INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Parallel State Space Construction for Model-Checking

Hubert Garavel — Radu Mateescu — Irina Smarandache

N° 4341

Mars 2001 (révisé en décembre 2001)

THÈME 1

Rapport de recherche

ISSN 0249-6399 ISRN INRIA/RR--4340--FR

Parallel State Space Construction for Model-Checking

Hubert Garavel* , Radu Mateescu† , Irina Smarandache‡

Thème 1 — Réseaux et systèmes
Projet VASY

Rapport de recherche n° 4341 — Mars 2001 (révisé en décembre 2001) — 18 pages

Abstract: The verification of concurrent finite-state systems by model-checking often requires to generate (a large part of) the state space of the system under analysis. Because of the state explosion problem, this may be a resource-consuming operation, both in terms of memory and CPU time. In this report, we aim at improving the performances of state space construction by using parallelization techniques. We present parallel algorithms for constructing state spaces (or Labeled Transition Systems) on a network or a cluster of workstations. Each node in the network builds a part of the state space, all parts being merged to form the whole state space upon termination of the parallel computation. These algorithms have been implemented within the CADP verification tool set and experimented on various concurrent applications specified in LOTOS. The results obtained show linear speedups and a good load balancing between network nodes.

Key-words: distributed algorithm, labeled transition system, LOTOS, model-checking, state space construction, verification

This research report is a revised version of an article published as "Parallel State Space Construction for Model-Checking," in Moshe Y. Vardi and Matthew B. Dwyer, editors, Proceedings of the 8th International SPIN Workshop on Model Checking of Software SPIN'2001 (Toronto, Canada), May 19-20, 2001.

* `Hubert.Garavel@inria.fr`
† `Radu.Mateescu@inria.fr`
‡ `Irina.Sturm@st.com`

Unité de recherche INRIA Rhône-Alpes
655, avenue de l'Europe, 38330 Montbonnot-St-Martin (France)
Téléphone : +33 4 76 61 52 00 — Télécopie +33 4 76 61 52 52

Construction parallèle des espaces d'états pour la vérification basée sur les modèles

Résumé : La vérification énumérative des systèmes distribués à nombre fini d'états nécessite souvent la génération (d'une large partie) de l'espace d'états du système à vérifier. A cause du problème de l'explosion d'états, cette opération peut être coûteuse en mémoire et en temps de calcul. Ce rapport propose d'optimiser les performances de la construction des espaces d'états au moyen de techniques de parallélisation. Nous présentons des algorithmes parallèles pour construire les espaces d'états (ou les systèmes de transitions etiquetées) sur un réseau ou une grappe de stations de travail. Chaque noeud du réseau construit une partie de l'espace d'états, toutes les parties étant rassemblées pour former l'espace d'états complet après terminaison du calcul parallèle. Ces algorithmes ont été implémentés dans la boîte à outils Cadp et expérimentés sur différentes applications distribuées spécifiées en Lotos. Les résultats obtenus mettent en évidence des gains en vitesse linéaires et un bon équilibrage de charge entre les noeuds du réseau.

Mots-clés : algorithme distribué, construction d'espace d'états, Lotos, model-checking, système de transitions étiquetées, vérification

1 Introduction

As formal verification becomes increasingly used in the industry as a part of the design process, there is a constant need for efficient tool support to deal with real-size applications. Model-checking [Hol91, CGP00] is a successful verification method based on reachability analysis (state space exploration) and allows an automatic detection of early design errors in finite-state systems. Model-checking works by constructing a model (state space) of the system under design, on which the desired correctness properties are verified.

There are essentially two approaches to model-checking: *symbolic* verification [CCGR00, CGP00] represents the state space in comprehension, by using various encoding techniques (e.g., BDDs), and *enumerative* verification [RS97, Dil96, FGK+96, Hol97] represents the state space in extension, by enumerating all reachable states. Enumerative model-checking techniques can be further divided in *global* techniques, which require to entirely construct the state space before performing the verification, and *local* (or *on-the-fly*) techniques, which allow to construct the state space simultaneously with the verification.

In this report, we focus on enumerative model-checking, which is well-adapted to asynchronous, nondeterministic systems containing complex data types (records, sets, lists, trees, etc.). More precisely, we consider the problem of constructing a *Labeled Transition System* (LTS), which is the natural model for high-level, action-based specification languages, especially process algebras such as CCS [Mil89], CSP [Hoa85], ACP [BK84], or LOTOS [ISO88]. An LTS is constructed by exploring the transition relation starting from the initial state (*forward reachability*). During this operation, all explored states must be kept in memory in order to avoid multiple exploration of a same state. Once the LTS is constructed, it can be used as input for various verification procedures, such as bisimulation/preorder checking and temporal logic model-checking. Moreover, when the verification requires to explore the entire LTS (e.g., when verifying invariant temporal properties or checking bisimulation), since the state contents is abstracted away in a constructed LTS, the memory consumed is generally much smaller than for on-the-fly verification on the initial specification.

State space construction may be very consuming both in terms of memory and execution time: this is the so-called *state explosion* problem. During the last decade, different techniques for handling state explosion have been proposed, among which partial orders and symmetries; however, for industrial-scale systems, these optimizations are not always sufficient. Moreover, most of the currently available verification tools work on sequential machines, which limits the amount of memory (between 0.5 and 2 GBytes on usual configurations), and therefore the use of clusters or networks of workstations is desirable.

In this report, we investigate an approach to parallelize state space construction on several machines, in order to benefit from all the local memories and CPU resources of each machine. This allows to reduce both the amount of memory needed on each machine and the overall execution time. We propose algorithms for parallel construction of LTSs, developed using the generic environments BCG and OPEN/CÆSAR [Gar98] for LTS manipulation provided by the CADP verification tool set [FGK+96]. Since these environments are language independent, our algorithms can be directly used not only for LOTOS, but also for every language connected to the OPEN/CÆSAR application programming interface, such as UML [JHGP99].

The implementation is based on standard sockets, available everywhere, and was experimented on two different configurations: a typical network of workstations (Sparc workstations running Solaris and PCs running Linux, connected using 100 Mb/s ETHERNET), and a cluster of PCs (with 450 MHz processor and 512 MBytes main memory) connected using SCI (*Scalable Coherent Interface*). Each machine in the network is responsible for constructing a part of the LTS, this part being determined using a static partition function. Upon termination of the parallel computation, which is detected by means of a virtual ring-based distributed algorithm, all parts are merged to form the complete LTS.

We experimented with our algorithms on three non-trivial protocols specified in LOTOS: the home audio-video (HAVi) protocol of Philips [Rom99], the TOKENRING leader election protocol [GM97], and the SCSI-2 bus arbitration protocol [ANS94].

Related work Distributed state space construction has been studied in various contexts, mostly for the analysis of low-level models such as Petri nets, stochastic Petri nets, discrete-time and continuous-time Markov chains [CCBF94, CCM95, AKH97, ADK97, CGN98, NC97, MCC97, HBB99, KH99].

