The Annals of Applied Probability
1994, Vol. 4, No. 4, 1161–1176

PHASE TRANSITION IN A LOAD SHARING LOSS MODEL

By Vadim Malyshev and Philippe Robert

Moscow State University and INRIA

In this paper we analyze the following loss network: When a customer arrives at a node of the network, it is served by this node if the node is not occupied; otherwise it is transmitted to some empty node where it will be served at a different rate. For the simplest systems of this type with a very large number of nodes and with global sharing, we show the existence of second order phase transitions and present explicit formulas for probability characteristics. For local sharing, we study the case of an infinite network and present some convergence results. Formulas for small and large loads are obtained.

1. Introduction. We consider a finite set of Λ of nodes of a (large) network such that for $x \in \Lambda$ there is a Poisson arrival stream, called the x-stream. These streams are independent and have a constant rate λ. A customer of the x-stream arriving at time t is served at x if this node is empty. Otherwise, depending on the assignment policy chosen, the customer can be either lost or dispatched to another empty node. The service of a customer from the x-stream and served at node y is exponentially distributed with parameter $\mu(x, y) \neq 0$. If all the possible servers are busy, then the customer is lost. Let $\mathscr{R}$ denote the policy chosen to assign (or discard) the customers who arrive at an occupied node. Let us give some examples of possible choices for the policy $\mathscr{R}$.

Independent policy. A customer who finds his server busy is discarded and $\mu(x, x) = \mu$. This means that we have independent nodes. The loss probability here is

$$\frac{\lambda}{\mu + \lambda}, \tag{1}$$

which is the simplest case of Erlang's formula.

Local sharing—processes with local interaction. If all Λ's belong to some countable metric space $\mathscr{L}$ with the distance $\rho(x, y)$, local sharing means that a customer arriving at x is lost if $\rho(x, y) > d$ for all empty sites y. Then we have a process with local interaction in the standard sense. A custome: can be served only in the vicinity of its arrival node.

Global sharing. If all nodes are available for any stream, then we say that sharing is global. In particular a customer is lost only if all the servers are busy. In this case, for most of our results, we do not need to specify the particular rule used to allocate the customers who do not find an empty site.

Received September 1992; revised February 1994.

AMS 1991 *subject classifications.* 60K25, 60K20.

Key words and phrases. Loss network, stationary measures, loss probabilities, thermodynamic limit.

If we think of the nodes as the routes of a network, using a policy $\mathscr{R}$ amounts to saying that a customer who finds a busy route may be rerouted, that is, another route may be offered to him. Usually rerouting implies an increased occupancy of the network; instead of using one node, the rerouted customer will occupy at least two nodes (alternative route) with the same service rate (see [3], Section 4.3, and the references contained therein for a complete survey of these problems). Here, the situation is quite different: The rerouted customer will occupy another route but will be served at a different rate (smaller in general). The cost of a rerouted customer is a longer service for this customer, hence, a larger load for the network.

In this paper we are interested in two kinds of behavior of this network:

1. The convergence to a thermodynamic limit, that is, the limit of the stationary measures when the size of the network tends to infinity.
2. The behavior after the thermodynamic limit, that is, the analysis of a network with an infinite number of nodes.

Throughout this paper, we assume that there are two possible values for the service rates of the customers who are not discarded: one for the customers who arrive in an empty site (the directly routed customers) and one for the others (alternatively routed customers). In Section 2 we analyze this network with the global sharing policy. When the size of the network goes to infinity, we prove convergence results for the stationary number of directly (resp. alternatively) routed customers. In particular, for some values of the parameters, it is shown that when the size of the network gets large, the alternatively routed customers will saturate the network. However, even in this case, we prove that the stationary loss probability converges to a quantity which is less than 1 and, moreover, the number of idle nodes converges in distribution to a geometrically distributed random variable. As a consequence, we obtain the condition under which rerouting is worthwhile to minimize the loss probability of this network. In Section 3 we consider an infinite network on $\mathbb{Z}$ for which is a customer is assigned to the first empty place on the right of its arriving node (the arrival node included). This process may be described as an infinite particle system with a long range interaction. We give the conditions under which it is possible to construct such a process and analyze its asymptotic behavior when time goes to infinity. Section 4 is devoted to the local sharing policy which is the model analyzed in Section 3, but with a local interaction: A customer is lost if it cannot find an empty place among the d nodes on the right of its arrival node. For this process we derive asymptotic expansions of the loss probabilities, taking the arrival rate and the service rates as variables. Finally we conclude with some open problems concerning these networks.

We will denote by $\xi_t^\Lambda(x)$, $x \in \Lambda, 0 \leq t < +\infty$, the corresponding continuous time homogeneous Markov process with the state space $\mathscr{A}^\Lambda$, where $\mathscr{A}$ consists of 0 and of the set $\{1, 2\}$ if there are two possible values for $\mu(x, y)$. So we will have $\xi_t(y) = 0$ if the node y is empty at time t and $\xi_t(y) = 1$ [resp. 2] if the node y is occupied by a directly [resp. alternatively] routed customer at time t. We put

$$p_t^{\Lambda}(s_{\Lambda}) = P^{\Lambda}\big(\xi_t^{\Lambda}(x) = s_x, x \in \Lambda\big) \quad \text{with } s_{\Lambda} = (s_x)_{x \in \Lambda},$$

where s_x is equal to 0, 1 or 2. Let π^{Λ} be the corresponding stationary probability. For any $A \subset \Lambda$ we define the correlation functions $p_t^{\Lambda}(s_{\mathrm{A}}) = p_t^{\Lambda}(A; s_A) = P^{\Lambda}(\xi_t(x) = s_x,\ x \in A)$. We denote by q_{Λ} the stationary loss probability of an arriving customer for the network defined by μ_1, μ_2 and by $\mathscr{R}$.

2. Global sharing. Let $N = |\Lambda|$ be the number of nodes of the network and η_t^{Λ} (resp. ζ_t^{Λ}) the number of directly routed customers (resp. alternatively routed customers) at time t in Λ, and let η_{Λ} (resp. ζ_{Λ}) be the corresponding random variables in the stationary state. Let us begin with a well known result (see [7]).

