (s, e_d).$$

So the chain cannot couple in a state $s$ with $0 < s < C$. Furthermore we have

$$\phi(C, e_a) = C = \phi(C - 1, e_a),$$

$$\phi (0, e_d) = 0 = \phi (1, e_d).$$
Hence, forward coupling can only occur in 0 or $C$.

\[ \mathbb{E} [\tau^b] = \mathbb{E} [\min\{h_{0\rightarrow C}, h_{C\rightarrow 0}\}] . \]  

\[ (4.1) \]

4.1. Explicit calculation of $\mathbb{E} [\tau^b]$. In order to compute $\min\{h_{0\rightarrow C}, h_{C\rightarrow 0}\}$ we have to run two copies of the Markov chain for a $M/M/1/C$ queue simultaneously. One copy starts in state 0 and the other one starts in state $C$. We stop when either the chain starting in 0 reaches state $C$ or when the copy starting in state $C$ reaches state 0.

\[ \psi ((x, y), e_a) = ((x + 1) \land C, (y + 1) \land C) \]

\[ \psi ((x, y), e_d) = ((x - 1) \lor 0, (y - 1) \lor 0) . \]

Without any loss of generality, we may assume that $i \leq j$. This system corresponds with the Markov chain $X(q)$ displayed in Figure 4.1.

Since we can only couple in 0 or $C$, this coupling occurs as soon as the chain $X(q)$ reaches states $(0, 0)$ or $(C, C)$. Define

\[ H_{i,j} := \text{number of steps to reach state } (0, 0) \text{ or } (C, C) \text{ starting from state } (i, j) \]

with $(i, j) \in S \times S$. By definition, $\min\{h_{0\rightarrow C}, h_{C\rightarrow 0}\} = H_{0,C}$. Now $H_{i,j}$ represents the hitting time of the coupling states $(0, 0)$ and $(C, C)$ (also called absorption time) in a product Markov chain. Using a one step analysis, we get the following system of equations for $\mathbb{E}[H_{i,j}]$:

\[ \begin{cases} 
\mathbb{E}[H_{i,j}] = 1 + p\mathbb{E}[H_{(i+1)\land C,(j+1)\land C}] + q\mathbb{E}[H_{(i-1)\lor 0,(j-1)\lor 0}], & i, j \in S, i \neq j, \\
\mathbb{E}[H_{i,i}] = 0, & i = j
\end{cases} \]

\[ (4.2) \]
Two states \((i, j)\) and \((i', j')\) are said to be at the same level if \(|j - i| = |j' - i'|\). In Figure 4.1 we can distinguish \(C + 1\) levels. Because of monotonicity of \(\phi\), \(|j - i|\) cannot increase. Hence, starting at a level with \(|j - i|\), the chain will gradually pass all intermediate levels to reach finally the level with \(|j - i| = 0\) in state \((0, 0)\) or \((C, C)\). Thus, starting in state \((0, C)\), the chain will run through all levels to end up at the level with \(|j - i| = 0\) in state \((0, 0)\) or \((C, C)\). So, \(H_{0,C} = \min\{h_{0\to C}, h_{C\to 0}\}\). To determine \(\mathbb{E}[H_{0,C}]\) we determine the mean time spent on each level and sum up over all levels.

A state \((i, j)\) belongs to level \(m\) if \(|j - i| = C + 2 - m\). Then state \((0, C)\) belongs to level 2 and the states \((0, 0)\) and \((C, C)\) belong to level \(C + 2\). To get from \((0, C)\) into either \((0, 0)\) or \((C, C)\), the chain \(X(q)\) needs to cross all levels between the levels 2 and \(C + 2\). Let \(T_m\) denote time it takes to reach level \(m + 1\), starting in level \(m\). Then

\[
H_{0,C} = \sum_{m=2}^{C+1} T_m. \tag{4.3}
\]