All these approaches share a common idea: each machine in the network explores a subset of the state space. However, they differ on a number of design principles and implementation choices such as: the choice between a shared memory architecture and a message-passing one, the use of hash tables or B-trees to store states on each machine, the way of partitioning the state space using either static hash functions or dynamic ones that allow dynamic load balancing, etc.

As regards high-level languages for asynchronous concurrency, a distributed state space exploration algorithm [LS99] derived from the SPIN model-checker [Hol97] has been implemented for the PROMELA language. The algorithm performs well on homogeneous networks of machines, but it does not outperform the standard, sequential implementation of SPIN, except for problems that do not fit into the main memory of a single machine. Several SPIN-specific partition functions are experimented, the most advantageous one being a function that takes into account only a fraction of the state vector.

Another distributed state enumeration algorithm has been implemented in the MURφ verifier [SD97]. The speedups obtained are close to linear and the hash function used for state space partition provides a good load balancing. However, experimental data reported concerns relatively small state spaces (approximatively 1.5 M states) on a 32-node UltraSPARC Myrinet network of workstations.

There also exist approaches, such as [KMHK98], in which parallelization is applied to "partial" verification, i.e., state enumeration in which some states can be omitted with a low probability. In the present report, we only address exact, exhaustive verification issues.

For completeness, we can also mention an alternative approach [HGGS00] in which symbolic reachability analysis is distributed over a network of workstations: this approach does not handle states individually, but sets of states encoded using BDDs.

Report outline Section 2 gives some preliminary definitions and specifies the context of our work. Section 3 describes the proposed algorithms for parallel construction of LTSs. Section 4 discusses implementation issues and presents various experimental results. Finally, Section 5 gives some concluding remarks and directions for future work.

2 Definitions

A (monolithic) Labeled Transition System (LTS) is a tuple $M = (S, A, T, s_0)$, where S is the set of states, A is the set of actions, $T \subseteq S \times A \times S$ is the transition relation, and $s_0 \in S$ is the initial state. A transition $(s, a, s') \in T$ indicates that the system can move from state s to state s' by performing action a. All states in S are assumed to be reachable from s_0 via (sequences of) transitions in T.

In the model-checking approach by state enumeration, there are essentially two ways to represent an LTS:

explicitly, by enumerating all its states and transitions. In this case, the contents of states becomes irrelevant, since the essential information is given by actions (transition labels). Therefore, when storing an LTS as a computer file, it is sufficient to encode states as natural numbers. An explicit representation of LTSs is provided by the BCG (*Binary Coded Graph*) file format of the CADP verification tool set [FGK+96]. The BCG format is based upon specialized compression algorithms, allowing compact encodings of LTSs.

implicitly, by giving its initial state s_0 and its successor function $succ : S \rightarrow 2^T$ defined by $succ(s) = \{(s, a, s') \mid (s, a, s') \in T\}$. An implicit representation of LTSs is provided by the generic, language independent environment OPEN/CÆSAR [Gar98] of CADP. OPEN/CÆSAR offers primitives for accessing the initial state of an LTS and for enumerating the successors of a given state, as well as various data structures (state tables, stacks, etc.), allowing straightforward implementations of on-the-fly verification algorithms.

Our objective is to translate LTSs from an implicit to an explicit representation by using parallelization techniques.

In order to represent a monolithic LTS $M = (S, A, T, s_0)$ on N machines (numbered from 0 to $N-1$), we introduce the notion of *partitioned* LTS $(B_0, \ldots, B_{N-1}, s_0)$, where each B_i is a triple (S_i, A_i, T_i) such that:

- $S = \cup_{i=0}^{N-1} S_i$ and $S_i \cap S_j = \emptyset$ for all $0 \leq i, j < N$ (the state set is split[1] into N parts S_i, one part per machine),
- $A = \cup_{i=0}^{N-1} A_i$ (the sets A_i being not necessarily pairwise disjoint),
- $T = \cup_{i=0}^{N-1} T_i$ and $T_i \subseteq S \times A_i \times S_i$ for all $0 \leq i < N$ (consequently, the sets T_i are pairwise disjoint and each T_i contains all transitions (s, a, s') of T whose target state s' belongs to S_i),
- $s_0 \in S_i$ for some $0 \leq i < N$.

$B_0, ..., B_{N-1}$ are called the *components* (or *component* LTSs) of M, although each B_i is not (strictly speaking) an LTS, because $T_i \not\subseteq S_i \times A_i \times S_i$ and because the notion of initial state is irrelevant for all B_i but the one containing s_0. Yet, a partitioned LTS can be represented as a collection of BCG files encoding its components.

3 Parallel generation of LTSs

In this section we present two complementary algorithms allowing to convert an implicit LTS (defined using the OPEN/CÆSAR interface) to an explicit one (represented as a BCG file) using N machines connected by a network. These algorithms operate in two steps:

- Construction of a partitioned LTS represented as a collection of BCG files. This is done by using an algorithm called DISTRIBUTOR, which is executed on every machine in order to generate a BCG file encoding a component of the partitioned LTS.
- Conversion to a monolithic LTS represented as a single BCG file. This is done using an algorithm called BCG_MERGE, which is executed on a sequential machine in order to generate a single BCG file containing all the states and transitions of the partitioned LTS.

[1]Strictly speaking, this is not a partition of S because some S_i may be empty.

Once the Bcg file encoding the initial Lts has been constructed, it can be used as input for other Cadp verification tools, such as the Evaluator 3.0 model-checker [MS00], which allows linear-time verification of temporal formulas expressed in regular alternation-free μ-calculus.

3.1 Construction of partitioned LTSs

We consider a network of N machines numbered from 0 to $N-1$ and an Lts $M = (S, A, T, s_0)$ given implicitly by its initial state s_0 and its successor function *succ*. Each machine executes an instance of the parallel generation algorithm Distributor shown on Figure 1, which does a (partial) forward exploration of M and produces a component Lts $B_i = (S_i, A_i, T_i)$ stored as a Bcg file. The state set S_i constructed by machine i is determined using a static partition function $h : S \to [0, N-1]$. This set is defined as $S_i = \{s \in S \mid h(s) = i\}$, which also determines the sets T_i and A_i according to the above definitions.

Machine i keeps in its local memory the states of S_i, whilst the transitions of T_i are written (after being kept temporarily in a work list L_i) to the Bcg file B_i stored on the local disk of the machine. The states visited and explored by machine i are stored in two disjoint sets V_i ("visited") and E_i ("explored"), which are implemented using the state table library provided by the Open/Cæsar environment.

Machine i can send a message m to machine j by invoking a primitive named Send (j, m) and can receive a message by invoking another primitive Receive (m). There are four kinds of messages: *Arc*, *Rec*, *Snd*, and *Trm*, the first one being used for sending Lts transitions and the others being related to termination detection. Send and Receive are assumed to be non-blocking. Receive returns a boolean answer indicating whether a message has been received or not.

