EURANDOM PREPRINT SERIES

2018-007

March 21, 2018

Transition time asymptotics of queue-based activation protocols in random-access networks

S. Borst, F. den Hollander, F. Nardi, M. Sfragara

ISSN 1389-2355

Transition time asymptotics of queue-based activation protocols in random-access networks

Sem Borst [1] [3]
Frank den Hollander [2]
Francesca R. Nardi [1] [4]
Matteo Sfragara [2]

March 21, 2018

Abstract

We study queue-based activation protocols in random-access networks. The network is modelled as an interference graph. Each node of the graph represents a server with a queue. Packets arrive at the nodes as independent Poisson processes and have independent exponentially distributed sizes. Each node can be either active or inactive. When a node is active, it deactivates at unit rate. When a node is inactive, it activates at a rate that depends on its current queue length, provided none of its neighbouring nodes is active. Thus, two nodes that are connected by a bond cannot be active simultaneously. This situation arises in random-access wireless networks where, due to interference, servers that are close to each other cannot use the same frequency band. In the limit as the queue lengths at the nodes become very large, we compute the transition time between the two states where one half of the network is active and the other half is inactive.

We compare the transition time with that of a network in which the activation rates are not controlled by the queue length but are externally driven, a situation that was dealt with in an earlier paper. Namely, we first sandwich the transition time between that of two networks in which the activation rates are small perturbations of a certain prescribed function of the mean queue length. After that we show that, as the perturbation tends to zero, the two transition times become asymptotically similar. We focus on a *complete bipartite network*: we identify the scale of the transition time in terms of the model parameters and we show that its law on the scale of its mean has a *trichotomy* depending on the aggressiveness of the activation rates. Our aim in future work is to use similar comparison techniques for more general bipartite networks and for more complicated queue-based activation protocols.

Keywords: Random-access networks, comparison of activation protocols, transition time, metastability.
MSC2010: 60K25, 60K30, 90B15, 90B18.
Acknowledgment: The research in this paper was supported through NWO Gravitation Grant 024.002.003–NETWORKS.

[1] Department of Mathematics and Computer Science, Eindhoven University of Technology, The Netherlands
[2] Mathematical Institute, Leiden University, The Netherlands
[3] Nokia Bell Labs, Murray Hill, USA
[4] Department of Mathematics, University of Florence, Italy

Contents

1 Introduction

Section 1.1 provides motivation and background. Section 1.2 formulates the mathematical model. Section 1.3 states the main theorems. Section 1.4 offers a brief discussion of these theorems, as well as an outline of the remainder of the paper.

1.1 Motivation and background

In the present paper we investigate metastability properties and transition time asymptotics of *queue-based random-access protocols in wireless networks*. Specifically, we consider a stylised stochastic model for a wireless network (see Fig. 1 below), represented in terms of an undirected graph $G = (N, B)$, referred to as the *interference graph*. The set of nodes N labels the servers and the set of bonds B indicates which pairs of servers interfere and are therefore prevented from simultaneous activity. We denote by $X(t) \in \mathcal{X}$ the joint activity state at time t, with state space

$$\mathcal{X} = \big\{x \in \{0,1\}^N \colon\ x_i x_j = 0 \ \ \forall\, (i,j) \in B\big\}, \tag{1.1}$$

where $x_i = 0$ means that node i is inactive and $x_i = 1$ that it is active.

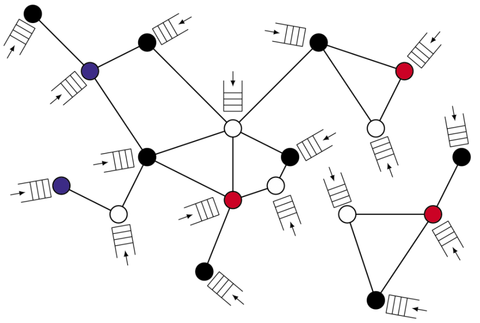

Figure 1: A random-access network. Each node represents a server with a queue. Packets arrive that require a random service time.

We assume that packets arrive at the nodes as independent Poisson processes and have independent exponentially distributed sizes. When a packet arrives at a node, it joins the queue at that node and the queue length undergoes an instantaneous jump equal to the size of the arriving packet. The queue decreases at a constant rate c (as long as it is positive) when the node is active. We denote by $Q(t) \in \mathbb{R}_+^N$ the joint queue size vector at time t, with $Q_i(t)$ representing the queue size at node i at time t. When node i is inactive at time t, it activates at a time-dependent exponential rate $r_i(Q_i(t))$, *provided none of its neighbours is active*, where $q \mapsto r_i(q)$ is some increasing function. Activity durations are exponentially distributed with unit mean, i.e., when a node is active it deactivates at rate 1. Thus, $(X(t), Q(t))_{t\geq 0}$ evolves as a time-inhomogeneous Markov process with state space $\mathcal{X} \times \mathbb{R}_+^N$.

The activity process $(X(t))_{t\geq 0}$ may be viewed as a *hard-core interaction model with state-dependent activation rates*. The state not only depends on the history of the stochastic process of the packet arrivals (which cause upward jumps in the queue sizes), but also on the past evolution of the activity process itself (through the gradual reduction in queue sizes during activity periods). The state-dependent nature of the activation rates raises interesting and challenging issues from a methodological perspective. We will in particular examine *metastability properties* and transition times of the activity process in an asymptotic regime where the initial queue sizes $Q_i(0)$, $i \in N$, grow large in some suitable sense.

The metastable behaviour is not only of significant methodological interest, it is also relevant for analysing the *spatio-temporal dynamics of random-access algorithms in wireless networks*, in particular, so-called queue-based Carrier Sense Multiple-Access (CSMA) policies. In *conventional* CSMA policies, the various nodes activate at fixed rates, which gives rise to classical hard-core interaction models. Metastability properties of the latter models provide fundamental insight into performance characteristics of wireless networks. In particular, for high activation rates, the stationary distribution of the activity process concentrates on states where the maximum number of nodes is simultaneously active, with extremely slow transitions between them, a situation that was studied in [1] and [8]. This ensures high overall efficiency, but from the perspective of an individual node it induces prolonged periods of starvation, possibly interspersed with long sequences of transmissions in rapid succession, resulting in severe build-up of queues and long delays.

In *queue-based* CSMA policies, the *activation rates are chosen to be functions of the queue lengths at the various nodes*, with the aim to provide greater transmission opportunities to nodes with longer queues. Specifically, the activation rate is an increasing function of the queue length of the node itself, and possibly a decreasing function of the queue lengths of its neighbours. Thus, these rates vary over time as queues build up or drain when packets are generated or transmitted. For suitable activation rate functions, queue-based CSMA policies have been shown to achieve maximum stability, i.e., provide stable queues whenever it feasible to do so at all (see [9, 5, 7, 10] and reference therein). Hence these policies have the capability to match the optimal throughput performance of centralised scheduling strategies, while requiring less computation and operating in a distributed fashion. On the downside, the very activation rate functions required for ensuring maximum stability tend to result in long queues and poor delay performance [2, 4]. As alluded to above, metastability properties play a pivotal role, and analysing transition times for the activity process $(X(t))_{t\geq 0}$ is critical in understanding, and possibly improving, the delay performance of queue-based CSMA policies.

In the present paper we focus on *complete bipartite* interference graphs G: the node set can be partitioned into two nonempty sets U and V such that the bond set is the product of U and V, i.e., two nodes interfere if and only if one belongs to U and the other belongs to V. Thus, the collection of all independent sets of G consists of all the subsets of U and all the subsets of V. For convenience, we assume that the activation rate functions are of the form

$$r_i(t) = \begin{cases} g_U(Q_i(t)), & i \in U, \\ g_V(Q_i(t)), & i \in V, \end{cases} \tag{1.2}$$

where $q \mapsto g_U(q)$ and $q \mapsto g_V(q)$ are increasing functions such that $\lim_{q\to\infty} g_U(q) = \infty$, $\lim_{q\to\infty} g_V(q)$ and $g_U(q) = g_V(q) = 0$ when $q < 0$. We denote by $u \in \mathcal{X}$ and $v \in \mathcal{X}$ the joint activity states where all the nodes in either U or V are active, respectively.

We will examine the distribution of the time until state v is reached,

$$\tau_v = \inf\{t \geq 0\colon\ X(t) = v\}, \tag{1.3}$$

when the system starts from state u at time $t = 0$. We consider an asymptotic regime where the initial queue sizes $Q_i(0)$, $i \in U$, grow large in some suitable sense. As it turns out, the metastable behaviour and asymptotic distribution of τ_v are closely related to those in a scenario where the *activation rates are not governed by the random queue sizes*, but are *deterministic* and of the form

$$r_i(t) = \begin{cases} h_U(t), & i \in U, \\ h_V(t), & i \in V, \end{cases} \tag{1.4}$$

for suitable functions $t \mapsto h_U(t)$ and $t \mapsto h_V(t)$.

Specifically, when the initial activity state is u and the initial queue sizes $Q_i(0)$ are large for all $i \in U$, all the nodes in U will initially be active virtually all the time, preventing any of the nodes in V to become active. Consequently, the queue sizes of the nodes in U will tend to decrease at rate $c - \rho_U > 0$, while the queue sizes of the nodes in V will tend to increase at rate $\rho_V > 0$, where ρ_U and ρ_V denote the common traffic intensity of the nodes in U and V, respectively.

While the packet arrivals and activity periods are governed by random processes, the trajectories of the queue sizes will be *roughly linear when viewed on the long time scales of interest*. This suggests that if we assume identical initial queue sizes $Q_i(0) \equiv Q_U(0)$, $i \in U$, and $Q_i(0) \equiv Q_V(0)$, $i \in V$, within the sets U and V, respectively, then the asymptotic distribution of τ_v in (1.3) in the model with queue-dependent activation rates defined in (1.2) should be close to that in the model with deterministic activation rates defined in (1.4) when we choose

$$h_U(t) = g_U\big(Q_U(0) - (c - \rho_U)t\big), \qquad h_V(t) = g_V\big(Q_V(0) + \rho_V t\big). \tag{1.5}$$

The asymptotic distribution of τ_v in the latter scenario was characterised in [1], with the help of the metastability analysis for hard-core interaction models developed in [6].

1.2 Mathematical model

We consider the case where $G = (N, B)$ is a *complete bipartite* graph, i.e., $N = U \cup V$ and B is the set of all bonds that connect a node in U to a node in V (see Fig. 2).

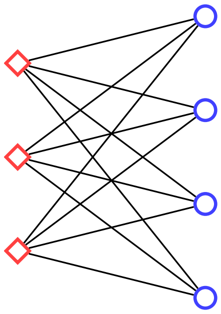

Figure 2: A complete bipartite graph with $|U| = 3$ and $|V| = 4$. At time $t = 0$, square-shaped nodes are active and circle-shaped nodes are inactive.

Definition 1.1 (**State of a node).** A node in the network can be *active* or *inactive*. The *state of node i* at time t is described by a Bernoulli random variable $X_i(t) \in \{0,1\}$, defined as

$$X_i(t) = \begin{cases} 0, & \text{if } i \text{ is inactive at time } t, \\ 1, & \text{if } i \text{ is active at time } t. \end{cases} \tag{1.6}$$

The joint activity state $X(t)$ at time t is an element of the set $\mathcal{X}$ defined in (1.1): the feasible configurations of the network correspond to the collection of independent sets of G. We denote by $u \in \mathcal{X}$ ($v \in \mathcal{X}$) the configuration where all the nodes in U are active (inactive) and all the nodes in V are inactive (active). □

Definition 1.2 (**Pre-transition time and transition time).** The following two times are the main objects of interest in the present paper.

- The *pre-transition time* $\bar{\tau}_v$ is the first time a node in V turns active, i.e.,

$$\bar{\tau}_v = \inf\{t > 0\colon\ X_i(t) = 1\ \exists\, i \in V\}. \tag{1.7}$$

- The *transition time* τ_v is the first time configuration v is hit, i.e.,

$$\tau_v = \min\big\{t \geq 0\colon\ X_i(t) = 0\ \forall\, i \in U,\ X_i(t) = 1\ \forall\, i \in V\big\}. \tag{1.8}$$

□

We are interested in the distribution of $\bar{\tau}_v$ and τ_v given that $X(0) = u$. The pre-transition time plays an important role in our analysis of the transition time, because the evolution of the network is simpler on the interval $[0, \bar{\tau}_v]$ than on the interval $[\bar{\tau}_v, \tau_v]$. However, we will see that $\tau_v - \bar{\tau}_v \ll \bar{\tau}_v$ when the initial queue lengths are large, so that both times have the same asymptotic scaling behaviour.

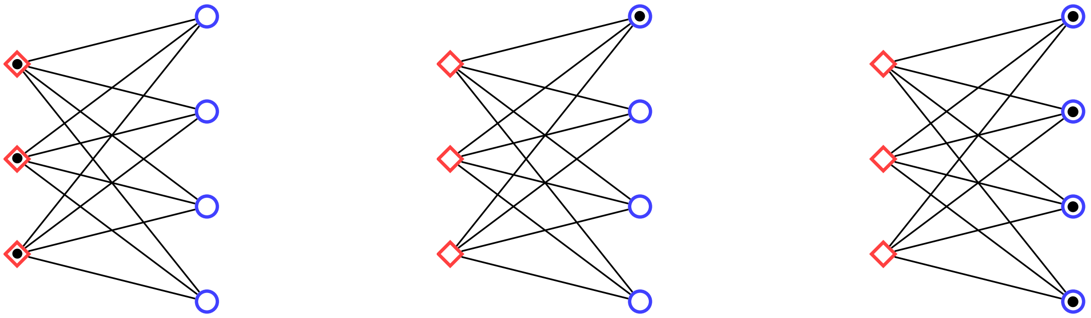

Figure 3: *Left*: initial configuration u. *Center*: pre-transition configuration. *Right*: final configuration v.

An active node i turns inactive according to a deactivation Poisson clock: when the clock ticks, the node switches itself off. Vice versa, an inactive node i attempts to become active at the ticks of an activation Poisson clock: an attempt at time t is successful when no neighbours of i are active at time t^-. Different models can be studied depending on the choice of the activation and deactivation rates of the clocks. Models where these rates are deterministic functions of t are called *external models*. In the present paper we are interested in what are called *internal models*, where the clock rates at node i depend on the queue length at node i at time t.

