The Annals of Applied Probability
1998, Vol. 8, No. 2, 337–353

THE LOAD TRANSFER MODEL

BY M. BEGUIN, L. GRAY AND B. YCART

LMC / IMAG, University of Minnesota and LMC / IMAG

An interacting particle model for load transferring in parallel architectures is defined. In the case of an infinite lattice the model is proved to be ergodic and to converge exponentially fast to its equilibrium. When the architecture is that of a complete graph, the total number of loads behaves as a birth and death process, and explicit upper bounds on the benefits that can be expected from a transferring policy are derived. Experimental results for different types of architectures are presented and compared to the solution of the mean field equations. There is fairly good agreement between the two for quantities of practical interest.

1. Introduction. A parallel architecture is a set of processors arranged into an unoriented graph pattern by communication lines that connect them. Currently used architectures include cycles, grids, toruses, hypercubes and complete graphs. One advantage of such an arrangement is to allow an even distribution of tasks over the whole set of processors. Typically, a situation where one processor has a long file of tasks waiting to be executed while one of its neighbors remains idle will be avoided by transferring part of the load from the first processor onto the latter. Since such a transferring policy may be costly in time and resources, it is of interest to evaluate the gain one can expect from its implementation (cf. [3] for a general reference).

Assuming immediate transfers, we propose to compute upper bounds on the best possible gains, based on an interacting particle model that can be described informally as follows. Each processor taken individually is seen as a Markovian queue with limited capacity. Tasks are supposed to arrive one by one according to a Poisson process with intensity λ, and they are completed by processors in exponential time, independently of each other. The time unit can be arbitrarily fixed to equal the average completion time of a task. The state of a given processor is the number of tasks either being treated or waiting to be treated by that processor. Since buffers have limited capacity, the state of a processor is bounded above by some fixed integer K. Any task that arrives at a processor at level K is rejected from the system. We will be interested in the probabilities $P_i(\lambda)$ that, in the stationary regime, a processor is in state $i = 0, \dots, K$. If the processors are not connected, they

Received February 1997; revised July 1997.
AMS 1991 *subject classifications*. 60K35, 68M20.
Key words and phrases. Load transferring, interacting particle system, ergodicity.

behave as independent $M/M/1/K$ queues, and these probabilities are

$$P_i(\lambda) = \begin{cases} \dfrac{1-\lambda}{1-\lambda^{K+1}}\lambda^i, & \text{if } \lambda \neq 1, \\ \dfrac{1}{K+1}, & \text{if } \lambda = 1. \end{cases} \tag{1.1}$$

The performance criteria usually considered in the computer science literature [3] can be expressed in terms of the probabilities $P_i(\lambda)$. These criteria are as follows:

1. The rejection probability
$$P_K(\lambda).$$
2. The mean number of tasks treated per processor and per time unit
$$\lambda(1 - P_K(\lambda)).$$
3. The mean load per processor
$$\sum_{i=0}^{K} iP_i(\lambda).$$
4. The mean response time
$$\frac{\sum_{i=0}^{K} iP_i(\lambda)}{\lambda(1 - P_K(\lambda))}.$$

The last expression comes from Little's formula (cf., e.g., [17]). One objective is to compute the same equilibrium probabilities under a policy of immediate transfer and to compare them with the expressions (1.1) above. We shall assume that immediate transfers occur between neighboring processors as soon as the difference of load between them exceeds 1. Suppose that processor x has level $i < K$ and some of its neighbors have levels less than i, thus being available to receive a possible transfer from x. Any incoming new task on x will immediately be transferred with equal probability onto any of the possible targets. However, these neighbors may in turn have other neighbors with less tasks. The incoming task may thus cascade down from its arrival processor onto another processor at further distance. Notice that, due to the limited capacity hypothesis, the number of successive transfers of an arriving task, or the distance between the initial arrival point and the final destination, can be at most $K - 1$ (from a processor at level $K - 1$ to one at level 0). Also, according to this rule, the load of a processor can increase only if it is less than or equal to that of its neighbors. So far, the model we have described is akin to sandpile models, widely studied in the physics literature (see [5] for a recent reference or [14] for a model closer to the one considered here). There is also an analogy with the crystal growth models of Gates and Westcott [10], in the sense that the rate of increase of a configuration at a given site depends on the difference of height between that site and its neighbors.

Unlike in sandpile or crystal growth models, tasks leave the system at each site as they are completed. The situation after the completion of a task is symmetric to the one described above after the arrival of a task. Assume a task is ready to leave processor x and some neighbors of x have higher load levels. The departure of that task will immediately be compensated by a transfer from any of the more loaded neighbors, chosen with equal probability. This in turn may provoke transfers from other neighbors. As before, the number of transfers caused by the completion of a task at x and the distance to the last such transfer from x are at most $K - 1$. Due to the obligation of transferring, an actual decrease in load can only occur at a site whose neighbors have as many loads or less. Notice that in this setting it is implicit that the limited capacity K should be at least 2.

The model described above is a spin system with values in $\{0, \dots, K\}$ (cf. [12] for a general reference) and its formal description will be given in Section 2, together with some basic properties.

In the case of massively parallel architectures, several thousands of processors have to be connected and, mainly for technical reasons, the most frequently used architecture is that of a torus. It is therefore of practical interest to study the load transfer model on $\mathbb{Z}^2$; it is also of mathematical interest. Section 3 contains our main result: the load transfer model is proved to be ergodic for any lattice structure on $\mathbb{Z}^d$ and any values of K and λ, and to converge at exponential speed to its equilibrium state. The proof uses a specific coupling, together with classical techniques of stochastic comparison.

The architecture for which optimal gains can be expected is obviously the complete graph. In that case, the total number of loads in the system evolves as a birth and death process. Its stationary measure and thus all quantities of interest can be computed explicitly. This will be treated in Section 4. The complete graph case has already been considered with different modeling hypotheses by Malyshev and Robert [13] and Spies [16].

The theoretical results of Section 3 and 4 will be complemented by approximate computations of the probabilities $P_i(\lambda)$ of the different levels of loads. Two types of approximations have been considered, using computer simulation and the mean field heuristics. The results turn out to be in close agreement. Numerical results will be presented in Section 5.