The computation is started by the machine called *initiator*, having the index $h(s_0)$, which explores the initial state of the Lts. The Distributor algorithm consists of a main loop, which performs various actions, among which:

(a) A state $s \in V_i$ is explored by enumerating all its successor transitions $(s, a, s') \in succ(s)$. If a target state s' belongs to machine i (i.e., $h(s') = i$), the corresponding transition is kept in the list L_i and will be processed later. Otherwise, the transition is sent to machine $h(s')$ as a message $Arc(n_i(s), a, s')$, where $n_i(s)$ is the number associated by machine i to s; note that only the number $n_i(s)$ is sent, but not the contents of state s itself. Machine $h(s')$ will be responsible for exploring state s' and for writing the transition to its local Bcg file, which will contain all Lts transitions whose target states are explored by this machine.

(b) A transition is taken from L_i and is written to the Bcg file B_i after computing an appropriate number for its target state. In order to obtain a bijective numbering of Lts states across the N Bcg files, each state s explored by machine i is assigned a number $n_i(s)$ such that $n_i(s)\ mod\ N = i$. This is done using a counter c_i, which is initialized to i and incremented by N every time a new state is visited.

(c) An attempt is made to receive a message m from another machine. If m has the form $Arc(n, a, s)$, it denotes a transition (s', a, s), where n is the source state number $n_j(s')$ assigned by the sender machine of index $j = n\ mod\ N$. In this case, the contents of m is stored in the list L_i; otherwise, m is related to termination detection (see below).

In order to detect the termination of the parallel Lts generation, we use a virtual ring-based algorithm inspired by [Mat87]. According to the general definition, (global) termination is reached when all local computations are finished (i.e., each machine i has neither remaining states to explore, nor transitions

```
procedure Distributor (in i, s_0, succ, h, N; out S_i, A_i, T_i) is
  initiator_i := (h(s_0) = i); L_i := ∅; E_i := ∅; A_i := ∅; T_i := ∅; c_i := i;
  if initiator_i then
    n_i(s_0) := c_i; V_i := {s_0}; S_i := {n_i(s_0)}
  else
    V_i := ∅; S_i := ∅
  endif;
  terminating_i := false; terminated_i := false; nbsent_i := 0; nbrecd_i := 0;
  while ¬terminated_i do
(a)    if V_i ≠ ∅ then
        choose s ∈ V_i; V_i := V_i \ {s}; E_i := E_i ∪ {s};
        forall (s, a, s') ∈ succ(s) do
          if h(s') = i then
            L_i := L_i ∪ {(n_i(s), a, s')}
          else
            Send (h(s'), Arc(n_i(s), a, s')); nbsent_i := nbsent_i + 1
          endif
        endfor
(b)    elsif L_i ≠ ∅ then
        choose (n, a, s) ∈ L_i; L_i := L_i \ {(n, a, s)};
        if s ∉ E_i ∪ V_i then
          c_i := c_i + N; n_i(s) := c_i; V_i := V_i ∪ {s}; S_i := S_i ∪ {n_i(s)};
        endif;
        A_i := A_i ∪ {a}; T_i := T_i ∪ {(n, a, n_i(s))}
(c)    elsif Receive (m) then
        case m is
          Arc(n, a, s) → L_i := L_i ∪ {(n, a, s)}; nbrecd_i := nbrecd_i + 1
          Rec(k) → if ¬initiator_i then
                       Send ((i + 1) mod N, Rec(k + nbrecd_i))
                     else
                       totalrecd_i := k; Send ((i + 1) mod N, Snd(nbsent_i))
                     endif
          Snd(k) → if ¬initiator_i then
                       Send ((i + 1) mod N, Snd(k + nbsent_i))
                     elsif totalrecd_i = k then
                       Send ((i + 1) mod N, Trm)
                     else
                       terminating_i := false
                     endif
          Trm →      if ¬initiator_i then
                       Send ((i + 1) mod N, Trm)
                     endif;
                     terminated_i := true
        endcase
      elsif initiator_i ∧ ¬terminating_i then
        terminating_i := true; Send ((i + 1) mod N, Rec(nbrecd_i))
      endif
  endwhile
end
```

Figure 1: Parallel generation of an Lts as a collection of Bcg files

to write in its BCG file B_i) and all communication channels are empty (i.e., all sent transitions have been received).

The principle of the termination detection algorithm used in DISTRIBUTOR is the following. All machines are supposed to be on an unidirectional virtual ring that connects every machine i to its successor machine $(i+1)$ *mod* N. Every time the initiator machine finishes its local computations, it checks whether global termination has been reached by generating two successive "waves" [Mat87] of *Rec* and *Snd* messages on the virtual ring to collect the number of messages received and sent by all machines, respectively. A message $Rec(k)$ (resp. $Snd(k)$) received by machine i indicates that k messages have been received (resp. sent) by the machines located on the ring between the initiator and the machine numbered $(i-1)$ *mod* N. Each machine i counts the messages it has received and sent using two integer variables $nbrecd_i$ and $nbsent_i$, and adds their values to the numbers carried by *Rec* and *Snd* messages. Upon receipt of the $Snd(k)$ message ending the second wave, the initiator machine checks whether the total number k of messages sent is equal to the total number $totalrecd_i$ of messages received (the result of the *Rec* wave). If this is the case, it will inform the other machines that termination has been reached by sending a *Trm* message on the ring. Otherwise, the initiator concludes that termination has not been reached yet and will generate a new termination detection wave later. The boolean variables $terminating_i$ and $terminated_i$ respectively indicate that a termination detection wave has started ($terminating_i$ can only be true for the initiator) and that global termination has been detected.

In our algorithm, each machine propagates the current wave only when its local computations are finished. Experimental results show that this approach strongly reduces the overhead caused by unsuccessful termination waves. This distributed termination detection scheme seems to use less messages than the centralized termination detection schemes used in the parallel versions of SPIN [LS99] and MURφ [SD97], which in all cases require several broadcast message exchanges between a coordinator machine and all other machines.

3.2 Merging of partitioned LTSs into monolithic LTSs

The DISTRIBUTOR algorithm allows to generate, from an implicit LTS $M = (S, A, T, s_0)$, a partitioned LTS $(B_0, ..., B_{N-1}, s_0)$ represented as a collection of BCG files. The next step is to merge these BCG files into a unique BCG file that can be processed using the verification tools of CADP (e.g., those tools based on bisimulations or modal μ-calculus).

