

Applying Multi-Core Model Checking to Hardware-Software Partitioning in Embedded Systems

Alessandro Trindade · Hussama Ismail · Renato Degelo · Edilson Galvão · Lucas Cordeiro

the date of receipt and acceptance should be inserted later

Abstract We present an alternative approach to solve the hardware (HW) and software (SW) partitioning problem, which uses Bounded Model Checking (BMC) based on Satisfiability Modulo Theories (SMT) in conjunction with a multi-core support using Open Multi-Processing. The multi-core SMT-based BMC approach allows initializing many verification instances based on processors cores numbers available to the model checker. Each instance checks for a different optimum value until the optimization problem is satisfied. The goal is to show that multi-core model-checking techniques can be effective, in particular cases, to find the optimal solution of the HW-SW partitioning problem using an SMT-based BMC approach. We compare the experimental results of our proposed approach with Integer Linear Programming and the Genetic Algorithm.

Keywords hardware-software co-design · hardware-software partitioning · optimization · model checking · multi-core · OpenMP

1 Introduction

Nowadays, with the strong development of embedded systems, the design phase plays an important role. At early stages, the design is split into separated flows: hardware and software. Consequently, the partitioning decision process, which deals with the decisions upon which parts of the application have to be designed in hardware (HW) and which in software (SW), must be supported by any well-structured methodology. If not, this leads to a number of issues (design flow interruptions, redesigns, and undesired iterations) which affects

the overall development process, the quality and the lifecycle of the final system. Starting at the 1990s, intensive research was performed, and several approaches proposed, as shown in [1] and [2].

In any HW and SW design of complex systems, more time is spent on verification than on construction [3]. Formal methods based on model checking offer great potential to obtain a more effective and faster verification in the design process. Programs may be viewed as mathematical objects with behavior that is, in principle, well determined. This makes it possible to specify programs using mathematical logic, which constitutes the intended (correct) behavior. Then, one can try to give a formal proof or otherwise establish that the program meets its specification [4]. Research in formal methods has led to the development of very promising verification techniques, which facilitate the early detection of errors. Model-based verification techniques use models that describe the possible system behavior in a mathematically precise and unambiguous manner. The system models are accompanied by algorithms that systematically explore all the states of the system model.

In [5] and [6] was shown that it is possible to use Bounded Model Checking (BMC) based on Satisfiability Modulo Theories (SMT) to perform HW-SW partitioning in embedded systems. The present work extends those studies since there is a substantial improvement in terms of the genetic algorithm and the SMT-based verification method, which has been extended with a multi-core architecture. Multi-core processors have been used in all segments of industry to implement high-performance computing [7]. In particular, hardware platforms, together with multi-processing platforms, have allowed verification algorithms to distribute tasks executions across multiple processors, which generate an increase in performance if compared to single-

core solution. However, most verification algorithms still disregard the limitations of the CMOS technology, which limits the increase of the chip's frequency after it reaches 4 GHz.

Here, we exploit the availability of multi-core processors. In particular, a multi-core SMT-based BMC method is applied to the HW-SW partitioning and then is compared to the results with classical integer linear programming (ILP) and genetic algorithm (GA) using a multi-core tool as well. To the best of our knowledge, this is the first work to use a multi-core SMT-based verification to solve a HW-SW partitioning problem in embedded systems. We implement our ideas with the Efficient SMT-based Bounded Model Checker (ES-BMC) tool [14]. As its main contribution, this paper shows that it is possible to take advantage of an SMT-based BMC tool in a multi-core architecture to solve optimization problems.

This paper is organized as follows: Section II gives a background on optimization, model checking, and multi-core support with Open MultiProcessing. Section III describes informal and formal mathematical modeling. Section IV describes briefly the binary integer programming and GA algorithms. The SMT-based BMC method is presented in Section V. Section VI presents the experimental evaluation. Section VII discusses related work. Section VIII presents the conclusion and future work.

2 BACKGROUND

2.1 Optimization

Optimization is the act of obtaining the best result (i.e., the optimal solution) under given circumstances [9]. In the design, construction, and maintenance of any engineering system, engineers have to make many technological and managerial decisions at several stages. The ultimate goal of all such decisions is either to minimize the effort required or to maximize the desired benefit. Because the effort required or the benefit desired in any practical situation can be expressed as a function of certain decision variables, optimization can be defined as the process of finding the conditions that give the maximum or minimum value of a function [9].