THEOREM 1. *Under global sharing, if $\mu_1 = \mu_2$, then, in distribution,*

$$\frac{1}{N}(\eta_{\Lambda} + \zeta_{\Lambda}) \xrightarrow{N \to +\infty} p \equiv \begin{cases} 1, & \text{if } \rho > 1, \\ \rho, & \text{if } \rho \le 1, \end{cases} \tag{2}$$

with $\rho = \lambda/\mu_1$. Moreover, the following limit exists:

$$q_{\Lambda} \xrightarrow{N \to +\infty} q \equiv \begin{cases} 1 - \dfrac{1}{\rho}, & \text{if } \rho > 1, \\ 0, & \text{if } \rho \le 1. \end{cases} \tag{3}$$

So p is the proportion of the busy servers. This means that q (resp. p) as a function of ρ, is continuous but not differentiable at the critical point $\rho_{\mathrm{cr}} = 1$. In other words, we have a phase transition of the second kind.

PROOF. In this case it is easy to see that the number of occupied buffers at time t, $\eta_t^{\Lambda} + \zeta_t^{\Lambda}$, is a Markov chain. The expression for the stationary probability of this Markov chain is given by

$$\pi_{\Lambda}\{j\} = \frac{\theta_j}{1 + \theta_1 + \cdots + \theta_N}, \qquad \theta_j = \left(\frac{\lambda}{\mu_1}\right)^j \frac{N^j}{j!}. \tag{4}$$

It is straightforward to get (2) and also (3) [as $q = \pi\{N\}$ from (4)]. □

We get a similar picture for the case $\mu_1 \neq \mu_2$.

THEOREM 2. *Under global sharing, if $\mu_1 \neq \mu_2$, then, in distribution,*

$$(5)\quad \frac{1}{N}(\eta_\Lambda, \zeta_\Lambda) \xrightarrow{N\to+\infty} \begin{cases} \left(\dfrac{\lambda}{\mu_1+\lambda+\lambda^2/(\mu_2-\lambda)}, \dfrac{\lambda^2}{(\mu_2-\lambda)(\mu_1+\lambda)+\lambda^2}\right), & \\ & \text{if } \mu_2 > \lambda, \\ (0,1), & \text{if } \mu_2 \leq \lambda, \end{cases}$$

and when $\lambda > \mu_2$, the number of idle servers converges in distribution to a geometric random variable with parameter μ_2/λ. In particular, the loss probability converges as $N \to +\infty$ and

$$(6)\qquad \lim_{N\to+\infty} q_\Lambda = \begin{cases} 0, & \mu_2 \geq \lambda, \\ 1-\mu_2/\lambda, & \mu_2 \leq \lambda \end{cases}$$

PROOF. It is easy to verify that the pair $(\eta_t^\Lambda/N, \zeta_t^\Lambda/N)$ is a Markov chain with the state space $[0,1]^2$ and values in $\mathscr{A}_N = \{(k/N, l/N) \mid k+l \leq N\}$. Its generator Ω_N is defined by

$$(7)\qquad \begin{aligned} \Omega_N(f)(x,y) &= \lambda N(1-x-y)(f(x+1/N,y)-f(x,y))1_{\{x+y<1\}} \\ &\quad + \lambda N(x+y)(f(x,y+1/N)-f(x,y))1_{\{x+y<1\}} \\ &\quad + \mu_1 Nx(f(x-1/N,y)-f(x,y))1_{\{x>0\}} \\ &\quad + \mu_2 Ny(f(x,y-1/N)-f(x,y))1_{\{y>0\}}. \end{aligned}$$

Let π_N be the stationary probability distribution of this Markov chain and

$$(8)\qquad \begin{aligned} x_0 &= \frac{\lambda}{\mu_1+\lambda+\lambda^2/(\mu_2-\lambda)}, & y_0 &= \frac{\lambda}{\mu_2-\lambda}x_0 \quad \text{for } \mu_2 > \lambda, \\ x_0 &= 0, & y_0 &= 1, \quad \text{for } \mu_2 \leq \lambda. \end{aligned}$$

We have to prove that the sequence $(\pi_N)_{N\in\mathbb{N}}$ converges weakly to $\delta_{(x_0,y_0)}$ (with δ_a the Dirac measure at a).

Define Ω by

$$\Omega(f)(x,y) = (\lambda(1-x-y)-\mu_1 x)\frac{\partial f}{\partial x} + (\lambda(x+y)-\mu_2 y)\frac{\partial f}{\partial y}, \qquad x,y \in [0,1].$$

For any polynomial function f there exists some real K such that for x, y in the interior of $\mathscr{A} = \{(x,y) \in [0,1]^2 \mid x+y \leq 1\}$,

$$(9)\qquad |\Omega(f)(x,y) - \Omega_N(f)(x,y)| \leq \frac{K}{N}.$$

In the case $\lambda < \mu_2$ the point (x_0, y_0) is in the interior of $\mathscr{A}$. We can write $\Omega(f)(x,y)$ as

$$(-(\lambda+\mu_1)(x-x_0)-\lambda(y-y_0))\frac{\partial f}{\partial x} + (\lambda(x-x_0)+(\lambda-\mu_2)(y-y_0))\frac{\partial f}{\partial y}.$$

If $f(x, y) = (x - x_0)^2/2 + (y - y_0)^2/2$, then for $x, y \in \mathscr{A}$,

$$\begin{aligned}\Omega(f)(x,y) &= -(\lambda+\mu_1)(x-x_0)^2 + (\lambda-\mu_2)(y-y_0)^2 \\ &\leq -2\min\{\mu_2-\lambda, \lambda+\mu_1\} f(x,y).\end{aligned} \tag{10}$$

Let us forget for a moment the boundary conditions of (7).