In the formal definition of partitioned LTSs given in Section 2, each component B_i has the form (S_i, A_i, T_i). However, when implemented as a BCG file, states of S_i are represented by their numbers rather than their semantic contents. This is possible because DISTRIBUTOR assigns a unique number to each state across all the machines. For this reason, each B_i is represented as a BCG file encoding an LTS $(Q_i \cup Q_i^*, A_i, R_i, q_i)$, where:

- $Q_i \subseteq \mathbb{N}$ contains the numbers of all states of S_i (note that $q \in Q_i$ implies q *mod* $N = i$).
- $Q_i^* \subseteq \bigcup_{j=0}^{N-1} Q_j$ contains the numbers of all states s of S for which there exists an outgoing transition (s, a, s') in T_i (note that s' belongs to S_i, and that Q_i and Q_i^* are not necessarily disjoint).
- $R_i \subseteq Q_i^* \times A \times Q_i$ is the transition relation derived from T_i by replacing each state by its number.
- $q_i \in Q_i$ is the number assigned by DISTRIBUTOR to the initial state s_0: if $i = h(s_0)$ (i.e., if i is the index of the initiator machine) then $q_i = h(s_0)$ as s_0 is the first state processed by the initiator; for other values of i, q_i is undefined.

Merging the N BCG files B_i into a single one could be done simply by taking the union of all states sets Q_i and the union of all transition relations R_i. However, since the partition function h is not perfect, this simple approach might result in a BCG file with an initial state number different from 0 (when $h(s_0) \neq 0$) and with "holes" in the numbering of states (when $|Q_i| \neq |Q_j|$ for two BCG files B_i and B_j). For example, given an LTS with 7 states and $N = 2$, DISTRIBUTOR could produce $Q_0 = \{0, 2, 4, 6, 8\}$, $Q_1 = \{1, 3\}$, and $h(s_0) = 1$; taking the union of state sets would lead to a "sparse" state set $Q = \{0, 1, 2, 3, 4, 6, 8\}$, whereas one would normally expect $Q = \{0, 1, 2, 3, 4, 5, 6\}$. Moreover, a contiguous numbering of states is suitable to achieve a better compaction of the final BCG file.

The conversion algorithm BCG_MERGE that we propose (see Figure 2) takes as inputs a partitioned LTS $(B_0, \ldots, B_{N-1}, s_0)$, where each B_i is represented as a BCG file $(Q_i \cup Q_i^*, A_i, R_i, q_i)$, and constructs a corresponding BCG file $B = (Q, A, R, q_{init})$ such that $q_{init} = 0$ and $Q = \{0, \ldots, |Q| - 1\}$.

```
procedure BCG_MERGE (in B_0, ..., B_{N-1}, s_0; out Q, A, R, q_init) is
    i_0 := h(s_0);
    c := 0;
    for k = 0 to N − 1 do
        i := (i_0 + k) mod N;
        c_i := c;
        c := c + |Q_i|
    end;
    Q := {0, ..., c − 1};
    A := ∅;
    R := ∅;
    q_init := 0;
    forall i ∈ {0, ..., N − 1} do
        forall (q, a, q') ∈ R_i do
            A := A ∪ {a};
            R := R ∪ {(c_{q mod N} + (q div N), a, c_{q' mod N} + (q' div N))}
        end
    end
end
```

Figure 2: Merging of a collection of BCG files into a single one

Informally, BCG_MERGE assigns to each BCG file B_i a contiguous range of state numbers $[c_i, c_i + |Q_i| - 1]$ such that $c_{h(s_0)} = 0$ and $c_{(i+1) \ mod \ N} = c_i + |Q_i|$. The values of $c_0, \ldots, c_{N-1}$ induce a state renumbering function mapping each state number q to a new number $r(q) = c_{q \ mod \ N} + (q \ div \ N)$, where *div* denotes integer division. As regards the initial state s_0, its former number associated by DISTRIBUTOR was $h(s_0)$ (where $h(s_0) < N$) and its new number will be: $q_{init} = r(h(s_0)) = c_{h(s_0) \ mod \ N} + (h(s_0) \ div \ N) = c_{h(s_0)} = 0$. The set of transitions R of the resulting BCG file is then obtained by taking all the transitions (q, a, q') of $R_0 \cup \ldots \cup R_{N-1}$ in which q and q' are replaced by $r(q)$ and $r(q')$, respectively.

It is worth noticing that the BCG_MERGE algorithm processes only one BCG file B_i at a time and does not require to load in memory the transition relation R_i of B_i. State renumbering is performed on-the-fly, resulting in a low memory cost, which is independent from the sizes of $B_0, \ldots, B_{N-1}$.

4 Experimental results

We implemented the DISTRIBUTOR and BCG_MERGE algorithms within the CADP verification tool set [FGK+96] by using the OPEN/CÆSAR [Gar98] and BCG environments. To ensure maximal portability, the communication primitives of DISTRIBUTOR are built on top of TCP/IP using standard UNIX sockets. An alternative implementation using the MPI (*Message Passing Interface*) standard [GHLL+98] would have been possible; we chose sockets because they are built-in in most operating systems and because the DISTRIBUTOR algorithm was simple enough not to require the higher-level functionalities provided by MPI.

We experimented DISTRIBUTOR and BCG_MERGE on three industrial-sized protocols specified in LOTOS:

(a) The HAVi protocol [Rom99], standardized by several companies, among which Philips, in order to solve interoperability problems for home audio-video networks. HAVi provides a distributed platform for developing applications on top of home networks containing heterogeneous electronic devices and allowing dynamic plug-and-play changes in the network configuration. We considered a configuration of the HAVi protocol with 2 device control managers (1,039,017 states and 3,371,039 transitions, state size of 80 bytes).

(b) The correct TOKENRING leader election protocol [GM97] for unidirectional ring networks, which is an enhanced version of the protocols proposed by Le Lann [Lan77] and by Chang & Roberts [CR79]. This TOKENRING protocol corrects an error in Le Lann's and Chang & Roberts' protocols, by allowing to designate a unique leader station in presence of various faults of the system, such as message losses and station crashes. We considered a configuration of the TOKENRING protocol with 3 stations (12,362,489 states and 45,291,166 transitions, state size of 6 bytes).

(c) The arbitration protocol for the SCSI-2 bus [ANS94], which is designed to provide an efficient peer-to-peer I/O bus for interconnecting computers and peripheral devices (magnetic and optical disks, tapes, printers, etc.). We considered SCSI-2 configurations consisting of a controller device and several disks that accept data transfer requests from the controller. Two versions of the specification have been used: v1, with 5 disks (961,546 states and 5,997,701 transitions, state size of 13 bytes) and v2, with 6 disks (1,202,208 states and 13,817,802 transitions, state size of 15 bytes).

The experiments have been performed on a cluster of 450 MHz, 512 MBytes PCS connected via SCI (the DISTRIBUTOR and BCG_MERGE have been developed and debugged on an ETHERNET network of three SUN workstations; however, using a dedicated SCI network with more machines was more appropriate for performance measurement). Our performance measurements concern three aspects: speedup, partition function, and use of communication buffers.