There is no single method available for solving all optimization problems efficiently [9]. The most known technique is linear programming, which is a method applicable for the solution of problems in which the objective function and the constraints appear as linear functions of the decision variables. A particular case of linear programming is ILP, in which the variables can assume just integer values. Eq. (1) shows a typical

linear programming problem, where \mathbf{a} and \mathbf{b} are vectors or matrixes that describe the constraints

$$\min f^t x \text{ such that } = \begin{cases} A \cdot x \leq b, \\ Aeq \cdot x = beq, \\ x \geq 0. \end{cases} \quad (1)$$

In some cases, the time to find a solution using ILP is impractical. Even with the use of powerful computers, a problem can take hours running before an optimal solution is reached. If the optimization problem is complex, some heuristics can be used to solve the same problem faster, e.g., those used in the GA [9]. The only drawback is that the found solution may not be the global minimum or maximum.

2.2 Optimization with Vz

Um solucionador SMT é uma ferramenta com finalidade de verificar a satisfabilidade de formulas baseada em um conjunto de regras SMT. Z3 é um solucionador SMT da Microsoft Research, disponibilizado livremente para a comunidade que integra suporte para uma variedade de teorias. Z3 traz consigo um solucionador DPLL juntamente de uma máquina de correspondência para quantificadores. Z3 é implementado em C++.

vZ é uma das ferramentas implementadas no solucionador Z3, a função base desta ferramenta é otimizar objetivos dentro do contexto lógico das restrições, este se destaca pelas funções que dão a possibilidade de formular critérios de otimização. vZ inclui uma versão incremental de MaxRes para MaxSAT e um Simplex para solucionar números sem padrão definido, em seu núcleo MaxSat é responsável pelas restrições enquanto OptSMT tem como função otimizar objetivos de aritmética linear. Dentre todos os avanços do Z3 usando vZ destaca-se as funções: maximize, minimize e assert-soft.

- **maximize(T)** esta instrução diz ao solucionador que a variável T deverá ser maximizada, e possível maximizar valores reais, inteiros e vetores de bit.
- **minimize(T)** esta instrução diz ao solucionador que a variável T deverá ser minimizada, os tipos aceitos pelo minimize e similar ao maximize.
- **Assert-Soft F :weight n** a instrução Assert-soft é uma restrição sobre F, e possível adicionar um peso n, o padrão é 1.

Vamos otimizar $(K+W)$, com restrições em $(K < 2)$ e $(W-K < 1)$, o resultado da otimização descrita no código abaixo é 2, o modelo gerado pelo vZ mostra que $K = 1$ e $W = 1$.

```

1. (declare-const K Int)
2. (declare-const W Int)
3. (assert (< K 2))
4. (assert (< (- W K) 1))
5. (maximize (+ K W))
6. (check-sat)

```

De acordo com [34] a arquitetura do vZ pode ser descrita na imagem XXX, inicialmente as formulas e os objetivos SMT sao reescritos para variaveis do tipo 0-1, posteriormente os objetivos sao convertidos para problemas MaxSAT, se houver muitos objetivos vZ invoca MaxSAT ou OptSMT. Para restricoes sob valores reais usando “Box”, vZ combina todas a aritmetica linear e faz chamada para OptSMT. Quando usamos “soft constraints” no modo “lexicographic” vZ invoca MaxSAT.

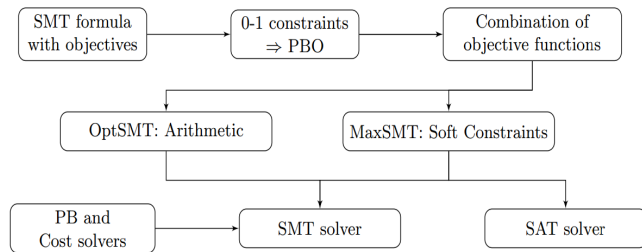


Fig. 1 vZ Architecture

Z3 is available for platforms like C, C++, Java, .NET and Python, it is possible to download the Z3 with vZ in <http://z3.codeplex.com>. API in python was used to formulate the optimization problem on the vZ tool. The pseudo code below describes the steps to achieve desired results.