The equilibrium equations give

$$\int \Omega_N(f)(x,y)\pi_N(dx,dy) = 0, \qquad N \geq 1. \tag{11}$$

Since the π_N, $N \geq 1$, are measures on the compact $[0,1]^2$, one can extract a convergent subsequence $\pi_{N_i} \to_{i \to +\infty} \pi$. By (9), (11) and the continuity of f we can write that

$$0 = \lim_{i \to +\infty} \int \Omega_{N_i}(f)(x,y)\pi_{N_i}(dx,dy) = \int \Omega(f)(x,y)\pi(dx,dy).$$

Since $\Omega(f)(x,y) < 0$ except at $(x,y) = (x_0, y_0)$, we have $\pi = \delta_{(x_0,y_0)}$. Hence $(\pi_N)_{N \in \mathbb{N}}$ converges weakly to $\delta_{(x_0,y_0)}$.

The boundary conditions: Let us consider the segment of the boundary $\Delta = \{(x,y) \in \mathscr{A} \mid x + y = 1\}$. On Δ we have

$$|\Omega_N(f)(x,y) - \Omega_\Delta(f)(x,y)| \leq \frac{K}{N}$$

with

$$\Omega_\Delta(f)(x,y) = -\mu_1 x \frac{\partial f}{\partial x} - \mu_2 y \frac{\partial f}{\partial y}.$$

If $f_\Delta(x,y) = x^2/2 + y^2/2$, then there exists a fixed α_0 such that $\Omega_\Delta(f_\Delta)(x,y) \leq -\alpha_0$ for $(x,y) \in \Delta$ and $\Omega(f_\Delta(x,y) \leq -\alpha_0$ for (x,y) in a strip S_ε of width ε around Δ for a sufficiently small ε. If we choose ε so that $\pi\{(x,y)| x + y = 1 - \varepsilon\} = 0$ [where π is the limit of some subsequence of $(\pi_N)_{N \in \mathbb{N}}$] and

$$g(x,y) = f_\Delta 1_{S_\varepsilon}(x,y) + f 1_{S_\varepsilon^c}(x,y), \tag{12}$$

then the previous proof can be carried out replacing f by g. The two other boundary conditions on the axes can be included in the same way to complete the proof.

In Figure 1, we represent the vector field associated with Ω_Δ on S_ε, $(-\mu_1 x, -\mu_2 y)$, and with Ω outside S_ε, $(\lambda(1 - x - y) - \mu_1 x, \lambda(x + y) - \mu_2 y)$.

The case $\lambda > \mu_2$ is analyzed in the same way. Here the function $f(x,y) = (\lambda - (\lambda + \mu_1)x - \lambda y)^2/2 + \alpha(1 - y)^2/2$ can be chosen as a Lyapounov function in the interior of $\mathscr{A}$ if α is sufficiently large. The vector field in this case is shown in Figure 2.

Convergence of the number of idle servers when $\lambda > \mu_2$: Let $Z_t = N - \eta_t^\Lambda - \zeta_t^\Lambda$ and let $Z_N = N - \eta_\Lambda - \zeta_\Lambda$ be the stationary version of Z. The station-

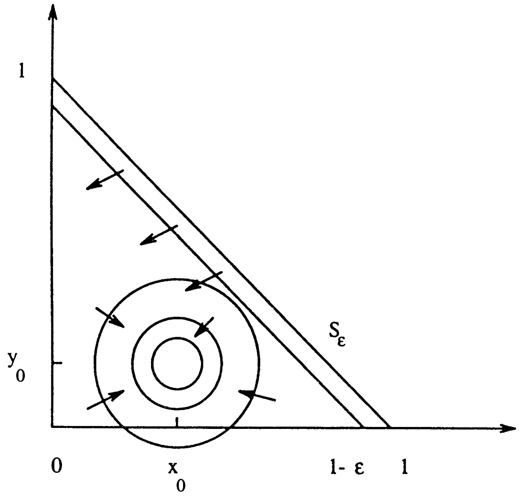

FIG. 1.

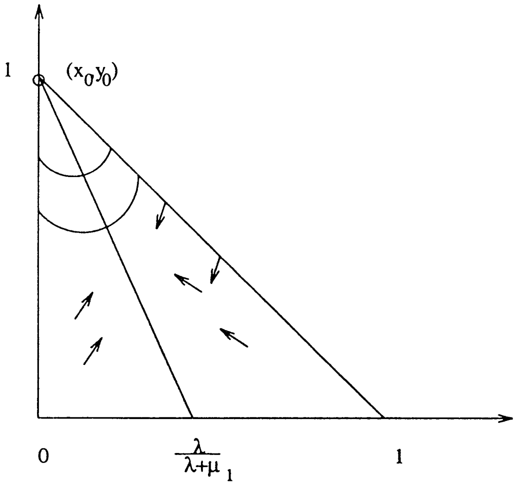

FIG. 2.

ary loss probability for the system with N nodes is simply $P(Z_N = 0)$. Using (7), for any function f, the equilibrium equations for Z_N can be written as

$$\begin{aligned}(13)\qquad &\int_{[0,1]^2} \lambda N(f(z-1) - f(z))1_{\{z>0\}}\pi_N(dx,dy)\\ &\quad+\int_{[0,1]^2}(\mu_1 Nx + \mu_2 Ny)(f(z+1) - f(z))1_{\{z<N\}}\pi_N(dx,dy) = 0,\end{aligned}$$

with $z = N(1 - x - y)$ in the expressions under the integrals. If $f(t) = (t - K)1_{\{t > K\}}$ and $\phi_N(K) = \pi_N(z \geq K) = P(Z_N \geq K)$, then (13) becomes, for $0 < K < N$,

$$\lambda\phi_N(K+1) = \mu_2\phi_N(K) + \int_{[0,1]^2}\big[(\mu_2 y + \mu_1 x)1_{\{K \leq z < N\}} - \mu_2 1\{z \geq K\}\big]\pi_N(dx, dy),$$

$$\lambda\phi_N(K+1) = \mu_2\phi_N(K) + \int_{[0,1]^2}(\mu_2(y-1) + \mu_1 x)1_{\{K \leq z\}}\pi_N(dx, dy).$$

Using our convergence result in the case $\lambda > \mu_2$, it is easy to see that the integral in the above expression converges to 0. If $\overline{\phi}(K) = \limsup_{N \to +\infty} \phi_N(K)$, we get $\overline{\phi}(K) \leq (\mu_2/\lambda)\overline{\phi}(K - 1)$; hence, $\overline{\phi}(K) \leq (\mu_2/\lambda)^k\, \overline{\phi}(0)$. In particular, this means that the sequence $(Z_N)_{N \in \mathbb{N}}$ is tight. Using again the above equation, we obtain that any limit point π satisfies $\pi\{k\} = (\mu_2/\lambda)^k(1 - \mu_2/\lambda)$.