In order to determine \(T_m\), we associate to each level \(m\) a random walk \(R_m\) on \(0, \ldots, m\) with absorbing barriers at state 0 and state \(C\). In the random walk, the probability of going up is \(p\) and of going down is \(q = 1 - p\). We have the following correspondence between the states of the random walk \(R_m\) and the states of \(X(q)\) (see Figure 4.2):

State 0 of \(R_m\) corresponds with state \((0, C - m + 1)\) of \(X(q)\),
State \(i\) of \(R_m\) corresponds with state \((i - 1, C - m + 1 + i)\) of the \(X(q)\),
\[1 \leq i \leq m - 1,\]
State \(m\) of \(R_m\) corresponds with state \((m - 1, C)\) of \(X(q)\),

Now the time spent on level \(m\) in \(X(q)\) is the same as the time spent in a random walk \(R_m\) before absorption. Therefore, we will use the following two results on random walks in determining \(T_m\), which are well known (see for example [8]).

Let \(\alpha_{m-0}^m\) denote the probability of absorption in state 0 of the random walk \(R_m\) starting in \(i\). Then:

\[
\alpha_{m-0}^m = \begin{cases} 
\frac{a^m - ai}{a^m - 1}, & p \neq \frac{1}{2}, \\
\frac{m-i}{m}, & p = \frac{1}{2},
\end{cases} \tag{4.4}
\]

where \(a = q/p\).
Now, absorption occurs in \( R_m \) once the state 0 or \( C \) has been achieved.

**Lemma 4.1.** Let \( \tilde{T}^m_i \) denote the mean absorption time of a random walk \( R_m \) starting in \( i \). Then:

\[
\mathbb{E}[\tilde{T}^m_i] = \begin{cases} 
\frac{i-m(1-a^m_{i-0})}{q-p}, & p \neq \frac{1}{2}, \\
 i(m-i), & p = \frac{1}{2}.
\end{cases}
\] (4.5)

Now, let \( \beta^m_i \) denote the probability that absorption occurs in \( i = 0, m, \ldots \). Then

\[
\beta^m_0 = \sum_{i=0}^{m} \alpha^m_{i-0} \mathbb{P}(R_m \text{ starts in state } i),
\] (4.6)

and \( \beta^m_m = 1 - \beta^m_0 \). From the structure of the Markov chain \( X(q) \) and the correspondence between \( X(q) \) and the random walks, we have that (see Figure 4.2):

\[
\mathbb{P}(\text{enter level } m + 1 \text{ at } (0, C - m + 1)) = \mathbb{P}(\text{absorption in } 0 \text{ in } R_m) = \beta^m_0.
\]

Now one has:

\[
\mathbb{E}[T_m] = \mathbb{E}[\tilde{T}^m_1] \beta^m_0 - 1 + \mathbb{E}[\tilde{T}^m_{m-1}] \beta^m_{m-1}
= \mathbb{E}[\tilde{T}^m_{m-1}] + \left( \mathbb{E}[\tilde{T}^m_1] - \mathbb{E}[\tilde{T}^m_{m-1}] \right) \beta^m_{m-1}
\] (4.7),

\[
\mathbb{E}[T_m] = \mathbb{E}[\tilde{T}^m_1] \beta^m_0 - 1 + \mathbb{E}[\tilde{T}^m_{m-1}] \beta^m_{m-1} = m - 1.
\] (4.8)

**4.1.1. Case \( q = p = 1/2 \).** \( \mathbb{E}[T_m] \) can be calculated explicitly for \( p = \frac{1}{2} \). Since the random walk is symmetric, we have \( \beta^m_0 = \beta^m_m = \frac{1}{2} \). So:

\[
\mathbb{E}[T_m] = \mathbb{E}[\tilde{T}^m_1] \beta^m_0 - 1 + \mathbb{E}[\tilde{T}^m_{m-1}] \beta^m_{m-1} = m - 1.
\] (4.9)

Hence,

\[
\mathbb{E}[H_{0,C}] = \sum_{m=2}^{C+1} \mathbb{E}[T_m] = \sum_{m=2}^{C+1} m - 1 = \frac{C^2 + C}{2}.
\]

