Stochastic Performance Prediction for Iterative Algorithms in Distributed Environments

Henri Casanova
University of California at San Diego
Department of Computer Science and Engineering
La Jolla, CA 9209-0114, U.S.A
e-mail: casanova@cs.ucsd.edu

Michael G. Thomason
University of Tennessee
Department of Computer Science, 104 Ayres Hall
Knoxville, TN 37996-1301, U.S.A
e-mail: thomason@cs.utk.edu

Jack J. Dongarra
University of Tennessee
Department of Computer Science, 104 Ayres Hall
Knoxville, TN 37996-1301, U.S.A
e-mail: dongarra@cs.utk.edu
and
Oak Ridge National Laboratory
Mathematical Science Section, PO Box 2008, Building 6012
Oak Ridge, TN 37821-6367, U.S.A
e-mail: dongarra@msr.epm.ornl.gov

Running Head:
Performance Prediction for Iterative Algorithms

Author to whom proofs should be sent:
Henri Casanova
UCSD, Dept. 0114
9500 Gilman Dr.
LA JOLLA, CA 92093-0114
e-mail: casanova@cs.ucsd.edu
ph: (619)-534-5913
fax: (619)-534-7029

Abstract:
The parallelization of iterative algorithms is an important issue for efficient solution of large numerical problems. Several theoretical results concerning sufficient conditions for, and speed of, convergence of parallel iterative algorithms are available. However, those results usually do not take into account the processor workloads and network communications at the application level. The approach in this paper develops a Markov chain based on random variables which describe aspects of the multi-user, distributed-memory environment and the phases of the algorithm. The performance characterization addresses stochastic characteristics of the algorithmic execution time such as mean values and standard deviations. We present simulation results as well as experimental results over different time periods. The results provide information about the impact of distributed environment and implementation style on long-run, expected execution time characteristics.

Keywords:
convergence, distributed-memory, iterative algorithm, Markov chain, multi-user, parallel, performance prediction, stochastic modeling

1 Introduction

Iterative algorithms are widely used in different areas of science and engineering, e.g., medical imaging [22] and network flow in electrical networks, communication networks and financial models [5]. A broad class of iterative algorithms aims at finding a fixed point of a given operator. Many well-know numerical methods use such an algorithm with linear or non-linear operators. For problems with large dimension and/or extensive numerical computation for each component of the solution vector at each iteration (e.g., gradient approximations or Hessian computations for non-linear operators), it is natural to consider parallel implementations of iterative algorithms.

Analyzing the behavior and thus the performance of a parallelized iterative algorithm in a distributed environment is not an easy task. The amount of computation performed by a processor to update a component of the solution vector often is not known a priori and in non-linear cases may depend on the operator. For instance, in popular algorithms to compute approximations of the gradient of the operator at each iteration (e.g., gradient descent algorithms), the amount of computation required to approximate the value of a component of the gradient (say the i-th component) depends on the *shape* (or *geometry*) of the operator along direction i at the current solution vector. Furthermore, if distributed iteration is implemented on a non-dedicated system with other users, the computation and communication of those users impact the availability of processor and network resources to the iterative algorithm.

In this research, we focus on non-dedicated, distributed-memory environments such as clusters of processors on a network. There are existing convergence results that indirectly support a quantitative assessment of the parallel algorithm convergence rate, but almost all these results are purely theoretical and do not take into account the nature of the multi-user, distributed environment itself. Commonly available results are lower bounds on the algorithm theoretical rate of convergence.

Due to non-determinism (randomness) both in network communications and in computational workloads at processors, stochastic methods appear to be a natural way to move towards more comprehensive models. These models should capture the fluctuations of the distributed environment and the algorithm, and their impact on the user's implementation, at least in terms of average or "expected" characteristics of execution time conditioned on the assumptions made to obtain tractable models. In the following sections, we give definitions and assumptions concerning parallel iterative algorithms in multi-user, distributed-memory environments; introduce application-level models of the distributed environment and the algorithm which lead to a finite-state, time-homogeneous Markov chain; discuss performance characterizations related to convergence; and describe results of simulations and experiments.

2 Parallel Iterative Algorithms

We use two, well-established paradigms for parallel implementation of iterative algorithms: *synchronous* and *asynchronous*. In [27], it is shown that asynchronous implementations have "good" communication complexity as compared to synchronous ones, but it is difficult to use these results to obtain quantitative estimates of actual performance in a given distributed environment. A performance comparison between the two paradigms is provided in [8, Section 6.3.5], but the model is non-random and does not describe non-deterministic systems.

A unique reference that proposes a stochastic approach is [30] which gives an analysis of asynchronous iteration with expected values based on [18]. This work is of interest for shared-memory homogeneous implementations and an *age scheduling* strategy for which processors' execution times are described by a specific family of probability distributions (increasing failure rate (IFR) functions) with simulation used to approximate some parameters. In contrast, this paper focuses on distributed-memory systems that may be non-homogeneous, and assumes *static scheduling* (which simulations in [30] show to be superior) to develop a Markov chain model.

We consider iterative algorithms in which an operator is applied repetitively to a vector of real-valued data until some convergence criteria are met. Let $\mathbb{R}$ denote the reals and $\mathbb{N}$ the non-negative integers $\{0, 1, 2, \dots\}$. The computation of the algorithm is a sequence of vectors $\{x(t)\}$ in $\mathbb{R}^m$, and the iteration can be written as:

$$\begin{cases} x(0) \in \mathbb{R}^m \\ x(t+1) = Op(x(t)) \in \mathbb{R}^m \quad \text{for all } t \in \mathbb{N}. \end{cases} \tag{1}$$

If the algorithm converges, the sequence $\{x(t)\}$ converges to a fixed point of operator Op. Much work has been devoted to finding useful operators for specific problems and finding operators that provide the highest convergence rates (cf. [24]).

In this research, we consider the general iterative equation Equ. (1) without looking at details of the numerical method(s) implemented to compute Op. The iteration to update the components of the solution vector $x(t)$ is distributed among a collection of processors. As previously stated, we consider only *static scheduling*, meaning that each processor updates one *piece* of $x(t)$ which is a preassigned, fixed subset of the components of vector $x(t)$.

2.1 Synchronism vs. Asynchronism

Synchronous implementations of iterative algorithms are straightforward parallelizations of sequential implementations. This makes them attractive as their convergence properties are thus well known. It is often easy to convert a sequential implementation of a given algorithm into a synchronous parallel implementation. Assume that the distributed environment used to execute the algorithm consists of p processors. Each processor can access its local memory and communicate with any other processor via a network. Each processor starts each iteration with the entire current solution vector in its memory and updates its subset of $x(t)$ by applying part of the operator Op to the entire vector. The processors then perform an *all-to-all* communication, exchange their up-to-date subsets of the solution vector, and proceed to the next iteration. More formally, if the components of the solution vector $x(t)$ are denoted by $x_i(t)$, $i = 1, ..., m$ and if the components of $Op(x(t))$ are denoted by $Op_i(x_1(t), ..., x_m(t))$, the synchronous iteration can be written as:

$$x_i(t+1) = Op_i(x_1(t), ..., x_m(t)) \quad \text{for } 1 \leq i \leq m \text{ and all } t \in \mathbb{N}. \tag{2}$$

The most obvious performance bottleneck in synchronous implementations is the all-to-all communication phase. First, for slow networks, having to exchange $(p-1)^2$ messages at each iteration can be prohibitive [22]. Improved network technologies make this less a concern in many applications today, e.g., for implementations that transmit reasonably small messages on a fast local area network, but it remains a factor. Second, and more importantly for this paper, the possible lack of synchronization among the processors [25, 19, 10, 15] can lead to serious performance losses because relatively fast processors may be idle for large percentages of real-time while awaiting slower processors. This lack of synchronization tends to become particularly prominent when the iterative algorithm is run on a non-dedicated cluster of workstations with multiple users. This phenomenon is a clear motivation for studying asynchronous implementations.