Let us finish by the convergence of the loss probabilities $q_\Lambda = P(\eta_\Lambda + \zeta_\Lambda = N)$ when $N \to +\infty$. For $\lambda < \mu_2$, we have shown that $(\eta_\Lambda + \zeta_\Lambda)/N$ converges in distribution to $x_0 + y_0$, with the above notation. It is easily verified that in this case $x_0 + y_0 < 1$; hence, $\lim_{N \to +\infty} q_\Lambda = 0$. For the case $\lambda \geq \mu_2, q_\Lambda = P(Z_N = 0)$, and according to the convergence of Z_N, we obtain $\lim_{N \to +\infty} q_\Lambda = 1 - \mu_2/\lambda$. The proof of Theorem 2 is thus complete. □

REMARK 1. (a) The function g (partially) defined by (12) is usually called a Lyapounov function. It is surprising that a Lyapounov function approach can be used here for a problem which is completely different from the conventional type of stability problems with infinite buffers (see also [5] and [2], Section 4).

(b) It is important to understand that in the case $\lambda > \mu_2$, the limiting value of the loss probability could not be obtained even if we had complete information about the equilibrium measure after the thermodynamic limit. For the system with an infinite number of servers it follows from (5) that a customer is lost with probability 1, but the stationary probability of loss converges to $1 - \mu_2/\lambda$.

From Theorem 2 we can derive the following result: From (3) it follows that for $\mu_1 = \mu_2$, global sharing is better than the independent policy. The loss probability is always smaller for global sharing. However, for $\mu_2 < \mu_1$ this is the case iff

$$\mu_1 - \mu_2 - \frac{\mu_1\mu_2}{\lambda} < 0. \tag{14}$$

As a corollary we get the following result about all correlation functions in the case when the rule $\mathscr{R}$ is chosen in the following way: If an arriving customer finds its own station is occupied, then it is assigned to any empty station with equal probability.

THEOREM 3. *For a fixed Λ, under the above rule, the invariant measure is exchangeable, that is, invariant with respect to all the permutations of Λ and for any fixed subset of stations $A = \{1, \dots, s\}$,*

$$\lim_{N \to +\infty} \pi^{\Lambda}(s_A) = \prod_{x \in A} \pi(s_x),$$

with $\pi(s) = \lim_{N \to +\infty} \pi^{\Lambda}(\xi_{\Lambda}(1) = s)$.

PROOF. If we start the network with an exchangeable distribution, then according to the assignment policy, the property of exchangeability will be preserved at any time, and thus, at infinity. We deduce that π^{Λ} is exchangeable.

Let us now consider the case when $|A| = 2$:

$$\left(\sum_{i \in \Lambda} 1_{\{\xi_{\Lambda}(i) = s_1\}} \right) \left(\sum_{i \in \Lambda} 1_{\{\xi_{\Lambda}(i) = s_2\}} \right) = \sum_{i \in \Lambda} 1_{\{\xi_{\Lambda}(i) = s_1,\, \xi_{\Lambda}(i) = s_2\}} + \sum_{i \neq j \in \Lambda} 1_{\{\xi_{\Lambda}(i) = s_1,\, \xi_{\Lambda}(j) = s_2\}}.$$

From Theorem 2, the ratio of the left-hand side and N^2 converges in distribution to $\pi(s_1)\pi(s_2)$. By symmetry the expected value of the right-hand side is

$$N\pi^{\Lambda}(\xi_{\Lambda}(0) = s_1, \xi_{\Lambda}(0) = s_2) + N(N-1)\pi^{\Lambda}(\xi_{\Lambda}(0) = s_1, \xi_{\Lambda}(1) = s_2);$$

hence, $\lim_{|\Lambda| \to +\infty} \pi^{\Lambda}(s_A) = \pi(s_1)\pi(s_2)$. The proof of the general case is done in a similar way. □

3. Load sharing on an infinite network.

In this section we consider the one-dimensional network $\mathbb{Z}$ with the following rule to assign the customers: If the arrival site of a customer is occupied, then it is moved to the first empty place on its right. If the customer has not been moved (resp. has been moved), it is served with rate μ_1 (resp. μ_2) and the state of the site is 1 (directly routed customer) [resp. 2 (alternatively routed)]. (We set $\mu_0 = 0$ for convenience.) Because of the nonlocal interaction, an additional problem in this case is the construction of the associated Markov process with generator Ω:

$$\Omega(f)(\mathbf{x}) = \sum_{i \in \mathbb{Z}} \Big[\mu_{x_i}(f(\mathbf{x} - x_i\delta_i) - f(\mathbf{x})) + \lambda(f(\mathbf{x} + \delta_i) - f(\mathbf{x}))1\{x_i = 0\}$$
$$+ \lambda\big(f(\mathbf{x} + 2\delta_{\nu_i(\mathbf{x})}) - f(\mathbf{x})\big)1\{x_i \neq 0\}\Big], \qquad \mathbf{x} \in \{0,1,2\}^{\mathbb{Z}},$$

where f is some function belonging to the domain of Ω (see [4]), δ_i is the Dirac function at i and $\nu_i(\mathbf{x}) = \inf\{j > i : x_j = 0\}$. Let us denote by $\mathscr{E}$ the set of translation invariant and ergodic measures ϕ on $\{0,1,2\}^{\mathbb{Z}}$ such that $\phi\{\mathbf{x} : x_0 = 0\} > 0$. In the following, we will consider Markov processes $\mathbf{X}(t)$ with generator Ω and initial distribution in $\mathscr{E}$.