**Lemma 4.2.** For a M/M/1/C with \( \lambda = \mu \), \( \mathbb{E}T^b = \frac{C^2 + C}{2} \).

**4.1.2. Case \( p \neq 1/2 \).** Since the random walks are not symmetric, we cannot apply the same reasoning as for the case \( q = 1/2 \) to determine \( \beta^m_0 \). Therefore, we need to calculate \( \beta^m_0 \). Entering the random walk \( R_m \) corresponds to entering level \( m \) in \( X(q) \). Since we can only enter level \( m \) in the state \( (0, C - m + 2) \) and \( (m - 2, C) \) this means we can only start the random walk in state 1 or \( m-1 \). Therefore 4.6 becomes:

\[
\beta^m_0 = \sum_{i=0}^{m} \alpha^m_{i-0} \mathbb{P}(R_m \text{ starts in state } i)
= \alpha^m_{1-0} \mathbb{P}(R_m \text{ starts in } 1) + \alpha^m_{m-2} \mathbb{P}(R_m \text{ starts in } m - 1)
= \alpha^m_{m-1-0} + (\alpha^m_{1-0} - \alpha^m_{m-1-0}) \beta^m_{m-1}
= \frac{\alpha^m - \alpha^m_{m-1}}{\alpha^m - 1} \beta^m_{m-1}.
\]
So this gives the recurrence:

\[
\begin{align*}
\beta_0^m &= \frac{a^m - a^{m-1}}{a^{m-1}} \beta_{0}^{m-1} + \frac{a^{m-1}}{a^{m-1}} \beta_0^{m-1}, \quad m \geq 2; \\
\beta_0^2 &= 2.
\end{align*}
\tag{4.10}
\]

Thus we obtain,

**Proposition 4.3.** For a M/M/1/C queue, using the foregoing notations,

\[
\mathbb{E}r^b = \mathbb{E}[H_{0,c}] = \sum_{m=2}^{C+1} \mathbb{E}[\tilde{T}_m] + \left( \mathbb{E}[\tilde{T}_1] - \mathbb{E}[\tilde{T}_{m-1}] \right) \beta_0^{m-1},
\]

with \( \beta_0^m \) defined by (4.10) and \( \mathbb{E}[\tilde{T}_{m-1}] \) and \( \mathbb{E}[\tilde{T}_1] \) defined by (4.5).

**4.1.3. Comparison between the cases** \( q = 1/2 \) **and** \( q \neq 1/2 \).

**Proposition 4.4.** The coupling time in a M/M/1/C queue is maximal when the input rate \( \lambda \) **and** the service rate \( \mu \) are equal.

**Proof.** By definition, \( \lambda = \mu \) corresponds to \( p = q = 1/2 \). The proof holds by induction on \( C \). The result is obviously true when \( C = 0 \), because whatever \( q \), \( \mathbb{E}[H_{0,c}] = 0 \).

For \( C + 1 \), let \( q \) be an arbitrary probability with \( q > 1/2 \) (the case \( q < 1/2 \) is symmetric). We will compare the expected time for absorption of three Markov chains. The first one is the Markov chain \( X := X(1/2) \) displayed in Figure 4.1, with \( q = p = 1/2 \). The second one is the Markov chain \( X' := X(q) \) displayed in Figure 4.1 and the last one \( X'' \) is a mixture between the two previous chains: The first \( C \) levels are the same as in \( X \) while the last level \( (C + 1) \) is the same as in \( X' \).

The expected absorption time for the first \( C \) levels is the same for \( X \) and for \( X'' \): \( \sum_{m=2}^{C} \mathbb{E}T_m = \sum_{m=2}^{C} \mathbb{E}T_m' \). By induction, this is larger than for \( X' \): we have \( \sum_{m=2}^{C} \mathbb{E}T_m = \sum_{m=2}^{C} \mathbb{E}T_m' \geq \sum_{m=2}^{C} \mathbb{E}T_m' \). Therefore, we just need to compare the exit times out of the last level, namely \( \mathbb{E}T_{C+1} \), \( E_{C+1} \) and \( \mathbb{E}T_{C+1}'' \) to finish the proof.