Definition 1.3 (Queue length at a node). Let $Q_i(t) \in \mathbb{R}_+$ denote the queue length at node i at time t, defined as

$$Q_i(t) = Q_i(0) + Q_i^+(t) - Q_i^-(t) = Q_i(0) + \sum_{j=0}^{N_i(t)} Y_{ij} - cT_i(t), \tag{1.9}$$

where $Q_i(0)$ is the *initial queue length*, the *input process* $t \mapsto Q_i^+(t)$ describes packets arriving according to a Poisson process $t \mapsto N(t) = \text{Poisson}(\lambda t)$ and having i.i.d. exponential service times $Y_j = \text{Exp}(\mu)$, $j \in \mathbb{N}$, and the *output process* $t \mapsto Q_i^-(t)$ represents the cumulative amount of work that is processed in the time interval $[0, t]$ at rate c. □

In order to ensure that the queue length remains non-negative, we let a node switch itself off when its queue length hits zero. The initial queue lengths are assumed to be

$$Q_i(0) = \begin{cases} \gamma_U r, & i \in U, \\ \gamma_V r, & i \in V, \end{cases} \tag{1.10}$$

where $\gamma_U \geq \gamma_V > 0$, and r is a parameter that tends to infinity. Thus, the initial queue lengths are of order r, i.e., $Q_i(0) \asymp r$, and the ones at the nodes in U are larger than the ones at the nodes in V. Note that both the pre-transition and the transition time grow to infinity with r, since the larger the initial queue lengths are, the longer it takes for the transition to occur.

For each node i, the *input process* $t \mapsto Q_i^+(t) = \sum_{j=0}^{N_i(t)} Y_{ij}$ is a compound Poisson process. In the time interval $[0, t]$ packets arrive according to a Poisson process $t \mapsto N_i(t)$ with a rate λ_U or λ_V, depending on whether the node is in U or V. Moreover, each packet j brings the information of its service time: the service time Y_{ij} of the j-th packet at node i is exponentially distributed with parameter μ. Hence the expected value of $Q_i^+(t)$ for a node in U is the product of the expected value $\mathbb{E}[N_i(t)] = \lambda_U t$ and the expected value $\mathbb{E}[Y_j] = 1/\mu$, i.e., $\mathbb{E}[Q_i^+(t)] = \frac{\lambda_U}{\mu} t = \rho_U t$. Analogously, for a node in V we have $\mathbb{E}[Q_i^+(t)] = \rho_V t$. We assume that all the service times are i.i.d. random variables, and are independent of the Poisson process $t \mapsto N_i(t)$.

For each node i, the *output process* is $t \mapsto Q_i^-(t) = cT_i(t) = c\int_0^t X_i(u)du$, where the *activity process* $t \mapsto T_i(t)$ represents the cumulative amount of active time of node i in the time interval $[0, t]$. This is not independent of the input process. Intuitively, the average queue length increases when the node is inactive and decreases when the node is active, which means that packets are being served at a rate c larger than their arrival rate, i.e., $c > \rho_U, \rho_V > 0$. Since all nodes in V are initially inactive, for some time the queue length of these nodes in V is not affected by their output process. However, as soon as a vertex in V turns active, we have to consider its output process as well.

The choice of functions g_U, g_V in (1.2) determines the transition time of the network, since the activation rates of the nodes depend on them. We will assume that g_U, g_V fall in the following class of functions:

$$\mathcal{G} = \Big\{ g \colon \mathbb{R} \to \mathbb{R}_{\geq 0} \colon g \text{ non-decreasing and globally Lipschitz}, g(\mathbb{R}_{\leq 0}) = 0, \lim_{x \to \infty} g(x) = \infty \Big\}. \tag{1.11}$$

Definition 1.4 (Models). Let $g_U, g_V \in \mathcal{G}$ and $\delta > 0$. Assume (1.10). The four models of interest in the present paper are the following:

- In the *internal model* the deactivation Poisson clocks tick at rate 1, while the activation Poisson clocks tick at rate

$$r_i^{\text{int}}(t) = \begin{cases} g_U(Q_i(t)), & i \in U, \\ g_V(Q_i(t)), & i \in V, \end{cases} \qquad t \geq 0. \tag{1.12}$$

- In the *external model* the deactivation Poisson clocks tick at rate 1, while the activation Poisson clocks tick at rate

$$r_i^{\text{ext}}(t) = \begin{cases} g_U(\gamma_U r - (c - \rho_U)t), & i \in U, \\ g_V(\gamma_V r + \rho_V t), & i \in V, \end{cases} \qquad t \geq 0. \tag{1.13}$$

- In the *lower external model* the deactivation Poisson clocks tick at rate 1, while the activation Poisson clocks tick at rate

$$r_i^{\text{low}}(t) = \begin{cases} g_U(\gamma_U r - (c - \rho_U)t - \delta r), & i \in U, \\ g_V(\gamma_V r + \rho_V t + \delta r), & i \in V, \end{cases} \qquad t \geq 0. \tag{1.14}$$

- In the *upper external model* the deactivation Poisson clocks tick at rate 1, while the activation Poisson clocks tick at rate

$$r_i^{\text{upp}}(t) = \begin{cases} g_U(\gamma_U r - (c - \rho_U)t + 2\delta r), & i \in U, \\ g_V(\gamma_V r + \rho_V t - \delta r), & i \in V, \end{cases} \qquad t \geq 0. \tag{1.15}$$

□

Note that in the external models the rates depend on time via certain fixed parameters, while in the internal model the rates depend on time via the actual queue lengths at the nodes. In the lower external model the activation rates in U tend to be *less aggressive* than in the internal model (i.e., the activation clocks tick less frequently), while the activation rates in V tend to be *more aggressive*. In the upper external model the reverse is true: the activation rates in U are more aggressive and the activation rates in V are less aggressive. For simplicity, when considering the external model we sometimes write

$$r_U(t) \text{ and } r_V(t) \tag{1.16}$$

for the activation rate at time t of a node in U and a node in V. We will see that the upper external model is actually defined only for $t \in [0, T_U]$ with $T_U = \frac{\gamma_U}{c - \rho_U} r$ (see Section 2 for details). However, the transition occurs with high probability before time T_U.

1.3 Main theorems

The main goal of the present paper is to compare the transition time of the internal model with that of the two external models. Through a large-deviation analysis of the queue length process at each of the nodes, we define a notion of *good behaviour* that allows us to define perturbed models with externally driven rates that *sandwich* the queue lengths of the internal model and its transition time. We show with the help of *coupling* that with high probability the asymptotic behaviour of the mean transition time for the internal model is the same as for the external model.

The metastable behaviour and the transition time τ_v of a network in which the activation rates are time-dependent in a deterministic way was characterised in [1], and the asymptotic distribution of τ_v was studied in detail. For $s \geq 0$, let

$$\nu(s) = \frac{1}{\mathbb{E}_u[\tau_v](s)} \tag{1.17}$$

be the inverse mean transition time of the time-homogeneous model where the activation rates are equal to

$$r_i^{\text{ext}}(s) = \begin{cases} r_U(s), & i \in U, \\ r_V(s), & i \in V. \end{cases} \tag{1.18}$$

Then, for any time scale $M = M(r)$ and any threshold $x \in [0, \infty)$,

$$\lim_{r\to\infty} \mathbb{P}_u\left(\frac{\tau_v}{M} > x\right) = \begin{cases} 0, & \text{if } M\nu(Mx) \succ 1, \\ e^{-\int_0^x M\nu(Ms)ds}, & \text{if } M\nu(Mx) \asymp 1, \\ 1, & \text{if } M\nu(Mx) \prec 1. \end{cases} \tag{1.19}$$

(Here, $a \succ b$ means $b = o(a)$, $a \prec b$ means $a = o(b)$, while $a \asymp b$ means $a = \Theta(b)$.) If we let M_c be the unique solution of the equation

$$M\nu(M) = 1, \tag{1.20}$$

then the transition occurs on the time scale M_c, in the sense that $\mathbb{P}_u(\tau_v > t) \approx 1$ for $t \prec M_c$ and $\mathbb{P}_u(\tau_v > t) \approx 0$ for $t \succ M_c$. On the critical time scale M_c, the transition time follows an exponential law with time-varying rate. It was proven in [6] that, for a complete bipartite graph and $s \in [0, \infty)$,

$$\mathbb{E}_u[\tau_v](s) = \frac{1}{|U|}\, r_U(s)^{|U|-1}\, [1 + o(1)], \qquad r \to \infty. \tag{1.21}$$

We want the nodes in V to be more aggressive than the nodes in U, so that the transition from u to v can be viewed as the crossover from a "metastable state" to a "stable state". Therefore we assume from now on that

$$\lim_{x\to\infty} \frac{g_V(x)}{g_U(x)} = \infty, \tag{1.22}$$

and we focus on activation rates for nodes in U of the form $g_U(x) \sim Gx^\beta$ when $x \to \infty$ (see Remark 4.1 in Section 4 for more general activation rates g_U). The following two theorems will be proven in Sections 4.1–4.2 with the help of (1.17)–(1.21).

Theorem 1.5 (Critical time scale in the external model). *Suppose that $g_U(x) \sim Gx^\beta$ with $\beta, G \in (0, \infty)$. Then*

$$M_c = F_c r^{1 \vee \beta(|U|-1)}\, [1 + o(1)], \qquad r \to \infty, \tag{1.23}$$

with

$$F_c = \begin{cases} \frac{\gamma_U^{\beta(|U|-1)}}{|U| G^{-(|U|-1)}}, & \textit{if } \beta \in (0, \frac{1}{|U|-1}), \\ \frac{\gamma_U}{|U| G^{-(|U|-1)} + (c - \rho_U)}, & \textit{if } \beta = \frac{1}{|U|-1}, \\ \frac{\gamma_U}{c - \rho_U}, & \textit{if } \beta = (\frac{1}{|U|-1}, \infty). \end{cases} \tag{1.24}$$

Theorem 1.6 (Transition time in the external model). *Suppose that $g_U(x) = Gx^\beta$ with $\beta, G \in (0,\infty)$. Then*

$$\mathbb{E}_u[\tau_v^{\mathrm{ext}}] = F_c r^{1\vee\beta(|U|-1)}\,[1+o(1)], \qquad r\to\infty. \tag{1.25}$$

and

$$\lim_{r\to\infty}\mathbb{P}_u\left(\frac{\tau_v^{\mathrm{ext}}}{\mathbb{E}_u[\tau_v^{\mathrm{ext}}]} > x\right) = \mathcal{P}(x), \qquad x\in[0,\infty), \tag{1.26}$$

with

$$\mathcal{P}(x) = \begin{cases} e^{-x}, & \text{if } \beta\in(0,\frac{1}{|U|-1}),\ x\in[0,\infty),\\ (1-Cx)^{\frac{1-C}{C}}, & \text{if } \beta=\frac{1}{|U|-1},\ x\in[0,\frac{1}{C}),\\ 0, & \text{if } \beta=\frac{1}{|U|-1},\ x\in[\frac{1}{C},\infty),\\ 1, & \text{if } \beta\in(\frac{1}{|U|-1},\infty),\ x\in[0,1),\\ 0, & \text{if } \beta\in(\frac{1}{|U|-1},\infty),\ x\in[1,\infty),\end{cases} \tag{1.27}$$

and $C = \frac{F_c(c-\rho_U)}{\gamma_U} \in (0,1)$.

In other words, the mean transition time scales like M_c, while the distribution of the transition time divided by its mean is exponential, truncated polynomial, respectively, deterministic (see Fig. 4).

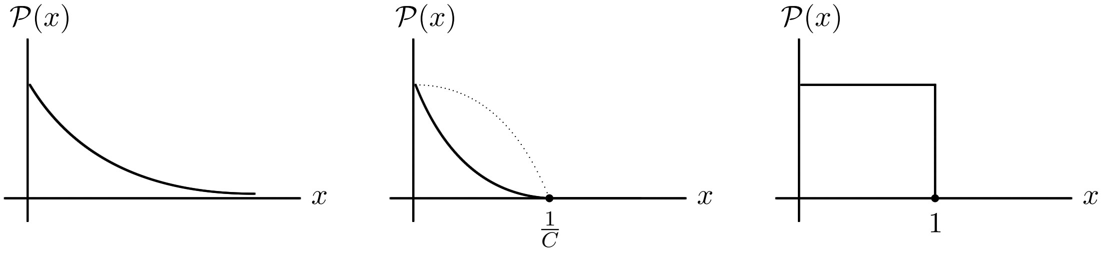

Figure 4: Trichotomy for $x\mapsto\mathcal{P}(x)$: $\beta\in(0,\frac{1}{|U|-1}]$ (left); $\beta=\frac{1}{|U|-1}$ (middle); $\beta\in(\frac{1}{|U|-1},\infty)$ (right). The curve in the middle is convex when $C\in(0,\frac{1}{2})$ and concave when $C\in(\frac{1}{2},1)$. The curve on the right is the limit of the curve in the middle as $C\uparrow 1$.

As shown in Remark 4.1, we can even include the case $\beta=0$, and get that if $g_U(x)=\hat{\mathcal{L}}(x)$ with $\lim_{x\to\infty}\hat{\mathcal{L}}(x)=\infty$, then

$$\mathbb{E}_u[\tau_v^{\mathrm{ext}}] = M_c\,[1+o(1)], \qquad M_c = \frac{1}{|U|}\hat{\mathcal{L}}(\gamma_U r)^{|U|-1}\,[1+o(1)], \qquad r\to\infty, \tag{1.28}$$

and $\mathcal{P}(x)=e^{-x}$, $x\in[0,\infty)$. Similar properties hold for the lower and the upper external model, with perturbed $F_{c,\delta}^{\mathrm{low}}$ and $F_{c,\delta}^{\mathrm{upp}}$ satisfying

$$\lim_{\delta\downarrow 0}F_{c,\delta}^{\mathrm{low}} = \lim_{\delta\downarrow 0}F_{c,\delta}^{\mathrm{upp}} = F_c. \tag{1.29}$$

The main result in the present paper is the following sandwich of τ_v^{int} between τ_v^{low} and τ_v^{upp}, for which we already know the asymptotic behaviour. Because of this sandwich we can deduce the asymptotics of the transition time in the internal model.