THEOREM 4. *For any $\phi \in \mathscr{E}$, there exists some $h > 0$ and a stochastic process $(\mathbf{X}(t))_{0 \le t \le h}$ such that:*

(i) *the distribution of $\mathbf{X}(0)$ is ϕ,*
(ii) *$\mathbf{X}(t)$ is a Markov process on $[0, h]$ with generator Ω.*

The method to construct our process is quite simple. We will show that the real line can be cut into noninteracting (random) intervals between 0 and some $h > 0$. For $i \in \mathbb{Z}$, we denote by N_i the Poisson process (with rate λ) of arrivals at site i and

$$\nu(k,h) = \inf\left\{ i \geq 0 \,\middle|\, \sum_{j=k}^{i+k} N_j[0,h] - 1_{\{x_j \neq 0\}} \leq 0 \right\},$$

where $\mathbf{X}_0 = (x_i)_{i \in \mathbb{Z}}$ is some random variable with distribution ϕ and $N_j[0,h]$ denotes the number of points of N_j between 0 and h.

Lemma 5. (a) *If* $\lambda h < \phi\{\mathbf{x} \mid x_0 = 0\} > 0$, *then* $\nu(k,h)$ *is finite a.s. for* $k \in \mathbb{Z}$.

(b) $H(i) = \sum_{-\infty}^{j=i-1} 1_{\{j+\nu(j,h) \geq i\}}$ *is almost surely finite and* $P(H(i) = 0) > 0$.

Proof. According to the ergodic theorem,

$$\frac{1}{i} \sum_{j=k}^{i+k} N_j[0,h] - 1_{\{x_j \neq 0\}} \xrightarrow{i \to +\infty} \lambda h - \phi\{\mathbf{x} \mid x_0 = 0\} < 0 \quad \text{a.s.}$$

Thus $\nu(k,h)$ is finite a.s. Using that $\{\sum_k^i N_j[0,h] - 1_{\{x_j \neq 0\}} \leq 0\} \subset \{k + \nu(k,h) \leq i\}$, we get that $H(i)$ is also a.s. finite. The maximal lemma of ergodic theory (see [1], Appendix 3) gives

$$\int_{\sup_{k \leq 0}\{\sum_{j=k}^{-1} N_j[0,h] - 1\{_{x_j \neq 0}\}\} > 0} \left(N_0[0,h] - 1_{\{x_0 \neq 0\}}\right) dP \geq 0,$$

with $\sum_0^{-1} = 0$. Using that $E(N_0[0,h] - 1_{\{x_0 \neq 0\}}) < 0$, we obtain

$$P\left(\sup_{k<0} \left\{ \sum_{j=k}^{-1} N_j[0,h] - 1_{\{x_j \neq 0\}} \right\} < 0 \right) > 0;$$

hence, $P(H(i) = 0) = P(H(0) = 0) > 0$. The lemma is proved. □

The statement of Lemma 5 (b) has the following consequence: Even if the customers never leave the network, there is a positive probability that none of the customers arriving during $[0,h]$ on the sites with negative index is assigned to a site with a positive index. This implies that on this event during $[0,h]$, one can discard all the sites on the left of 0. However, if we discard the sites $i < 0$, the process $(x_0(t), \ldots, x_n(t), \ldots)_{t \in [0,h]}$ is trivial to construct. Using that, almost surely, there is a nondecreasing sequence $(t_i)_{i \in \mathbb{Z}}$ such that $H(t_i) = 0$, we can thus construct our process on the interval $[0,h]$. Theorem 4 is thus proved. □

Remark 2. Because of our construction, it is easy to prove that for $0 \leq t \leq h$, the distribution of $\mathbf{X}(t)$ is in $\mathscr{S}$ (by verifying that the ergodic theorem holds, for example).

So, our process can be constructed as long as $P(x_0(t) = 0) > 0$. We will denote by † the (cemetery) state of $\mathbf{X}(t)$ whenever this condition is not satisfied and denote by τ the (deterministic) hitting time of †. We now turn to the asymptotic behavior of our process.

THEOREM 6. (a) *If $\mu_2 < \lambda$, the process almost surely dies, that is, for any $\phi \in \mathscr{E}$ there exists some t_0 such that $\mathbf{X}(t_0) = \dagger$.*

(b) *If $\lambda < \min\{\mu_1, \mu_2\}$, then the Markov process $\mathbf{X}(t)_{t \geq 0}$ is almost surely defined.*

This result is a consequence of the following lemma which is related in some sense to the fact that $(\eta_t, \zeta_t)_{t \in \mathbb{R}}$ is a Markov process (Section 2).

LEMMA 7. *If $P_i(t) = P(x_0(t) = i)$ for $i \in \{0, 1, 2\}$ and $t < \tau$, then*

$$\frac{\partial P_1(t)}{\partial t} = \lambda - \lambda P_2(t) - (\lambda + \mu_1) P_1(t),$$

$$\frac{\partial P_2(t)}{\partial t} = (\lambda - \mu_2) P_2(t) + \lambda P_1(t).$$

PROOF. Using the definition of the generator Ω, we have

$$P_2(t + dt) = (1 - \mu_2\, dt) P_2(t) + \lambda\, dt \sum_{i \leq -1} P(x_i(t) \neq 0, \dots, x_{-1}(t) \neq 0, x_0(t) = 0) + o(dt).$$

The invariance under translation gives

$$\frac{\partial P_2(t)}{\partial t} = -\mu_2 P_2(t) + \lambda \sum_{i \geq 1} P(x_0(t) \neq 0, \dots, x_{i-1}(t) \neq 0, x_i(t) = 0).$$

The sequence $(x_i(t))_{i \geq 1}$ hits 0 with probability 1; hence,

$$\frac{\partial P_2(t)}{\partial t} = -\mu_2 P_2(t) + \lambda P(x_0(t) \neq 0).$$

Replacing $P(x_0(t) \neq 0)$ by $P_1(t) + P_2(t)$ in the above equation gives the second equation of our lemma. The other equation is proved in the same way. □

PROOF OF THEOREM 6. The limiting values of the solutions of the differential equation of Lemma 7 are given by

$$(15) \quad P_1 = \frac{\lambda}{\mu_1 + \lambda + \lambda^2/(\mu_2 - \lambda)}, \qquad P_2 = \frac{\lambda^2}{(\mu_2 - \lambda)(\mu_1 + \lambda) + \lambda^2}.$$

(a) If $\mu_2 < \lambda$, then it is easily checked that $P_1 + P_2 > 1$. Hence, there exists some t for which $P_0(t) = 0$ and τ cannot be infinite in this case.