The study of asynchronous implementations started as early as 1969 [12] and has been the object of many extensions and generalizations. As for the synchronous case, we assume that there are p processors in the distributed environment and that the solution vector is segmented in pieces (subsets of components of $x(t)$) assigned to each processor, i.e., static scheduling. By contrast with the synchronous implementation, there is no all-to-all communication phase to synchronize the processors. Instead, a processor may perform more than one update between communications, possibly using *out-of-date data* for the subsets from the other processors. Each processor must at times communicate its most up-to-date values for its subset to other processors.

A formal description of the asynchronous iteration is given in [3] and is inspired by the definition of chaotic relaxations in [12]. The definition we give here is very similar: for $1 \leq i \leq m$ and $t = 1, 2, \ldots,$

$$x_i(t) = \begin{cases} x_i(t-1) & \text{if } i \notin J_t \\ Op_i(x_1(s_1(t)), ..., x_m(s_m(t))) & \text{if } i \in J_t \end{cases} \tag{3}$$

where J_t is a subset of $\{1, \ldots, m\}$, and $s_i(t)$ is in $\mathbb{N}$. We adopt three additional conditions for asynchronous iteration proposed in [3]:

Condition 2.1 *For* $1 \leq i \leq m$*:*

(i) $s_i(t) \leq t$ *for all* $t = 1, 2, ...$

(ii) $\lim_{t\to\infty}(s_i(t)) = \infty.$

(iii) i *occurs infinitely often in the sets* J_t, $t = 1, 2, ...$

Condition **(i)** states that when a processor updates a component of the solution vector, it can only use previously computed components. Condition **(ii)** states that the same value for a component cannot be used indefinitely when computing updates. Condition **(iii)** requires that a processor not abandon a component forever. In the formal definition of asynchronous iterations that we have given so far, there is no limit on the amount by which a component used in an update can be out-of-date. If there is no upper bound on this amount, the implementation is referred to as *totally asynchronous* [8]; otherwise, the implementation is said to be *partially asynchronous*. Actual implementations are often partially asynchronous since it is often practical to fix some kind of bound on the asynchronism for implementation purposes.

2.2 Convergence

The definition of the asynchronous iteration shows clearly that the algorithm can be as "asynchronous as needed" to take advantage of the very phenomena that are performance bottlenecks for a synchronous implementation. However, convergence of the algorithm is no longer implied by the same conditions as for a sequential implementation and its convergence rate must also be reexamined.

Work in analyzing the convergence of asynchronous parallel iterative algorithms includes [8, 7, 12, 21, 20, 3, 6, 4, 26, 14, 28, 29]. Some of the earliest work focused on specific iterative algorithms or on specific implementations. A sufficient condition for convergence for linear operators is available in [12], only for partially asynchronous implementations. In [21, 20], this sufficient condition is generalized to the case of certain non-linear operators, still in a partially asynchronous setting. A recent and general theorem in [8] gives a sufficient condition for convergence of asynchronous iterative algorithms based on a sequence of subsets of $\mathbb{R}^m$ (a "box condition"). The applications given in [8] are contractions or pseudo-contractions with respect to a weighted maximum norm (traditional "Lipschitz-like" properties detailed in [8]), and it may be difficult to fully exploit the generality of the theorem in practical situations. A lower bound on rate of convergence is obtained under additional assumptions.

A fundamental reference on which to develop a stochastic approach is Baudet's work [3]. That work contains a theorem establishing the convergence of asynchronous iterations for *contracting operators* defined by the following "Lipschitz-like" condition:

Definition 2.1 *An operator Op from $\mathbb{R}^m$ to $\mathbb{R}^m$ is said to be contracting on a subset D of $\mathbb{R}^m$ if there exists a nonnegative $m \times m$ matrix A such that*

$$\forall x, y \in D \quad |Op(x) - Op(y)| \leq A|x - y|, \qquad \textit{component-wise}$$

and $\rho(A) \leq 1$ where $\rho(A)$ denotes the spectral radius of A (i.e. the magnitude of A's largest magnitude eigenvalue).

Furthermore, [3] provides a lower bound on the convergence rate of the algorithm defined traditionally as:

$$\mathcal{R} \stackrel{\Delta}{=} \liminf_{t\to\infty}[(-\log\|x(t) - \xi\|)/t] \tag{4}$$

where $\|.\|$ denotes a norm of $\mathbb{R}^m$ and ξ the fixed point of the operator. This definition of the rate of convergence has an immediate interpretation: if the logarithm is in base 10, then $1/\mathcal{R}$ measures the asymptotic number of *iterations* required to divide the initial error by a factor of 10 where an iteration is the computation described by Equ. (3) for all i. Without any additional assumptions, it is shown that

$$\mathcal{R} \geq -[\liminf_{t\to\infty}(k_t/t)] \log \rho(A). \tag{5}$$

where $\{k_t\}$ is a sequence of integers defined in [3]. With reference to [3] for details, we note here that this sequence $\{k_t\}$ is increasing, and the more asynchronous the implementation, the less rapid the increase.

The insight into performance provided by Equ. (5) is in the form of a lower bound on the theoretical rate of convergence. Experiments in [3] indicate that the bound is very conservative, and the possibility for stochastic approaches is mentioned. Our goal is to introduce tractable stochastic models and obtain convergence results relevant to the user for a practical purpose: deciding which implementation is the best for a given multi-user, distributed-memory environment in terms of long-run, average performance.

3 Application-level Models

We introduce two models to describe the distributed environment and the algorithm at an application-level. Some of the low-level elements of the computer system are ignored or approximated to develop and analyze the models.

3.1 Modeling the Distributed Environment

We assume that the distributed environment is a computer network of p nodes connected by a communication facility. A node is composed of a processor, memory and a network interface. Each node has its own memory accessed only by its processor. In this distributed-memory setting, nodes can exchange data via the communication facility, thanks to their network interfaces. We do not require that all the nodes be identical, i.e., the environment may be *heterogeneous*, but do make several strong assumptions below. The communication facility is seen as an abstract device that allows reliable point-to-point communication between any two nodes of the network and we do not make any assumptions about the network topology. Our model is therefore applicable in diverse environments, from a Massively Parallel Processor (MPP) system to an Internet-wide collection of machines.

The performance of the network in terms of transmission speed is modeled by a random variable (RV) for each point-to-point data path (for a total of $p(p-1)$ RVs). Similarly, the performance of each node in terms of local computation is modeled by a RV that describes the time that node spends to perform one update of its subset of the solution vector (for a total of p RVs). The distributions of all these RVs describe the behavior of the algorithm execution in the distributed environment. In order to introduce tractable stochastic models, these RVs are assumed independent and stationary during the run of the iterative algorithm; however, any finitely-specified, discrete probability distributions can be used—empirically estimated, analytically specified, or arbitrarily chosen.

3.2 Modeling the Algorithm

The algorithm is partitioned into *phases*. Figure 1 depicts one phase for three processors *i, i+1, i+2*. Note that real-time intervals for the same phase of the algorithm do not generally coincide at different processors. Each phase is composed of two *sub-phases*, α and β. During the α *sub-phase*, a processor performs successive updates on its subset of the solution vector. All updates performed by a processor during its α sub-phase, beyond the first update, use out-of-date data until the next messages are received from the other processors. At the end of the α sub-phase, a processor broadcasts its subset of the current solution vector to all the other processors. Just after this broadcast, its β *sub-phase* starts and the processor expects $p-1$ messages from the other processors. During its β sub-phase, a processor may perform additional updates (up to a preset number) on its subset of the solution vector. If any updates are performed during this sub-phase, they will also use out-of-date data. A processor finishes its β sub-phase when it has received all the $p-1$ messages; it then moves onto the next algorithm phase.

For each processor, the user must choose the number of updates to be performed during the α sub-phase and the maximum number of updates allowed during the β sub-phase. The larger those numbers, the more asynchronous the algorithm. A more asynchronous implementation usually implies a lower convergence rate but better use of computational resources (less processor idle-time). A method to evaluate this trade-off in terms of the expected execution time is an objective of this research.

Note that the model includes synchronism as the special case that, for each processor, there is a single update during its α sub-phase and no update during its β sub-phase. In the next section, we give several definitions and assumptions, then define a Markov chain (cf. [16]) of interest.