2. The load transfer model.

Our model being a spin system with values in $\{0, \dots, K\}$, we shall stay as close as possible to Chapter III of [12] regarding terminology and notations. The set of processors or sites is denoted by S. The communication channels are seen as edges an their set is denoted by E. If x is a site, its successive sets of neighbors are defined as

$$\mathscr{N}(x) = \mathscr{N}_1(x) = \{x\} \cup \{y \in S; \{x, y\} \in E\}$$

and

$$\forall k \geq 2, \qquad \mathscr{N}_k(x) = \bigcup_{y \in \mathscr{N}_{k-1}(x)} \mathscr{N}(y).$$

A configuration η is a mapping that describes the load level of each processor:

$$\forall x \in S, \qquad \eta(x) \in \{0, \dots, K\}.$$

The set of all configurations, $X = \{0, \dots, K\}^S$, should a priori be the state space of the process. However, it will be convenient to reduce it to the subset $\mathscr{A}$ of configurations such that the load difference between neighbors is at most 1:

$$\mathscr{A} = \{\eta \in X \text{ s.t. } \forall \{x, y\} \in E, |\eta(x) - \eta(y)| \leq 1\}.$$

In our model, configurations are allowed to change at one site at a time at most (spin system dynamics). Any such movement is either an arrival or a departure of a task at some site x (possibly as the result of a series of load transfers ending at x), and the configuration will be increased or decreased by 1 at x. This corresponds to a transition from configuration η to one of the two configurations denoted by η_x^+ and η_x^- and defined by

$$\begin{aligned} \eta_x^+(x) &= \eta(x) + 1, & \eta_x^-(x) &= \eta(x) - 1, \\ \eta_x^+(y) &= \eta(y), \quad \forall y \neq x, & \eta_x^-(y) &= \eta(y), \quad \forall y \neq x. \end{aligned}$$

The transition rate from η to η_x^+ (arrival at x) will be denoted by $a(x, \eta)$; that from η to η_x^- (departure at x) will be denoted by $d(x, \eta)$.

The definitions of $a(x, \eta)$ and $d(x, \eta)$ have to take transfers into account. The arrival of a task at site x can happen only if x is a local minimum for configuration η, but it can correspond either to a direct arrival at x from outside or to an arrival at some other site in $\mathscr{N}_{K-1}(x)$ that cascades down to x through successive transfers. We denote by $\overline{n}(x, \eta)$ [respectively, $\underline{n}(x, \eta)$] the number of neighbors of x with a strictly higher (respectively, lower) load level under configuration η. We define the *virtual arrival rate* $\alpha(x, \eta)$ of *any* site x under configuration η recursively as follows.

DEFINITION 1. (i) If $\overline{n}(x, \eta) = 0$, then

$$\alpha(x, \eta) = \begin{cases} \lambda, & \text{if } \eta(x) < K, \\ 0, & \text{if } \eta(x) = K. \end{cases}$$

(ii) If $\overline{n}(x, \eta) > 0$, then

$$\alpha(x, \eta) = \lambda + \sum_{\substack{y \in \mathscr{N}(x) \\ \eta(y) > \eta(x)}} \frac{\alpha(y, \eta)}{\underline{n}(y, n)}.$$

The actual arrival rate at x can be nonzero only if x is a local minimum for η:

$$a(x, \eta) = \begin{cases} \alpha(x, \eta), & \text{if } \underline{n}(x, \eta) = 0, \\ 0, & \text{if } \underline{n}(x, \eta) > 0. \end{cases}$$

The above definitions can be read as follows. If x is a local maximum for configuration η ($\overline{n}(x, \eta) = 0$), then no incoming task can be transferred onto

x. A task may arrive directly from outside, provided x is not at the maximal capacity K. If x is not a local maximum, then at least one site y among its neighbors has a higher load. Any task received by y (either directly or after transfer from another neighbor) will be transferred to x with a probability $1/(\underline{n}(y,\eta))$ depending on the number of neighbors that can receive such a transfer (at least 1 since x is included).

There is a symmetry in the model that makes the definition of departure rates completely analogous to that of arrival rates. We first define recursively the *virtual departure rate* of *any* site x under configuration η.

DEFINITION 2. (i) If $\underline{n}(x,\eta) = 0$, then

$$\delta(x,\eta) = \begin{cases} 1, & \text{if } \eta(x) > 0,\\ 0, & \text{if } \eta(x) = 0.\end{cases}$$

(ii) If $n(x,\eta) > 0$, then

$$\delta(x,\eta) = 1 + \sum_{\substack{y\in\mathscr{N}(x)\\ \eta(y)<\eta(x)}} \frac{\delta(y,\eta)}{\overline{n}(y,\eta)}.$$

The actual departure rate of x under configuration η can be nonzero only if x is a local maximum for η:

$$d(x,\eta) = \begin{cases} \delta(x,\eta), & \text{if } \overline{n}(x,\eta) = 0,\\ 0, & \text{if } \overline{n}(x,\eta) > 0.\end{cases}$$

If x is a local minimum, there is no site to which x could transfer a task. One task can leave x after completion, provided $\eta(x) > 0$. If x is not a local minimum, then at least one task y among its neighbors has a lower level. If this site y loses a task—either by completion or transfer—x will be able to transfer one of its tasks onto y, with probability $1/(\overline{n}(y,\eta))$ depending on the number of neighbors of y that can benefit from such a transfer.

The rates that have just been defined correspond obviously to finite range interactions, since the arrival and departure rates at x depend only on the values of the configuration in $\mathscr{N}_K(x)$. When $S = \mathbb{Z}^d$, endowed with a lattice structure, the rates are translation invariant and the process can be constructed explicitly using Harris' argument (cf. [6], page 119). The interacting particle system so defined will be referred to as the load transfer model (LTM).

So far, the rates $a(x,\eta)$ and $d(x,\eta)$ have been defined for any configuration η in $X = \{0,\dots,K\}^S$. However, the set $\mathscr{A}$ of those configurations such that the load difference between neighbors is at most 1, is absorbing. Indeed the only possible movements are either arrivals at local minima or departures at local maxima. If $\eta \in \mathscr{A}$ and x is a local minimum for η, then the load differences between x and its neighbors can only be 0 or -1. An arrival at x will change these differences to $+1$ or 0, and the new configuration η_x^+ will still be in $\mathscr{A}$. The situation is symmetric for a departure at a local maximum. From now on we shall consider that the state space of the LTM is $\mathscr{A}$.