We first compare \( \mathbb{E}T_{C+1} \) and \( \mathbb{E}T_{C+1}'' \). In both cases, the Markov chain enters level \( C + 1 \) in state \( (0, 1) \) with probability 1/2.

Equation (4.9) says that \( \mathbb{E}T_{C+1} = C \) and Equation (4.5) gives after straightforward computations, \( \mathbb{E}T_{C+1}'' = 1/2q^{C(1-aC-1)} + 1/2q^{C(1-aC-1) - 1} = 2q^{aC-1} \leq C/(2q) = C = \mathbb{E}T_{C+1}'' \).

In order to compare \( \mathbb{E}T_{C+1}'' \) and \( \mathbb{E}T_{C+1}'' \), let us first show that \( \beta_0^m \) is larger than 1/2, for all \( m \geq 2 \). This is done by an immediate induction on Equation (4.10). If \( \beta_0^{m-1} > 1/2 \), then \( \beta_0^m > 2q^{aC-1} \). Now, \( 2q^{aC-1} \) is 1/2 if \( 2a^m - a^{m-1} - a > a^m - 1 \), i.e. after recombining the terms, \( (a - 1)(a^m - 1) \geq 0 \). This is true as soon as \( q \geq 1/2 \).

To end the proof, it is enough to notice that for the chain \( X' \), time to absorption starting in 1, \( \mathbb{E}T_{m-1}' \) is smaller that time to absorption starting in \( m - 1 \), \( \mathbb{E}T_{m-1}' \) for all \( m \). The difference \( \mathbb{E}T_{m-1}' - \mathbb{E}T_{1}' \) is

\[
\frac{ma^m - ma^{m-1} + ma - m - 2a^m + 2}{p(a^m - 1)(a - 1)} = \frac{m(a - 1)\left(\frac{a^m - 1}{2} - \frac{1 + a + \ldots + a^{m-1}}{m}\right)}{p(a^m - 1)(a - 1)} \geq 0,
\]

by convexity of \( x \mapsto ax^2 \).
Finally, \( ET'_{C+1} = \beta C^{C+1} E \tilde{T}_1^{C+1'} + (1 - \beta C^{C+1}) E \tilde{T}_C^{C+1'} \)
\[ < \frac{1}{2} E \tilde{T}_1^{C+1'} + \frac{1}{2} E \tilde{T}_C^{C+1'} \]
\[ = ET''_{C+1}. \]

\[ \square \]

4.2. Explicit Bounds. Equation (4.11) provides a quick way to compute \( E[H_{0,C}] \)
using recurrence equation (4.10). However, it may also be interesting to get a sim-
ple closed form for an upper bound for \( E[H_{0,C}] \). This can be done using the last
inequality in Equation 3.5 that gives an upper bound for \( E[H_{0,C}] \) amenable to direct
computations.

\[ E[H_{0,C}] = E[\min\{h_{0-C}, h_{C-0}\}] \leq \min\{E[h_{0-C}], E[h_{C-0}]\}. \tag{4.12} \]

The exact calculation of \( E[h_{0-C}] \) can be done with an approach derived from [8]
by conditioning on the next event. Let \( T_i \) denote the time to go from state \( i \) to \( i + 1 \).
Then

\[ E[h_{0-C}] = \sum_{i=0}^{C-1} E[T_i] \tag{4.13} \]

To get an expression for \( T_i \), with \( 0 < i \leq C \), we condition on the first event. Therefore
let \( E[T_i|e] \) denote the conditional expectation of \( T_i \) knowing that the next event is \( e \).
Since \( E[T_i|e_a] = 1 \) and \( E[T_i|e_d] = 1 + E[T_{i-1}] + E[T_i] \), conditioning delivers the
following recurrent expression for the \( E[T_i] \):

\[ E[T_i] = \begin{cases} \frac{1}{p}, & i = 0, \\ \frac{1}{p} + \frac{2}{p} E[T_{i-1}], & 0 < i < C. \end{cases} \tag{4.14} \]