3.3 Underlying Markov Chain

Let us define two non-negative, integer-valued constants and three non-negative, real-valued RVs:

Definition 3.1

(i) *User-specified, integer constant $A_i > 0$ denotes the number of updates performed by processor i during its α sub-phase of each algorithm phase.*

(ii) *User-specified, integer constant $B_i \geq 0$ denotes the maximum number of updates that processor i is allowed to perform during its β sub-phase of each algorithm phase.*

(iii) *RV $\alpha^i(k) \in \mathbb{R}$ is the duration in seconds of the α sub-phase of the k^{th} algorithm phase on processor i.*

(iv) *RV $n_{i\to j}(k) \in \mathbb{R}$ is the duration in seconds of the message transfer from processor i to processor j during the k^{th} algorithm phase. By convention, processor i sends a message to itself at each phase and $n_{i\to i}(k) = 0$ for all i and k.*

(v) *RV $T^i(k) \in \mathbb{R}$ denotes the time of the beginning of the k^{th} algorithm phase on processor i.*

We assume independence of RVs but do not require identical distributions for all processors i. We first assume that, given i, the RVs $\{\alpha^i(k)\}$ are independent and identically distributed (i.i.d) for all k. Independence of RVs means that the computational time $\alpha^i(k)$ for processor i to perform one update does not exhibit a dependence on update-times $\{\alpha^j(k), j \neq i\}$ at other processors. (This assumption is violated, e.g., when workloads of two or more processors are correlated, due for instance to other parallel applications sharing the resources. Relaxing this independence assumption is a topic for future development.) CHECK STATEMENTS HERE!

Similarly for given i and j, the RVs $\{n_{i\to j}(k)\}$ are i.i.d for each k; but it is not required that the distribution of $n_{i\to j}(k)$ be the same as that of $n_{h\to \ell}(k)$ for $i \neq h$ or $j \neq \ell$.

Although these are strong assumptions, independence and identical distribution of RVs are widely used in computational models (cf. [30, 1]). The experimental results presented in Section 6 illuminate some aspects of validity and limitations and Section 7 introduces new research directions to loosen the assumptions. We also assume that network times are bounded, that messages are sent at exactly the same time during a broadcast, and that message-sends are free in terms of CPU cycles on the sending processor. The last two simplifying assumptions could be removed by the use of additional RVs at the cost of more complicated modeling.

Recall that $T^i(k) \in \mathbb{R}$ denotes the time of the beginning of the k^{th} phase of the algorithm on processor i. Taking arbitrary processor 1 as reference, we define the *wavefront* $X(k)$ as

$$\begin{aligned} X(k) &= (X_1(k), X_2(k), ..., X_p(k)) \\ &= (0, T^2(k) - T^1(k), ..., T^p(k) - T^1(k)) \in \mathbb{R}^p. \end{aligned} \tag{6}$$

$X(k)$ describes the shape of the line joining the starting times of each processor in the k^{th} phase. $X(k)$ is represented on figure 1 as a thick line, and it is shown in Appendix A that for each processor i,

$$\begin{aligned} X_i(k+1) = &\max_{j\in\{1,..,p\}} [X_j(k) + \alpha^j(k) + n_{j\to i}(k)] - \\ &\max_{j\in\{1,..,p\}} [X_j(k) + \alpha^j(k) + n_{j\to 1}(k)]. \end{aligned} \tag{7}$$

The wavefront is the key to the stochastic model because, based on the assumption of i.i.d. RVs, Equ. (7) defines a Markov process; specifically, the wavefront $X(k)$ is a time-homogeneous Markov chain. It is shown in Appendix B that, with a few technical assumptions, the Markov chain is finite-state. Equ. (7) can be used to compute the transition probability matrix $\mathbf{P}$ of the Markov chain by using the discrete distributions of the different RVs. Examples of the transition matrix $\mathbf{P}$ can be found in [11].

Since the finite-state chain is recurrent, it has a stationary distribution which describes the behavior of the chain in the long-run (cf [16, 2, 9]). We refer to the probabilities in this distribution as the π-values. π_s (> 0) for state s is the long-run, relative-frequency of occurrence of state s in realizations of the chain.

4 Performance Characterization

The wavefront Markov chain is exploited to obtain performance results for the parallel iterative algorithm. The goal is to obtain information on the average execution time. The execution time can be computed as the ratio of the number of iterations to perform over the number of iterations performed per time unit. The number of iterations required can be approximated using the asymptotic convergence rate of the algorithm. This is the topic of the next section.

4.1 Asymptotic Rate of Convergence

The challenge here is to refine the lower bound on the asymptotic rate of convergence in [3] by modified estimates that take into account randomness at the application-level. One can compute three estimates (see Appendix D) respectively called *worst-*, *average-*, and *best-*cases and denoted respectively as $\underline{\mathcal{R}}$, $\widehat{\mathcal{R}}$, and $\overline{\mathcal{R}}$. The usefulness of the average- and best-case estimates is yet to be demonstrated; however, the worst-case estimate is a straightforward improvement over the lower bound in [3] as (i) it is higher than Baudet's and (ii) it is still a lower bound on the asymptotic convergence rate.

Let $\mathcal{R}^*$ denote one of the three estimates for this rate. Then a user who wants the initial error on the solution vector to be divided by a factor of 10^ω may approximate the number of iterations needed as $\omega/\mathcal{R}^*$ assuming that ω is reasonably large. The value of ω chosen by the user is the *convergence criterion*: the larger ω, the smaller the final error. Of course, the choice of $\mathcal{R}^*$ is crucial, and we expect that the three estimates will provide information about this choice.

4.2 Execution Speed

We can use $\mathcal{R}^*$ to estimate the speed of the execution in terms of number of iterations performed per time unit. Let $\Phi(k)$ denote the duration in seconds of the k^{th} algorithm phase on processor 1. Let $N(k)$ denote the number of iterations performed during that phase. Both $\Phi(k)$ and $N(k)$ are RVs and their long-run probability distributions can be approximated by making use of the wavefront π-values. Indeed, the probability distributions of RVs of interest in the model are all conditioned on the wavefront state. The π-values are used to replace those conditional probabilities by unconditional probabilities, and a simple convolution is used to obtain the long-run estimate.

The speed of the execution is entirely described by sums of the random vector $\binom{N(k)}{\Phi(k)}$ for successive values of k. This vector can be used in different ways, as described in the next section.

4.3 Performance Characterization Levels

<u>Level 1:</u> This level provides an estimate of the execution time mean value. Using the Strong Law of Large Numbers [16], it is possible to compute the asymptotic algorithm speed as the ratio of the limiting expected values of $N(k)$ and $\Phi(k)$. Denoting this ratio as $\mathcal{S}$, one obtains Θ_1, an asymptotic estimate for the execution time expected value:

$$\Theta_1 = \frac{\omega}{\mathcal{S}\mathcal{R}^*}.$$

<u>Level 2:</u> This level provides an estimate of the execution time standard deviation. The derivation of the estimate is detailed in Appendix C.

5 Simulation

Simulation results are summarized here for a Gradient algorithm for a multi-polynomial cost function with thirty variables (see also [11, Section 5.1]). The distributed environment consists of three processors with the different update-time distributions depicted in Figure 2 obtained by sampling the distributions of actual workstations. The three processors are interconnected by a network that delivers constant, non-random performance, i.e., given i and j, $n_{i \to j}(k)$ is constant for all k and, by convention in Definition 3.1(**iv**), is 0 for $i = j$. Three implementations are simulated: *synchronous* with $A_i = 1$ and $B_i = 0$ for all processors i; *asynchronous* (Async. 1) with $A_i = B_i = 1$ for all processors i; and *more asynchronous* (Async. 2) with $A_i = 1$ and $B_i = 2$ for all processors i. The assumptions set forth above are not violated in these simulations.