2.3 Bounded Model Checking with ESBMC

2.4 Multi-core ESBMC with OpenMP

2.5 Multi-core ESBMC with OpenMP using Binary Search

The original parallelization was made by continuously forking the instances of the ESBMC until the first violation is find.

Since OpenMP only returns from a parallelized loop when every forked thread finishes, some processor could remain idle for some time. So the first approach was to remove the Idle time from the parallel loops by creating workers inside the threads to always execute the next step until a violation is find. At first though it seemed to be a great improvement but as ESBMC executes almost the same time the steps, there was almost no optimization.

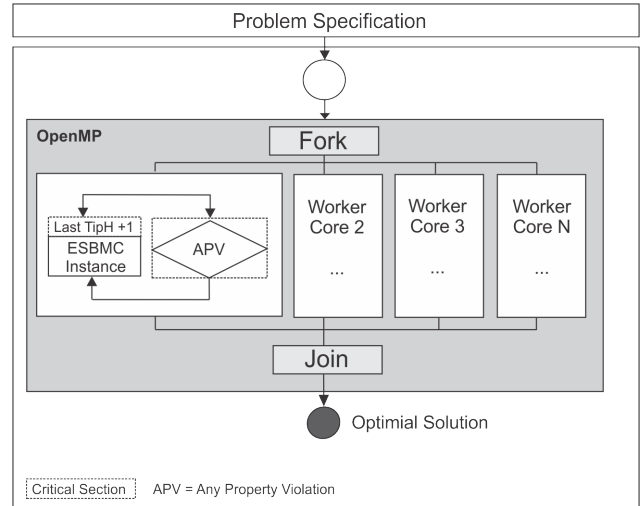


Fig. 2 gestaucht

The most optimized approach used was to use a parallelized binary search to reduce the complexity of the tasks. So was made a controller to return the best step to execute in order to reduce the domain as maximum as possible. This is done by dividing the domain in chunks and then returning the middle of the biggest chunk existent, and creating two new chunks.

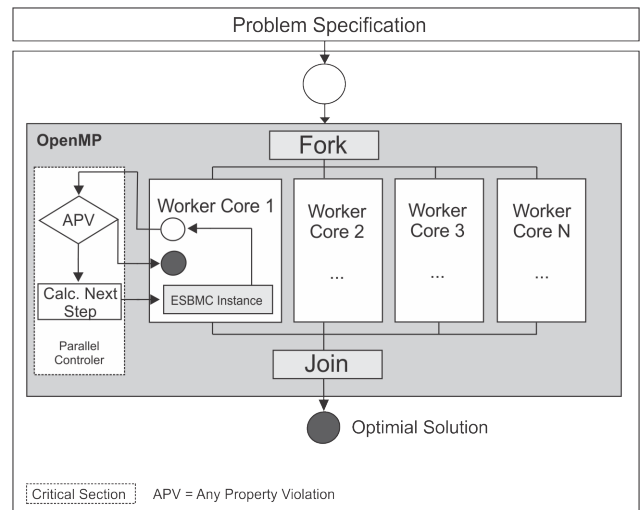


Fig. 3 gestaucht

To exemplify, given a problem of domain 1 to 20, first we create an initial chunk 1 – 20. When the next available core asks for a step to execute, the controller gets the biggest chunk which in this case is 1 – 20, divide it by 2, creates two new chunks and return the middle of the original chunk. So it would be 2 new chunks 1 – 9 and 11 – 20, and the returned step will be 10. Additional checks are needed to avoid the creation of empty or invalid chunks.

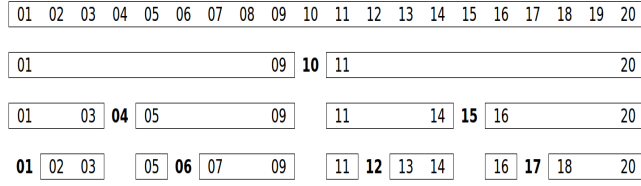


Fig. 4 gestaucht

Additionally, since there will be a gap between the steps, produced by the customized binary search algorithm, it is possible that a lot of steps will not be needed to execute. For instance, on the example above, if a step 10 return false, it is obvious that every step after 10 will also be a violation. On the other hand, if the same step was successful, we can assume that every step before 10 will also be a valid.