Two basic properties of the LTM can immediately be derived from the definition of the rates. First, the roles of arrivals and departures are exactly symmetric. Suppose one associates to each processor its free space $K - \eta(x)$ instead of its load level $\eta(x)$. The process of free spaces has exactly the same dynamics as above, with the roles of λ and 1 exchanged.

Another basic property is attractiveness.

PROPOSITION 1. *Consider on $\mathscr{A}$ the natural componentwise ordering*

$$\eta \le \zeta \quad \Leftrightarrow \quad \forall x \in S,\, \eta(x) \le \zeta(x).$$

The LTM is attractive on $\mathscr{A}$ in the following sense

$$\forall \eta \le \zeta \in \mathscr{A}, \qquad \eta(x) = \zeta(x) \quad \Rightarrow \quad \begin{cases} a(x,\eta) \le a(x,\zeta), \\ d(x,\eta) \ge d(x,\zeta). \end{cases} \tag{2.1}$$

In other terms, the dynamics of the LTM tend to have each site agree with its neighbors.

PROOF. Assume $\eta \le \zeta \in \mathscr{A}$. Let $x \in S$ be such that $\eta(x) = \zeta(x)$. We shall check the result for arrivals only; the situation for departures is symmetric. If x is not a local minimum for η, the inequality is trivially satisfied. If x is a local minimum for η, it is also a local minimum for ζ. Call a *descending path* for η any path in the sense of the graph structure of S, ending at x, along which the values of η decrease exactly by 1 at each step. The key observation is that along any descending path for η, the values of ζ must coincide with those of η (since $\zeta \in \mathscr{A}$). The arrival rate $a(x,\eta)$ can be seen as the product by λ of the sum of all descending paths of the probability for a load arriving at the beginning of that path, to follow it through successive transfers, all the way to site x. If γ is such a path, the probability to follow it is larger on ζ than on η. Since ζ has at least as many descending paths as η, one obtains the desired inequality. □

It is easy to extend Theorem 2.2 (page 134 of [12]) to check that attractiveness in the sense of (2.1) above is equivalent to stochastic monotonicity for spin systems on $\mathscr{A}$ ([12], Definition 2.3, page 72). One can view it also as a particular case of Theorem 3.3 (page 596 of [9]). Stochastic monotonicity will be an important ingredient for the proof of ergodicity in Theorem 1.

3. The load transfer model on lattices.

Throughout this section, the set of sites is $S = \mathbb{Z}^d$. The only restriction on the graph structure is that it should be translation invariant and locally finite. For instance,

$$\forall x \in S, \qquad \mathscr{N}(x) = \{y \text{ s.t. } \|x - y\| \le r\},$$

where $\|\cdot\|$ is any of the usual norms and r is a fixed integer.

THEOREM 1. *The LTM on $\mathbb{Z}^d$ is ergodic and converges exponentially fast to its stationary measure.*

PROOF. For an attractive spin system with values in $\{0,1\}$, the classical technique of proof for ergodicity is described in [12] (Chapter III, Theorem 2.3 and Corollary 2.4, page 136). It can easily be extended to attractive spin systems with values in $\{0,\dots,K\}$. The general idea is the following. Consider two copies of the process: one starts at time 0 from the minimal configuration; the other starts from the maximal configuration. The two copies are coupled in such a way that the initial order is preserved at any instant. If the difference between the two copies tends to zero in distribution, then the process is ergodic. The coupling traditionally used is the so-called basic or Vasershtein coupling ([12], page 124). We shall use a stronger coupling, which not only preserves the order between configurations, but is also such that the overall difference of loads between the two copies can only decrease in time. Moreover, our coupling has the advantage that it fits the intuitive description of the model that was given in the Introduction. In the basic coupling, the two copies evolve independently at each site until they first meet. We shall allow much less independence, and decide that arrivals and completions of tasks are simultaneous for both copies.

To each site x, associate two Poisson processes $\{A_t(x);\ t \geq 0\}$ and $\{D_t(x)\colon t \geq 0\}$ with respective intensities λ and 1; all these processes are independent. The process $\{A_t(x)\}$ counts the arrivals of tasks at site x (accepted or not) and the process $\{D_t(x)\}$ counts the completions of tasks at site x (effective or not). We want to construct a Feller process $\{(\eta_t, \zeta_t);\ t \geq 0\}$ on $\mathscr{A} \times \mathscr{A}$ starting from the initial couple of configurations

$$\forall x \in S, \qquad (\eta_0(x), \zeta_0(x)) = (0, K).$$

Since both initial configurations are in the set $\mathscr{A}$, which is absorbing for the dynamics of the LTM, both coordinates will stay in $\mathscr{A}$ at any time. Therefore, in the construction below we can assume that all configurations are such that the difference of load between neighbors is at most 1. The main requirement is that the initial ordering should be preserved, that is,

$$\forall x \in S, \forall t \geq 0, \qquad \eta_t(x) \leq \zeta_t(x).$$

We shall describe only the evolution of both coordinates upon an arrival at site x, that is, an instant of jump for $A_t(x)$. The evolution for a departure is exactly symmetric and is obtained by replacing $\eta(x)$ by $K - \eta(x)$ and $A_t(x)$ by $D_t(x)$.

Assume that at some instant t, $\eta_t = \eta$ and $\zeta_t = \zeta$ with $\eta \leq \zeta$, and a task arrives at site x. Three cases are possible.

Case 1. If $\eta(x) = K$ and $\zeta(x) = K$, the task is rejected and both configurations stay the same.