(b) Using Lemma 7, we get

$$\frac{\partial (P_1 + P_2)(t)}{\partial t} = \lambda - \mu_1 P_1(t) - \mu_2 P_2(t).$$

A finite extremum of $(P_1 + P_2)(t)$ (if any) satisfies $\lambda - \mu_1 P_1(t_0) - \mu_2 P_2(t_0) = 0$; hence,

$$(P_1 + P_2)(t_0) \le \frac{\lambda}{\min\{\mu_1, \mu_2\}} < 1.$$

At infinity, it is easy to check that $P_1 + P_2 < 1$. Our process is thus always alive. □

REMARK. Trite calculations with the explicit solution of the equation of Lemma 7 would show that when $\mu_1 < \lambda < \mu_2$ and if $\phi\{\mathbf{x} \mid x(1) \neq 0\}$ is sufficiently small, then $\mathbf{X}(t)_{t \ge 0}$ is almost surely defined; otherwise it dies.

In the case, where $\tau = +\infty$, there exists at least one invariant measure. Our next proposition shows that there is a unique one.

PROPOSITION 8. *If $\mu_1 = \mu_2 = \mu$ and $\lambda < \mu$, then for any $\phi \in \mathcal{E}$, $(\mathbf{X}(t))_{t \in \mathbb{R}}$ converges in distribution to a unique invariant measure.*

PROOF. In this case, the state space can be reduced to $\{0, 1\}^{\mathbb{Z}}$. Let $(\mathbf{X}^0(\mathrm{t}))_{t \in \mathbb{R}}$ (resp. $(\mathbf{X}(t))_{t \in \mathbb{R}}$) denote the Markov process with the empty network (resp. ϕ) as starting point. Using our construction of Theorem 4, it is easy to couple the two processes so that $\mathbf{X}^0(t) \preccurlyeq \mathbf{X}(t)$, that is, $x_i^0(t) \le x_i(t)$ for all $i \in \mathbb{Z}$. In particular, $(\mathbf{X}^0(t))_{t \in \mathbb{R}}$ is stochastically nondecreasing, and hence is converging in distribution to some measure π. Any limiting distribution ψ of $(\mathbf{X}(t))$ satisfies $\pi \preccurlyeq_{\mathrm{st}} \psi$(where $\preccurlyeq_{\mathrm{st}}$ is the stochastic order on measures associated with $\preccurlyeq$), but according to Lemma 7, $\pi(x_0 = i) = \psi(x_0 = i)$ for $i = 0, 1$, which implies $\pi = \psi$ (see [4], Section 2, for example). □

REMARK 3. In the constant case $\mu_1 = \mu_2 = \mu$ when $\lambda = \mu$, it is easy to show that $\tau = +\infty$. The process $(\mathbf{X}(t))_{t \in \mathbb{R}}$ is completely defined, but its limiting distribution is the Dirac mass at †.

4. Local sharing. We consider now the model on the one-dimensional lattice with $\Lambda = [-L, L] \subset \mathbb{Z}^1$. We assume that a customer is lost if it cannot find an empty site within a distance d to the right of its arrival point. As before, a customer is served at rate μ_1 if it has not been rerouted and at rate μ_2 otherwise.

From (14) it is seen that important parameters for understanding the qualitative behavior of the loss probability are μ_2/μ_1 and λ. The following theorem is standard. It gives a satisfactory description of the process in the thermodynamic limit.

THEOREM 9. *There exists $\lambda_0 > 0$ such that for any $\lambda < \lambda_0$ the following limits exist*:

$$p(s_A) = \lim_{L\to+\infty}\lim_{t\to+\infty} p_t^{\Lambda}(s_A) = \lim_{t\to+\infty}\lim_{L\to+\infty} p_t^{\Lambda}(s_A). \tag{16}$$

The $p(s_A)$ are analytic in λ (including $\lambda = 0$) and define a translation invariant measure on $\mathscr{A}^{\mathbb{Z}^d}$ with exponential decay of correlations. For example, for the two-point correlation functions we have

$$|p(s_x, s_y) - p(s_x)p(s_y)| \le Ce^{-\alpha|x-y|}$$

for some C, $\alpha > 0$.

This is known from the cluster expansion theory for the processes with local interaction (see [6] and references therein). Now we proceed to the calculation of the loss probability. Contrary to Remark 1(b), here the loss probability is defined by the thermodynamical limit of the finite volume distributions. Moreover, it is equal to

$$q_\Lambda(x) = p^{\Lambda}(\xi(y) \ne 0 \text{ for all } y \in [x, x+d]). \tag{17}$$

To provide explicit calculations, we use methods from mathematical physics: the correlation equations for correlation functions. The reader is not assumed to be acquainted with these techniques, so we give a detailed exposition. We start by explaining how to get equations for the correlation functions.