Theorem 1.7 (Transition time in the internal model). *For $\delta > 0$ small enough, there exists a coupling such that*

$$\lim_{r\to\infty} \hat{\mathbb{P}}_u\big(\tau_v^{\text{low}} \le \tau_v^{\text{int}} \le \tau_v^{\text{upp}}\big) = 1, \tag{1.30}$$

where $\hat{\mathbb{P}}_u$ is the joint law induced by the coupling, with all three models starting from the configuration u. Consequently, if $g_U(x) \sim Gx^\beta$ with $\beta, G \in (0,\infty)$, then

$$\mathbb{E}_u[\tau_v^{\text{int}}] = F_c r^{1\vee\beta(|U|-1)}\,[1+o(1)], \qquad r\to\infty, \tag{1.31}$$

and

$$\lim_{r\to\infty} \mathbb{P}_u\left(\frac{\tau_v^{\text{int}}}{\mathbb{E}_u[\tau_v^{\text{int}}]} > x\right) = \mathcal{P}(x), \qquad x\in[0,\infty). \tag{1.32}$$

1.4 Discussion and outline

Theorems. Theorem 1.6 gives us the *leading-order asymptotics* of the transition time in the external model, including the lower and the upper external model. Theorem 1.7 is the main result of our paper and provides the *leading-order asymptotics* of the transition time in the internal model, via the coupling in (1.30) and the continuity property in (1.29). Equations (1.24)–(1.25) identify the scaling of the transition time in terms of the model parameters. The trichotomy between $\beta \in (0, \frac{1}{|U|-1})$, $\beta = \frac{1}{|U|-1}$ and $\beta \in (\frac{1}{|U|-1}, \infty)$ is particularly interesting, and leads to different limit laws for the transition time on the scale of its mean.

Interpretation of trichotomy. In order to interpret the above trichotomy, observe first of all that the activation rates of each of the nodes in U remain of order r^β almost all the way up T_U. Specifically, in the absence of the nodes in V, by time yT_U, $y\in[0,1)$, the queue lengths of the nodes in U have decreased by roughly a fraction y, and their activation rates are approximately $G(1-y)^\beta r^\beta$. Hence the fraction of joint inactivity time of the nodes in U is of order $(1/r^\beta)^{|U|} = r^{-\beta|U|}$, and all nodes in U are simultaneously inactive for the first time after a period of order $r^{-\beta}/r^{-\beta|U|} = r^{\beta(|U|-1)}$, which is $o(r)$ when $\beta < \frac{1}{|U|-1}$. With the nodes in V actually present, these then all activate and the transition occurs almost immediately with high probability (see Section 4.3). Note that the queue lengths of the nodes in U have only decreased by an amount of order $r^{\beta(|U|-1)} = o(r)$, and hence are still of order r. In contrast, when $\beta = \frac{1}{|U|-1}$, the probability that all nodes in U become simultaneously inactive before time yT_U is approximately $\pi(y)$ with $\pi(y) = 1-(1-y)^{(1-C)/C}$, $y\in[0,1)$ (see (1.27)). Again, the nodes in V then all activate and the transition occurs almost immediately with high probability. Note that the queue lengths in the nodes in U have then dropped by a non-negligible fraction, but are still of order r. A potential scenario is that the nodes in U are not all simultaneously inactive until their activation rates have become of a smaller order than r^β, due to the queue lengths no longer being of order r just before time T_U. However, the fact that $\pi(y)\uparrow 1$ as $y\uparrow 1$ implies that this scenario has negligible probability in the limit. In contrast, this scenario does occur when $\beta > \frac{1}{|U|-1}$, implying that the crossover occurs in a narrow window around T_U (see Sections 4.1–4.2 for details). We will see that this window has size $O(r^{1/\beta(|U|-1)}) = o(r)$. In particular, the *window gets narrower as the activation rate for nodes in U increases.*

Proofs. We look at a single-node queue length process $t\mapsto Q(t)$ and prove that with high probability it follows a path that lies in a *narrow tube around its mean path* (see Fig. 5). We study separately the input process $t \mapsto Q^+(t)$ and the output process $t \mapsto Q^-(t)$: we use Mogulskii's theorem (a pathwise large-deviation principle) for the first, and Cramér's theorem (a pointwise large-deviation principle) for the second. We derive upper and lower bounds for the queue length process and we use these bounds to construct two couplings that allow us to compare the different models.

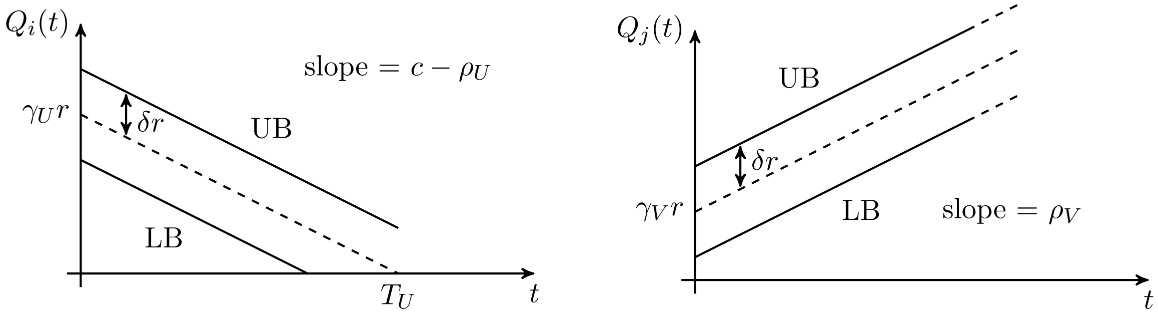

Figure 5: Sketches of the tubes around the mean of the queue length processes, respectively, for a node $i \in U$ and a node $j \in V$.

Dependent packet arrivals. Our large-deviation estimates are so sharp that we can actually allow the Poisson processes of packet arrivals at the different nodes to be *dependent*. Indeed, as long at the marginal processes are Poisson, our large-deviation estimates are valid at every single node, and since the network is finite a simple union bound shows that they are also valid for all nodes simultaneously, at the expense of a negligible factor that is equal to the number of nodes. For modelling purposes independent arrivals are natural, but it is interesting to allow for dependent arrivals when we want to study activation protocols that are more involved.

Open problems. If we want to understand how small the term $o(1)$ in (1.31) actually is, then we need to derive sharper estimates in the coupling. One possibility would be to study moderate deviations for the queue length processes and to look at shrinking tubes. We do not pursue such refinements here. Our main focus for the future will be to extend the model to more complicated settings, where the activation rate at node i depends also on the queue length at the neighbouring nodes of i. We want to be able to compare models with (externally driven) time-dependent rates and models with (internally driven) queue-dependent rates, and show again that their metastable behaviour is similar. We also want to move away from the complete bipartite interference graph and consider more general graphs that capture more realistic wireless networks.

Other models. There are other ways to define an internal model. We mention a few examples.

(i) A simple variant of our model is obtained by fixing the activation rates, but letting the rate at time t of the Poisson deactivation clock of node i depend on the reciprocal of the queue length at time t, i.e., $1/g_i(Q_i(t))$ for some $g_i \in \mathcal{G}$. This can be equivalently seen as a unit-rate Poisson deactivation clock, where node i either deactivates with a probability reciprocal to $g_i(Q_i(t))$, or starts a second activity period. Nodes with a large queue length are more likely to remain active for a long time before switching off, while nodes with a short queue length have extremely short activity periods. If at time t the activation clock of an inactive node with $Q_i(t) = 0$ ticks, then the node does not become active. On the other hand, if during an activity period the queue length of an active node hits zero, then the server switches itself off independently of the deactivation rate. For fixed activation and deactivation rates, this model and our internal model are equivalent up to a time scaling factor. In particular, they have similar stationary distributions.

(ii) An alternative approach is to use a discrete notion of queue length, namely, $Q_i(t) = N_i(t) - S_i(t)$, where $N_i(t)$ is a Poisson process with rate λ, denoting the number of packets arriving at node i during $[0, t]$, while $S_i(t)$ indicates the total number of times node i turns active (or inactive) during $[0, t]$ (we may use λ_U and λ_V to represent different arrival rates for the two sets U and V). The processes $t \mapsto S_i(t)$ and $t \mapsto N_i(t)$ are assumed to be independent. We can define a model where each time a node turns active it serves exactly one packet and then switches off again. The activation clocks still have rates $g_i(Q_i(t))$ with $g_i \in \mathcal{G}$. We can establish results similar to our internal model by adapting the arguments to the discrete setting.

Outline. The remainder of the paper is organised as follows. In Section 2 we state large-deviation bounds for the input and the output process, which allow us to show that the queue length process at every node has specific lower and upper bounds that hold with very high probability. The proofs of these bounds are deferred to Appendices A and B. In Section 3 we use the bounds to couple the lower and the upper external model (with rates (1.14) and (1.15), respectively) to the internal model (with rates (1.12)). In Section 4 we derive the scaling results for the external model, and combine these with the coupling to derive Theorem 1.7 (as stated in Section 1.3).

2 Bounds for the input and output processes

In this section we state the main results of our analysis of the input process and the output process at a fixed node. With the help of path-large-deviation techniques, we show that with high probability the input process lies in a narrow tube around the deterministic path $t \mapsto \frac{\lambda}{\mu} t$ (Proposition 2.1). For simplicity, we suppress the index for the arrival rates λ_U and λ_V, and consider a general rate λ. The same holds for $\rho = \frac{\lambda}{\mu}$. We study the output process only for nodes in U, and we give lower and upper bounds (Equation (2.4) and Proposition 2.5). We look at a single node and suppress its index, since the queues are independent of each other as long as the servers remain active or inactive. The proofs of the propositions below for the input process and the output process are given in Appendices A and B, respectively.

Proposition 2.1 (Tube for the input process). *For $\delta > 0$ small enough and time horizon $S > 0$, let*

$$\Gamma_{S,\delta S} = \left\{ \gamma \in L_\infty([0, S])\colon \ \frac{\lambda}{\mu} s - \delta S < \gamma(s) < \frac{\lambda}{\mu} s + \delta S \ \ \forall\, s \in [0, S] \right\}. \tag{2.1}$$

With high probability the input process lies inside $\Gamma_{S,\delta S}$ as $S \to \infty$, namely,

$$\mathbb{P}\big(Q^+([0, S]) \notin \Gamma_{S,\delta S}\big) = e^{-K_\delta S\,[1+o(1)]}, \qquad S \to \infty. \tag{2.2}$$

$$K_\delta = (\lambda + \delta\mu) + \lambda - 2\sqrt{\lambda(\lambda+\delta\mu)} \in (0,\infty). \tag{2.3}$$

(Note that $\Gamma_{S,\delta S}$ contains negative values. This is of no concern because the path is always non-negative.)

We want to derive lower and upper bounds for the output process for a node in U. The upper bound is trivial by definition, namely,

$$Q^-(t) \le ct, \qquad t \ge 0. \tag{2.4}$$

For the lower bound there are some complications, which is why we need to introduce an auxiliary model.

Definition 2.2 (Isolated model). In the *isolated model* the activation of nodes in U is not affected by the activity states of nodes in V, i.e., they behave as if they were in isolation. On the other hand, nodes in V are still affected by nodes in U, i.e., they cannot activate until every node in U has become inactive. Nodes in V have zero output process.

We will see later that the internal model behaves in exactly the same way as the isolated model up to the pre-transition time, in particular, the pre-transition times in the internal and the isolated model coincide in distribution.

We next define an auxiliary time that will be useful in our analysis.

Definition 2.3 (Auxiliary time T_U). Given the initial queue length $Q(0) = \gamma_U r$ for a node in U, define T_U to be the expected time at which the queue length hits zero. We can write

$$T_U = T_U(r) \sim \alpha r, \qquad r \to \infty, \tag{2.5}$$

with

$$\alpha = \frac{\gamma_u}{c - \rho_U}. \tag{2.6}$$

Remark 2.4. The quantity αr is the expected time at which the queue length hits zero when the node is always active. Since the total inactivity time of a node in U before time T_U will turn out to be negligible compared to αr, we have $T_U \sim \alpha r$ as $r \to \infty$.

We study the output process for the isolated model up to time T_U. We will see later in Corollary 4.3 that the transition time in the internal model occurs with high probability before T_U, so it is enough to look at the time interval $[0, T_U]$. In the rare case when the transition does not occur before T_U, we expect it to occur in a very short time after T_U.

We are now ready to give the lower bound for the output process.

Proposition 2.5 (Lower bound for the output process in the isolated model). *Consider a node in U. For $\delta, \epsilon, \epsilon_1, \epsilon_2 > 0$ small enough, the output process satisfies*

$$\begin{aligned}\mathbb{P}\big(Q^-(t) < ct - \epsilon r \ \ \forall t \in [0, T_U]\big) \le\, & e^{-K_\delta \alpha r\,[1+o(1)]} + e^{-K_1 r\,[1+o(1)]} \\ & + e^{-\left(K_2 r + K_3 \frac{r}{g_U(r)} + K_4 r \log g_U(r)\right)\,[1+o(1)]}, \qquad r \to \infty,\end{aligned} \tag{2.7}$$

with

$$
\begin{aligned}
K_1 &= \Big(\gamma_U - \frac{2\delta\alpha}{c-\rho_U}\Big)\frac{\epsilon_1 - \log(1+\epsilon_1)}{1+\epsilon_1},\\
K_2 &= \Big(\gamma_U - \frac{2\delta\alpha}{c-\rho_U}\Big)(1+\epsilon_1)\bigg(-1-\log\bigg(\frac{\epsilon_2}{\big(\gamma_U - \frac{2\delta\alpha}{c-\rho_U}\big)(1+\epsilon_1)}\bigg)\bigg),\\
K_3 &= \epsilon_2,\\
K_4 &= \Big(\gamma_U - \frac{2\delta\alpha}{c-\rho_U}\Big)(1+\epsilon_1),
\end{aligned}
\tag{2.8}
$$

satisfying $K_1, K_2, K_3, K_4 \in (0,\infty)$.

By combining the bounds for the input process and the output process, and picking $\delta = \epsilon$ and $S = r$, we obtain lower and upper bounds for the queue length process $Q(t)$ of a node in U.