The simulations provide information about accuracy and sensitivity of the two levels for the environment described. First, we observe that the new estimates of asymptotic convergence rate are improvements on estimates in [3]. Table 1 shows the relative errors between the different estimates and the observed convergence rate for different implementations in the simulated environment. The gaps between the four estimates increase with asynchronicity. It is to be noted that no estimate exactly predicts the observed convergence rate. A primary reason is that the estimates depend only on the spectral radius of the matrix associated to the contracting operator, but not on the actual shape of that operator. Therefore, the same estimates will be computed for different operators that happen to have the same contracting matrix.

The simulations also show that level 2 characterization is rather sensitive. Indeed, as it is based on a binomial Gaussian approximation, its accuracy as an approximation generally increases with a larger number of samples. It is shown in [11] that the error between level 2 characterization and the observed standard deviation decreases for increasing values of ω (recall that larger ω implies more iterations performed).

Figure 3 shows the simulation versus the characterization for an asynchronous implementation in the simulated heterogeneous distributed environment. On each graph, the empirical distribution of the execution time is shown as a bar diagram labeled "simulation." The empirical mean is shown as a vertical solid line and the empirical standard deviation is represented as a horizontal line segment inside bold vertical lines on each side of the empirical mean. Level 1 characterization is shown as a dashed vertical line. Level 2 characterization is shown as two horizontal dashed line segments in bold vertical lines on each side of level 1. We show four characterizations — Baudet [3], $\underline{\mathcal{R}}$, $\widehat{\mathcal{R}}$, $\overline{\mathcal{R}}$ — each corresponding to a different asymptotic convergence rate estimate. Note that the average-case estimate is a dramatic improvement over the estimate in [3] for the execution time distribution (22% vs. 214% error).

6 Experiment

This section presents experimental results [11] with a gradient algorithm for a real multi-polynomial cost function with thirty variables. The algorithm runs in parallel on three Sun Sparc Ultra 1 workstations interconnected by a standard 10 Mbps Ethernet. Those workstations are being used by students for course-work as well as for personal research. Measurements were obtained throughout one week (Nov. 17-24, 1997).

6.1 Preliminary Remarks

Figure 4 shows the execution times observed throughout the whole week for the synchronous implementation and for the first asynchronous implementation (*mildly* asynchronous: at most one update performed in β sub-phase for each processor). This corresponds to 862 observations for each implementation. The measurements for the second asynchronous implementation (*fairly* asynchronous: at most two update performed in β sub-phase for each processor) are not shown on Figure 4 because they would be difficult to distinguish from the ones of the first asynchronous implementation on that time scale. The three different implementations were run in a round-robin fashion, each run using the three workstations and taking about 2 minutes.

The first observation is that the asynchronous implementations are generally more efficient than the synchronous one. The first asynchronous implementation is up to 150 seconds faster than the synchronous implementation, and 30 seconds faster on average. On average the second implementation is faster than the first one by about 1.9 seconds. But in only 15% of the observations is the absolute difference between between the two implementations more than 10 seconds.

It seems that, in this environment, a good choice is an asynchronous implementation as opposed to a synchronous one. However, a *mild* asynchronicity is sufficient to obtain improvement over a synchronous implementation. This can be explained both by the nature of the distributed environment and by the nature of the iterative algorithm. Several other references also include examples for which asynchronous implementations outperform synchronous ones [3, 5, 22].

A fundamental observation on Figure 4 is that the execution time is *bursty*. In fact, the distributed environment, and therefore the algorithm, behaves very differently at different times in the time period, for the system is in use for a variety purposes during the experimental runs. In order to illustrate these different behaviors, Figures 5(a), (b) and (c) show three close-ups of the execution times for each implementation during three short sub-periods about two hours long. The distributed environment appears to exhibit distinct modes during the week, wich violates our i.i.d. assumption.

Figure 7 shows the execution times for the parallel iterative algorithm throughout a 24-hour time period at the end of the week. The distributed environment exhibited a fairly stable behavior, leading to relatively smoother observations. We expect the stochastic models to yield better results for the 24-hour time period than for the entire week as burstiness indicates violation of the stationarity assumption. We do not claim that any 24-hour time period would lead to better result (as the execution time may be bursty on many time scales). We merely chose to highlight a subset of the time line that exhibited close-to-stationary behavior to evaluate how the model would perform in a more stable environment. In the following sections, we present and comment on some of the results for both time periods.

6.2 Applying the Model

6.2.1 The One Week Time Period

The workload distributions of the three processors were sampled throughout the week (roughly every 20 seconds) and are shown in figure 6. Figure 8 shows the results for the synchronous implementation. The empirical distribution is clearly multi-modal as already seen in Section 6.1. The level 1 characterization makes an error of about 17% in predicting the mean of the execution time. The level 2 characterization underestimates the observed standard deviation by a factor of 50. In fact, level 2 is very sensitive to the violation of the stationary assumption as a Gaussian approximation is involved.

Figure 9 shows the results for the mildly asynchronous implementation. Four characterizations are shown, one of each estimate of the asymptotic convergence rate. If one uses the average-case estimate (see Section 4.1), then the error on the mean prediction is only 4% (as opposed to 116% with the estimate in [3]). The observations made on Figure 8 about level 2 are still valid. The results for the second asynchronous implementation are not shown here as they are fairly similar to the results for the first asynchronous implementation. The next section reduces the time period to 24 hours and should lead to improvements, especially for the level 2 characterization as the stationary assumption should be less violated.

6.2.2 The 24 Hours Time Period

Figure 10, for a synchronous implementation, demonstrates that the level 2 characterization becomes much more accurate for this shorter, more nearly stationary time period. Its error is here about 27% whereas it was a factor of 50 for the whole week. Similar improvements were observed for all implementations. Furthermore, the level 1 characterization is more informative than for the one week time period. Even though the level 1 characterizations do not exactly agree with the observed mean, they are sufficient for purposes of comparisons. In this experiment, the level 1 characterizations computed in [11] imply that an asynchronous implementation will outperform a synchronous one on average by 40 seconds (which agrees with the experimental results on Figure 7).

7 Conclusion and Future Work

Parallelizing iterative algorithms for the solution of complex problems is a crucial issue. The amount of computation required to solve such problems may be prohibitive for a sequential implementation, especially in non-linear cases. We use synchronous and asynchronous paradigms for parallel iteration and develop stochastic models that take characteristics of multi-user, distributed-memory environments into account. The models are used to obtain performance characterizations that are directly meaningful to the end-user. In practice, a user must provide the probability distributions for network and CPU loads (as finitely-specified, discrete distributions obtained through sampling or from tools such as the NWS [31], for instance) and employ the model for comparisons among degrees of asynchronism in implementations in terms of the estimates on algorithmic convergence rates, in particular, for the worst-case.

An additional level of performance characterization based on Large Deviation Theory is developed in [11] and will be reported in the future. Further research may also incorporate Markov-modulated random processes to model bursty behaviors in more detail [23, 13, 17] and may investigate more complex models by introducing dependences among RVs.

Acknowledgement

The authors thank the reviewers for their constructive criticism and suggestions.

Appendix

A Wavefront Equation

Let $\beta^i(k) \in \mathbb{R}$ be the the duration in seconds of the β sub-phase of the k^{th} algorithm phase on processor i. $\beta^i(k)$ is a RV and it can be computed as follows. The β sub-phase of the k^{th} algorithm phase on processor i clearly starts at time $\beta^i_{start}(k) = T^i(k) + \alpha^i(k)$. It ends when the last expected message has been received by processor i. The message expected from processor j is received by processor i at time $\beta^j_{start}(k) + n_{j\to i}(k)$. Therefore,

$$\begin{aligned} \beta^i(k) &= \max_{j\in\{1,..,p\}}[\beta^j_{start}(k) + n_{j\to i}(k)] - \beta^i_{start}(k) \\ &= \max_{j\in\{1,..,p\}}[\beta^j_{start}(k) + n_{j\to i}(k) - \beta^i_{start}(k)]. \end{aligned}$$

Since $n_{i\to i}(k) = 0$, one obtains:

$$\beta^i(k) = \max[0, \max_{j\in\{1,..,p\}-\{i\}}(T^j(k) + \alpha^j(k) - T^i(k) - \alpha^i(k) + n_{j\to i}(k))].$$