Case 2. If $\eta(x) < K$ and $\zeta(x) = K$, the task is rejected by the second coordinate and ζ stays the same. It is accepted by the first coordinate and provokes an increase of configuration η at a certain site. This site is x itself if it was a local minimum. If not, the increase of η happens at some other site y in $\mathscr{N}_K(x)$, chosen according to the transferring rules described in the Introduction. For this to happen, there has to be some chain of successive neighbors linking x to y, along which η decreases exactly by 1 at each step. Along the same sequence ζ can decrease by at most 1 at each step. Since $\eta(x)$ is strictly less than $\zeta(x)$, the same has to be true for $\eta(y)$ and $\zeta(y)$, so the ordering of configurations is preserved by the increase of η at site y.

Case 3. If $\eta(x) < K$ and $\zeta(x) < K$, the task is accepted and provokes for each configuration an increase at some site in $\mathscr{N}_K(x)$. Assume first that $\eta(x) < \zeta(x)$. The incoming task will either stay at x or go elsewhere according to whether x is a local minimum or not. The random choices in successive transfers are made independently. Let y be the final destination of the task for configuration η. By the argument already used in the previous case, $\eta(y)$ had to be strictly less than $\zeta(y)$. So the increase of η at y preserves the ordering. Assume now that $\eta(x) = \zeta(x)$. To preserve the ordering, one has to have transfers agree as much as possible. Let us examine the first move. If x was a local minimum for η, then it had to be the same for ζ. In this case, both configurations increase at x. If x was a local minimum for ζ and not for η, then the increase of ζ is at x and the increase for η happens at some site $y \neq x$. By the already used argument $\eta(y)$ had to be strictly less than $\zeta(y)$ and the ordering is preserved. The last case is when x is not a local minimum for either η or ζ. Since $\eta(x) = \zeta(x)$ and $\eta \leq \zeta$, necessarily the incoming task has at least the same possibilities of transfer on η as on ζ. Let us choose randomly the first destination y of the task for η. If y is also a possible destination for ζ, let the task to go to that same place on both configurations. If the chosen destination y was not possible for ζ, then send the task there for η and send the task on ζ elsewhere with another random choice, independent from the first. Saying that y was a possible destination for η and not for ζ implies that $\eta(y) < \zeta(y)$, so that an increase at y for η preserves the ordering. Now iterate the same procedure if the chosen destinations were the same; go on independently if they were different. The final destination of the task on η will necessarily be either the same as for ζ or at a site where η is strictly less than ζ; thus the ordering will be preserved.

Let $R \subset S$ be a finite subset of sites. Define the *load difference* over R at time t, denoted by $C_t(R)$, as the cumulated difference between configurations:

$$C_t(R) = \sum_{x \in R} \zeta_t(x) - \eta_t(x).$$

From the above construction, it is clear that upon an arrival at site x, the load difference over any set R containing $\mathscr{N}_K(x)$ can only decrease by 1 (the task is rejected by ζ_t, not by η_t) or stay the same (any other case). The coupling being exactly symmetric for a departure at x, this property of decreasing load differences holds at any instant. This will be a key ingredient in the rest of the proof.

Let $\{(\eta_t, \zeta_t); t \geq 0\}$ be the coupled process defined above. To prove Theorem 1, we have to find a positive constant δ such that for any function f from $\mathscr{A}$ into $\mathbb{R}$, depending on a finite number of coordinates, there exists a constant γ such that

$$\left|\mathbb{E}\left[f(\zeta_t) - f(\eta_t)\right]\right| \leq \gamma e^{-\delta t}.$$

By a standard argument involving translation invariance, it is sufficient to prove that

$$\mathrm{Prob}\left[\zeta_t(x) - \eta_t(x) > 0\right] = \mathrm{Prob}\left[C_t(\{x\}) > 0\right] \leq \gamma e^{-\delta t}$$

for some positive constants γ and δ. Let x be a site at which $\zeta_t(x) > \eta_t(x)$. The next step consists in showing that over a unit interval of time the load difference at x will actually decrease by 1 with positive probability:

$$\mathrm{Prob}\left[C_{t+1}(\{x\}) - C_t(\{x\}) = -1\right] \geq \delta > 0.$$

To do this, we consider two configurations $\eta \leq \zeta$ with $\eta(x) < \zeta(x)$ and construct an event depending on x, η and ζ according to which a task arriving at x will be rejected by the higher configuration and not by the lower; no site outside $\mathscr{N}_{2K}(x)$ is affected. To construct this event, the idea is to ask for enough arrivals in $\mathscr{N}_K(x)$ [and no other event in $\mathscr{N}_{2K}(x)$] so as to achieve the following criteria:

1. To raise configuration ζ up to K at site x.
2. To maintain configuration η strictly below K at site x.

Then one more arrival on x will be rejected by ζ and not by η.

We call a *descending path* of ζ any chain of neighbors, starting from x and ending at a local minimum of configuration ζ, along which configuration ζ decreases strictly. There is only a finite number of these paths, bounded above by $|\mathscr{V}(x)|^K$, and all descending paths are included in $\mathscr{N}_K(x)$. First ask for one arrival of a task on all ends of descending paths (local minima for ζ). After these arrivals, the configuration ζ has been changed into another configuration for which all descending paths are shorter by exactly 1. After at most $K - 1$ iterations of this procedure, $\zeta(x)$ has not been changed and x becomes a local minimum of configuration ζ. Meanwhile, the other coordinate η has been changed only at sites in $\mathscr{N}_{2K}(x)$. Some of the incoming loads might have been transferred to x itself for configuration η, but this can only happen when there exists a descending path for η, from the arrival site to x, which implies that the difference between both configurations at x remains positive. If $\zeta(x) = K$, then one more arrival at x will be rejected at ζ and not by η, and the desired event occurs. Otherwise, one more arrival at x increases ζ by 1 at x, still preserving the property $\zeta(x) > \eta(x)$. Repeating again the whole procedure at most $K - 1$ times leads to the desired event. Whatever η and ζ, the probability for that event to occur between t and $t + 1$ is strictly greater than some $\delta > 0$.