4.1 Speedup

Figure 3 shows the speedups obtained by generating the LTSS of the aforementioned LOTOS specifications in parallel on a cluster with up to 10 PCS. For the TOKENRING and HAVi protocols, the speedups observed on N machines are given approximately by the formulas $S_N = t_1/t_N = 0.4N$ and $S_N = 0.3N$ (t_k being the execution time on k machines). For the v1 and v2 versions of the SCSI-2 protocol, the speedups obtained are close to ideal.

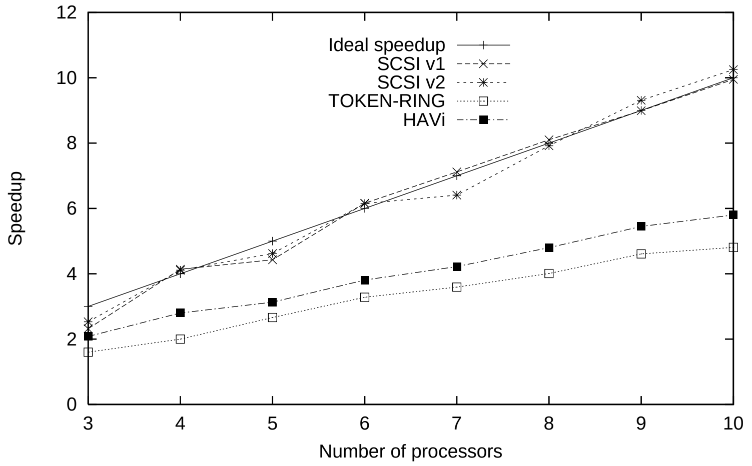


Figure 3: Speedup measurements for the HAVi, TOKENRING, and SCSI-2 protocols

These results can be explained by examining the implementation of the DISTRIBUTOR algorithm. The state sets explored by each machine in the network are stored locally using generic hash tables provided by the OPEN/CÆSAR library. Since these hash tables use open hashing (with a fixed number of hash entries), the number of states contained in each table is not bounded (except by the amount of available memory on each machine) and the search time in the hash table grows linearly with the number of states already inserted in the table. Therefore, splitting the state set among N machines is likely to reduce by N the overall search time. Also, parallelization becomes efficient when the time spent in generating state successors is important, which happens for LOTOS specifications having many parallel processes and complex synchronization patterns. This explains why the speedup obtained for the SCSI-2 is better than for the TOKENRING: the SCSI-2 example involves complex data computations (handling of disk buffers and of device status kept by the controller) and synchronizations (multiple rendezvous between 6 or 7 devices to gain bus access), whereas the TOKENRING example has very simple computations and only binary synchronizations between stations and communication links.

The speedups obtained show a good overlapping between computations and communications during the execution of DISTRIBUTOR. This is partly due to a buffered communication scheme with well-chosen dimensions of transmission buffers (see Section 4.3).

4.2 Choosing a good partition function

In order to increase the performance of the parallel generation algorithm, it is essential to achieve a good load balancing between the N machines, meaning that the N parts of the distributed LTS should contain (nearly) the same number of states. As indicated in Section 3.1, we adopted a static partition scheme, which avoids the potential communication overhead occurring in dynamic load balancing schemes. Then, the problem is to choose an appropriate partition function $h : S \to [0, N-1]$ associating to each state a machine index.

Because we target at language independent state space construction, we cannot assume that state contents exhibit structural properties (e.g., constant fields, repeated patterns, etc.) particular to a given language. However, assuming that state contents are uniformly distributed bit strings might not be true in practice: for instance, if the last field of the state vector is a pointer to a value of type

t, it is likely that, for each reachable state $s \in S$, the integer value of s (computed by interpreting s as an arbitrarily long integer number) will be a multiple of some constant (e.g., the alignment factor of t).

To address this problem, we chose a partition function of the form $h(s) = f(s, P)\ mod\ N$, where P is a prime number and $f(s, P)$ is the hash function that computes the remainder modulo P of the integer value of state s[2].

The value of P must be carefully chosen for h to distribute states among the N machines uniformly. An appropriate value for P would be N, which is only suitable if N is prime (and different from 2 because of the above alignment problem). If N is not prime, we choose for P a large value such that $P\ mod\ N = 1$: if state contents are uniformly distributed, the condition $P\ mod\ N = 1$ assigns to all machines the same number of states, except for machine 0, which is assigned more states than the other machines. Choosing P sufficiently big reduces the number of exceeding states (approximately $|S|\ div\ P$) assigned to machine 0. For the experiments presented in this report, we chose P around 1,600,000.

Figures 4 and 5 show the distribution of the states on 10 machines for the three protocols described above. In order to evaluate the quality of the distribution, we calculated the standard deviation $\sigma = \sqrt{(\sum_{i=0}^{N-1}(|S_i| - |S|/N)^2)/N}$ between the sizes $|S_i|$ of the state sets explored by each machine i in the network. For all examples considered, the values obtained for σ are very small (less than 1% of the mean value $|S|/N$), which indicates a good performance of the partition function h.

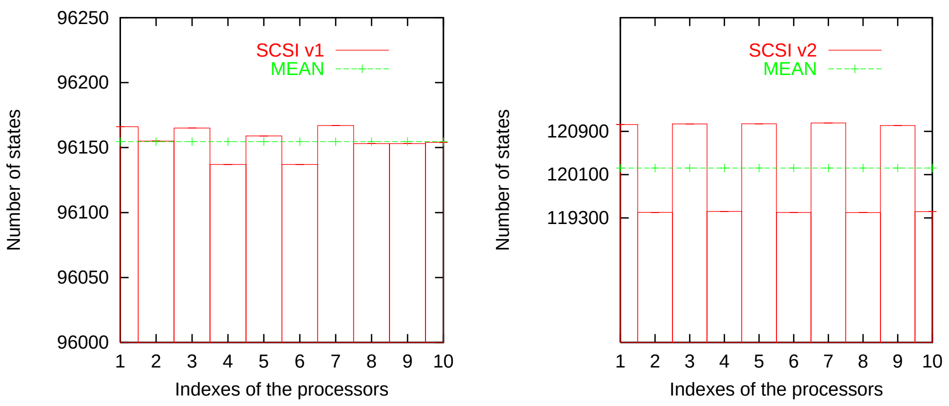
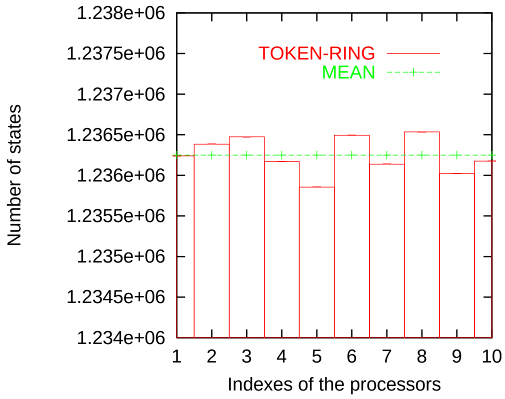


Figure 4: State distributions for the SCSI-2 protocol on 10 machines

The quality of a partition function could also be estimated according to the number of "cross-border" transitions of the partitioned LTS (i.e., transitions having the source state in a component and the target state in another component). This number should be as small as possible, since it is equal to the number of *Arc* messages sent over the network during the execution of DISTRIBUTOR. However, in practice, reducing the number of cross-border transitions would require additional information about the structure of the program, and therefore must be language dependent. Since DISTRIBUTOR is built using the language independent OPEN/CÆSAR environment, we did not focused on developing language dependent (e.g., LOTOS-specific) partition functions. This might be done in the future, by extending the OPEN/CÆSAR application programming interface to provide more information about the internal structure of program states.