By definition 3.1(**v**) and Equ. (6):

$$X^i(k+1) = X^i(k) + \alpha^i(k) + \beta^i(k) - \alpha^1(k) - \beta^1(k).$$

Replacing $\beta^i(k)$ and $\beta^j(k)$ by their expression and using the fact that

$$\forall x, y \in \mathbb{R} \quad \max(0, x-y) + y = \max(x, y),$$

one obtains Equ. (7). ∎

B Wavefront State-Space

Lemma B.1 $\exists M, \forall k \geq 1 \quad \|X(k)\|_\infty \leq M.$

Proof. Let us consider processor 1 and processor $i \neq 1$ during the k^{th} algorithm phase. The times at which these two processors receive a message from a processor h are apart by $|n_{h\to 1}(k) - n_{h\to i}(k)|$ seconds since we assume that processor h sends all messages at the exact same time. Therefore, the times at which processors 1 and i receive the last messages they were expecting are apart by at most $\max_{h\in\{1,..,p\}}(|n_{h\to 1}(k) - n_{h\to i}(k)|)$. The communication times are assumed to be bounded above and below as:

$$\forall s, d \quad \exists \underline{n_{s\to d}}, \overline{n_{s\to d}}, \quad \forall k \quad \underline{n_{s\to d}} \leq n_{s\to d}(k) \leq \overline{n_{s\to d}}.$$

One can then write:

$$\begin{aligned}\forall h \quad |n_{h\to 1}(k) - n_{h\to i}(k)| &\leq \max(\overline{n_{h\to 1}} - \underline{n_{h\to i}}, \overline{n_{h\to i}} - \underline{n_{h\to 1}}) \\ &\leq \max(\overline{n_{h\to 1}}, \overline{n_{h\to i}}) \\ &\leq \max_{j\in\{1,..,p\}}(\overline{n_{h\to j}}).\end{aligned}$$

This implies that the times at which processors 1 and i receive the last message they were expecting are apart by at most $\max_{h,j\in\{1,..,p\}}(\overline{n_{h\to j}})$. But those times are also apart by $X_i(k)$ according to definition 3.1(**v**) and Equ. (6). Since $\|X(k+1)\|_\infty \equiv \max_{i\in\{1,..,p\}}|X_i(k+1)|$, the proof is complete. ∎

The preceding establishes that the wavefront vector is in a closed ball of $\mathbb{R}^p$. If one assumes that $\alpha^i(k)$, $n_{i\to j}(k)$, and the components of $X(0)$ are rational (in $\mathbb{Q}$), then for each k, $X(k)$ is in a finite subset of $\mathbb{R}^p$ that does not depend on k. Those assumptions are really purely technical: the data being manipulated is in $\mathbb{Q}$ since it is processed by computers with finite arithmetic. The size of the state-space of the wavefront Markov chain depends on the finitely-specified, discrete distributions of $\alpha^i(k)$ and $n_{i\to j}(k)$ [11]. The π-values for the chain are the unique values in the solution of the linear system $\Pi = \Pi\mathbf{P}$ where $\mathbf{P}$ is the transition probability matrix, $\sum_s \pi_s = 1$, and $\pi_s > 0$ for every state s [16, 2, 9].

C Level 2 Characterization

One can make a binomial Gaussian approximation of the distribution of the random vector $\binom{N(k)}{\Phi(k)}$ (with covariance matrix C). The covariance matrix of the sum of those vectors for each algorithm phase until convergence, C', can then be estimated as

$$C' = \frac{\omega}{\mathbb{E}\{N(k)\}\mathcal{R}^*}C$$

where $\mathbb{E}\{N(k)\}$ denotes the expected value of $N(k)$ (this expected value does not depend on k). Using C', it is then easy to obtain an estimate of the standard deviation of the execution time. Indeed, if

$$C' = \begin{bmatrix}\sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2\end{bmatrix},$$

then the standard deviation estimate is computed as [16]:

$$\sigma = \sigma_Y\sqrt{1 - (\frac{\sigma_{XY}}{\sigma_X\sigma_Y})^2}.$$

D Asymptotic Convergence Rate Estimates

To compute estimates of the algorithm asymptotic rate of convergence, we must extend Equ. (5). In [3], the sequence $\{t_k\}$ is defined as:

$$\begin{cases} t_0 & = 0 \\ t_k & = t_k + a_k + b_k \end{cases}$$

where $\{a_k\}$ and $\{b_k\}$ are defined as:

(i) starting with the $(t_k + a_k)^{th}$ iteration, no solution vector update makes use of values of components corresponding to iterates with indices smaller than t_k.

(ii) all solution vector components are updated at least once between the $(t_k + a_k)^{th}$ and the $(t_k + a_k + b_k)^{th}$ iterations.

The sequence $\{k_t\}$ of Equ. (5) is then defined as:

$$k_t \stackrel{\Delta}{=} \sup\{k \in \mathbb{N} | a_0 + b_0 + ... + a_{k-1} + b_{k-1} \leq t\}$$

for non-negative integers $\mathbb{N}$.

In our setting,

$$\forall k = 0, 1, ... \begin{cases} a_k & = N(k) \\ b_k & = 0 \end{cases}$$

where $N(k)$ denotes the number of iterations performed during the k^{th} algorithm phase. One can then compute:

$$k_t = \sup\{k | \sum_{l=0}^{k-1} \max_{i \in \{1,..,p\}} (A_i + N(l)) \leq t\},$$

The long-term probability distribution of the RV $N(k)$ can be approximated using the π-values for the wavefront Markov chain, leading to the probability distribution of k_t for each t. It is then possible to compute three estimates for the asymptotic rate of convergence by replacing k_t in Equ. (5) by its minimal observable value, its expectation, or its maximal observable value. A formal proof of the convergence of the limit in Equ. (5) for each estimate is left for future work. A finite limit has been obtained in all simulations and experiments.

References

[1] V. Adve and M. Vernon. The influence of Random Delays on Parallel Execution Times. In *Proceedings of the 1993 ACM Sigmetrics Conference on Measurement and Modeling of Computer Systems*, pages 61–73, May 1993.

[2] R. Ash. *Information Theory.* Dover Publications, Mineola, N.Y., 1990.

[3] G. Baudet. Asynchronous Iterative Methods for Multiprocessors. *Journal of the Association for Computing Machinery*, 25:226–244, April 1978.

[4] D. El Baz. M-functions and parallel asynchronous algorithms. *SIAM Journal of Numerical Analysis*, 27:136–140, 1990.

[5] D. El Baz, P. Spiteri, J.C. Miellou, and D. Gazen. Asynchronous Iterative Algorithms with Flexible Communication for Nonlinear Network Flow Problems. *Journal of Parallel and Distributed Computing*, 38:1–15, 1996.

[6] D. P. Bertsekas. Distributed asynchronous computation of fixed point. *Math. Programming*, 27:107–120, 1983.

[7] D. P. Bertsekas and J. N. Tsitsiklis. Convergence rate and termination of asynchronous iterative algorithms. In *Proceedings of the Int. Conf. on Supercomputing*, pages 461–470, 1989.

[8] D. P. Bertsekas and J. N. Tsitsiklis. *Parallel and Distributed Computation*. Prentice-Hall, Englewood Cliffs, NJ, 1989.

[9] A. Bharucha-Reid. *Elements of the Theory of Markov Processes and their Applications*. Dover Publications, Mineola, N.Y., 1997.

[10] L. Brochard, J.-P. Prost, and F. Fauire. Synchronization and load unbalance effects of parallel iterative algorithms. In *Proceedings of the International Conference on Parallel Processing (ICPP)*, volume III, pages 153–160, 1989.

[11] H. Casanova. *Stochastic Models for Performance Anlyses of Iterative Algorithms in Distributed Environments*. PhD thesis, Dept. of Computer Science, University of Tennessee, Knoxville, TN, 1998. Available as TR ut-cs-98-386.

[12] D. Chazan and W. Miranker. Chaotic relaxation. *Linear Algebra and Applications*, 2:199–222, 1969.