Consider now the cube $R = [-n, +n]^d \cap \mathbb{Z}^d$ and denote by R' its closure with respect to neighborhoods at distance $2K$:

$$R' = \bigcup_{x \in R} \mathscr{N}_{2K}(x).$$

The load difference over R', $C_t(R')$, is bound to decrease due to arrivals and departures at those sites in R such that $\zeta_t(x) - \eta_t(x) > 0$. It can also increase due to edge effects. These increases can only come from arrivals and departures at sites whose distance (in the sense of the graph structure) is at most K from the boundary of R. The number of such sites is bounded by some constant multiplied by n^{d-1}. For each site x in R, the probability of an actual decrease due to the rejection of an incoming task at x by ζ and not by η is larger than

$$\delta \operatorname{Prob}[\zeta_t(x) - \eta_t(x) > 0].$$

Hence

$$\mathbb{E}[C_{t+1}(R) - C_t(R)] \leq -(2n+1)^d \delta \operatorname{Prob}[\zeta_t(x) - \eta_t(x) > 0] + o(n^d).$$

However, the left-hand side of the above inequality is also

$$\mathbb{E}[C_{t+1}(R) - C_t(R)] = (2n+1)^d \mathbb{E}[C_{t+1}(\{x\}) - C_t(\{x\})].$$

For n large enough, one gets

$$\mathbb{E}[C_{t+1}(\{x\}) - C_t(\{x\})] \leq -\delta' \operatorname{Prob}[C_t(\{x\}) > 0].$$

However, $C_t(\{x\})$ is a random variable with values in $\{0, \dots, K\}$. Hence

$$\frac{1}{K}\mathbb{E}[C_t(\{x\})] \leq \operatorname{Prob}[C_t(\{x\}) > 0].$$

Thus

$$\mathbb{E}[C_{t+1}(\{x\})] - \mathbb{E}[C_t(\{x\})] \leq -\frac{\delta'}{K}\mathbb{E}[C_t(\{x\})].$$

For any positive integer m, one gets

$$\mathbb{E}[C_m(\{x\})] \leq \left(1 - \frac{\delta'}{K}\right)^m \mathbb{E}[C_0(\{x\})];$$

hence, there exist two positive constants γ and δ'' such that for all $t \geq 0$,

$$\mathbb{E}[C_t(\{x\})] \leq \gamma e^{-\delta'' t}$$

and also

$$\operatorname{Prob}[C_t(\{x\}) > 0] \leq \gamma e^{-\delta'' t}. \qquad \square$$

4. The complete graph case. In this section the set of sites S is finite with n elements. It turns out that the best possible benefits that one can expect from transferring are obtained in the complete graph case. This situation was studied in full detail in [2]. Let $E = \{\{x, y\};\ x, y \in S\}$ and let

$\{\eta_t,\ t \geq 0\}$ be the LTM on (S, E) with rate λ. Let L_t denote the total load of the system at time t:

$$L_t = \sum_{x \in S} \eta_t(x).$$

The main observation is that, due to the symmetry of the system, the process $\{L_t,\ t \geq 0\}$ satisfies the classical lumpability conditions of Rosenblatt [15]. Hence it is a birth and death process (cf., e.g, [1]) with values in $\{0, \dots, Kn\}$ and birth and death rates as follows:

Birth rate (from j to $j+1$):

$$\lambda(j) = \begin{cases} n\lambda, & \text{for } j = 0, \dots, (K-1)n, \\ (Kn - j)\lambda, & \text{for } j = (K-1)n, \dots, Kn - 1. \end{cases}$$

Death rate (from j to $j-1$):

$$\mu(j) = \begin{cases} j, & \text{for } j = 1, \dots, n, \\ n, & \text{for } j = n+1, \dots, Kn. \end{cases}$$

Indeed, as long as there are no more than $n(K-1)$ tasks in the system, all processors have at most $K-1$ tasks, and any new task is accepted. If there are $j = n(K-1) + l$ tasks in the system, l processors have K tasks and only $n - l$ processors can accept new tasks. The situation is symmetric for departures, as already noticed. In other words, in the range $\{0, \dots, (K-1)n\}$, the process $\{L_t\}$ behaves as an $M/M/n$ queue, with n servers, arrival rate $n\lambda$ and service rate 1 for each server. The stationary distribution $(p_j)_{0 \leq j \leq Kn}$ of $\{L_t\}$ is easy to compute. As in the Introduction, we denote by $P_i(\lambda)$ the probability for each site to be at level i, for $i = 0, \dots, K$. The $P_i(\lambda)$s are related to the p_js as

$$P_0(\lambda) = \sum_{j=0}^{n-1} \frac{n-j}{n} p_j,$$

$$P_K(\lambda) = \sum_{j=0}^{n-1} \frac{n-j}{n} p_{Kn-j},$$

$$\forall 0 < j < K, \qquad P_i(\lambda) = \sum_{j=0}^{n} \frac{j}{n} p_{(i-1)n+j} + \sum_{j=1}^{n-1} \frac{n-j}{n} p_{in+j}.$$

Indeed, when there are $(i-1)n + j$ tasks in the system, j processors have i tasks and $n - j$ processors have $j - 1$ tasks. The explicit expressions of these quantities are easily derived and we shall not reproduce them here (cf. [2]). Their asymptotics as n tends to infinity are more interesting. Indeed the benefits one can expect from transfer are an increasing function of n; therefore, the asymptotics given in Proposition 2 yield upper bounds of these benefits, whatever the criterion of evaluation. Moreover, numerical evidence shows that these limits give good approximations for finite ns even for relatively low values.

TABLE 1

$\lambda < 1$	$\lambda = 1$	$\lambda > 1$
$P_0(\lambda) \to 1 - \lambda$	$P_0(\lambda) \to 0$	$P_0(\lambda) \to 0$
$P_1(\lambda) \to \lambda$	$P_1(\lambda) \to 1/(2K - 4)$	$P_1(\lambda) \to 0$
$\vdots$	$\vdots$	$\vdots$
$P_i(\lambda) \to 0$	$P_i(\lambda) \to 1/(K - 2)$	$P_i(\lambda) \to 0$
$\vdots$	$\vdots$	$\vdots$
$P_{K-1}(\lambda) \to 0$	$P_{K-1}(\lambda) \to 1/(2K - 4)$	$P_{K-1}(\lambda) \to 1/\lambda$
$P_K(\lambda) \to 0$	$P_K(\lambda) \to 0$	$P_K(\lambda) \to 1 - 1/\lambda$

PROPOSITION 2. *Assume the capacity of each processor is $K \geq 3$. As n tends to infinity, the probabilities of the different levels of load for the LTM on the complete graph with n vertices have the limits shown in Table* 1.