By mathematical induction we can show that \( E[T_i] = \frac{1}{p} \sum_{k=0}^{i} \left( \frac{q}{p} \right)^k \). Hence, \( E[T_i] = \frac{1-(\frac{q}{p})^{i+1}}{p-q} \). and from 4.13 it follows that

\[ E[h_{0-C}] = \sum_{i=0}^{C-1} \frac{1 - \left( \frac{q}{p} \right)^{i+1}}{p-q} = \frac{C}{p-q} - \frac{q(1 - \left( \frac{q}{p} \right)^C)}{(p-q)^2}. \tag{4.15} \]

By reasons of symmetry, we have

\[ E[h_{C-0}] = \frac{C}{q-p} - \frac{p(1 - \left( \frac{p}{q} \right)^C)}{(q-p)^2} \tag{4.16} \]

The curves of \( E[h_{0-C}] \) and \( E[h_{C-0}] \) intersect in \( C^2 + C \) when \( p = q \). If \( p > q \) then \( E[h_{0-C}] < E[h_{C-0}] \) and because of symmetry, if \( p < q \) then \( E[h_{0-C}] > E[h_{C-0}] \). Since also \( \frac{C^2 + C}{2} \) is an upper bound on the mean coupling time \( E[x] \), as is shown before, we can state:
**Proposition 4.5.** The mean coupling time $\tau^b$ of a queue $M/M/1/C$ queue with arrival rate $\lambda$ and service rate $\mu$ is bounded using $p = \lambda/(\lambda + u)$ and $q = 1 - p$.

**Heavy traffic Bound:**

If $p \geq q$, 

$$\mathbb{E}\tau^b \leq \min \left\{ \frac{C}{p - q} - \frac{q(1 - \left(\frac{q}{p}\right)^C)}{(p - q)^2}, \frac{C^2 + C}{2} \right\}.$$ 

**Light traffic bound:**

If $p < q$, 

$$\mathbb{E}\tau^b \leq \min \left\{ \frac{C}{q - p} - \frac{p(1 - \left(\frac{q}{p}\right)^C)}{(q - p)^2}, \frac{C^2 + C}{2} \right\}.$$ 

![Figure 4.3](image_url)  

**Figure 4.3.** Expected coupling time in an $M/M/1/10$ queue when $q$ varies from 0 to 1 and the three explicit bounds given in Proposition 4.5.

Figure 4.3 displays both the exact expected coupling time for a queue with capacity 10 as given by Equation (4.11) as well as the three explicit bounds given in Proposition 4.5. Note that the bounds are very accurate under light or heavy traffic ($q \leq 0.4$ and $q > 0.6$). In any case, the ratio is never larger than 1.2.

**5. Coupling in acyclic queueing networks.** This section is dedicated to the effective computation of a bound of the coupling time in acyclic networks.

If one gives a close look to the coupling time for the acyclic network given in Figure 3.1, one may see in Figure 3.2 that the coupling time has a peak when $\lambda = 0.4$. This corresponds to the case when the input rate and service rate in Queue 4 are equal. This should not be surprising regarding the result for a single queue, which says that coupling maximal when the rates are equal. Then a second peak occurs around $\lambda = 1.4$ when coupling in Queue 1 is maximal. The rest of the curve shows a linear increase of the coupling time which may suggest an asymptotic linear dependence in $\lambda$.

In this part, an explicit bound on the coupling time which exhibits these two features will be derived.

The first result concerns an extension of inequality 3.5 to distributions. The second part shows how the results for a single $M/M/1/C$ queue can be used to get an effective computation of bounds for acyclic networks on queues.
5.1. The distribution of the coupling time in acyclic networks. In the following, the queues $Q_1, \ldots, Q_K$ are numbered according to the topological order of the network. Thus, no event occurring in queue $Q_i$ has any influence on the state of queue $Q_j$ as long as $i > j$.