[2]Function f is implemented by function `CAESAR_STATE_3_HASH()` of the OPEN/CÆSAR library.



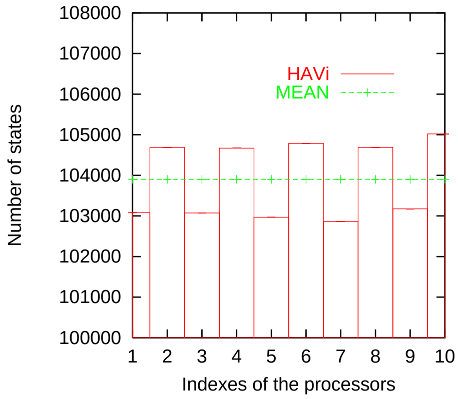


Figure 5: State distribution for the TokenRing and HAVi protocols on 10 machines

4.3 Using communication buffers

To reduce the overhead of message transmission and to increase the overlapping between communications and computations, we chose an asynchronous, non-blocking implementation of the Send and Receive primitives used in the Distributor algorithm. Also, to reduce communication latency, these primitives actually perform a buffering of messages (Lts transitions) instead of sending them one by one as indicated in Figure 1.

The implementation is based on Tcp/Ip and standard Unix communication primitives (sockets). In practice, for each machine $0 \leq i \leq N-1$, there is a virtual channel (i, j) to every other machine $j \neq i$ with a corresponding logical buffer of size L used for storing messages transmitted on the channel. The $N-1$ virtual channels associated with each machine share the same physical channel (socket), which has an associated buffer of size L_p. For a given size d of messages (which depends on the application), we observed that the optimal length L_{opt} of the logical transmission buffer is given by the formula $L_{opt} = L_p/d(N-1)$. Experiments show that for L_{opt}, all transitions accumulated in the logical transmission buffers can be sent at the physical level by the next call to Send. Figure 6 illustrates the effect of buffering on Distributor's speedup for the Scsi-2 and the TokenRing protocols. A uniform increase of speedup is observed between the variants $L = 1$ (no buffering) and $L = L_{opt}$. The difference in speedup is greater for the TokenRing protocol because the percentage of communication time w.r.t. computation time is more important than for the Scsi-2 protocol. Therefore, the value L_{opt} seems a good choice for ensuring a maximal overlapping of communications and computations.

5 Conclusion and future work

We presented a solution for constructing an Lts in parallel using N machines connected by a network. Each machine constructs a part of the Lts using the Distributor algorithm, all resulting parts being combined using the Bcg_Merge algorithm to form the complete Lts. These algorithms have been implemented within the Cadp tool set [FGK+96] using the generic environments Open/Cæsar [Gar98] and Bcg for implicit and explicit manipulation of Ltss.

Being independent from any specification language is a difference between our approach and other related work. To our knowledge, all published algorithms but [CGN98] are dedicated to a specific low-

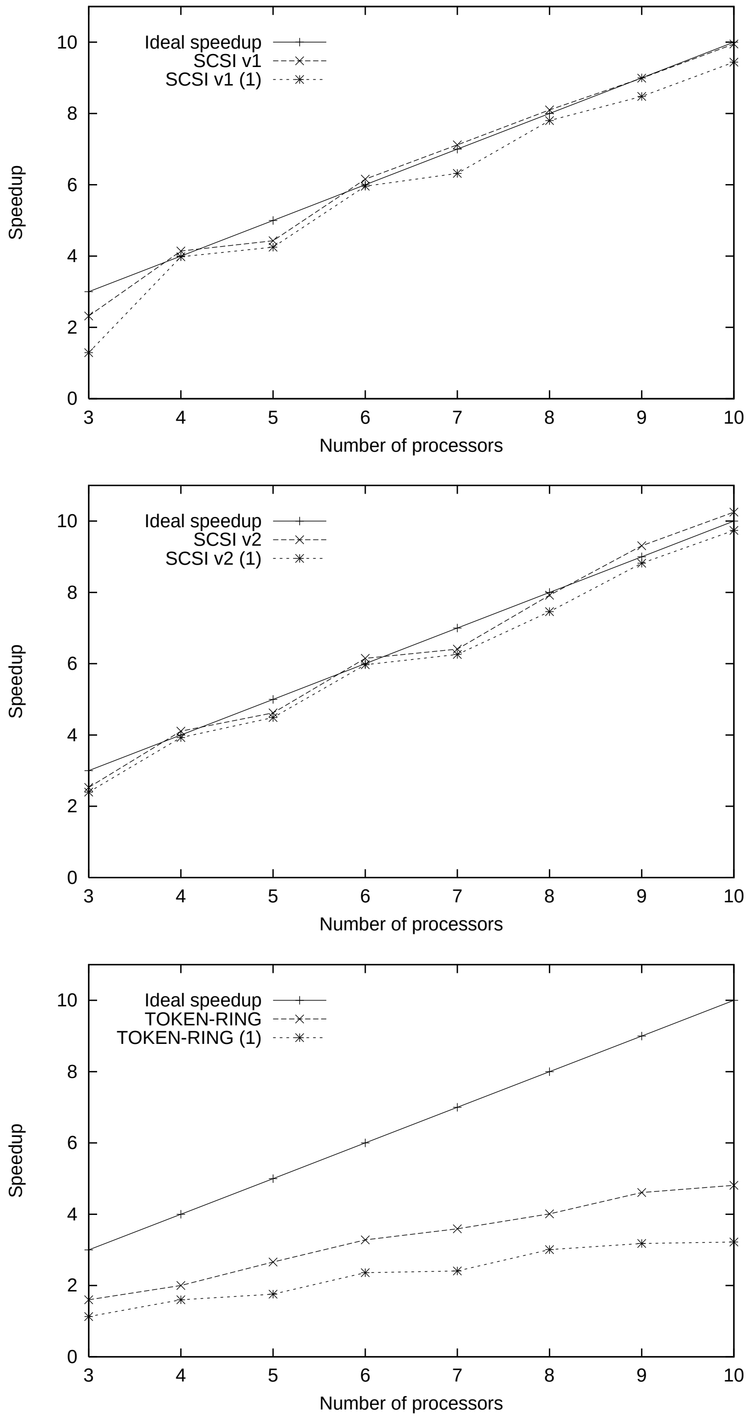


Figure 6: Speedup measurements for the Scsi-2 and TokenRing protocols for transmission buffers of size 1 and L_{opt}

level model (Petri nets, Markov chains, etc.) or high-level language (MURφ, PROMELA, etc.). On the contrary, as the OPEN/CÆSAR and BCG environments are language independent, the DISTRIBUTOR and BCG_MERGE tools can be used not only for LOTOS, but also for every language having a connection to the OPEN/CÆSAR interface, such as the UMLAUT compiler for UML [JHGP99].