The results in Table 1 hold also for $K = 2$ except when $\lambda = 1$, in which case $P_1(\lambda)$ tends to 1. For $K > 2$, the limit of the stationary distribution of the load levels has a discontinuity at $\lambda = 1$. Such an irregular behavior was already observed in a simpler model by Malyshev and Robert [13]. The technical proof of Proposition 2 (cf. [2]) requires some analytical computations that we shall not reproduce here.

5. Mean field heuristics and simulation experiments. As was explained in the Introduction, the usual performance evaluation criteria of practical interest can be expressed in terms of the probabilities $P_i(\lambda)$ of the different load levels in the stationary limit. Our main objective in this section is the numerical evaluation of the $P_i(\lambda)$s, using the mean field heuristics, and the comparison with simulation experiments.

The idea of the mean field heuristics is to treat the stationary measure of an interacting particle system as if it was the product measure of local distribution at each site [in our case $(P_i(\lambda))_{0 \leq i \leq K}$]. This idea is widely used in the physics literature. For models similar to the one considered here, examples are reported, for instance, in [11] and [8]. De Masi, Ferrari and Lebowitz gave a rigorous justification by proving that if fast stirring is added to an interacting particle system, then the distribution stays close to a product measure (see [4] and [7] for precise definitions).

Consider the LTM on a finite set of sites or on a lattice of $\mathbb{Z}^d$. In both cases, there exists a unique stationary measure. That probability measure will be denoted by π and the corresponding expectation will be denoted by $\mathbb{E}_\pi$. If R is a subset of S, the set of configurations on R will be denoted by $X(R)$ and the restriction of π to R will be denoted by $\pi_{|R}$:

$$\forall \zeta \in X(R), \qquad \pi_{|R}[\,\zeta\,] = \pi[\{\eta \in \mathscr{A} \text{ s.t. } \eta(x) = \zeta(x)\ \forall x \in R\}].$$

We shall assume that some group of transformations operates transitively on (S, E), which is true for all cases of practical interest. Then

$$\forall x \in S, \forall i = 0, \dots, K, \qquad P_i(\lambda) = \pi_{|\{x\}}[i].$$

Applying the definition of the generator of the LTM to the indicator function of site x being in state i, $\mathbb{1}_{\eta(x)=i}$, yields the following local balance equation:

$$\mathbb{E}_\pi\left[a(x,\eta)\left(\mathbb{1}_{\eta(x)=i-1} - \mathbb{1}_{\eta(x)=i}\right) + d(x,\eta)\left(\mathbb{1}_{\eta(x)=i+1} - \mathbb{1}_{\eta(x)=i}\right)\right] = 0.$$

Since the rates depend only on the coordinates of η in $\mathscr{N}_K$, the expectation above can be written as a finite sum over configurations on $\mathscr{N}_K(x)$, the set of which is denoted by X':

$$\begin{aligned}&\sum_{\substack{\zeta\in X'\\ \zeta(x)=i}} a(x,\zeta)\pi_{|\mathscr{N}_K(x)}[\zeta] + \sum_{\substack{\zeta\in X'\\ \zeta(x)=i}} d(x,\zeta)\pi_{|\mathscr{N}_K(x)}[\zeta]\\ &\quad = \sum_{\substack{\zeta\in X'\\ \zeta(x)=i-1}} a(x,\zeta)\pi_{|\mathscr{N}_K(x)}[\zeta] + \sum_{\substack{\zeta\in X'\\ \zeta(x)=i+1}} d(x,\zeta)\pi_{|\mathscr{N}_K(x)}[\zeta].\end{aligned} \tag{5.1}$$

The mean field heuristics consists of replacing π by a product measure in the local balance equation (5.1):

$$\forall \zeta \in X', \qquad \pi_{|\mathscr{N}_K(x)}[\zeta] \leftrightarrow \prod_{y\in\mathscr{N}_K(x)} P_{\zeta(y)}(\lambda).$$

Thus a system of $K+1$ nonlinear equations in the $P_i(\lambda)$s is obtained (the mean field equations). For high values of K, this system will not be tractable except in very particular cases. For $K = 2$, it turns out to be quite simple. Actually, in that case the equations do not depend on the geometry of the graph except through the cardinality of a neighborhood. Therefore, we shall assume only that each site has exactly k neighbors. The equation for level 0 is

$$\begin{aligned}P_1(\lambda)(1-P_2(\lambda))^k &= \lambda P_0(\lambda)\\ &\quad + \lambda k P_1(\lambda)\sum_{l=1}^{k}\frac{1}{l}\binom{k-1}{l-1}(1-P_0(\lambda))^{k-l}P_0(\lambda)^l.\end{aligned} \tag{5.2}$$

To derive it, observe that a site can drop down from 1 to 0 only when there is a departure at that site and no neighbor can benefit from it (left-hand side). The site x can go from 0 to 1, either by a direct arrival on x or by an arrival on one of its k neighbors, say y, if it is at level 1. If that neighbor has l possibilities of transfer, it will choose x with probability $1/l$, which gives the right-hand side. The equation (5.2) is easily simplified into

$$P_1(\lambda)(1-P_2(\lambda))^k = \lambda(1-P_2(\lambda)) - \lambda P_1(\lambda)(1-P_0(\lambda))^k. \tag{5.3}$$

The equation for level 2 is similar:

$$\lambda P_1(\lambda)(1-P_0(\lambda))^k = (1-P_0(\lambda)) - P_1(\lambda)(1-P_2(\lambda))^k. \tag{5.4}$$

Combining these two equations, one gets

$$\lambda(1 - P_2(\lambda)) = (1 - P_0(\lambda)). \tag{5.5}$$

This is the global balance equation of the system: the left-hand side is the average number of tasks accepted per processor and per time unit, whereas the right-hand side is the average number of tasks treated. Setting $x = 1 - P_0(\lambda)$ leads to

$$(\lambda + \lambda^{-k})x^{k-1}(x(1 + 1/\lambda) - 1) = 1.$$

This equation has a single solution $x_k(\lambda)$ in the interval $]0, 1[$, which can be computed numerically. As k tends to infinity, $x_k(\lambda)$ tends to λ if $\lambda < 1$ and to 1 if $\lambda \geq 1$. Not surprisingly, the limit distribution obtained through the mean field equations tends to the asymptotics of the complete graph case (Proposition 2) as the number of neighbors per site tends to infinity.