[13] N. Duffield, J. Lewis, N. O'Connel, R. Russell, and F. Toomey. Entropy of ATM Traffic Streams: A Tool for Estimating QoS Parameters. *IEEE Journal on Selected Areas in Communications*, 13(6):981–989, August 1995.

[14] A. Frommer. On asynchronous iterations in partially ordered spaces. *Numerical Funct. Anal. Optimization*, 12(3 & 4):315–325, 1991.

[15] A. Greenbaum. Synchronization costs on multiprocessors. *Parallel Computing*, 10:3–14, 1989.

[16] S. Karlin and H. Taylor. *A First Course in Stochastic Processes*. Academic Press, New York, NY, second edition, 1975.

[17] K. Kawahara, Y. Oie, M. Murata, and H. Miyahara. Performance Analysis of Reactive Congestion Control for ATM Networks. *IEEE Journal on Selected Areas in Communications*, 13(4):651–661, May 1995.

[18] C. P. Kruskal and A. Weiss. Allocating independent subtasks on parallel processors. In *Proceedings of the International Conference on Parallel Processing (ICPP)*, pages 236–240, 1984.

[19] B. Lubachevsky and D. Mitra. Chaotic Asynchronous Algorithm for Computing the Fixed Point of a Nonnegative Matrix of Unit Spectral Radius. *Journal of the ACM*, 33(1):130–150, January 1986.

[20] J.C. Miellou. Algorithmes de relaxation à retards. *Revue d'Automatique, Informatique et Recherche Opérationnelle*, 9:55–82, 1970.

[21] J.C. Miellou. Itérations chaotiques à retards. *Comptes Rendus de l'Acad, Sci. Paris*, 278:957–960, 1974.

[22] S. P. Olesen, J. Gregor, M. G. Thomason, and G. T. Smith. EM-ML PET reconstruction on multiple processors with reduced communications. *Int'l. Journal of Imaging Systems and Technology*, 7(3):215–223, 1996.

[23] R. Onvural. *Asynchronous Transter Mode Networks, Performance Issues.* Artech House, Inc., Norwood, MA, second edition, 1995.

[24] J. M. Ortega and W. C. Rheinboldt. *Iterative solution of nonlinear equations in several variables.* Academic Press, New York, 1970.

[25] J. T. Robinson. Some Analysis Techniques for Asynchronous Multiprocessor Algorithms. *IEEE Transactions on Software Engineering*, SE-5(1):24–31, January 1979.

[26] E. Tarazi. Some convergence results for asynchronous algorithms. *Numerical Mathematics*, 39:325–340, 1982.

[27] J. N. Tsitsiklis and G.D. Stamoulis. On the average communication complexity of asynchronous distributed algorithms. Technical Report LIDS-P-1986, MIT Laboratory for Information and Decision Systems, 1990.

[28] A. Üresin and M. Dubois. Sufficient conditions for the convergence of asynchronous distributed algorithms. *Parallel Computing*, 10:83–92, November 1989.

[29] A. Üresin and M. Dubois. Parallel asynchronous algorithms for discrete data. *Journal of the ACM*, 37(3):588–606, 1990.

[30] A. Üresin and M. Dubois. Effects of Asynchronism on the Convergence Rate of Iterative Algorithms. *Journal of Parallel and Distributed Computing*, 34:66–81, 1996.

[31] R. Wolski. Dynamically forecasting network performance using the network weather service. Technical Report TR-CS96-494, U.C. San Diego, October 1996.

Henri Casanova received the BS from l'Ecole Nationale Supérieure d'Electrotechnique, d'Informatique et d'Hydraulique de Toulouse (ENSEEIHT) in 1993, MS from l'Université Paul Sabatier, Toulouse, in 1994, and PhD from the University of Tennessee, Knoxville, in 1998, and is currently a project scientist at the University of California at San Diego. His research interests include metacomputing, parallel/distributed computing, performance modeling, and stochastic models.

Michael G. Thomason received the BS from Clemson University in 1965, MS from Johns Hopkins University in 1970, and PhD from Duke University in 1973. He worked for Westinghouse (Baltimore) and currently is Professor of Computer Science at the University of Tennessee, Knoxville. His research interests include pattern/image analysis, parallel/distributed computation, and stochastic models in computer science. He is a member of Association for Computing Machinery (ACM) and senior member of the Institute of Electrical and Electronics Engineers (IEEE).

Jack Dongarra Jack Dongarra received the BS in Mathematics from Chicago State University in 1972, the MS in Computer Science from the Illinois Institute of Technology in 1973, and the PhD in Applied Mathematics from the University of New Mexico in 1980. Dongarra is a Distinguished Scientist specializing in numerical algorithms in linear algebra at the University of Tennessee's Computer Science Department and Oak Ridge National Laboratory's Mathematical Sciences Section. Professional activities include membership in the Society for Industrial and Applied Mathematics and in the Association for Computing Machinery (ACM).

Table 1: Simulation: Convergence rate errors.

Impl.	$\underline{\mathcal{R}}$	$\widehat{\mathcal{R}}$	$\overline{\mathcal{R}}$	$\mathcal{R}_{Baudet}$
Sync.	7.69%	7.69%	7.69%	7.69%
Async. 1	31.96%	17.53%	36.08%	54.64%
Async. 2	9.55%	17.86%	57.14%	69.05%