**Proposition 5.1.** The coupling time for an acyclic network is bounded in the stochastic sense by the sum of the forward coupling time of all queues:

$$
\tau^b \leq_{st} \tau^f_K + \cdots + \tau^f_1.
$$

**Proof.** The proof is based on the following idea: construct a trajectory of a backward simulation over which the comparison holds. This will imply the stochastic comparison using Strassen Theorem.

Consider a backward simulation of the network starting at time 0 until coupling occurs for the last queue, at time $-\tau^b_K$. From time $-\tau^b_K$, run a backward simulation until Queue $Q_{K-1}$ couples. From time $-\tau^b_K - \tau^b_{K-1}$, run the backward simulation again until Queue $Q_{K-2}$ couples. Continue this construction until the first queue has coupled at time $-\tau^b_K + \cdots + \tau^b_1$. Now, on this trajectory, the state in queue 1 has coupled between times $-(\tau^b_K + \cdots + \tau^b_1)$ and $-(\tau^b_K + \cdots + \tau^b_2)$. From this time on, queue 1 will remain coupled since no event in other queues may alter its state. The same property holds for queue $Q_i$, between times $-(\tau^b_K + \cdots + \tau^b_i)$ and $-(\tau^b_K + \cdots + \tau^b_{i+1})$, and at time 0, all queues have coupled. Finally, note that the intervals of this simulation are independent of each other so that $\tau^b = \tau^f$ in distribution and one gets $\tau^b \leq_{st} \tau^f_K + \cdots + \tau^f_1$. $\blacksquare$

This result calls for several comments. First, note that acyclicity is essential in the proof above. For networks with cycles, one would need some kind of association properties of the states of the queues to assess something about the comparison of the distribution of $\tau^b$ and the $\tau^f_i$’s.

Second, the technique used in Section 4 to get explicit bounds for the expectation of the forward coupling time can also be used to derive an explicit form for the generating function of bounds of the distribution of the forward coupling time for each queue. Using the theorem above, this gives a generating function of a bound of the coupling time for an acyclic network of queues, up to a convolution formula.

5.2. Computation of an upper bound on the coupling time. Here, an acyclic network of /M/1/C queues with an arbitrary topology and Bernoulli routings is considered. The events here are of only two types: exogenous arrivals (Poisson with rate $\gamma_i$ in queue $i$) and routing of one customer from queue $i$ to queue $j$ after service completion in queue $i$ (with rate $\mu_{ij}$). Queue $K + 1$ is a dummy queue representing exits: routing a customer to queue $K + 1$ means that the customer exits the network forever. In case of overflow, the new customer trying to enter the full queue is lost. The service rate at queue $i$ is also denoted $\mu_i = \sum_{i=1}^{K+1} \mu_{ij}$.

In the following, we will show how one can find upper bounds for each $\tau^f_i$, $i = 1, \ldots, K$.

Now consider the first queue of the network. By acyclicity, the behavior of this queue is independent of the rest of the network. For $Q_1$ as well as any input queue (the incoming traffic is exogenous only), the expectation of $\tau^f_i$ can be computed as in Section 4 with the following adaptation. Here, one needs to use $p = \gamma_1/(\gamma_1 + \mu_1)$ and $q = 1 - p$. In the one step analysis, all events not involving queue 1 occur with probability $d = (\Lambda - \gamma_1 - \mu_1)/\Lambda$. One gets
Bounds for the Coupling Time

\[ E[H_{i,j}] = 1 + dE[H_{i,j}] + \frac{\lambda_i}{\Lambda} E[H_{(i+1)\land C,(j+1)\land C}] + \frac{\mu_i}{\Lambda} E[H_{(i-1)\lor 0,(j-1)\lor 0}]
\]
\[ = \frac{1}{1-d} + pE[H_{(i+1)\land C,(j+1)\land C}] + qE[H_{(i-1)\lor 0,(j-1)\lor 0}]. \]