Let $h(s'_\Lambda, s_\Lambda)$, $s_\Lambda \ne s'_\Lambda$, be transition rates from s'_Λ to s_Λ for our Markov chain in Λ so that (we omit for a while the upper index Λ, having in mind that we consider Markov chain $\mathscr{A}^{\Lambda}$)

$$h(s'_\Lambda, s_\Lambda)\,dt = P(\xi_{t+dt}(x) = s_x,\ x \in \Lambda \mid \xi_t(x) = s'_x,\ x \in \Lambda). \tag{18}$$

Then the Kolmogorov equations are

$$p_{t+dt}(s_\Lambda) = p_t(s_\Lambda) + \sum_{s'_\Lambda} h(s'_\Lambda, s_\Lambda)p_t(s'_\Lambda)\,dt,$$

where

$$h(s_\Lambda, s_\Lambda) = -\sum_{s'_\Lambda:\ s'_\Lambda \ne s_\Lambda} h(s'_\Lambda, s_\Lambda).$$

For our model only one node can change at a time, that is, for $s_\Lambda \ne s'_\Lambda$ the rates $h(s_\Lambda, s'_\Lambda)$ can be different from 0 only if $s_x \ne s'_x$ exactly at one point $x \in \Lambda$. One can put

$$h(s_\Lambda, s'_\Lambda) = h_x^s(s'_\Lambda)$$

when s_Λ is different from s'_Λ only in the point x and $s_x = s$. From (18) one gets, using Kronecker symbols,

$$P(\xi_{t+dt}(x) = s \mid \xi_t) = \Bigl(1 - \delta_s(\xi_t(x))h_x^s(\xi_t)\,dt + \delta_s(\xi_t(x))\Bigl(1 - dt\sum_{s':\ s'\ne s} h_x^{s'}(\xi_t)\Bigr)\Bigr).$$

Taking expectations, we get

$$\begin{aligned} P_{t+dt}(x;s) = & \sum_{s_\Lambda:\, s_x \neq s} h_x^s(s_\Lambda) p_t(\Lambda; s_\Lambda)\, dt + p_t(x;s) \\ & - dt \sum_{s_\Lambda:\, s_x = s} \sum_{s' \neq s} h_x^{s'}(s_\Lambda) p_t(\Lambda; s_\Lambda), \end{aligned}$$

or for the stationary measure,

$$0 = \sum_{s_\Lambda:\, s_x \neq s} h_x^s(s_\Lambda) p(\Lambda; s_\Lambda) - \sum_{s_\Lambda:\, s_x = s} \sum_{s' \neq s} h_x^{s'}(s_\Lambda) p(\Lambda; s_\Lambda). \tag{19}$$

For any subsets A, B, C, D of λ, we shall use the shorthand notation $1_A 2_B 0_C *_D$ to denote the subset of the elements of $\mathscr{A}^\Lambda$ for which the nodes of C are empty, there is a directly routed customer at each node of A, there is an alternatively routed customer at each node of B and all the nodes of D are occupied.

Then (19) becomes [if we put $s = 0$ for the first equation in (20) and $s = 1$ for the second one]

$$\begin{aligned} 0 &= \lambda p(0_x) - \mu_1 p(1_x), \\ 0 &= \lambda \sum_{k=0}^{d-1} p(*_{[x-k-1,x-1]} 0_x) - \mu_2 p(2_x) \\ &= \lambda \left(\sum_{k=0}^{d-1} p(*_{[x-k-1,x-1]}) - p(*_{[x-k-1,x]}) \right) - \mu_2 p(2_x). \end{aligned} \tag{20}$$

REMARK 4. If in (20) one puts $\Lambda = \mathbb{Z}$, $d = +\infty$, then the process is the one we analyzed in Section 3. If we solve the above equations, one gets the solutions given by the equations (15), which are also the expressions obtained in Theorem 2.

To get equations for higher order correlation functions, we consider

$$\begin{aligned} & P(\xi_{t+dt}(x) = s_x, x \in A \mid \xi_t) \\ & \quad = \sum_{x \in A} (1 - \delta_{s_x}(\xi_t(x))) \Bigl(\prod_{y:\, y \neq x,\, y \in A} \delta_{s_y}(\xi_t(y)) \Bigr) h_x^{s_x}(\xi_t)\, dt \\ & \qquad + \Bigl(\prod_{x \in A} \delta_{s_x}(\xi_t(x)) \Bigr) \Bigl(1 - dt \sum_{x \in A} \sum_{s'_x \neq s_x} h_x^{s'_x}(\xi_t) \Bigr). \end{aligned}$$

Then for $A \cup B = \varnothing$,

$$\begin{aligned} \frac{\partial p_t(2_A 1_B)}{\partial t} = & \lambda \sum_{x \in B} p_t(2_A 1_{B-\{x\}} 0_x) \\ & + \lambda \sum_{x \in A} \sum_{k=0}^{d-1} p_t(2_{A-\{x\}} 1_B 0_x *_{[x-k-1,x-1]}) \\ & - (\mu_2 |A| + \mu_1 |B|) p_t(2_A 0_B) \end{aligned}$$

or for the stationary measure,

$$\begin{aligned} & p(2_A 1_B)(\mu_2|A| + \mu_1|B|) \\ (21)\qquad & = \lambda \sum_{x\in A} \sum_{k=0}^{d-1} \big(p(2_{A-\{x\}}1_B *_{[x-k-1,x-1]}) - p(2_{A-\{x\}}1_B *_{[x-k-1,x]})\big) \\ & + \lambda \sum_{x\in B} \big(p(2_A 1_{B-\{x\}}) - p(2_A 1_{B-\{x\}} *_x)\big). \end{aligned}$$

Coming to the problem of minimizing the loss probability, we note first that for $\mu_1 = \mu_2$ the independent policy is always the worst: If there is an empty available node, a customer should be put into it. For $\mu_1 > \mu_2$ this is not always the case.