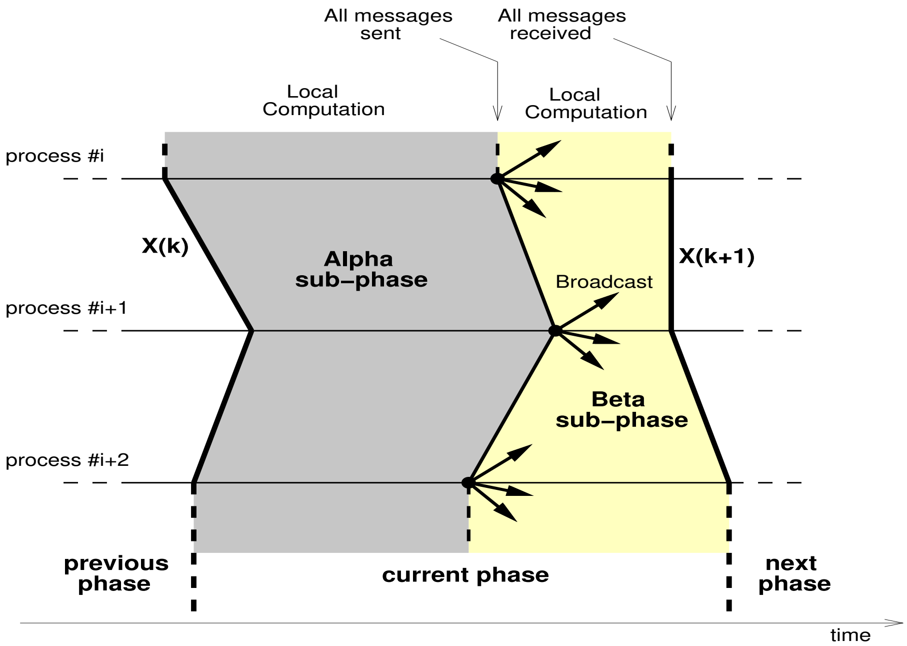

Figure 1: Decomposition of the algorithm into phases

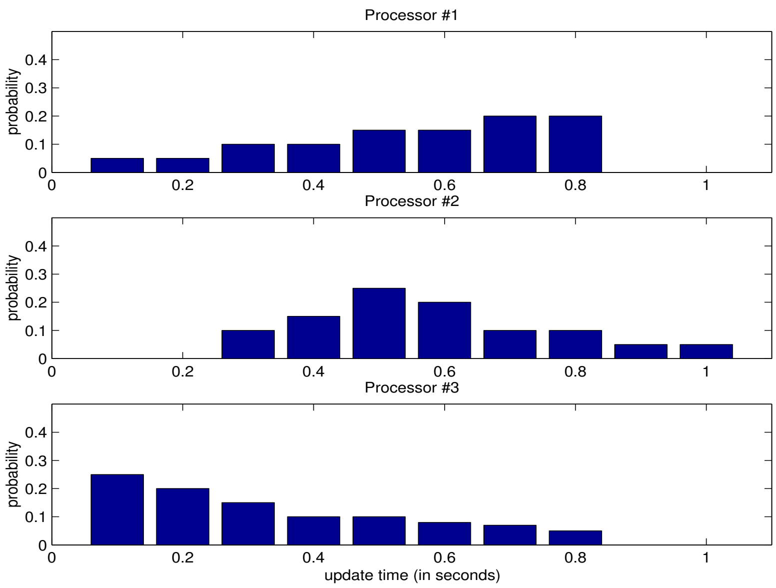

Figure 2: Update time distributions for the three processors

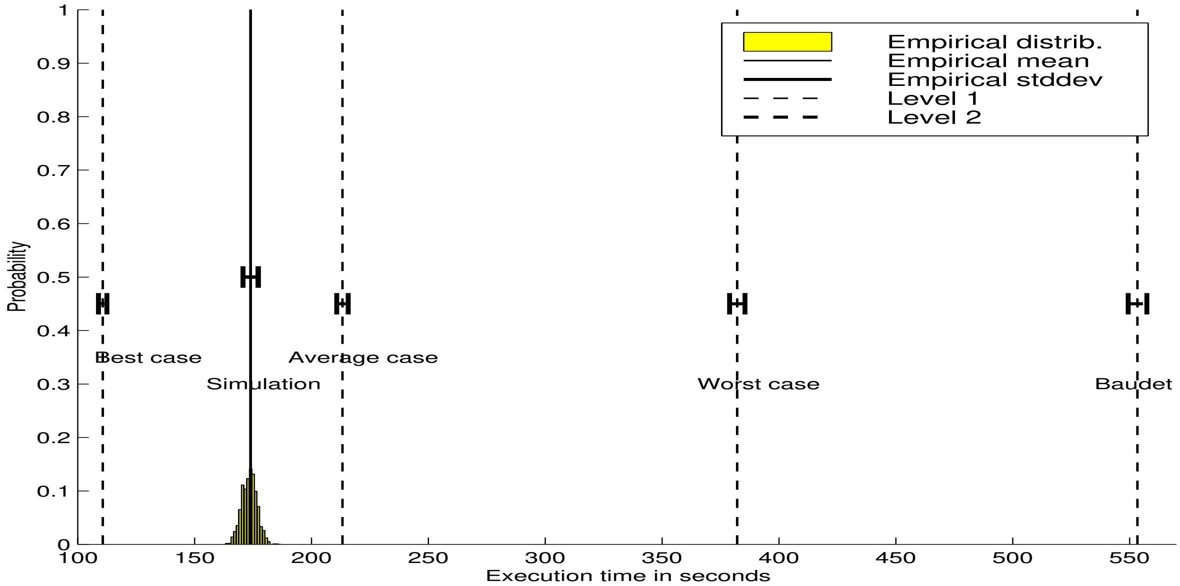

Figure 3: Simulation vs. characterizations for an asynchronous implementation

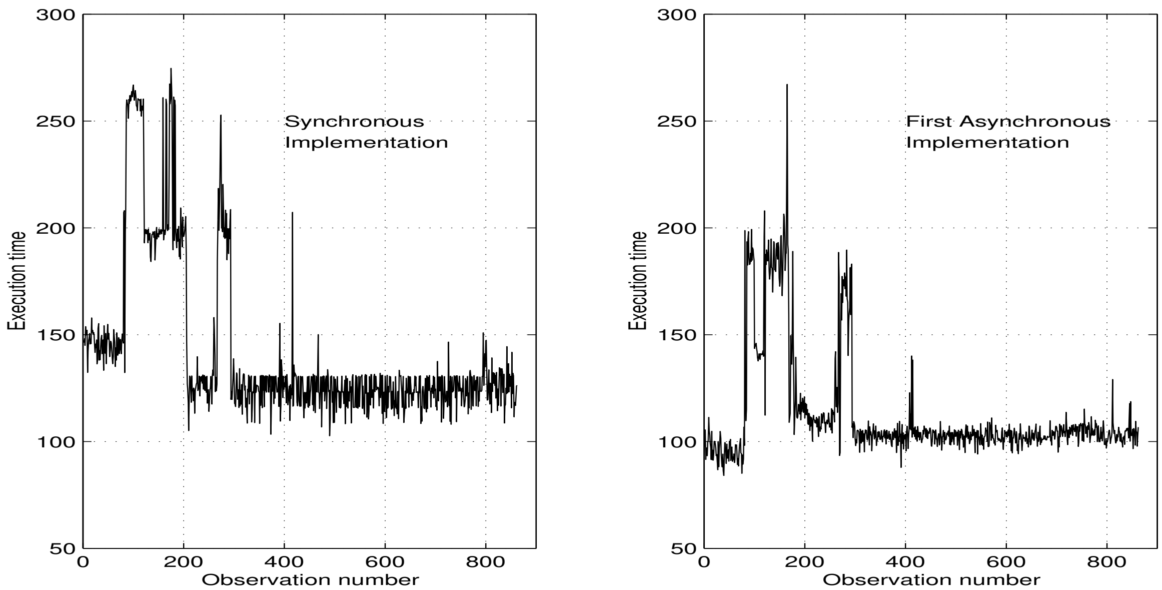

Figure 4: Execution time measurements over a week

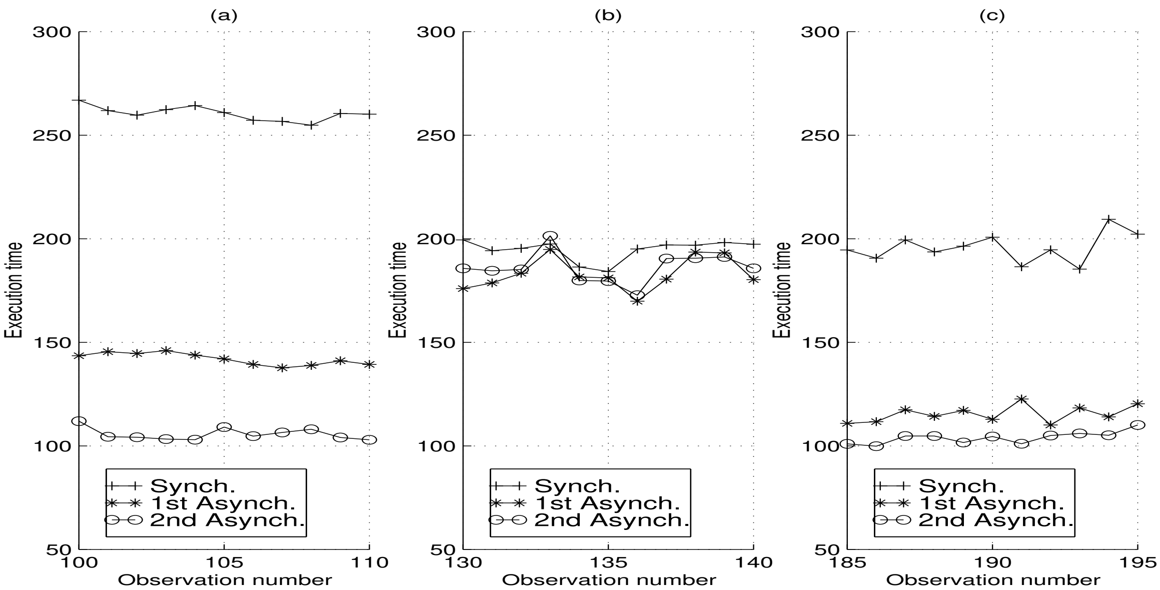

Figure 5: Different experimental behaviors throughout one week

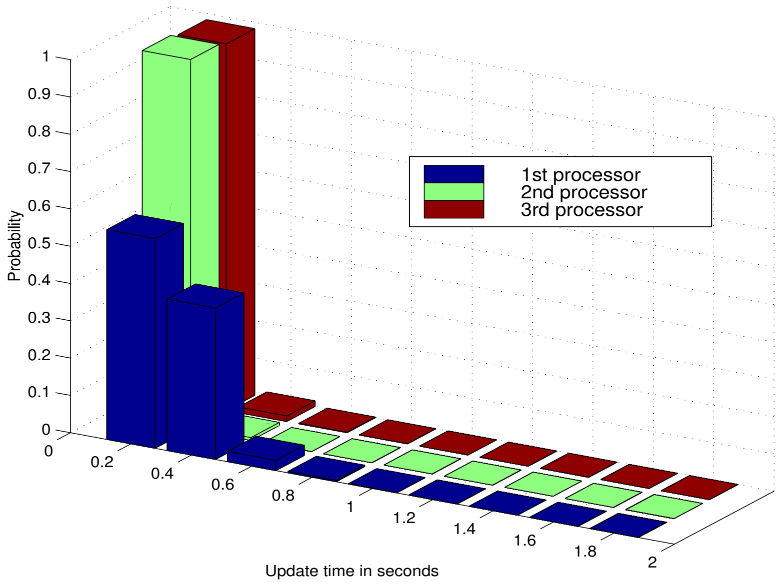