Corollary 2.6 (**Bounds for the queue length process in the isolated model).** *For $\delta > 0$ small enough, the following bounds hold with high probability when $r \to \infty$ for a node in U:*

$$
\begin{aligned}
(LB)_U\colon &\quad Q(t) \geq Q_U^{\mathrm{LB}}(t) = \gamma_U r - (c-\rho_U)t - \delta r, \quad t \geq 0,\\
(UB)_U\colon &\quad Q(t) \leq Q_U^{\mathrm{UB}}(t) = \gamma_U r - (c-\rho_U)t + 2\delta r, \quad t \in [0, T_U].
\end{aligned}
\tag{2.9}
$$

Similarly, the following bounds hold with high probability when $r \to \infty$ for a node in V:

$$
\begin{aligned}
(LB)_V\colon &\quad Q(t) \geq Q_V^{\mathrm{LB}}(t) = \gamma_V r + \rho_V t - \delta r, \quad t \geq 0,\\
(UB)_V\colon &\quad Q(t) \leq Q_V^{\mathrm{UB}}(t) = \gamma_V r + \rho_V t + \delta r, \quad t \geq 0.
\end{aligned}
\tag{2.10}
$$

Proof. The claims follow directly from Propositions 2.1 and 2.5 in combination with (2.4). □

3 Coupling the internal and the external model

In Sections 3.1 and 3.2 we use the bounds defined in Section 2 to construct two couplings that allow us to compare the internal and the external model (Proposition 3.5, respectively, Proposition 3.7 and Corollary 3.8). Throughout the sequel we assume that the deactivation rates are fixed, i.e., the deactivation Poisson clocks ring at rate 1. A node becomes active if and only if all its neighbours are inactive. If a node is inactive, then the activation Poisson clocks ring at rates that vary over time in a deterministic way, or as functions of the queue lengths.

We are interested in coupling the models in the time interval $[0, T_U]$ and on the following event.

Definition 3.1 (**Good behaviour).** Let $\mathcal{E}_\delta$ be the event that the queue length processes all have *good behaviour* in the interval $[0, T_U]$, in the sense that

$$
\begin{aligned}
\mathcal{E}_\delta = &\big\{Q_U^{\mathrm{LB}}(t) \leq Q_i(t) \leq Q_U^{\mathrm{UB}}(t)\ \forall\, t \in [0, T_U]\ \forall\, i \in U\big\}\\
&\cup \big\{Q_V^{\mathrm{LB}}(t) \leq Q_i(t) \leq Q_V^{\mathrm{UB}}(t)\ \forall\, t \in [0, T_U]\ \forall\, i \in V\big\},
\end{aligned}
\tag{3.1}
$$

i.e., the paths lie between their respectively lower and upper bounds for nodes in U and V. This event depends on the perturbation parameter δ.

Lemma 3.2. *For $\delta > 0$ small enough,*

$$\lim_{r\to\infty} \mathbb{P}(\mathcal{E}_\delta) = 1. \tag{3.2}$$

Proof. This follows directly from Corollary 2.6. □

In what follows we couple on the event $\mathcal{E}_\delta$ only. The coupling can be extended in an arbitrary way off the event $\mathcal{E}_\delta$. The way this is done is irrelevant because of Lemma 3.2.

3.1 Coupling the internal model and the lower external model

The lower external model defined in (1.14) can also be described in the following way. At time $t \geq 0$ the activation rates are

$$r_i^{\mathrm{low}}(t) = \begin{cases} g_U(Q_U^{\mathrm{LB}}(t)), & i \in U, \\ g_V(Q_V^{\mathrm{UB}}(t)), & i \in V. \end{cases} \tag{3.3}$$

Remark 3.3. *Note that when the lower bound $Q_U^{\mathrm{LB}}(t)$ becomes negative the activation rate g_U is zero by definition. In this way we are able extend the coupling to any time $t \geq 0$, even though we consider only the interval $[0, T_U]$.*

Lemma 3.4. *With high probability when $r \to \infty$, the transition time in the lower external model is smaller than T_U, i.e.,*

$$\lim_{r\to\infty} \mathbb{P}_u(\tau_v^{\mathrm{low}} \leq T_U) = 1. \tag{3.4}$$

Proof. As we will see in Section 4.2, with high probability the transition time in the external model is smaller than T_U. Since the lower external model is defined for an arbitrarily small perturbation $\delta > 0$, we conclude by using the continuity of g_U, g_V. □

We introduce a system $\mathcal{H}^{\mathrm{low}}$ that allows us to couple the internal model with the lower external model.

- ($\mathcal{H}^{\mathrm{low}}$) Suppose that $h_i(t) \geq \max\{Q_U^{\mathrm{UB}}(t), Q_V^{\mathrm{UB}}(t)\}$ for all $i \in U \cup V$ and all $t \in [0, T_U]$. Consider a system $\mathcal{H}^{\mathrm{low}}$ where clocks are associated with each node in the following way:
  - A Poisson deactivation clock ticks at rate 1. Both the nodes in the lower external model and in the internal model are governed by this clock:
    - if both nodes are active, then they become inactive together;
    - if only one node is active, then it becomes inactive;
    - if both nodes are inactive, then nothing happens.
  - A Poisson activation clock ticks at rate $g_U(h_i(t))$ at time t for a node $i \in U$. Both the nodes in the lower external model and in the internal model are governed by this clock:
    - if both nodes are active, or both are inactive but have active neighbours, then nothing happens;

- if the node in the internal model is active and the node in the lower external model is not, then the latter node becomes active (if it can) with probability

$$\frac{r_i^{\mathrm{low}}(t)}{g_U(h_i(t))}; \tag{3.5}$$

- if both nodes are inactive but can be activated, then this happens with probabilities

$$\begin{aligned} &\frac{r_i^{\mathrm{low}}(t)}{g_U(h_i(t))} \quad \text{for the lower external model,} \\ &\frac{r_i^{\mathrm{int}}(t)}{g_U(h_i(t))} \quad \text{for the internal model,} \end{aligned} \tag{3.6}$$

  where

$$\frac{r_i^{\mathrm{low}}(t)}{g_U(h_i(t))} \leq \frac{r_i^{\mathrm{int}}(t)}{g_U(h_i(t))}, \tag{3.7}$$

  in such a way that if the node in the lower external model activates, then it also activates in the internal model.

- A Poisson activation clock ticks at rate $g_V(h_i(t))$ at time t for a node $i \in V$. The same happens as for the nodes in U, but the activation probabilities are

$$\begin{aligned} &\frac{r_i^{\mathrm{low}}(t)}{g_V(h_i(t))} \quad \text{for the lower external model,} \\ &\frac{r_i^{\mathrm{int}}(t)}{g_V(h_i(t))} \quad \text{for the internal model,} \end{aligned} \tag{3.8}$$

  where

$$\frac{r_i^{\mathrm{low}}(t)}{g_U(h_i(t))} \geq \frac{r_i^{\mathrm{int}}(t)}{g_U(h_i(t))}, \tag{3.9}$$

  in such a way that if the node in the internal model activates, then it also activates in the lower external model.

With the constructions above, we are now able to compare the transition times of the two models.

Proposition 3.5 (Comparison between internal and lower external models).

(i) Under the coupling $\mathcal{H}^{\mathrm{low}}$, the joint activity processes in the internal and in the lower external model are ordered for all $t \in [0, T_U]$, i.e.,

$$\begin{aligned} X_i^{\mathrm{low}}(t) \leq X_i^{\mathrm{int}}(t), \quad i \in U, \\ X_i^{\mathrm{int}}(t) \leq X_i^{\mathrm{low}}(t), \quad i \in V. \end{aligned} \tag{3.10}$$

(ii) With high probability when $r \to \infty$, the transition time τ_v^{int} in the internal model is at least as large as the transition time τ_v^{low} in the lower external model, i.e.,

$$\lim_{r\to\infty} \hat{\mathbb{P}}_u(\tau_v^{\mathrm{low}} \leq \tau_v^{\mathrm{int}}) = 1, \tag{3.11}$$

where $\hat{\mathbb{P}}_u$ is the joint law induced by the coupling with starting configuration u.

Proof. (i) For each node $i \in U$ and for all $t \in [0, T_U]$, $Q_i^{\mathrm{LB}}(t) \leq Q_i(t)$ and $g_U(Q_i^{\mathrm{LB}}(t)) \leq g_U(Q_i(t))$ by the monotonicity of the function g_U. On the other hand, for each node $i \in V$, $Q_i(t) \leq Q_i^{\mathrm{UB}}(t)$ and $g_V(Q_i(t)) \leq g_V(Q_i^{\mathrm{UB}}(t))$ by the monotonicity of the function g_V. Under the system $\mathcal{H}^{\mathrm{low}}$, at any moment the random variable describing the state of a node $i \in U$ in the lower external model is dominated by the one in the internal model, i.e., by (3.7) for all $t \in [0, T_U]$,

$$X_i^{\mathrm{low}}(t) \leq X_i^{\mathrm{int}}(t). \tag{3.12}$$

On the other hand, the random variable describing the state of a node $j \in V$ in the lower external model dominates the one in the internal model, i.e., by (3.9) for all $t \in [0, T_U]$,

$$X_i^{\mathrm{int}}(t) \leq X_i^{\mathrm{low}}(t). \tag{3.13}$$

Hence the joint activity processes in the two models are ordered.

(ii) By construction of the coupling and the ordering above, on the event $\mathcal{E}_\delta$ the nodes in U in the lower external model turn off earlier than in the internal model, and also nodes in V turn on earlier in the lower external model. Hence the transition occurs earlier in the lower external model.

Note that we are able to compare the transition times only when $\tau_v^{\mathrm{low}} \leq T_U$, so we look at the coupling also on the event $\{\tau_v^{\mathrm{low}} \leq T_U\}$, which has high probability when $r \to \infty$ (Lemma 3.4). On this event we have $\tau_v^{\mathrm{low}} \leq \tau_v^{\mathrm{int}}$. Therefore

$$1 = \lim_{r\to\infty} \hat{\mathbb{P}}_u(\mathcal{E}_\delta, \tau_v^{\mathrm{low}} \leq T_U, \tau_v^{\mathrm{low}} \leq \tau_v^{\mathrm{int}}) = \lim_{r\to\infty} \hat{\mathbb{P}}_u(\tau_v^{\mathrm{low}} \leq \tau_v^{\mathrm{int}}). \tag{3.14}$$

□

3.2 Coupling the isolated model and the upper external model

The upper external model defined in (1.15) can also be described in the following way. At time $t \in [0, T_U]$ the activation rates are

$$r_i^{\mathrm{upp}}(t) = \begin{cases} g_U(Q_U^{\mathrm{UB}}(t)), & i \in U, \\ g_V(Q_V^{\mathrm{LB}}(t)), & i \in V. \end{cases} \tag{3.15}$$

Lemma 3.6. *With high probability when $r \to \infty$, the transition time in the upper external model is smaller than T_U, i.e.,*

$$\lim_{r\to\infty} \mathbb{P}_u(\tau_v^{\mathrm{upp}} \leq T_U) = 1. \tag{3.16}$$

This statement is to be read as follows. Let δ be the perturbation parameter in the upper external model appearing in (1.15)*. Then for every $\delta > 0$ there exists a $\delta'(\delta) > 0$, satisfying $\lim_{\delta\downarrow 0} \delta'(\delta) = 0$, such that $\lim_{r\to\infty} \mathbb{P}_u(\tau_v^{\mathrm{upp}} \leq [1 + \delta'(\delta)]T_U) = 1$.*

Proof. Analogous to the proof of Lemma 3.4. □

We introduce a system $\mathcal{H}^{\mathrm{upp}}$ that allows us to couple the isolated model with the upper external model up to time $\bar{\tau}_v^{\mathrm{iso}}$.

- ($\mathcal{H}^{\text{upp}}$) Suppose that $h_i(t) \geq \max\{Q_U^{\text{UB}}(t), Q_V^{\text{UB}}(t)\}$ for all $i \in U \cup V$ and all $t \in [0, \bar{\tau}_v^{\text{iso}}]$. Couple the processes in the same way as for $\mathcal{H}^{\text{low}}$, but with different activation probabilities. The probabilities for the isolated model and for the upper external model are such that

$$\begin{aligned} \frac{r_i^{\text{iso}}(t)}{g_U(h_i(t))} &\leq \frac{r_i^{\text{upp}}(t)}{g_U(h_i(t))}, \quad i \in U, \\ \frac{r_i^{\text{upp}}(t)}{g_V(h_i(t))} &\leq \frac{r_i^{\text{iso}}(t)}{g_V(h_i(t))}, \quad i \in V, \end{aligned} \tag{3.17}$$

where for $t \in [0, \bar{\tau}_v^{\text{iso}}]$

$$r_i^{\text{iso}}(t) = \begin{cases} g_U(Q_i(t)), & i \in U, \\ g_V(Q_i(t)), & i \in V. \end{cases} \tag{3.18}$$

Note that when $\bar{\tau}_v^{\text{iso}} \leq T_U$, the isolated model behaves exactly as the internal model in the interval $[0, \bar{\tau}_v^{\text{iso}}]$, as shown in Appendix B.2. Moreover, the coupling is defined only when $\bar{\tau}_v^{\text{iso}} \leq T_U$. We look then at the coupling also on the event $\{\tau_v^{\text{upp}} \leq T_U\}$, which has high probability when $r \to \infty$ (Lemma 3.6). In the following proposition we see how this ensures that the coupling is well defined, and we compare the pre-transition times of the two models.

Proposition 3.7 (Comparison between isolated and upper external model).

(i) Under the coupling $\mathcal{H}^{\text{upp}}$, the joint activity processes in the isolated model and in the upper external models are ordered up to time $\bar{\tau}_v^{\text{iso}}$, i.e., for all $t \in [0, \bar{\tau}_v^{\text{iso}}]$,

$$\begin{aligned} X_i^{\text{iso}}(t) &\leq X_i^{\text{upp}}(t), \quad i \in U, \\ X_i^{\text{upp}}(t) &\leq X_i^{\text{iso}}(t), \quad i \in V. \end{aligned} \tag{3.19}$$

(ii) With high probability when $r \to \infty$, the pre-transition time $\bar{\tau}_v^{\text{upp}}$ in the upper external model is at least as large as the pre-transition time $\bar{\tau}_v^{\text{iso}}$ in the isolated model, i.e.,

$$\lim_{r\to\infty} \hat{\mathbb{P}}_u(\bar{\tau}_v^{\text{iso}} \leq \bar{\tau}_v^{\text{upp}}) = 1, \tag{3.20}$$

where $\hat{\mathbb{P}}_u$ is the joint law induced by the coupling with starting configuration u.

Proof. (i) The proof is analogous to that of Proposition 3.5, but this time we use the system $\mathcal{H}^{\text{upp}}$ up to time $\bar{\tau}_v^{\text{iso}}$ and all the inequalities are reversed.