Another distinctive feature of our approach relies in the scheme used by DISTRIBUTOR and BCG_MERGE to assign unique numbers to states. Although the DISTRIBUTOR algorithm is similar to the *ExploreDistributed* algorithm of [CGN98], we manage to represent states with mere integers, whereas [CGN98] uses pairs of the form $\langle$*processor number*, *local state number*$\rangle$.

We experimented our approach on several real-size LOTOS specifications, for which we generated large LTSs (up to 12 million states and 45 million transitions). Compared to the data reported for other high-level languages such as MURφ [SD97] and PROMELA [LS99], respectively, we were able to generate larger (11 times and 4.2 times, respectively) state spaces.

We believe that the memory overhead required by distribution (i.e., hash table auxiliary data structures, communication buffers, etc.) is negligible. Moreover, our experimental results show that parallel construction of LTSs provides linear speedups. This is due both to the good quality of the partition function used to distribute the state space among different machines, and to well-dimensioned communication buffers. The speedups obtained are more important for the specifications involving complex data computations and synchronizations, because in this case the traversal of LTS transitions becomes time expensive and can be distributed profitably across different machines.

In this report, we focused on the problem of constructing LTSs in parallel, with a special emphasis on resource management issues such as state storage in distributed memories and transition storage in distributed filesystems. For a proper separation of concerns, we deliberately avoided to mix parallel state space constructions with other issues such as on-the-fly verification. Obviously, it would be straightforward to enhance the parallel algorithms with on-the-fly verification capabilities such as deadlock detection, invariant checking, or more complex properties. However, this was not suitable to obtain meaningful experimental results (especially, the sizes of the largest state spaces that can be constructed using the parallel approach), because on-the-fly verification may either terminate early without exploring the entire state space, or explore a larger state space when relying on automata product techniques.

This work can be continued in several directions. Firstly, we plan to pursue our experiments on new examples and assess the scalability of the approach using a more powerful parallel machine, a cluster of 200 PCs that is currently under construction at INRIA Rhône-Alpes.

Secondly, we plan to extend the DISTRIBUTOR tool in order to handle specifications containing dynamic data structures, such as linked lists, trees, etc. This will require the transmission of variable length, typed data values over a network, contrary to the current implementation of DISTRIBUTOR, which uses messages of fixed length.

Finally, we will seek to determine at which point the sequential verification algorithms available in CADP (for model-checking of temporal logic formulas on LTSs, comparison and minimization of LTSs according to equivalence/preorder relations) will give up. As the sizes of LTSs constructed by DISTRIBUTOR will increase, it will be necessary to parallelize the verification algorithms themselves. Two approaches can be foreseen: parallel algorithms operating on-the-fly during the exploration of the LTS, or sequential algorithms working on (already constructed) partitioned LTSs.

Acknowledgements

We are grateful to Xavier Rousset de Pina and to Emmanuel Cecchet for interesting discussions and for providing valuable assistance in using the PC cluster of the SIRAC project of INRIA Rhône-Alpes. We also thank Adrian Curic, Frédéric Lang, and particularly Christophe Joubert for their careful reading and comments on this report.

References

[ADK97] S. Allmaier, S. Dalibor, and D. Kreische. Parallel Graph Generation Algorithms for Shared and Distributed Memory Machines. In E. H. D'Hollander, G. R. Joubert, F. J. Peters, and U. Trottenberg, editors, *Parallel Computing: Fundamentals, Applications and New Directions, Proceedings of the Conference ParCo'97 (Bonn, Germany)*, volume 12, pages 581–588. Elsevier, North-Holland, 1997.

[AKH97] S. Allmaier, M. Kowarschik, and G. Horton. State Space Construction and Steady-State Solution of GSPNs on a Shared-Memory Multiprocessor. In *Proceedings of the 7th IEEE International Workshop on Petri Nets and Performance Models PNPM'97 (Saint Malo, France)*, pages 112–121. IEEE Computer Society Press, 1997.

[ANS94] ANSI. Small Computer System Interface-2. Standard X3.131-1994, American National Standards Institute, January 1994.

[BK84] J. A. Bergstra and J. W. Klop. Process Algebra for Synchronous Communication. *Information and Computation*, 60:109–137, 1984.

[CCBF94] S. Caselli, G. Conte, F. Bonardi, and M. Fontanesi. Experiences on SIMD Massively Parallel GSPN Analysis. In G. Haring and G. Kotsis, editors, *Computer Performance Evaluation: Modelling Techniques and Tools*, volume 794 of *Lecture Notes in Computer Science*, pages 266–283. Springer Verlag, 1994.

[CCGR00] A. Cimatti, E. Clarke, F. Giunchiglia, and M. Roveri. NUSMV: a New Symbolic Model Checker. *Springer International Journal on Software Tools for Technology Transfer (STTT)*, 2(4):410–425, April 2000.

[CCM95] S. Caselli, G. Conte, and P. Marenzoni. Parallel State Space Exploration for GSPN Models. In G. De Michelis and M. Diaz, editors, *Applications and Theory of Petri Nets 1995*, volume 935 of *Lecture Notes in Computer Science*, pages 181–200. Springer Verlag, 1995.

[CGN98] G. Ciardo, J. Gluckman, and D. Nicol. Distributed State Space Generation of Discrete-State Stochastic Models. *INFORMS Journal on Computing*, 10(1):82–93, 1998.

[CGP00] E. Clarke, O. Grumberg, and D. Peled. *Model Checking*. MIT Press, 2000.

[CR79] Ernest Chang and Rosemary Roberts. An Improved Algorithm for Decentralized Extrema-Finding in Circular Configurations of Processes. *Communications of the ACM*, 22(5):281–283, may 1979.

[Dil96] D. Dill. The Murφ Verification System. In R. Alur and T. Henzinger, editors, *Proceedings of the 8th International Conference on Computer-Aided Verification CAV'96*, volume 1102 of *Lecture Notes in Computer Science*, pages 390–393. Springer Verlag, July 1996.

[FGK+96] Jean-Claude Fernandez, Hubert Garavel, Alain Kerbrat, Radu Mateescu, Laurent Mounier, and Mihaela Sighireanu. CADP (CÆSAR/ALDEBARAN Development Package): A Protocol Validation and Verification Toolbox. In Rajeev Alur and Thomas A. Henzinger, editors, *Proceedings of the 8th Conference on Computer-Aided Verification (New Brunswick, New Jersey, USA)*, volume 1102 of *Lecture Notes in Computer Science*, pages 437–440. Springer Verlag, August 1996.