We get the same equation as 4.2 except for the additional constant which is now \( \frac{1}{1-d} \) instead of 1, so that the solution is the same as before up to a multiplicative factor of \( \frac{1}{1-q} = \frac{\Lambda}{\gamma_1 + \mu_1} \). Using Proposition 4.4, one gets

\[ E[\tau^b_1] \leq \frac{\Lambda}{\gamma_1 + \mu_1} \left( C_i^2 + C_i \right). \]

As for any internal queue, say \( Q_i \), we will bound \( \tau^b_i \) using an auxiliary network. In order to do so, we consider a new circuit with one difference from the original one: all queues are replaced by infinite queues, except for queue \( Q_i \) which stays the same. In the following, all the notations related to this new network will be expressed by primes. This new circuit up to Queue \( i \) is product form and using Burke’s Theorem, the input stream in Queue \( i \) is Poisson. The rate of the input stream in queue \( i \) is given by the \( i - th \) coordinate \( \ell_i \) of the solution \( \ell \) to the flow equation: \( \ell = \ell R + \gamma \) where \( R \) is a \( K + 1 \times K + 1 \) tridiagonal matrix: \( R_{i,j} = \mu_{i,j}/\mu_i \). The network is said to be stable for Queue \( i \) as soon as \( i, \ell_i \leq \mu_i \).

When running a backward simulation of the original network, one may also run a backward simulation of the network with infinite capacities queues using the same event realizations with the same initial number of customers. It is not difficult to see that because of the monotonicity properties of all events, the number of customers in all queues of the new network is always larger than in the original one. This means that in the new network, the coupling time to 0 in \( Q_i \), \( h'_{C_i \rightarrow 0} \), is stochastically larger than in this original one. Therefore one gets

\[ \tau^b_i \leq_{st} h_{C_i \rightarrow 0} \leq_{st} h'_{C_i \rightarrow 0}, \quad \forall i \in \{1, \ldots, K\}. \quad (5.1) \]

Now, computing \( E[h'_{C_i \rightarrow 0}] \) can be done as in Section 4, up to the same modifications used for queue \( Q_1 \).

Setting \( p = \ell_i/(\ell_i + \mu_i) \) and \( q = 1 - p \), one gets

\[ E[h'_{C_i \rightarrow 0}] = \frac{\Lambda}{\ell_i + \mu_i} \left( \frac{C_i}{q - p} - \frac{p(1 - \left( \frac{q}{p} \right)^C)}{(q - p)^2} \right). \quad (5.2) \]

Furthermore, when the queue is stable \( (\ell_i \leq \mu_i) \), one gets from the results in Section 4.2,

\[ E[h'_{C_i \rightarrow 0}] \leq \frac{\Lambda}{\ell_i + \mu_i} (C_i^2 + C_i). \]

The result of this part is summarized in the following theorem.

**Theorem 5.2.** In an acyclic network of \( K / M / 1 / C_i \) queues with Bernoulli routing and losses in case of overflow, the expected coupling time from the past satisfies:

\[ E[\tau^b] \leq \sum_{i=1}^{K} \frac{\Lambda}{\ell_i + \mu_i} (C_i + C_i^2), \]
as soon as all the queues are stable.

Note that this bound is actually linear in the rate of any event in the system. This behavior was also noticeable for $E[\tau^b]$ itself. Also note that this first approximation is a very simple formula where the rates of the event do not play a crucial role.

### 5.3. Some numerical experiments

In the construction of the bound given in Theorem 5.2, several factors may be responsible for the inaccuracy of the bound.

1. The first factor is the replacement of the max by the sum. We believe that it may be a hard task to get rid of this first approximation because of the intricate dependencies between the queues. Furthermore, experiments reported below show that this may not even be possible in many cases (see Figure 5.2).

2. Another factor which may increase the inaccuracy of our bounds is the fact that most events change the states of several queues at the same time, while the bound given here disregards this. In the network studied here, this may add a factor 2 between the true coupling time and the bound given in Theorem 5.2.

3. The most important factor which jeopardized the quality of the bound is the stability issue. If one queues is unstable, the bound provided by Equation (5.2), also called the light traffic bound in Proposition 4.5 is very bad (as seen in Figure 4.3). So far we have not been able to come up with a better bound for unstable queues. However, when all queues are stable (and even more so when the load is smaller than $2/3$), the bound tends to be very accurate. This is further verified in the experiments reported below.