(ii) By construction of the coupling and the ordering above, on the event $\mathcal{E}_\delta \cap \{\tau_v^{\text{upp}} \leq T_U\}$ the nodes in U in the isolated model turn off earlier than in the upper external model, and also the first node in V turns on earlier in the isolated model. Hence the pre-transition occurs earlier in the isolated model, and we have $\bar{\tau}_v^{\text{iso}} \leq \bar{\tau}_v^{\text{upp}} \leq \tau_v^{\text{upp}} \leq T_U$. Therefore the coupling is well defined and

$$1 = \lim_{r\to\infty} \hat{\mathbb{P}}_u(\mathcal{E}_{\delta,T_U}, \tau_v^{\text{upp}} \leq T_U, \bar{\tau}_v^{\text{iso}} \leq \bar{\tau}_v^{\text{upp}}) = \lim_{r\to\infty} \hat{\mathbb{P}}_u(\bar{\tau}_v^{\text{iso}} \leq \bar{\tau}_v^{\text{upp}}). \tag{3.21}$$

□

Corollary 3.8. *With high probability when $r \to \infty$, the transition time $\tau_v^{\rm upp}$ in the upper external model is at least as large as the pre-transition time $\bar{\tau}_v^{\rm int}$ in the internal model, i.e.,*

$$\lim_{r\to\infty} \hat{\mathbb{P}}_u(\bar{\tau}_v^{\rm int} \leq \tau_v^{\rm upp}) = 1. \tag{3.22}$$

Proof. Since $\lim_{r\to\infty} \mathbb{P}(\bar{\tau}_v^{\rm iso} \leq T_U) = 1$, we have, as shown in Proposition B.6 in Appendix B.2, that the pre-transition times in the isolated model and in the internal model coincide. Hence

$$1 = \lim_{r\to\infty} \hat{\mathbb{P}}_u(\bar{\tau}_v^{\rm iso} \leq \bar{\tau}_v^{\rm upp}) = \lim_{r\to\infty} \hat{\mathbb{P}}_u(\bar{\tau}_v^{\rm int} \leq \bar{\tau}_v^{\rm upp}) \leq \lim_{r\to\infty} \hat{\mathbb{P}}_u(\bar{\tau}_v^{\rm int} \leq \tau_v^{\rm upp}). \tag{3.23}$$

□

4 Transition times

The goal of this section is to identify the asymptotic behaviour of the transition time in the internal model. In Sections 4.1 and 4.2 we look at the external model and prove Theorems 1.5 and Theorem 1.6, respectively. In Section 4.3 we show that the difference between the transition time and the pre-transition time is negligible for all the models considered in the paper. In Section 4.4 we put these results together to prove Theorem 1.7.

4.1 Critical time scale in the external model

In this section we prove Theorem 1.5. Below, $a(r) \sim b(r)$ means that $\lim_{r\to\infty} a(r)/b(r) = 1$, while $a(r) \asymp b(r)$ means that $0 < \liminf_{r\to\infty} a(r)/b(r) \leq \limsup_{r\to\infty} a(r)/b(r) < \infty$.

Proof. In order to compute the critical time scale M_c, we must solve the equation $M\nu(M) = 1$ in (1.20). We know from (1.17) and (1.21) that

$$\nu(s) \sim |U| r_U(s)^{1-|U|}, \qquad r \to \infty. \tag{4.1}$$

We want to identify how the transition time is related to the choice of the activation function $g_U(x) = Gx^\beta$ with $\beta, G \in (0,\infty)$. Consider the time scale $M_c = F_c r^\gamma$, where $\gamma \in (0,1]$ and $F_c \in (0,\infty)$. For $r \to \infty$ we have

$$\begin{aligned} 1 = r^0 = M_c\nu(M_c) &= F_c r^\gamma\, \nu(F_c r^\gamma) \sim F_c r^\gamma |U| r_U(F_c r^\gamma)^{-(|U|-1)} \\ &= F_c r^\gamma |U| g_U\big(\gamma_U r - (c-\rho_U)F_c r^\gamma\big)^{-(|U|-1)} \\ &= F_c r^\gamma |U| G^{-(|U|-1)}\big(\gamma_U r - (c-\rho_U)F_c r^\gamma\big)^{-\beta(|U|-1)}. \end{aligned} \tag{4.2}$$

Recall from (2.6) that $\alpha = \frac{\gamma_U}{c-\rho_U}$. We consider three cases:

- **Case** $\gamma \in (0,1)$ **and** $F_c \in (0,\infty)$**.** For $r \to \infty$ the criterion in (4.2) reads

$$1 = r^0 \sim F_c r^\gamma |U| G^{-(|U|-1)}(\gamma_U r)^{-\beta(|U|-1)}. \tag{4.3}$$

  In order for the exponents of r to match, we need

$$\beta = \frac{\gamma}{|U|-1}. \tag{4.4}$$

Inserting (4.4) into (4.3), we get

$$F_c|U|G^{-(|U|-1)}\gamma_U^{-\beta(|U|-1)} = 1, \tag{4.5}$$

which gives

$$F_c = \frac{\gamma_U^{\beta(|U|-1)}G^{(|U|-1)}}{|U|}. \tag{4.6}$$

Hence

$$M_c = \frac{(G\gamma_U^{\beta})^{|U|-1}}{|U|}r^{\beta(|U|-1)}. \tag{4.7}$$

- **Case** $\gamma = 1$ **and** $F_c \in (0, \alpha)$**.** For $r \to \infty$ the criterion in (4.2) reads

$$1 = r^0 \sim F_c|U|G^{-(|U|-1)}(\gamma_U - (c - \rho_U)F_c)^{-\beta(|U|-1)}r^{1-\beta(|U|-1)}. \tag{4.8}$$

In order for the exponents of r to match, we need

$$\beta = \frac{1}{|U|-1}. \tag{4.9}$$

Inserting (4.9) into (4.8), we get

$$\frac{F_c|U|G^{-(|U|-1)}}{\gamma_U - (c - \rho_U)F_c} = 1, \tag{4.10}$$

which gives

$$F_c = \frac{\gamma_U}{|U|G^{-(|U|-1)} + (c - \rho_U)}. \tag{4.11}$$

Hence

$$M_c = \frac{\gamma_U}{|U|G^{-(|U|-1)} + (c - \rho_U)}r. \tag{4.12}$$

Recall from (2.5) that $T_U \sim \alpha r$ is the expected time at which the queue length at a node in U hits zero. We will see in Section 4.2 that the transition in the external model typically occurs before the queues are empty.

- **Case** $\gamma = 1$ **and** $F_c = \alpha - Dr^{-\delta}$**,** $\delta \in (0, 1)$**.** For $r \to \infty$ the criterion in (4.2) reads

$$1 = r^0 \sim \alpha r|U|G^{-(|U|-1)}\big((c - \rho_U)Dr^{1-\delta}\big)^{-\beta(|U|-1)}. \tag{4.13}$$

In order for the exponents of r to match, we need

$$\beta = \frac{1}{(1-\delta)(|U|-1)}. \tag{4.14}$$

Inserting (4.14) into (4.13), we get

$$\alpha|U|G^{-(|U|-1)}((c - \rho_U)D)^{-\beta(|U|-1)} = 1, \tag{4.15}$$

which gives

$$D = \frac{(\alpha|U|G^{-(|U|-1)})^{1/\beta(|U|-1)}}{c - \rho_U}. \tag{4.16}$$

Hence

$$M_c = \alpha r - \frac{(\alpha|U|G^{-(|U|-1)})^{1/\beta(|U|-1)}}{c-\rho_U} r^{1/\beta(|U|-1)}, \tag{4.17}$$

and so the crossover takes place in a window of size $O(r^{1/\beta(|U|-1)}) = o(r)$ around αr. Note that this window gets narrower as β increases, i.e., as the activation rate for nodes in U increases.

□

Remark 4.1 (Modulation with slowly varying functions)**.** We may consider the more general case where the activation function is $g_U(x) = x^\beta \hat{\mathcal{L}}(x)$ with $\beta \in (0,\infty)$ and $\hat{\mathcal{L}}(x)$ a slowly varying function (i.e., $\lim_{x\to\infty} \hat{\mathcal{L}}(ax)/\hat{\mathcal{L}}(x) = 1$ for all $a > 0$). Let $M_c = r^\gamma \mathcal{L}(r)$ with $\gamma \in (0,1)$ and $\mathcal{L}(r)$ a slowly varying function. When $r \to \infty$, we have

$$\begin{aligned} 1 = r^0 &\sim r^\gamma \mathcal{L}(r)|U|\big(\gamma_U r - (c-\rho_U)r^\gamma\mathcal{L}(r)\big)^{-\beta(|U|-1)}\hat{\mathcal{L}}(\gamma_U r - (c-\rho_U)r^\gamma \mathcal{L}(r))^{-(|U|-1)} \\ &\sim r^\gamma\mathcal{L}(r)|U|(\gamma_U r)^{-\beta(|U|-1)}\hat{\mathcal{L}}(\gamma_U r)^{-(|U|-1)}. \end{aligned} \tag{4.18}$$

In order for the exponents of r to match, we again need

$$\beta = \frac{\gamma}{|U|-1}. \tag{4.19}$$

We get

$$\mathcal{L}(r) = \frac{\gamma_U^{\beta(|U|-1)}}{|U|}\hat{\mathcal{L}}(\gamma_U r)^{|U|-1}. \tag{4.20}$$

Hence

$$M_c = \frac{\gamma_U^{\beta(|U|-1)}}{|U|} r^{\beta(|U|-1)}\hat{\mathcal{L}}(\gamma_U r)^{|U|-1}. \tag{4.21}$$

We can even include the case $\beta = 0$, and get that if $g_U(x) = \hat{\mathcal{L}}(x)$ with $\lim_{x\to\infty}\hat{\mathcal{L}}(x) = \infty$, then

$$M_c = \frac{1}{|U|}\hat{\mathcal{L}}(\gamma_U r)^{|U|-1}. \tag{4.22}$$

4.2 Transition time in the external model

In this section we prove Theorem 1.6. We already know that the transition occurs on the critical time scale M_c computed in Section 4.2.

Proof. Knowing the critical time scale M_c, we can compute the mean transition time from (1.19):

$$\begin{aligned} \mathbb{E}_u[\tau_v^{\mathrm{ext}}] &= \int_0^\infty \mathbb{P}_u(\tau_v^{\mathrm{ext}} > x)\,dx = M_c\int_0^\infty \mathbb{P}_u\Big(\frac{\tau_v^{\mathrm{ext}}}{M_c} > x\Big)\,dx \\ &\sim M_c\int_0^\infty e^{-\int_0^x M_c\nu(M_c s)\,ds}\,dx = M_c\int_0^\infty e^{-\int_0^x \frac{M_c\nu(M_c s)}{M_c\nu(M_c)}ds}\,dx \\ &= M_c\int_0^\infty e^{-\int_0^x \left(\frac{\gamma_U r-(c-\rho_U)M_c s}{\gamma_U r-(c-\rho_U)M_c}\right)^{-\beta(|U|-1)}ds}\,dx, \qquad r\to\infty, \end{aligned} \tag{4.23}$$

where the choice of β is important.

- **Case** $\beta \in (0, \frac{1}{|U|-1})$**,** $M_c = F_c r^\gamma$**,** $\gamma \in (0,1)$**.**
  We have
$$\left(\frac{\gamma_U r - (c-\rho_U)M_c s}{\gamma_U r - (c-\rho_U)M_c}\right)^{-\beta(|U|-1)} \to 1, \qquad r \to \infty. \tag{4.24}$$
  Hence
$$\mathbb{E}_u[\tau_v^{\text{ext}}] \sim M_c \int_0^\infty e^{-\int_0^x ds}\, dx = M_c \int_0^\infty e^{-x}\, dx = M_c, \qquad r \to \infty. \tag{4.25}$$
  The law of τ_v^{ext} is exponential, i.e.,
$$\lim_{r\to\infty} \mathbb{P}_u\left(\frac{\tau_v^{\text{ext}}}{\mathbb{E}_u[\tau_v^{\text{ext}}]} > x\right) = e^{-x}, \qquad x \in [0,\infty). \tag{4.26}$$

- **Case** $\beta = \frac{1}{|U|-1}$**,** $M_c = F_c r$**,** $F_c \in (0,\alpha)$**.** We have
$$\begin{aligned}\left(\frac{\gamma_U r - (c-\rho_U)F_c r s}{\gamma_U r - (c-\rho_U)F_c r}\right)^{-\beta(|U|-1)} &= \frac{\gamma_U - (c-\rho_U)F_c}{\gamma_U - (c-\rho_U)F_c s} = \frac{1-\frac{c-\rho_U}{\gamma_U}F_c}{1-\frac{c-\rho_U}{\gamma_U}F_c s}\\ &= \frac{1-\frac{F_c}{\alpha}}{1-\frac{F_c}{\alpha}s} = \frac{1-C}{1-Cs},\end{aligned} \tag{4.27}$$
  with $C = \frac{F_c}{\alpha}$. Hence
$$\begin{aligned}\mathbb{E}_u[\tau_v^{\text{ext}}] &\sim M_c \int_0^{1/C} e^{-\int_0^x \frac{1-C}{1-Cs} ds}\, dx = M_c \int_0^{1/C} e^{-\log(1-Cx)^{-\frac{1-C}{C}}}\, dx\\ &= M_c \int_0^{1/C} (1-Cx)^{\frac{1-C}{C}}\, dx = M_c \Bigg[(1-Cx)^{1+\frac{1-C}{C}} \frac{1}{(1+\frac{1-C}{C})(-C)}\Bigg]_0^{1/C}\\ &= M_c \Bigg[-(1-Cx)^{\frac{1}{C}}\Bigg]_0^{1/C} = M_c, \qquad r \to \infty.\end{aligned} \tag{4.28}$$
  Here, the integral must be truncated at $x = \frac{1}{C}$ because for larger x the integrand becomes negative. Indeed, note that when $x = \frac{1}{C} = \frac{\alpha}{F_c}$, which corresponds to time $T_U = \alpha r$, we have
$$\begin{aligned}\lim_{r\to\infty} \mathbb{P}_u\big(\tau_v^{\text{ext}} > T_U\big) = \lim_{r\to\infty} \mathbb{P}_u\left(\tau_v^{\text{ext}} > \frac{\alpha}{F_c}F_c r\right) &= \lim_{r\to\infty} \mathbb{P}_u\left(\frac{\tau_v^{\text{ext}}}{M_c} > \frac{\alpha}{F_c}\right)\\ = \left(1 - C\frac{\alpha}{F_c}\right)^{\frac{1-C}{C}} &= 0,\end{aligned} \tag{4.29}$$
  because $C = \frac{F_c}{\alpha}$. This means that, with high probability when $r \to \infty$, the transition occurs before time T_U. The law of τ_v^{ext} is truncated polynomial:
$$\lim_{r\to\infty} \mathbb{P}_u\left(\frac{\tau_v^{\text{ext}}}{\mathbb{E}_u[\tau_v^{\text{ext}}]} > x\right) = \begin{cases} (1-Cx)^{\frac{1-C}{C}}, & x \in [0, \frac{1}{C}),\\ 0, & x \in [\frac{1}{C}, \infty).\end{cases} \tag{4.30}$$

- **Case $\beta \in (\frac{1}{|U|-1}, \infty)$, $M_c = \alpha r$.** This case corresponds to the limit $C \uparrow 1$ of the previous case. In this limit, (4.30) becomes

$$\lim_{r\to\infty} \mathbb{P}_u\left(\frac{\tau_v^{\text{ext}}}{\mathbb{E}_u[\tau_v^{\text{ext}}]} > x\right) = \begin{cases} 1, & x \in [0,1), \\ 0, & x \in [1,\infty). \end{cases} \tag{4.31}$$

□

4.3 Negligible gap in the internal model

In this section we focus on the internal model and estimate the length of the interval $[\bar{\tau}_v^{\text{int}}, \tau_v^{\text{int}}]$, which turns out to be very small with high probability. This implies that the transition time has the same asymptotic behaviour as the pre-transition time.