[Gar98] Hubert Garavel. OPEN/CÆSAR: An Open Software Architecture for Verification, Simulation, and Testing. In Bernhard Steffen, editor, *Proceedings of the First International Conference on Tools and Algorithms for the Construction and Analysis of Systems TACAS'98 (Lisbon, Portugal)*, volume 1384 of *Lecture Notes in Computer Science*, pages 68–84, Berlin, March 1998. Springer Verlag. Full version available as INRIA Research Report RR-3352.

[GHLL+98] W. Gropp, S. Huss-Lederman, A. Lumsdaine, E. Lusk, B. Nitzberg, W. Saphir, and M. Snir. *MPI: The Complete Reference, Vol. 2 — The MPI-2 Extensions.* MIT Press, 1998.

[GM97] Hubert Garavel and Laurent Mounier. Specification and Verification of Various Distributed Leader Election Algorithms for Unidirectional Ring Networks. *Science of Computer Programming*, 29(1–2):171–197, July 1997. Special issue on Industrially Relevant Applications of Formal Analysis Techniques. Full version available as INRIA Research Report RR-2986.

[HBB99] B. R. Haverkort, A. Bell, and H.C. Bohnenkamp. On the Efficient Sequential and Distributed Generation of Very Large Markov Chains from Stochastic Petri Nets. In *Proceedings of the 8th International Workshop on Petri Nets and Performance Models PNPM'99 (Zaragoza, Spain)*, pages 12–21. IEEE Computer Society Press, 1999.

[HGGS00] T. Heyman, D. Geist, O. Grumberg, and A. Schuster. Achieving Scalability in Parallel Reachability Analysis of Very Large Circuits. In E. A. Emerson and A. P. Sistla, editors, *Proceedings of the 12th International Conference on Computer-Aided Verification CAV'2000 (Chicago, IL, USA)*, volume 1855 of *Lecture Notes in Computer Science*, pages 20–35. Springer Verlag, July 2000.

[Hoa85] C. A. R. Hoare. *Communicating Sequential Processes.* Prentice-Hall, 1985.

[Hol91] Gerard J. Holzmann. *Design and Validation of Computer Protocols.* Software Series. Prentice Hall, 1991.

[Hol97] G. Holzmann. The Model Checker SPIN. *IEEE Transactions on Software Engineering*, 23(5):279–295, May 1997.

[ISO88] ISO/IEC. LOTOS — A Formal Description Technique Based on the Temporal Ordering of Observational Behaviour. International Standard 8807, International Organization for Standardization — Information Processing Systems — Open Systems Interconnection, Genève, September 1988.

[JHGP99] J-M. Jézéquel, W.M. Ho, A. Le Guennec, and F. Pennaneac'h. UMLAUT: an Extendible UML Transformation Framework. In R.J. Hall and E. Tyugu, editors, *Proceedings of the 14th IEEE International Conference on Automated Software Engineering ASE'99.* IEEE, 1999. Also available as INRIA Technical Report RR-3775.

[KH99] W. J. Knottenbelt and P. G. Harrison. Distributed Disk-Based Solution Techniques for Large Markov Models. In B. Plateau, W. J. Stewart, and M. Silva, editors, *Proceedings of the 3rd International Meeting on the Numerical Solution of Markov Chains NSMC'99 (Zaragoza, Spain)*, pages 58–75. University of Zaragoza, September 1999.

[KMHK98] W. J. Knottenbelt, M. A. Mestern, P. G. Harrison, and P. Kritzinger. Probability, Parallelism and the State Space Exploration Problem. In R. Puigjaner, N. N. Savino, and B. Serra, editors, *Proceedings of the 10th International Conference on Computer Performance Evaluation - Modelling, Techniques and Tools TOOLS'98 (Palma de Mallorca, Spain)*, volume 1469 of *Lecture Notes in Computer Science*, pages 165–179. Springer Verlag, September 1998.

[Lan77] Gérard Le Lann. Distributed Systems — Towards a Formal Approach. In B. Gilchrist, editor, *Information Processing 77*, pages 155–160. IFIP, North-Holland, 1977.

[LS99] F. Lerda and R. Sista. Distributed-Memory Model Checking with SPIN. In D. Dams, R. Gerth, S. Leue, and M. Massink, editors, *Proceedings of the 5th and 6th International SPIN Workshops on Theoretical and Practical Aspects of SPIN Model Checking SPIN'99*, volume 1680 of *Lecture Notes in Computer Science*, pages 22–39. Springer Verlag, July 1999.

[Mat87] F. Mattern. Algorithms for Distributed Termination Detection. *Distributed Computing*, 2:161–175, 1987.

[MCC97] P. Marenzoni, S. Caselli, and G. Conte. Analysis of Large GSPN Models: a Distributed Solution Tool. In *Proceedings of the 7th IEEE International Workshop on Petri Nets and Performance Models PNPM'97 (Saint Malo, France)*, pages 122–131. IEEE Computer Society Press, 1997.

[Mil89] Robin Milner. *Communication and Concurrency*. Prentice-Hall, 1989.

[MS00] Radu Mateescu and Mihaela Sighireanu. Efficient On-the-Fly Model-Checking for Regular Alternation-Free Mu-Calculus. In Stefania Gnesi, Ina Schieferdecker, and Axel Rennoch, editors, *Proceedings of the 5th International Workshop on Formal Methods for Industrial Critical Systems FMICS'2000 (Berlin, Germany)*, GMD Report 91, pages 65–86, Berlin, April 2000. Also available as INRIA Research Report RR-3899.

[NC97] D. Nicol and G. Ciardo. Automated Parallelization of Discrete State-Space Generation. *Journal of Parallel and Distributed Computing*, 47(2):153–167, 1997.

[Rom99] Judi Romijn. Model Checking the HAVi Leader Election Protocol. Technical Report SEN-R9915, CWI, Amsterdam, The Netherlands, June 1999. submitted to Formal Methods in System Design.

[RS97] Y.S. Ramakrishna and S.A. Smolka. Partial-Order Reduction in the Weak Modal Mu-Calculus. In A. Mazurkiewicz and J. Winkowski, editors, *Proceedings of the 8th International Conference on Concurrency Theory CONCUR'97*, volume 1243 of *Lecture Notes in Computer Science*, pages 5–24. Springer Verlag, 1997.

[SD97] U. Stern and D. Dill. Parallelizing the Murφ Verifier. In O. Grumberg, editor, *Proceedings of the 9th International Conference on Computer-Aided Verification CAV'97 (Haifa, Israel)*, volume 1254 of *Lecture Notes in Computer Science*, pages 256–267. Springer Verlag, June 1997.

Unité de recherche INRIA Rhône-Alpes
655, avenue de l'Europe - 38330 Montbonnot-St-Martin (France)

Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399