Figure 6: Experimental update time distributions for the three processors

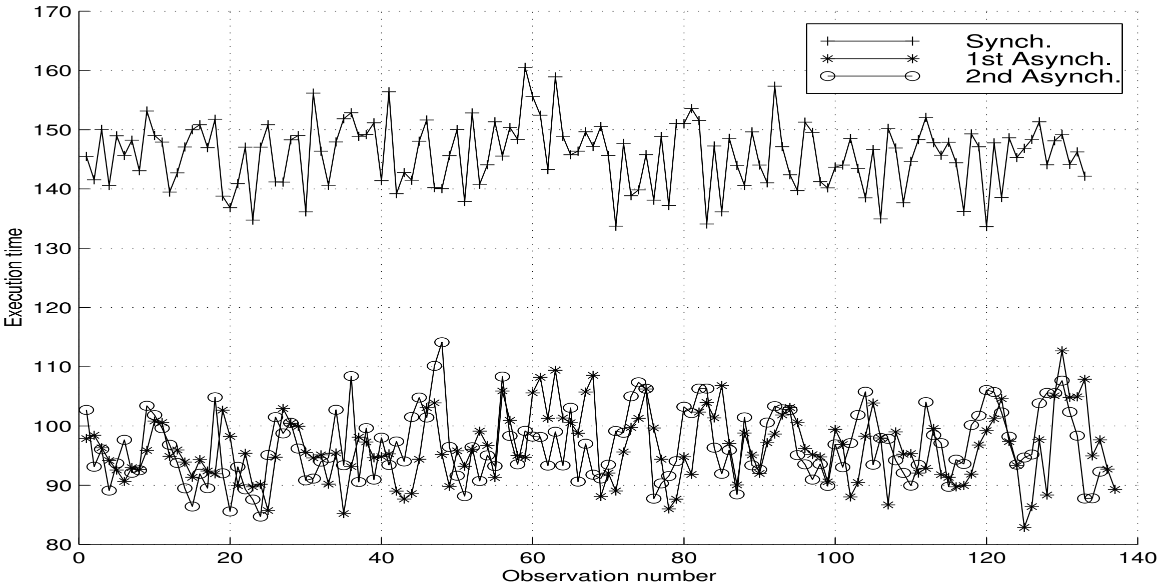

Figure 7: Measurements during 24 hours for the three implementations

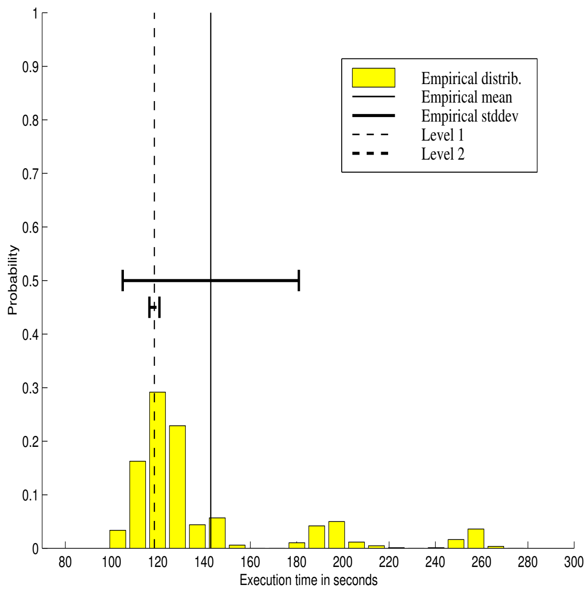

Figure 8: Experiment vs. Characterization for the synchronous implementation

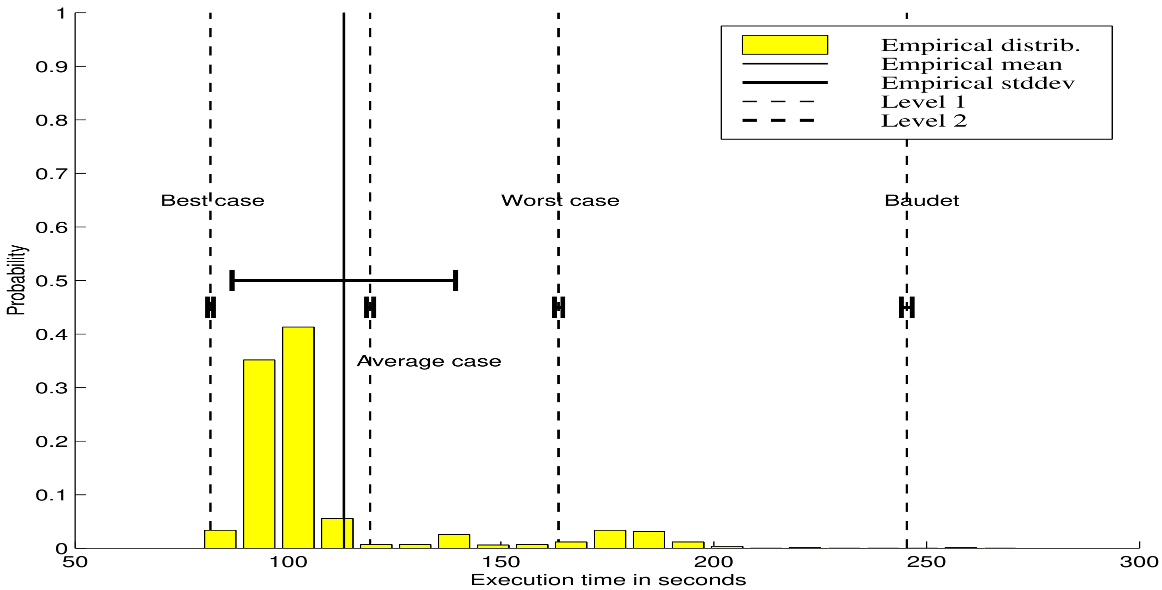

Figure 9: Experiment vs. Characterization for the first asynchronous implementation

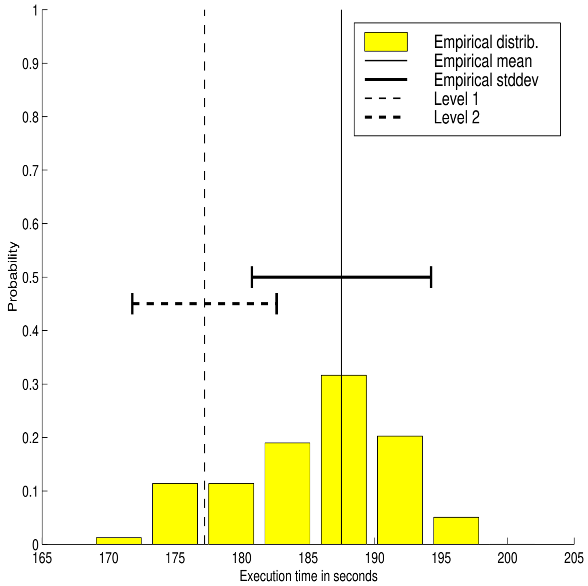

Figure 10: Experiment vs. Characterization for the synchronous implementation