We know that the queue at a node $i \in V$ is of order r at time $\bar{\tau}_v^{\text{int}}$, i.e., $Q_i(\bar{\tau}_v^{\text{int}}) \asymp r$, since it starts at $\gamma_V r$, with $\gamma_V > 0$, and only the input process is present until this time. Hence all the activation Poisson clocks at nodes in V tick at a very aggressive rate. The idea is that within the activation period (which has an exponential distribution with mean 1) of the first node activating in V, all the other nodes in V become active because they are not "blocked" by any node in U. Consequently, the system quickly reaches the configuration v.

Theorem 4.2 (Negligible gap). *In the internal model*

$$\lim_{r\to\infty} \mathbb{P}_u\left(\tau_v^{\text{int}} - \bar{\tau}_v^{\text{int}} = o\left(\frac{1}{g_V(r)}\right)\right) = 1. \tag{4.32}$$

Proof. Starting from $\bar{\tau}_v^{\text{int}}$, a node $x \in V$ remains inactive for an exponential period with mean $\frac{1}{r_x^{\text{int}}(\bar{\tau}_v)} = \frac{1}{g_V(Q(\bar{\tau}_v))} \asymp \frac{1}{g_V(r)}$. Denote by W_x the length of an inactivity period for a node $x \in V$. We have i.i.d. inactivity periods $W_x \simeq \text{Exp}(g_V(Q(\bar{\tau}_v)))$ for all $x \in V \setminus \{x_1\}$, where x_1 is the first node activating in V. We label the remaining nodes $x_2, \ldots, x_{|V|}$ in an arbitrary way. We also have i.i.d. activity periods $Z_x \simeq \text{Exp}(1)$ for all $x \in V$.

Consider a time $t_1 = o\big(\frac{1}{g_V(r)}\big)$. With high probability all the nodes in V activate before time t_1, i.e.,

$$\mathbb{P}\big(W_{x_i} < t_1, \forall\, i = 2, \ldots, |V|\big) = \mathbb{P}\big(W_{x_2} < t_1\big)^{|V|-1} = \big(1 - e^{-g_V(Q(\bar{\tau}_v))t_1}\big)^{|V|-1} \xrightarrow[r\to\infty]{} 1. \tag{4.33}$$

Moreover, with high probability, once activated, all nodes in V stay active for a period of length at least $t_2 \asymp \frac{1}{g_V(r)} > t_1$, i.e.,

$$\mathbb{P}\big(Z_{x_i} > t_2 \;\; \forall\, i = 1, \ldots, |V|\big) = \mathbb{P}\big(Z_{x_1} > t_2\big)^{|V|} = (e^{-t_2})^{|V|} \xrightarrow[r\to\infty]{} 1. \tag{4.34}$$

In conclusion, when $r \to \infty$, with high probability every node in V activates before time t_1 and remains active for at least a time $t_2 > t_1$. This ensure that the transition occurs before time t_2. In particular, it occurs when the last node in V activates (which occurs even before time t_1), so that $\tau_v^{\text{int}} - \bar{\tau}_v^{\text{int}} = o\big(\frac{1}{g_V(r)}\big)$. □

Note that this argument extends to any "external" model with activation rates that tend to infinity with r, in particular, to all the models considered in this paper. The transition always happens quickly after the pre-transition, due to the high level of aggressiveness of nodes in V.

Corollary 4.3. *With high probability when $r \to \infty$, the transition time in the internal model is smaller than T_U, i.e.,*

$$\lim_{r\to\infty} \mathbb{P}_u(\tau_v^{\mathrm{int}} \le T_U) = 1. \tag{4.35}$$

Proof. This follows immediately from Lemma 3.6, Corollary 3.8 and Theorem 4.2.

□

4.4 Transition time in the internal model

In this section we prove Theorem 1.7. First we derive the sandwich of the transition times in the lower external, the internal and the upper external model. After that we identify the asymptotics of the transition time for the internal model by using the results for the external models.

Proof. Using Proposition 3.5, Corollary 3.8 and Theorem 4.2, we have that there exists a coupling such that

$$\begin{aligned} 1 &= \lim_{r\to\infty} \hat{\mathbb{P}}_u\bigg(\tau_v^{\mathrm{low}} \le \tau_v^{\mathrm{int}}, \tau_v^{\mathrm{int}} = \bar{\tau}_v^{\mathrm{int}} + o\bigg(\frac{1}{g_V(r)}\bigg), \bar{\tau}_v^{\mathrm{int}} \le \tau_v^{\mathrm{upp}}\bigg) \\ &= \lim_{r\to\infty} \hat{\mathbb{P}}_u\bigg(\tau_v^{\mathrm{low}} \le \tau_v^{\mathrm{int}} \le \tau_v^{\mathrm{upp}} + o\bigg(\frac{1}{g_V(r)}\bigg)\bigg) \\ &= \lim_{r\to\infty} \hat{\mathbb{P}}_u\big(\tau_v^{\mathrm{low}} \le \tau_v^{\mathrm{int}} \le \tau_v^{\mathrm{upp}}\big), \end{aligned} \tag{4.36}$$

where $\hat{\mathbb{P}}_u$ is the joint law of the three models on the same probability space all three starting from configuration u.

By Theorem 1.6, we know the law of the transition time in the external model. By construction, we have $\mathbb{E}_u[\tau_v^{\mathrm{low}}] \le \mathbb{E}_u[\tau_v^{\mathrm{ext}}] \le \mathbb{E}_u[\tau_v^{\mathrm{upp}}]$. When considering the lower and the upper external models, the transition time asymptotics are controlled by the prefactors $F_{c,\delta}^{\mathrm{low}}$ and $F_{c,\delta}^{\mathrm{upp}}$, respectively, which are perturbations of the prefactor F_c due to the perturbations of the activation rates. In particular, we know from (1.29) that $\lim_{\delta\downarrow 0} F_{c,\delta}^{\mathrm{low}} = \lim_{\delta\downarrow 0} F_{c,\delta}^{\mathrm{upp}} = F_c$. Hence, for all $\epsilon > 0$,

$$\mathbb{E}_u[\tau_v^{\mathrm{int}}] = (F_c \pm \epsilon) r^{\beta(|U|-1)}\,[1 + o(1)], \qquad r \to \infty, \tag{4.37}$$

and since ϵ can be taken arbitrarily small, it may be absorbed into the $o(1)$-term, as

$$\mathbb{E}_u[\tau_v^{\mathrm{int}}] = F_c r^{\beta(|U|-1)}\,[1 + o(1)], \qquad r \to \infty. \tag{4.38}$$

The same kind of argument applies to the law of the transition time, since for any $x \in [0, \infty)$,

$$\lim_{r\to\infty} \mathbb{P}_u(\tau_v^{\mathrm{low}} > x) \le \lim_{r\to\infty} \mathbb{P}_u(\tau_v^{\mathrm{int}} > x) \le \lim_{r\to\infty} \mathbb{P}_u(\tau_v^{\mathrm{upp}} > x). \tag{4.39}$$

□

A Appendix: the input process

The main target of this appendix is to prove Proposition 2.1 in Section 2. We use path large-deviation techniques. For simplicity, we suppress the index for the arrival rates λ_U and λ_V, and consider a general rate λ. We show that with high probability the input process lies in a narrow tube around the deterministic path $t \mapsto \frac{\lambda}{\mu}t$.

Consider a single queue, and for simplicity suppress its index. For $T > 0$, define the scaled process

$$Q_n^+(t) = \frac{1}{n}Q^+(nt) = \frac{1}{n}\sum_{j=1}^{N(nt)} Y_j, \qquad t \in [0,T], \tag{A.1}$$

with $Q_n^+(0) = 0$. We have

$$\mathbb{E}[Q_n^+(t)] = \frac{1}{n}\frac{\lambda nt}{\mu} = \frac{\lambda}{\mu}t, \tag{A.2}$$

and, by the strong law of large numbers, $Q_n^+(t) \to \frac{\lambda}{\mu}t$ almost surely for every t as $n \to \infty$.

When studying the process $t \mapsto Q_n^+(t)$, we need to take into account that this is a combination of the arrival process $t \mapsto N(nt)$ and the service times Y_j, $j \in \mathbb{N}$. Two different types of fluctuations can occur: packets arrive at a slower/faster rate than λ, respectively, shorter/longer service times occur for each packet. Both need to be considered for a proper large-deviation analysis.

A.1 LDP for the two components

Definition A.1 (Space of paths). Consider the space $L_\infty([0,T])$ of *essentially bounded* functions in $[0,T]$, with the norm $\|f\|_\infty = \operatorname{ess\,sup}_{x\in[0,T]} |f(x)|$ called the essential norm. A function f is essentially bounded, i.e. $f \in L_\infty([0,T])$, when there is a measurable function g on $[0,T]$ such that $f = g$ except on a set of measure zero and g is bounded. Let $\mathcal{AC}_T \subset L_\infty([0,T])$ denote the space of *absolutely continuous* functions $f\colon [0,T] \to \mathbb{R}$ such that $f(0) = 0$. □

Given the Poisson arrival process $t \mapsto N(nt)$ with rate λ, define the scaled process $t \mapsto Z_n(t)$ by

$$Z_n(t) = \frac{1}{n}N(nt) = \frac{1}{n}\sum_{i=1}^{[nt]} X_i, \qquad t \in [0,T], \tag{A.3}$$

where $X_i = \text{Poisson}(\lambda)$ are i.i.d. random variables. Note that $N(nt) = \text{Poisson}(\lambda nt)$. Let ν_n be the law of $(Z_n(t))_{t\in[0,T]}$ on $L_\infty([0,T])$. Note that $Z_n(t)$ is asymptotically equivalent to $N(t)$ with mean $\mathbb{E}[Z_n(t)] = \lambda t$, and $(Z_n(t))_{t\in[0,T]} \to (\lambda t)_{t\in[0,T]}$ as $n \to \infty$.

Lemma A.2 (LDP for arrival process). *The family of probability measures $(\nu_n)_{n\in\mathbb{N}}$ satisfies the LDP on $L_\infty([0,T])$ with rate n and with good rate function I_N given by*

$$I_N(\eta) = \begin{cases} \int_0^T \Lambda_N^*(\dot\eta(t))\,dt, & \eta \in \mathcal{AC}_T, \\ \infty, & \textit{otherwise}, \end{cases} \tag{A.4}$$

where $\Lambda_N^*(x) = x\,\log(\frac{x}{\lambda}) - x + \lambda$, $x \in \mathbb{R}_+$.

Proof. Apply Mogulskii's theorem ([3, Theorem 5.1.2]). Use that Λ_N^* is the Fenchel-Legendre transform of the cumulant generating function Λ defined by $\Lambda(\theta) = \log \mathbb{E}(e^{\theta X_1})$, $\theta \in \mathbb{R}$. □

The LDP implies that if $\Gamma \subset L_\infty([0,T])$ is an I_N-continuous set, i.e., $I_N(\Gamma) = I_N(\mathrm{int}(\Gamma)) = I_N(\mathrm{cl}(\Gamma))$, then

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big(Z_n([0,T]) \in \Gamma\big) = -I_N(\Gamma). \tag{A.5}$$

Informally, the LDP reads as the approximate statement

$$\mathbb{P}\big(Z_n([0,T]) \approx \eta([0,T]\big) = e^{-nI_N(\eta)[1+o(1)]}, \qquad n \to \infty, \tag{A.6}$$

where $\approx$ stands for close in the essential norm. Informally, on this event we may approximate

$$Q_n^+(t) = \frac{1}{n} \sum_{j=1}^{N(nt)} Y_j = \frac{1}{n} \sum_{j=1}^{nZ_n(t)} Y_j \approx \frac{1}{n} \sum_{j=1}^{n\eta(t)} Y_j \approx \frac{1}{n} \sum_{j=1}^{[n\eta(t)]} Y_j, \qquad t \in [0,T], \tag{A.7}$$

where $\approx$ now stands for close in the Euclidean norm. Given $\eta \in L_\infty([0,T])$, let μ_n^η denote the law of $(Q_n^+(t))_{t\in[0,T]}$ on $L_\infty([0,T])$.

Lemma A.3 (**LDP for input process conditional on arrival process).** *Given $\eta \in L_\infty([0,T])$, the family of probability measures $(\mu_n^\eta)_{n\in\mathbb{N}}$ satisfies the LDP on $L_\infty([0,T])$ with rate n and with good rate function I_Q^η given by*

$$I_Q^\eta(\phi) = \begin{cases} \int_0^T \Lambda_Q^*\left(\frac{d\phi(t)}{d\eta(t)}\right) d\eta(t), & \phi \in \mathcal{AC}_T, \\ \infty, & \textit{otherwise}, \end{cases} \tag{A.8}$$

where $\Lambda_Q^*(x) = x\mu - 1 - \log(x\mu)$, $x \in \mathbb{R}_+$.

Proof. Again apply Mogulskii's theorem, this time with $\eta(t)$ as the time index. Use that Λ^* is the Fenchel-Legendre transform of the cumulant generating function Λ defined by $\Lambda(\theta) = \log \mathbb{E}(e^{\theta Y_1})$, $\theta \in \mathbb{R}$. □

A.2 Measures in product spaces

The rate function I_Q^η describes the large deviations of the sequence of processes $(Q_n^+(t))_{t\in[0,T]}$ given the path η. To derive the LDP averaged over η, we need a small digression into measures in product spaces.