THEOREM 10. *Under the condition* $d = 1$:

(i) *If* λ *is small enough and* μ_1, μ_2 *is fixed, then the loss probability* (*in the thermodynamic limit*) *is*

$$q = \frac{2\mu_1 + \mu_2}{\mu_1^2(\mu_1 + \mu_2)}\lambda^2 + O(\lambda^3).$$

(ii) *In the case when* λ *is fixed and*

$$(22)\qquad \mu_1 = o(\lambda), \qquad \mu_2 = o(\mu_1),$$

we have

$$q = 1 - 2\frac{\mu_2}{\lambda} + O\left(\frac{\mu_2^2}{\lambda\mu_1}\right).$$

PROOF. (i) We use the analyticity result of Theorem 9 and calculate the first terms of the expansion in λ for $d = 1$ and small λ with $\mu_1, \mu_2 > 0$ fixed. Using translation invariance of the limiting measure, we get from the formula (21) and the second equation of (20),

$$p(2) = p(2_x) = O(\lambda^2)$$

and from the first equation of (20),

$$p(1) \le \frac{\lambda}{\mu_1};$$

hence,

$$p(1) = \frac{\lambda}{\mu_1}(1 - p(1) - p(2)) = \frac{\lambda}{\mu_1} + O(\lambda^2).$$

From (21) we get

$$\begin{aligned} 2\mu_1 p(11) &= \lambda(p(10) + p(01)) = \lambda(2p(1) - p(1*) - p(*1)) \\ &= \lambda(2p(1) - 2p(11) - p(12) - p(21)) \end{aligned}$$

or

$$p(11) = \frac{\lambda}{\lambda + \mu_1}p(1) - \frac{\lambda}{2(\lambda + \mu_1)}(p(12) + p(21)) = \left(\frac{\lambda}{\mu_1}\right)^2 + O(\lambda^3).$$

Similarly one can show that $p(22)$, $p(21)$ are $O(\lambda^3)$ and

$$p(12) = \frac{\lambda^2}{\mu_1(\mu_1 + \mu_2)} + O(\lambda^3),$$

so

$$q = p(**) = \frac{2\mu_1 + \mu_2}{\mu_1^2(\mu_1 + \mu_2)}\lambda^2 + O(\lambda^3).$$

(ii) Now we pass to case (ii). Let us note that equations (22) have been written for correlation functions $p(2_A 1_B)$ as these are small quantities; more exactly, $p(2_A 1_B) = O(\lambda^{|A \cup B|})$. Now the correlation equations should be written for the correlation functions $p(0_A 1_B)$. We leave rewriting the equations and the proof of the following theorem to the reader.

THEOREM 11. *There exists $\varepsilon_0 > 0$ such that for $\mu_2/\mu_1, \mu_1/\lambda < \varepsilon_0$ the following limits exist*:

$$(23) \qquad p(s_A) = \lim_{L \to +\infty} \lim_{t \to +\infty} p_t^\Lambda(s_A) = \lim_{t \to +\infty} \lim_{L \to +\infty} p_t^\Lambda(s_A).$$

The $p(s_A)$ are analytic in $\mu_2/\mu_1, \mu_1/\lambda$ and define a translation invariant measure on $\mathscr{A}^{\mathbb{Z}}$ with exponential decay of correlations. Moreover, $p(0_A 1_B) = O(\varepsilon^{|A \cup B|})$, $\varepsilon = \max(\mu_1/\lambda, \mu_2/\mu_1)$.

To finish the proof of Theorem 10 for case (ii), we shall use (20) and (21), having in mind that using the previous theorem we can expand in the small parameters specified for case (ii). We can write

$$q = p(**) = 1 - 2p(0) + p(00).$$

From (20) we get

$$\lambda p(0) = \mu_1 p(1),$$

$$\mu_2\left(1 - p(0) - \frac{\lambda}{\mu_1}p(0)\right) = \mu_2(1 - p(0) - p(1)) = \mu_2 p(2)$$
$$= \lambda p(*0) = \lambda(p(0) - p(00)),$$

so

$$p(0) = \frac{\mu_2/\lambda + p(00)}{\mu_2/\lambda + 1 + \mu_2/\mu_1} = \frac{\mu_2}{\lambda} + O\left(\frac{\mu_2^2}{\lambda\mu_1}\right).$$

Case (ii) of Theorem 10 is thus proved. □

5. Some open problems. Load sharing models promise to be rich in different phenomena that could be common in more complicated models, and we want to indicate some further problems.

1. In the model of the Section 4 take $d \to +\infty$ or even equal to infinity; take even $\mu_1 = \mu_2$. The resulting process is not a process with local interaction. Does the series defining the solution of correlation equations converge for λ/μ_1 small? Is the solution analytic in λ/μ_1?

2. In the model on the one-dimensional lattice with $\Lambda = [-L, L] \subset \mathbb{Z}$, $\mu(x, y) = \mu_1$ for $|x - y| \le d$ and $\mu(x, y) = \mu_2$ for $|x - y| > d$. Is the invariant measure unique for all values of the parameters μ_1, μ_2, λ and d? Find the phase diagram for the loss probability.
3. On $\mathbb{Z}^1$, let $\mu(x, y) \sim \mu_1/(1 + |x - y|^{\alpha})$. For which α it is true that this policy is better than the independent one with the parameter μ_1?
4. What about higher order correlation functions in the cases of Section 2 for a one-dimensional lattice when the rule $\mathscr{R}$ is just choosing the closest node.
5. Try similar problems for waiting time and infinite buffers.

Acknowledgment. The authors wish to thank Elena Petrova for her careful reading of the first draft of that paper.

REFERENCES

[1] Cornfeld, I. P., Fomin, S. V. and Sinai, Y. G. (1982). *Ergodic Theory*. Springer, Berlin.

[2] Ethier, S. and Kurtz, T. (1986). *Markov Processes: Characterization and Convergence*. Wiley, New York.

[3] Kelly, F. (1991). Loss networks. *Ann. Appl. Probab.* **1** 319–378.

[4] Liggett, T. (1985). *Interacting Particle Systems*. Springer, New York.

[5] Malyshev, V. and Menshikov, M. (1981). Ergodicity, continuity, and analyticity of countable Markov chains. *Trans. Moscow Math. Soc.* **1** 1–47.

[6] Malyshev, V. and Minlos, R. (1991). *Gibbs Random Fields*. Kluwer, Amsterdam.

[7] Syski, R. (1960). *Introduction to Congestion Theory in Telephone Systems*. Oliver and Boyd, Edinburgh.

Laboratory of Large Random Systems
Moscow State University
Moscow 119899
Russia

INRIA
Domaine de Voluceau
B.P. 105
78153 Le Chesnay Cedex
France