More curiously, in the lattice case, the solution of the mean field equations turns out to be remarkably close to the distribution of levels estimated through simulation.

Simulations were run for three types of graphs (cycles, two dimensional toruses and hypercubes) and for different values of the parameters λ and K. The program uses the classical Metropolis algorithm. Our objective was to estimate the probabilities $P_i(\lambda)$ in the stationary regime. In practice, one has first to decide when the stationary regime is reached for a given simulation. We will define the *access time to equilibrium* to be the upper bound of the 99% confidence interval for the expectation of a certain random variable T. The definition of this random variable makes use of stochastic monotonicity (Section 2). Consider two independent copies of the LTM: $\{\eta_t, t \geq 0\}$ and $\{\zeta_t, t \geq 0\}$. One copy starts from the empty configuration and the other copy starts from the full configuration:

$$\forall x \in S, \qquad \eta_0(x) = 0, \qquad \zeta_0(x) = K.$$

Since the LTM is attractive, the total load (sum of coordinates) of η_t increases stochastically in time; that of ζ_t decreases. We define T as the first instant at which they meet:

$$T = \inf\left\{t > 0 \text{ s.t. } \sum_x \eta_t(x) \geq \sum_x \zeta_t(x)\right\}.$$

The expectation and standard deviation of T were first estimated over 1000 independent simulations. For subsequent experiments, the access time to equilibrium was fixed to the upper bound of the 99% confidence interval for the expectation $\mathbb{E}[T]$. For different types of graphs, the curves giving the access time to equilibrium as a function of λ turned out to be quite similar. They are of course unchanged through the transformation $\lambda \leftrightarrow 1/\lambda$. They increase very sharply as λ approaches 1. Figure 1 shows the access time to equilibrium as a function of λ for the hypercube with 32 vertices and $K = 5$. The probabilities $P_i(\lambda)$ in the stationary regime were then estimated using the classical procedure for Monte Carlo Markov chain methods. A simulation

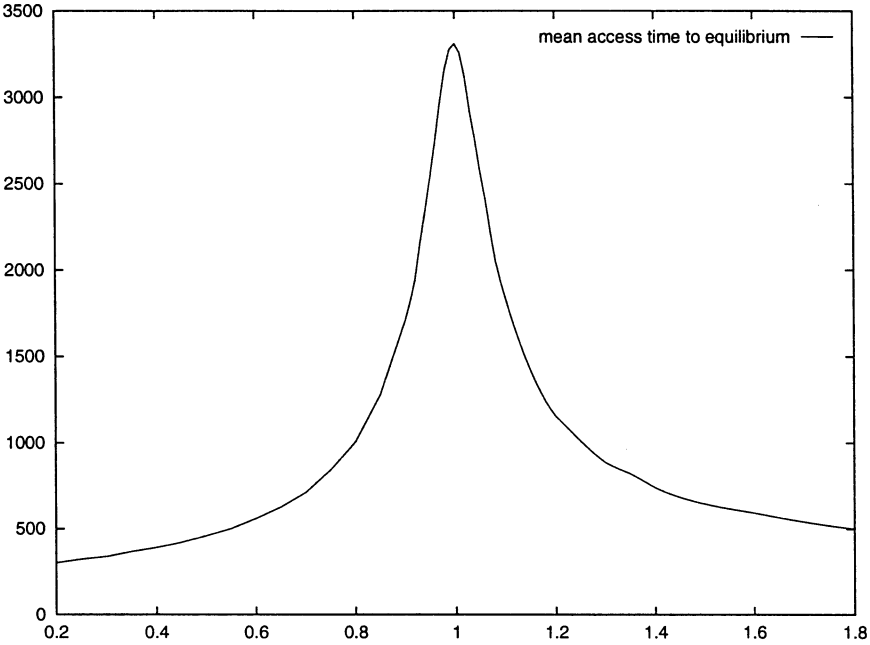

FIG. 1. *Access time to equilibrium as a function of λ for the hypercube $K = 5$, $d = 5$.*

was first run until the access time to equilibrium; then it was sampled at regular instants.

For the cycles and toruses, the values of $P_i(\lambda)$ converge very quickly as the number of sites tends to infinity, so the values to be presented below can be regarded as an estimate for the infinite lattice, since they were made based on simulations involving a large number of sites.

Figure 2 plots the rejection probabilities $P_K(\lambda)$ as a function of λ for different models and maximal capacity $K = 2$. The highest curve corresponds to the independent case; the lowest curve corresponds to the complete graph case (large size). The intermediate curves correspond to the cycle and the hypercube with 4096 sites. All of the curves nearly coincide for low and large values of λ. Figure 3 plots the mean response time as a function of λ for the same types of architectures and $K = 6$.

Some practical conclusions can be drawn from the theoretical and experimental studies. The benefits that one can expect from transferring can be approximated using the mean field equations or estimated by a Monte Carlo method on any type of architecture using the load transfer model. These benefits are maximal when the rate of arrival of tasks is close to their rate of treatment. Increasing the number of neighbors per processor improves the possibilities of transfer and, hence, increases the possible benefits. The differences between architectures decrease as the buffer capacity increases. This can be understood intuitively since a higher buffer capacity implies a