That said, was implemented in the controller a method to discard unnecessary steps by removing or shrinking existing chunks. It made a very big impact in the calculation time. However a step can be running and not be needed anymore, in this case the binary algorithm kills the forked process and starts a new one.

```

01. GetNextStep(){
02.  int largestChunk = -1;
03.  chunk largest;
04.  for each(chunk in chunks){
05.      if(chunk.right - chunk.left > largestChunk){
06.          largestChunk = chunk.right - chunk.left;
07.          largest = chunk;
08.      }
09.  }
10.  chunks.remove(largest);
11.  int median = largest.left +
    floor((largest.right - largest.left) / 2)
12.  if(median > 0){
13.      if(largest.right - largest.left > 1)
14.          chunks.add(new chunk(largest.left, median
    - 1));
15.      if(largest.right != largest.left)
16.          chunks.add(new chunk(median + 1,
    largest.right);
17.  }
18.  return median;
19. }
```

The execution is made by executing the procedure described pseudocode before and then starting the process of the ESBMC to execute the returned step. While the step is being executed the algorithm query the controller to check if this step is still needed. If not, the ESBMC instance is killed and the worker is free to start another step.

```

01. step = controller.GetNextStep();
02. int pid = ExecuteStep(step);
03. while(isRunning(pid)){
04.     if(!controller.isNeeded(step))
```

```

05.     kill(pid);
06. }
```

3 Mathematical modeling

The mathematical modeling was taken from [1], [2].

3.1 Informal Model (or Assumptions)

The informal model can be described by five characteristics. First, there is only one software context, i.e., there is just one general-purpose processor, and there is only one hardware context. The components of the system must be mapped to either one of these two contexts. Second, the software implementation of a component is associated with a software cost, which is the running time of the component. Third, the hardware implementation of a component has a hardware cost, which can be area, heat dissipation, and energy consumption. Fourth, based on the premise that hardware is significantly faster than software, the running time of the components in hardware is considered as zero. Finally, if two components are mapped to the same context, then there is no overhead of communication between them; otherwise, there is an overhead. The consequence of these assumptions is that scheduling does not need to be addressed in this work. Hardware components do not need scheduling, because the running time is assumed to be zero. Because there is only one processor, software components do not need to be scheduled as well. Therefore, the focus is only on the partitioning problem. That configuration describes a first-generation co-design, where the focus is on bipartitioning [17].

3.2 Formal Model

...

4 Partitioning problem using ILP-based, Genetic Algorithms

The ILP and GA were taken from [5] and [6]. Both use slack variables in order to be possible to represent the constraints and to use commercial tools. However, GA had improvements from the parameters of related studies in order to increase the solution accuracy without producing timeout. The tuning was performed by empirical tests and resulted in changing of three parameters, which are passed to function ga of MATLAB [18]: the population size was set from 300 to 500, the

Elite count changed from 2 (default value) to 50, and the number of Generations changed from 100* NumberOfVariables (default) to 75.

5 Analysis of the partitioning problem using ESBMC

6 Analysis of the partitioning problem using vZ

O pseudocódigo abaixo mostra as condições aplicadas em ILP, GA e ESBMC convertida para a ferramenta vZ descrita em [34] e [35].

Um contexto do vZ é criado, este recebe as restrições e devolve um modelo caso exista uma solução para o conjunto de restrições. Posteriormente são declarados a quantidade de nós, quantidade de arestas, o valor de , o custo de hardware, custo de software, custo de comunicação, e a matriz de incidência E.

As declarações aritméticas descritas nas linhas 10, 11,12 e 13 do pseudocódigo é uma transcrição da fórmula descrita na seção III.B, é referente ao custo de software (SF). Observe que na linha 11 uma variável temporária chamada TMP é criada para representar , é aplicado então um produto entre o valor absoluto de TMP com o valor do vetor C declarado na linha 07, logo, gera-se o resultado para o custo de comunicação (CMC). Na linha 13, a declaração Fobj (Função objetivo) é um produto entre o custo de hardware com um vetor que contém valores 0 ou 1. Fobj deve ser minimizado para alcançarmos a melhor solução para o hardware.