Definition A.4 (**Product measures).** Define the family of probability measures $(\rho_n)_{n\in\mathbb{N}}$ such that $\rho_n = \nu_n \mu_n^\eta$. These measures are defined on the product space $L_\infty([0,T]) \times L_\infty([0,T])$ given by the Cartesian product of the space $L_\infty([0,T])$ with itself, equipped with the product topology,

The open sets in the product topology are unions of sets of the form $U_1 \times U_2$ with U_1, U_2 open in $L_\infty([0,T])$. Moreover, the product of base elements of $L_\infty([0,T])$ gives a basis for the product space $L_\infty([0,T]) \times L_\infty([0,T])$. Define the projections $\mathrm{Pr}_i \colon L_\infty([0,T]) \times L_\infty([0,T]) \to L_\infty([0,T])$, $i = 1,2$, onto the first and the second coordinates, respectively. The product topology on $L_\infty([0,T]) \times L_\infty([0,T])$ is the topology generated by sets of the form $\mathrm{Pr}_i^{-1}(U_i)$, $i = 1,2$, where and U_1, U_2 are open subsets of $L_\infty([0,T])$.

Lemma A.5 (**Product LDP).** *The family of probability measures $(\rho_n)_{n\in\mathbb{N}}$ satisfies the LDP on $L_\infty([0,T]) \times L_\infty([0,T])$ with rate n and with good rate function I given by*

$$I(\phi,\eta) = \begin{cases} \int_0^T \Lambda_Q^*\left(\frac{d\phi(t)}{d\eta(t)}\right) d\eta(t) + \int_0^T \Lambda_N^*(\dot\eta(t))\, dt, & \phi,\eta \in \mathcal{AC}_T, \\ \infty, & \textit{otherwise.} \end{cases} \tag{A.9}$$

A.3 LDP for the input process

The Contraction Principle allows us to derive the LDP averaged over η. Indeed, let $\mathcal{X} = L_\infty([0,T]) \times L_\infty([0,T])$ and $\mathcal{Y} = L_\infty([0,T])$, let $(\rho_n)_{n\in\mathbb{N}}$ be a sequence of product measures on $\mathcal{X}$, and consider the projection Pr_1 onto $\mathcal{Y}$, which is a continuous map. Then the sequence of induced measures $(\mu_n)_{n\in\mathbb{N}}$ given by $\mu_n = \rho_n \mathrm{Pr}_1^{-1}$ satisfies the LDP on $L_\infty([0,T])$ with good rate function

$$\tilde{I}_Q(\phi) = \inf_{(\phi,\eta)\in Pr_1^{-1}(\{\phi\})} I(\phi,\eta) = \inf_{\eta\in L_\infty([0,T])} I(\phi,\eta). \tag{A.10}$$

We can now state the LDP for the input process $(Q_n^+(t))_{t\in[0,T]}$.

Proposition A.6 (LDP for the input process). *The family of probability measures $(\mu_n)_{n\in\mathbb{N}}$ satisfies the LDP on $L_\infty[0,T]$ with rate n and with good rate function $\hat{I}$ given by*

$$\hat{I}_Q(\Gamma) = \inf_{\phi\in\Gamma} \tilde{I}_Q(\phi). \tag{A.11}$$

In particular, if Γ is $\hat{I}_Q$-continuous, i.e., $\hat{I}_Q(\Gamma) = \hat{I}_Q(\mathrm{int}(\Gamma)) = \hat{I}_Q(\mathrm{cl}(\Gamma))$, then

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big(Q_n^+([0,T]) \in \Gamma\big) = -\hat{I}_Q(\Gamma). \tag{A.12}$$

It is interesting to look at a specific subset of $L_\infty([0,T])$ that gives good bounds for the input process. We are now in a position to prove Proposition 2.1.

Proof. Computing the Fenchel-Legendre transforms Λ_Q^* and Λ_N^*, and picking $\eta(t) = \lambda t$ and $\phi(t) = \frac{1}{\mu}\eta(t) = \frac{1}{\mu}\lambda t$, we easily check that the rate function attains its minimal value zero. Hence with high probability the input process is close to this deterministic path.

We can now estimate the probability of the scaled input process to go outside $\Gamma_{T,\delta}$, which represents a tube of width 2δ around the mean path in the interval $[0,T]$. More precisely,

$$\Gamma_{T,\delta} = \left\{\gamma \in L_\infty([0,T])\colon\ \frac{\lambda}{\mu}t - \delta < \gamma(t) < \frac{\lambda}{\mu}t + \delta \ \ \forall\, t \in [0,T]\right\}. \tag{A.13}$$

We may set $T = 1$ for simplicity and look at the scaled input process in the time interval $[0,1]$. We have

$$\hat{I}_Q((\Gamma_{1,\delta})^c) = \hat{I}_Q(\mathrm{int}((\Gamma_{1,\delta}))^c) = \hat{I}_Q(\mathrm{cl}((\Gamma_{1,\delta}))^c). \tag{A.14}$$

Hence $(\Gamma_{1,\delta})^c$ is $\hat{I}_Q$-continuous, and so according to (A.12),

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big(Q_n^+([0,1]) \notin \Gamma_{1,\delta}\big) = -\hat{I}_Q((\Gamma_{1,\delta})^c). \tag{A.15}$$

Since

$$\begin{aligned}
&\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big(Q_n^+([0,1]) \notin \Gamma_{1,\delta}\big)\\
&= \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\Bigg(\left\{\frac{\lambda}{\mu}t - \delta < Q_n^+(t) < \frac{\lambda}{\mu}t + \delta \ \ \forall\, t \in [0,1]\right\}^c\Bigg)\\
&= \lim_{S\to\infty} \frac{1}{S} \log \mathbb{P}\Bigg(\left\{\frac{\lambda}{\mu}s - \delta S < Q^+(s) < \frac{\lambda}{\mu}s + \delta S \ \ \forall\, s \in [0,S]\right\}^c\Bigg),
\end{aligned} \tag{A.16}$$

where we put $s = nt$ and $S = n$, we conclude that the probability to go out of $\Gamma_{S,\delta S}$ is

$$\mathbb{P}\bigg(\bigg\{\frac{\lambda}{\mu}s - \delta S < Q^+(s) < \frac{\lambda}{\mu}s + \delta S \;\; \forall\, s \in [0,S]\bigg\}^c\bigg) = e^{-S\,\hat{I}_Q((\Gamma_{1,\delta})^c)\,[1+o(1)]}, \qquad S \to \infty. \tag{A.17}$$

Because I_Q is convex, to compute $\hat{I}_Q((\Gamma_{1,\delta})^c)$ it suffices to minimise over the linear paths. The minimiser turns out to be one of the two linear paths that go from the origin $(0,0)$ to $(1, \frac{\lambda}{\mu} \pm \delta)$, i.e., $\gamma^*(t) = kt$ with $k = \frac{1}{\mu}(\lambda \pm \delta\mu)$. By construction, $\hat{I}_Q((\Gamma_{1,\delta})^c) = \tilde{I}_Q(\gamma^*) = \inf_{\eta \in L_\infty([0,1])} I(\gamma^*, \eta)$, where

$$I(\gamma^*, \eta) = \int_0^1 \Lambda_Q^*\left(\frac{d\gamma^*(t)}{d\eta(t)}\right) d\eta(t) + \int_0^1 \Lambda_N^*(\dot{\eta}(t))\, dt. \tag{A.18}$$

We want to minimise the sum over all paths η such that $\eta(0) = 0$. Both integrals are convex as a function of γ^* and η, hence they are minimised by linear paths. Our choice of $\gamma^*(t) = kt$ is linear, so we set $\eta(t) = ct$ with some constant $c > 0$. We can then write

$$\begin{aligned} I(\gamma^*, \eta) &= \int_0^{\eta(1)} \Lambda_Q^*\left(\frac{d\gamma^*(t)}{cdt}\right) cdt + \int_0^1 \Lambda_N^*(\dot{\eta}(t))\, dt \\ &= \int_0^c \Lambda_Q^*\Big(\frac{k}{c}\Big)\, cdt + \int_0^1 \Lambda_N^*(c)\, dt \\ &= c\left[\frac{k}{c}\mu - 1 - \log\left(\frac{k\mu}{c}\right)\right] + c\log\left(\frac{c}{\lambda}\right) - c + \lambda. \end{aligned} \tag{A.19}$$

The value of c that minimises the right-hand side is $c = \sqrt{\lambda k \mu}$. Substituting this into the formula above, we get

$$K_\delta = \tilde{I}_Q(\gamma^*) = k\mu + \lambda - 2\sqrt{\lambda k \mu} = (\lambda + \delta\mu) + \lambda - 2\sqrt{\lambda(\lambda + \delta\mu)}. \tag{A.20}$$

Note that $K_\delta > 0$ for all $\delta > 0$ and $\lim_{\delta \downarrow 0} K_\delta = 0$. This proves the claim. □

B Appendix: the output process

The main goal of this appendix is to prove Proposition 2.5 in Section 2. In Section B.1 we show a lower bound for the output process for the nodes in U, in a setting where the nodes in U are not influenced by the nodes in V. We study the system up to time T_U.

In Section B.2 we show that, until the pre-transition time, the system in the internal model behaves actually as we described.

B.1 The output process in the isolated model

Recall that in the isolated model a node in U keeps activating and deactivating independently of the nodes in V, until its queue length hits zero. We again consider a single queue for a node in U and for simplicity suppress its index. In order to show that the output process $t \mapsto Q^-(t) = cT(t)$ when properly rescaled is close to a deterministic path with high probability, we will provide a lower bound for the output process. The upper bound $Q^-(t) \leq ct$ is trivial and holds for any $t \geq 0$, by the definition of output process.

Lemma B.1 (Auxiliary output process). *For all $\delta > 0$ and T large:*

(i) With high probability the process

$$Q^{\mathrm{LB},T}(t) = \gamma_U r + \rho_U t - \delta T - ct, \qquad t \in [0,T], \tag{B.1}$$

is a lower bound for the actual queue length process $(Q(t))_{t\in[0,T]}$.

(ii) The probability of the lower bound in (i) failing is

$$\frac{1}{2} e^{-K_\delta T\,[1+o(1)]}, \qquad T \to \infty, \tag{B.2}$$

with $K_\delta = (\lambda + \delta\mu) + \lambda - 2\sqrt{\lambda(\lambda+\delta\mu)}$.

Proof. (i) By Proposition 2.1, with high probability we have $Q^+(t) \geq \rho_U t - \delta T$ for any $\delta > 0$. Trivially, $Q^-(t) \leq ct$. It is therefore immediate that with high probability $Q^{\mathrm{LB},T}(t) \leq Q(t)$.
(ii) The exponentially small probability of $Q^+(t)$ going below the lower bound is half of the probability given by Proposition 2.1, i.e.,

$$\frac{1}{2} e^{-K_\delta T\,[1+o(1)]}, \qquad T \to \infty, \tag{B.3}$$

with $K_\delta = (\lambda + \delta\mu) + \lambda - 2\sqrt{\lambda(\lambda+\delta\mu)}$. □

We study the system up to time T_U defined in Definition 2.3, the expected time a single node queue takes to hit zero. We will prove in Appendix B.2 that the pre-transition time in the internal model with high probability coincides in distribution with the pre-transition time in the isolated model, which occurs with high probability before T_U. Hence it is enough to study the isolated model up to T_U.

Definition B.2 (Auxiliary times). We next define two times that will be useful in our analysis.

(T_U^*) Consider the auxiliary output process $Q^{\mathrm{LB},T_U}(t)$ up to time T_U. We define T_U^* as the time needed for the process to hit zero, i.e.,

$$T_U^* = T_U^*(r) = \frac{\gamma_u r - \delta T_U}{c - \rho_U} = \frac{\gamma_u - \delta\alpha}{c - \rho_U} r = \alpha' r \asymp r, \tag{B.4}$$

with $\alpha' = \frac{\gamma_u - \delta\alpha}{c-\rho_U}$. The difference $T_U - T_U^* = \frac{\delta\alpha}{c-\rho_U} r$ is of order r. The queue length at time T_U^* is not zero, but still of order r.

(T_U^{**}) We define a smaller time T_U^{**} such that, not only $Q(T_U^{**}) \asymp r$, but also $Q^{\mathrm{LB},T_U}(T_U^{**}) \asymp r$, i.e.,

$$T_U^{**} = T_U^{**}(r) = T_U - 2(T_U - T_U^*) = \left(\frac{\gamma_U - 2\delta\alpha}{c - \rho_U}\right) r = \alpha'' r \asymp r, \tag{B.5}$$

with $\alpha'' = \frac{\gamma_U - 2\delta\alpha}{c-\rho_U}$.

Definition B.3 (Inactivity process). Define the *inactivity process* by setting $W(t) = t - T(t)$, which equals the total amount of inactivity time until time t. □

Recall that the service process $t \mapsto Q^-(t)$ with $Q^-(0) = 0$ is an alternating sequence of activity periods and inactivity periods. The activity periods Z_i, $i \in \mathbb{N}$, are i.i.d. exponential random variables with mean 1. The inactivity periods W_m, $m \in \mathbb{N}$, are exponential random variables with a mean that depends on the actual queue length at the time when each of these periods starts, namely, if $W_m = \left[t_m^{(i)}, t_m^{(f)}\right]$, then $W_m = \text{Exp}(g_U(Q(t_m^{(i)})) + O(\frac{1}{r}))$. The queue length during this inactivity intervals is actually increasing, but we are considering very small intervals, whose lengths are of order $\frac{1}{r}$, so that the queue length does not change much and the error is then $O(\frac{1}{r})$.

To state our lower bound on the output process, we need the following two lemmas.