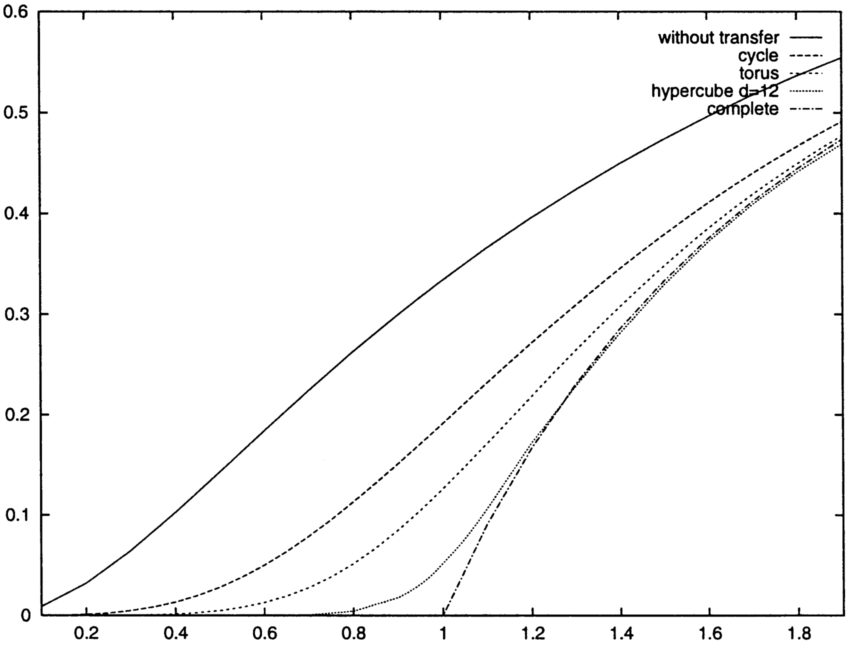

FIG. 2. *Rejection probabilities for $K = 2$.*

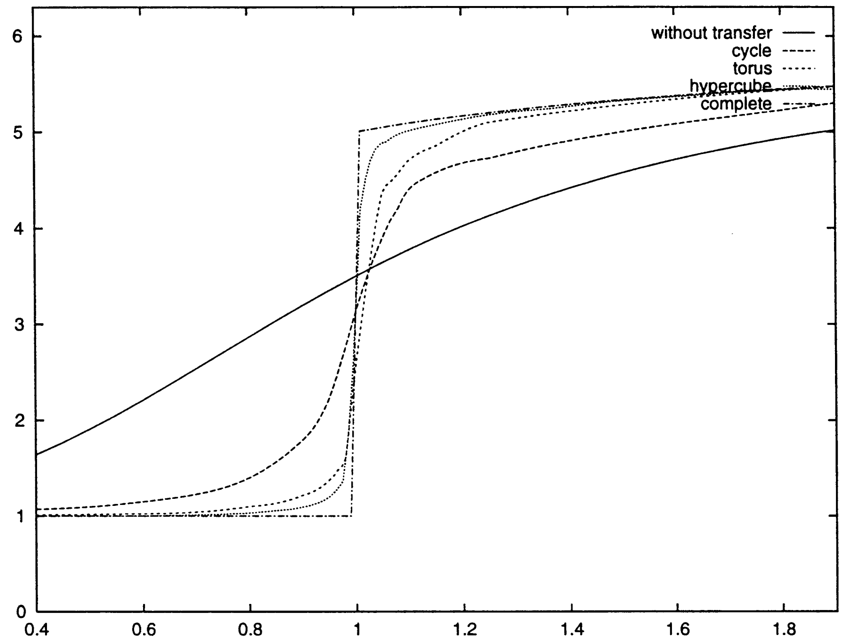

FIG. 3. *Mean response times for $K = 6$.*

longer range of interaction in the model and more possibilities of transfer. Further developments should include an evaluation of the cost of transfers by computing the expected number of transfers and by including the transfer time in the model. This will lead to different interacting particle models.

REFERENCES

[1] Barucha-Reid, A. T. (1960). *Elements of the Theory of Markov Processes and Their Applications*. McGraw-Hill, London.

[2] Béguin, M. (1997). Modèles markoviens de transferts de charge. Ph.D. thesis, Univ. Joseph Fourier, Grenoble, France, 1997.

[3] Bertzekas, D. P. and Tsitsiklis, J. N. (1989). *Parallel and Distributed Computation*. Prentice-Hall, Englewood Cliffs, NJ.

[4] De Masi, A. and Presutti, E. (1992). Mathematical methods and hydrodynamic limits. *Lecture Notes in Math*. **1501**. Springer, New York.

[5] Dhar, D., Ruelle, P., Sen, S. and Verma, D. N. (1995). Algebraic aspects of Abelian sandpile models. *J. Phys*. **28** 805–831.

[6] Durrett, R. T. (1995). Ten lectures on particle systems. *Ecole d'été de probabilité de Saint-Flour XXIII*. *Lecture Notes in Math*. **1608** 97–201. Springer, New York.

[7] Durrett, R. T. and Neuhauser, C. (1994). Particle systems and reaction-diffusion equations. *Ann. Probab*. **22** 289–333.

[8] Durrett, R. T. and Neuhauser, C. (1996). Coexistence results for some competition models. Unpublished manuscript.

[9] Forbes, F., François, O. and Ycart, B. (1996). Stochastic comparison for resource sharing models. *Markov Proc. Related Fields* **2** 581–605.

[10] Gates, D. J. and Westcott, M. (1993). Markov models of steady crystal growth. *Ann. Appl. Probab*. **3** 339–355.

[11] Janowski, S. A. and Laberge, C. A. (1993). Exact solution for a mean field Abelian sandpile. *J. Phys*. **26** 973–980.

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

[13] Malyshev, V. and Robert, P. (1994). Phase transition in a loss load sharing model. *Ann. Appl. Probab*. **4** 1161–1176.

[14] McNamara, B. and Wiesenfeld, K. (1990). Self-organized criticality in vector avalanche automata. *Phys. Rev*. **41** 1867–1873.

[15] Rosenblatt, M. (1959). Functions of a Markov process that are Markovian. *J. Math. and Mech*. **8** 585–596.

[16] Spies, F. (1996). Modeling of optimal load balancing strategy using queueing theory. *Microprocessing and Microprogramming* **41** 555–570.

[17] Walrand, J. (1989). *Introduction to Queuing Networks*. Prentice-Hall, Englewood Cliffs, NJ.

M. Beguin
B. Ycart
LMC / IMAG
BP 53
38041 Grenoble Cedex 9
France
E-mail: maryse.beguin@imag.fr
bernard.ycart@imag.fr

L. Gray
Department of Mathematics
University of Minnesota
Minneapolis, Minnesota 55455
E-mail: gray@math.umn.edu