Duas restrições são impostas, a primeira refere-se à soma entre custo de software com o custo de comunicação, a resultante da soma deve ser menor do que , a segunda restrição diz ao vZ que ele deverá minimizar Fobj. O modelo então é checado, uma vez que há uma solução que obedeça às restrições, o valor da função objetiva é impressa.

```

01.#Init Values
02. Create vZ context
03. Create binary vector (x)
04. Declare number of nodes, edges and S0
05. Declare hardware cost of each node as array (h)
06. Declare software cost of each node as array (s)
07. Declare communication cost of each edge (c)
08. Declare transposed incidence matrix graph G (E)
09.#Arithmetic
10. SF receive s(1-x)
11. TMP vector receive E|x|
12. CMC receive c*|EX|
13. Fobj receive x[i] * h[i]
14.#Asserts
15. Add restrictions (SF + CMC <= S0)
16. Add restrictions to minimize Fobj
17. Check Model

```

18. Print Result

7 Experimental Evaluation

8 Related Work

Since the second half of the first decade of the 2000s, three main paths have been tracked to improve or to present alternative solutions to the optimization of HW-SW partitioning, i.e., to find the exact solution [2], to use heuristics to speed up performance time [1], and hybrid ones [22].

In the first group, the exact solution to the HW-SW partitioning problem is found. The use of SMT-based verification presented in this paper can be grouped into this category, because the exact solution is found with the given algorithm. The difference is based only in terms of the technique chosen to solve the problem. Another path followed in past initiatives and which has had more studies is the creation of heuristics to speed up the running time of the solution. The difference between this kind of solutions and SMT-based verification is based on two facts: ESBMC is guaranteed to find the exact solution, but the heuristics are faster when the complexity is greater.

Finally, there are approaches that mixes heuristics with exact solution tools. The idea is to use a heuristic to speed up some phase of an exact solution tool. It worth mentioning that the final solution is not necessarily an optimal global solution. Only the SMT-based verification is guaranteed to find the exact solution, but hybrid algorithms are faster when complexity rises.

In terms of SMT-based verification, most work is restricted to present the model, its modification to programming languages (e.g., C/C++ and Java), and the application to multi-thread algorithms or to embedded systems to check for program correctness. In [16] it presents a bounded model checker for C++ programs, which is an evolution of dealing with C programs and [14] uses the ESBMC model checker for embedded ANSI-C software. In [5] and [6] it was proven that it is possible to use ESBMC to solve HW-SW partitioning, but in a single core way. There are related studies focused on decreasing the verification time of model checkers by applying Swarm Verification [29], and modifications of internal search engines to support parallelism [30], but there is still the need for initiatives related to parallel SMT solvers [31]. Recently, the SMT solver Z3 has been extended to pose and solve optimization problems modulo theories [32].

9 Conclusions

Concerning the comparative tests, with the four techniques presented in this paper to solve HW-SW partitioning, it was evident that none of them is indicated to partition problems with more than 400 nodes. The computing time to solve the optimization problem reached some hours of execution on a standard desktop computer. If we consider less than 400 nodes, then it is possible to use ILP as the best solution provider. If the problem to be solved has 150 nodes or less, then ESBMC represents a feasible alternative. GA had an intermediate result in terms of performance, but the error presented from exact solution made it not acceptable to that kind of application. This error may be reduced by changing some parameters.

If considering off-the-shelf tools, as MATLAB to ILP and GA, the coding is simpler. However, ESBMC has a BSD-style license and can be downloaded and used for free. Concerning the two versions of ESBMC, it is possible to conclude that Multi-core ESBMC had better performance results than pure ESBMC. Thus, considering that nowadays the processors have more and more cores, when modeling the problem, it is possible to consider multi-core ESBMC as an alternative to solve the partitioning problem. Future work can be done to decrease the processing time of ESBMC (solver included).

Finally, there is an issue about 150 nodes problem, since it seems to be the limit of ESBMC. It really depends on the modeling granularity of the problem. Some researchers propose fine-grained models, in which each instruction can be mapped to either HW or SW. This may lead to thousands of nodes or even more. Others defend coarse-grained models, where decisions are made for bigger components, thus even complex systems may consist of just some dozens of nodes to partition. In principle, a fine-grained approach may allow to obtain better partitions, but at the cost of an exponential increase of the size of the search space. In future work, we will exploit other search strategies in the multi-core ESBMC approach and address specifically more complex types of architectures in the HW-SW partitioning problem, including more than one CPU.

10 References