Lemma B.4 (Upper bound on number of activity periods). *Let $M(t)$ be the number of activity periods that end before time t. Then, for all $\epsilon_1 > 0$ and r large:*

(i) With high probability

$$M(T_U^{**}) \leq (1 + \epsilon_1) T_U^{**}. \tag{B.6}$$

(ii) The probability of the upper bound in (i) failing is

$$e^{-K_1 r\,[1+o(1)]} + \frac{1}{2} e^{-K_\delta \alpha r\,[1+o(1)]}, \qquad r \to \infty, \tag{B.7}$$

with $K_1 = \alpha'' \frac{\epsilon_1 - \log(1+\epsilon_1)}{1+\epsilon_1}$, K_δ as in Lemma B.1

Proof. (i) Note that $M(T_U^{**})$ counts the number of activity periods before time T_U^{**}, each of which has an average duration 1. Since activity periods alternate with inactivity periods, we expect $M(T_U^{**})$ to be less than T_U^{**}. Assume now, for small $\epsilon_1 > 0$, that $M(T_U^{**}) > (1+\epsilon_1) T_U^{**}$, which means that the number of activity periods before T_U^{**} is greater than the length of the interval $[0, T_U^{**}]$. This implies that the average length of each activity period before time T_U^{**} is strictly less than 1, namely, that $\frac{1}{T_U^{**}} \sum_{i=1}^{T_U^{**}} Z_i \leq \frac{1}{1+\epsilon_1}$. According to Cramér's theorem, we can compute the probability of this last event as

$$\mathbb{P}\left(\sum_{i=1}^{T_U^{**}} Z_i \leq \left(\frac{1}{1+\epsilon_1} \right) T_U^{**} \right) = e^{-T_U^{**} I\left(\frac{1}{1+\epsilon_1}\right)\,[1+o(1)]}, \qquad r \to \infty, \tag{B.8}$$

with rate function $I(x) = x \log(x) - x + 1$. Therefore, it occurs with exponentially small probability. Hence $M(T_U^{**}) > (1 + \epsilon_1) T_U^{**}$ must also occur with a probability which is also exponentially small. With high probability we then have that

$$M(T_U^{**}) \leq (1 + \epsilon_1) T_U^{**}. \tag{B.9}$$

Recall that $T_U^{**} = \alpha'' r$. The counting of alternating activity and inactivity periods gets affected when the queue length hits zero, since then the node is forced to switch itself off and the lengths of the activity periods are not regular anymore. Since at time T_U^{**} with high probability the queue length is still of order r, the probability that it hits zero at any time in the interval $[0, T_U^{**}]$ is very small, since this event would imply the node to have a queue length that is below the lower bound, $Q(T_U^{**}) \leq Q^{\text{LB},T_U}(T_U^{**}) = \gamma_U r + \rho_U T_U^{**} - \delta T_U - c T_U^{**}$, which happens with an exponentially small probability by Lemma B.1.

(ii) We can write

$$\begin{aligned}\mathbb{P}(M(T_U^{**}) > (1+\epsilon_1)T_U^{**}) &\le e^{-T_U^{**} I\left(\frac{1}{1+\epsilon_1}\right)[1+o(1)]} + \frac{1}{2}e^{-K_\delta T_U\,[1+o(1)]} \\ &= e^{-K_1 r\,[1+o(1)]} + \frac{1}{2}e^{-K_\delta \alpha r\,[1+o(1)]}, \qquad r \to \infty,\end{aligned} \tag{B.10}$$

with $K_1 = \alpha'' I\big(\frac{1}{1+\epsilon_1}\big) = \alpha'' \frac{\epsilon_1 - \log(1+\epsilon_1)}{1+\epsilon_1}$, K_δ as in Lemma B.1. □

Lemma B.5 (Upper bound on inactivity process). *For all $\delta, \epsilon_1, \epsilon_2 > 0$ small and r large:*

(i) With high probability

$$W(T_U^{**}) \le \epsilon_2 r. \tag{B.11}$$

(ii) The probability of upper bound in (i) failing is

$$\begin{aligned}\mathbb{P}\big(W(T_U^{**}) \le \epsilon_3 T_U^{**}\big) &\le e^{-K_\delta \alpha r\,[1+o(1)]} + e^{-K_1 r\,[1+o(1)]} \\ &+ e^{-\left(K_2 r + K_3 \frac{r}{g_U(r)} + K_4 r \log g_U(r)\right)[1+o(1)]}, \qquad r \to \infty,\end{aligned} \tag{B.12}$$

with $K_2 = \alpha''(1+\epsilon_1)\big(-1-\log\big(\frac{\epsilon_2}{\alpha''(1+\epsilon_1)}\big)\big), K_3 = \epsilon_2, K_4 = \alpha''(1+\epsilon_1)$.

Proof. (i) Since $M(t)$ counts the number of activity periods, and we start with an active node (in the starting configuration u all nodes in U are active), we have

$$W(T_U^{**}) \le \sum_{m=1}^{M(T_U^{**})} W_m \le \sum_{m=1}^{M(T_U^{**})} \hat{W}_m, \tag{B.13}$$

where $\hat{W}_m$ are i.i.d. exponential random variables with rate $g_U(Q^{\mathrm{LB},T_U}(T_U^{***}))$, and T_U^{***} is the starting point of the last inactivity period before time T_U^{**}. By the construction of T_U^{**}, we know that $Q^{\mathrm{LB},T_U}(T_U^{***})$ is of order r. The last inactivity period is expected to be longer than the previous ones, since the rates depend on the actual queue length, which is decreasing in time. To make the inactivity periods $\hat{W}_m$ longer, we consider the lower bound $Q^{\mathrm{LB},T_U}(t)$ for the actual queue length given in Lemma B.1.

By Lemma B.4, with high probability $M(T_U^{**}) \le (1+\epsilon_1)T_U^{**}$, and so

$$W(T_U^{**}) \le \sum_{m=1}^{M(T_U^{**})} \hat{W}_m \le \sum_{m=1}^{(1+\epsilon_1)T_U^{**}} \hat{W}_m. \tag{B.14}$$

Define $n = [(1+\epsilon_1)T_U^{**}]$. By Cramér's theorem, for small $\epsilon_3 > 0$,

$$\begin{aligned}\mathbb{P}\left(\sum_{m=1}^{(1+\epsilon_1)T_U^{**}} \hat{W}_m \ge \epsilon_3 T_U^{**}\right) &\le \mathbb{P}\left(\sum_{m=1}^{n} \hat{W}_m \ge \frac{\epsilon_3}{1+\epsilon_1} n\right) \\ &= e^{-nI\left(\frac{\epsilon_3}{1+\epsilon_1}\right)[1+o(1)]} = e^{-T_U^{**}(1+\epsilon_1)I\left(\frac{\epsilon_3}{1+\epsilon_1}\right)[1+o(1)]}, \qquad n \to \infty,\end{aligned} \tag{B.15}$$

where I is the rate function given by

$$I(x) = \frac{x}{g_U(Q^{\mathrm{LB},T_U}(T_U^{***}))} - 1 - \log x + \log g_U(Q^{\mathrm{LB},T_U}(T_U^{***})). \tag{B.16}$$

In order to apply Cramér's theorem, take $\epsilon_3 > (1+\epsilon_1)/g_U(Q^{\mathrm{LB},T_U}(T_U^{***})) \asymp 1/g_U(r)$ arbitrarily small. Combining (B.14)–(B.15), we obtain that with high probability

$$W(T_U^{**}) \leq \epsilon_3 T_U^{**} = \epsilon_3 \alpha'' r = \epsilon_2 r, \tag{B.17}$$

where $\epsilon_2 = \epsilon_3\alpha''$ can be taken arbitrarily small.
(ii) For large r,

$$\begin{aligned}\mathbb{P}\left(\sum_{m=1}^{(1+\epsilon_1)T_U^{**}} \hat{W}_m \geq \epsilon_3 T_U^{**}\right) &= e^{-T_U^{**}(1+\epsilon_1)I\left(\frac{\epsilon_3}{1+\epsilon_1}\right)[1+o(1)]}\\ &\asymp e^{-\alpha'' r(1+\epsilon_1)\left(\frac{\epsilon_3}{(1+\epsilon_1)g_U(r)}-1-\log\left(\frac{\epsilon_3}{1+\epsilon_1}\right)+\log g_U(r)\right)[1+o(1)]}\\ &= e^{-\left[\alpha''(1+\epsilon_1)\left(-1-\log\left(\frac{\epsilon_3}{1+\epsilon_1}\right)\right)r+\epsilon_3\alpha''\frac{r}{g_U(r)}+\alpha''(1+\epsilon_1)r\log(g_U(r))\right][1+o(1)]}\\ &= e^{-\left(K_2 r+K_3\frac{r}{g_U(r)}+K_4 r\log g_U(r)\right)[1+o(1)]}, \qquad r\to\infty,\end{aligned} \tag{B.18}$$

where $K_2 = \alpha''(1+\epsilon_1)\left(-1-\log\left(\frac{\epsilon_2}{\alpha''(1+\epsilon_1)}\right)\right)$, $K_3 = \epsilon_3\alpha'' = \epsilon_2, K_4 = \alpha''(1+\epsilon_1)$. We also have to consider the probabilities computed in (B.2) and (B.7). Hence we have

$$\begin{aligned}\mathbb{P}\big(W(T_U^{**}) \leq \epsilon_3 T_U^{**}\big) &\leq e^{-K_\delta \alpha r\,[1+o(1)]} + e^{-K_1 r\,[1+o(1)]}\\ &+ e^{-\left(K_2 r+K_3\frac{r}{g_U(r)}+K_4 r\log g_U(r)\right)[1+o(1)]}, \qquad r\to\infty.\end{aligned} \tag{B.19}$$

□

We are now in a position to prove Proposition 2.5.

Proof. The equation $Q^-(t) \geq ct - \epsilon r$ can be read as $T(t) \geq t - \frac{\epsilon r}{c}$. This is equivalent to saying that $W(t) \leq \frac{\epsilon r}{c}$ for all $t \in [0, T_U]$. By taking $\epsilon_2 = \frac{1}{3}\frac{\epsilon}{c}$ in Lemma B.5, we know that, for all $t \in [0, T_U^{**}]$, $W(t) \leq W(T_U^{**}) \leq \frac{1}{3}\frac{\epsilon r}{c}$. Moreover, in the interval $[T_U^{**}, T_U]$, the cumulative amount of inactivity time is trivially bounded from above by the length of the interval, which is $\frac{2\delta r}{c-\rho_U} \leq \frac{2}{3}\frac{\epsilon r}{c}$, and ϵ can be taken arbitrarily small, since δ can be taken arbitrarily small. Putting the two bounds together, we find that with high probability

$$W(t) \leq \epsilon_2 r + \frac{2\delta r}{c-\rho_U} \leq \frac{1}{3}\frac{\epsilon r}{c} + \frac{2}{3}\frac{\epsilon r}{c} = \frac{\epsilon r}{c}, \qquad t \in [0, T_U]. \tag{B.20}$$

It is immediate to see that the probability of this not happening is given by (B.12). □

The above lower bound $Q^-(t) \geq ct - \epsilon r$ and the trivial upper bound $Q^-(t) \leq ct$ imply that with high probability the output process $Q^-(t)$ stays close to the path $c \mapsto ct$ by sending ϵ to zero. In other words, the node stays almost always active all the time before T_U.

B.2 The output process in the internal model

In this section we want to couple the isolated model and the internal model and show that they have identical behaviour in the time interval $[0, \bar{\tau}_v^{\rm int}]$. Hence it follows that the output process in the internal model for nodes in U actually behaves as in the isolated model described in Section B.1, until the pre-transition time.

Proposition B.6. *Let $X_i^{\rm int}(t)$ and $X_i^{\rm iso}(t)$ denote the activity state of a node i at time t in the internal and the isolated model, respectively. Then*

$$\lim_{r\to\infty} \mathbb{P}_u\big(X_i^{\rm int}(t) = X_i^{\rm iso}(t) \ \ \forall\, i \in U \cup V \ \forall\, t \in [0, \bar{\tau}_v^{\rm int}]\big) = 1. \tag{B.21}$$

Consequently, with high probability the pre-transition times in the internal and the isolated model coincide, i.e.,

$$\lim_{r\to\infty} \mathbb{P}_u(\bar{\tau}_v^{\rm int} = \bar{\tau}_v^{\rm iso}) = 1. \tag{B.22}$$

Proof. In Section B.1 we determined upper and lower bounds for the output process for nodes in U in the isolated model up to time T_U. Assume now that $\bar{\tau}_v^{\rm int} \leq T_U$. When considering the internal model and the set of nodes in V, we immediately see that these bounds are not true for the whole interval $[0, T_U]$, since at time $\bar{\tau}_v^{\rm int}$ already some nodes in V start to activate and influence the behaviour of nodes in U.

If we look at the interval $[0, \bar{\tau}_v^{\rm int}]$, then we note that the queue length process for a node $i \in U$ is not affected by nodes in V, and so it behaves in exactly the same way as if the node were isolated. The activation and deactivation Poisson clocks at node i are synchronized, and are ticking at the same time in the isolated model and in the internal model, so that $X_i^{\rm int}(t) = X_i^{\rm iso}(t)$. Moreover, the activity states of nodes in V are always equal to 0 in both models. Hence we conclude that the activity states of every node coincide up to the pre-transition time $\bar{\tau}_v^{\rm int}$. Consequently, the pre-transition times in the internal and the isolated model coincide on the event $\{\bar{\tau}_v^{\rm int} \leq T_U\}$, which can then be written as the event $\{\bar{\tau}_v^{\rm iso} \leq T_U\}$. For the latter we know that it has a high probability when $r \to \infty$ (see proof of Proposition 3.7 in Section 3). □

References

[1] S.C. Borst, F. den Hollander, F.R. Nardi, S. Taati, Hitting-time asymptotics in bipartite hard-core interaction models with time-varying rates. In preparation.

[2] N. Bouman, S.C. Borst, J.S.H. van Leeuwaarden, Delay performance in random-access networks, Queueing Systems 77 (2014) 211–242.

[3] A. Dembo, O. Zeitouni, *Large Deviations Techniques and Applications* (2nd. Ed.), Springer, 1998.

[4] J. Ghaderi, S.C. Borst, P.A. Whiting, Queue-based random-access algorithms: fluid limits and stability issues, Stochastic Systems 4 (2014) 81–156.

[5] J. Ghaderi, R. Srikant, On the design of efficient CSMA algorithms for wireless networks. In: Proc. CDC 2010 Conf. 954–959.

[6] F. den Hollander, F.R. Nardi, S. Taati, Metastability of hard-core dynamics on bipartite graphs [arXiv:1710.10232], Preprint 2017.

[7] L. Jiang, D. Shah, J. Shin, J. Walrand, Distributed random access algorithm: scheduling and congestion control, IEEE Trans. Inf. Theory 56 (2010) 6182–6207.

[8] F.R. Nardi, A. Zocca, S.C. Borst, Hitting time asymptotics for hard-core interactions on grids, J. Stat. Phys. 162 (2016) 522–576.

[9] S. Rajagopalan, D. Shah, J. Shin, Network adiabatic theorem: An efficient randomized protocol for contention resolution, ACM SIGMETRICS Perf. Eval. Rev. 37 (2009) 133–144.

[10] D. Shah, J. Shin, Randomized scheduling algorithms for queueing networks, Ann. Appl. Prob. 22 (2012) 128–171.