Computations for the network displayed in Figure 3.1 are reported in Figures 5.1, 5.2, 5.3. We have used the following parameters.

The input rate is $\lambda_0 = 0.4$, the rates of the other events are $\lambda_1 = 1.4$, $\lambda_2 = 0.6$, $\lambda_3 = 0.8$, $\lambda_4 = 0.5$; Number of simulation runs : 10000; The service rate in the last queue $\lambda_5$ takes three values.

In the first case (Fig. 5.1), $\lambda_5 = 0.2$ so that queue $Q_4$ is unstable.
In the second case (Fig. 5.2), $\lambda_5 = 0.6$, and all queues are stable with a load smaller that $2/3$.
The last case (Fig. 5.3) is $\lambda = 0.4$, so that $Q_4$ is barely unstable. This would correspond to the maximal coupling time for $Q_4$ if it were alone.
The capacity $C$ is the same in all queues, and we let is vary from 1 to 20.

Figure 5.1 displays the bound given by formula 5.2 as well as the mean coupling time computed over 10000 simulation runs. As hinted before, the bound is indeed very bad for the unstable system. A ratio larger than 10 w.r.t the true coupling time is reached when $C = 5$. It should also be noticed that the bound is convex in $C$ while the coupling time is not.

Figure 5.2 shows the bound provided by (5.2) and the true coupling time computed by simulation runs. Both curves appear to be almost linear in $C$ (this is true for the bound: when $q/p$ is small, $Eh'_{G_i \rightarrow 0}$ is almost linear in $C_i$) and the ratio is smaller than 1.3. In that case, the curve $\max_{i \in \{1, \ldots, K\}} Eh_{G_i \rightarrow 0}$ is also displayed and is below the actual coupling time. This is to be related with the first item in the comments above.
Fig. 5.1. Coupling time (dots) and the bound given by Equation (5.2) when Queue $Q_4$ is unstable ($\lambda_5 = 2/10$) when the capacity $C$ varies from 1 to 20.

Fig. 5.2. Bound given by Equation (5.2), mean coupling time (dots) and maximal bound over all queues, when Queue $Q_4$ is stable ($\lambda_5 = 6/10$) when the capacity $C$ varies from 1 to 20.

Fig. 5.3. Coupling time (dots) and the bound given by Equation (5.2) when Queue $Q_4$ is barely unstable ($\lambda_5 = 4/10$) when the capacity $C$ varies from 1 to 20.
Figure 5.3 displays the backward coupling time and the bound provided by Equation (5.2). For queue $Q_4$, we use a bound in $C_4 + C_2^4$ which is a bad approximation because of the loss of the factor 2 when compared with the bound for isolated queues. Note that the total gap has a ratio which is almost 2. In that case the coupling time and the bound all exhibit a convex behavior w.r.t. $C$.

A ratio smaller than 2 is indeed interesting because efficient perfect simulation algorithm use a doubling window technique to reduce the complexity and their running time (see Equation (2.3)) so that our bound gives a good estimation of the mean running time of the algorithms.

5.3.1. Extension to more general networks. Actually, extensive simulation runs over many examples show that the bound given in Theorem 5.2 is robust and also holds for more general networks with blocking and with circuits, so that we conjecture

$$\mathbb{E}[\tau^b] \leq \sum_{i=1}^{K} \frac{\Lambda}{\ell_i + \mu_i} O(C_i^2),$$

for any monotone Markovian network of queues with a finite state space. Furthermore under light or heavy traffic in all queues, the bound should rather be linear:

$$\mathbb{E}[\tau^b] \leq \sum_{i=1}^{K} \frac{\Lambda}{\ell_i + \mu_i} O(C_i),$$

Acknowledgments. The authors would like to thank Jérôme Vienne who partially designed the psi2 perfect simulation software which was used to run all the simulations presented here.

REFERENCES