Meeting Deadlines Cheaply

Julien Legriel, Oded Maler
CNRS-Verimag
2, av. de Vignate
38610 Gieres, France
Email: {legriel,maler}@imag.fr

***Abstract*—We develop a computational framework for solving the problem of finding the cheapest configuration (in terms of the number of processors and their respective speeds) of a multiprocessor architecture on which a task graph can be scheduled within a given deadline. We then extend the problem in three orthogonal directions: taking communication volume into account, considering the case where a stream of instances of the task graph arrives periodically and reformulating the problem as a bi-criteria optimization for which we approximate the Pareto front.**

I. Introduction

This paper is motivated by several recent developments in computer technology, computer applications and algorithmics: 1) The shift toward multi-core computer architectures renewed the interest in efficient parallel execution of programs; 2) Mobile platforms such as tablets and phones are becoming central and consequently the problem of reducing power consumption has become crucial (also for data centers). One way of reducing the power consumption of a processor is to reduce its voltage and frequency to the minimal level in which it can still meet performance constraints; 3) New classes of applications characterized by ongoing streams of *structured* computational tasks that come from the outside environment (unlike the classical problems of loop parallelization) and where significant amounts of data are exchanged among tasks; 4) The recent progress in Boolean SAT solving, and in SAT Modulo Theories (SMT) has led to a new generation of powerful tools for solving mixed combinatorial and numerical constrained optimization problems.

This work is part of the French Minalogic Athole project, and inspired by the multi-core architecture xSTream designed by STMicroelectronics as a potential next-generation platform targeting mainly mobile and multi-media applications. This architecture is designed around a "streaming fabric" consisting of a set of processors or specialized hardware blocks connected via flow-control mechanisms on top of a network-on-chip (NoC) [21]. To facilitate power saving, the system has means to manage several power modes, for example, to change the state of the different computing nodes: on/off, idle, low-voltage/lower clock rates, etc. Energy-efficient scheduling of applications on this architecture is one of the primary goals of the Athole project: let the system work at the least energy-consuming configuration which can still meet the computational demands of the application.

In this work we formalize this problem as an extension of the well-known task-graph scheduling problem [23], [31], [29], where with each task we associate a *quantity of work* whose translation into *duration* depends on the speed of the processor on which it executes. A configuration of the architecture is characterized by the number of processors running in each of the frequencies, from which a cost function is derived, reflecting the (static) energy consumption of the configuration. Given such a task graph and an upper bound on its acceptable execution time (deadline) we ask the following question: *what is the cheapest configuration on which the task graph can be scheduled for execution while meeting the deadline?* We formulate the question as a mixed logical-numerical constrained optimization problem and solve it using the SMT solver Yices [30]. We develop a binary search algorithm over the cost space which uses the solver to provide solutions with guaranteed distance from the optimum. We manage to handle randomly-generated task-graph problems having around 40-50 tasks on execution platforms consisting of machines with 3 different speeds.

We then extend the problem in three directions. First, we annotate the task graph with quantities of data that have to be communicated among tasks and refine the model of the execution platform to include communication network topology. We then pose the same problem where the communication channels are considered as additional resources occupied for durations proportional to the amount of transmitted data, thus incorporating *data locality* considerations into the model. In this setting we can solve problems with 15-20 tasks on a *spidergon* topology [24] with up to 8 processors. Secondly we formulate a periodic extension of the problem where task instances arrive every ω time units and must be finished within a relative deadline $\delta > \omega$. Efficient resource utilization for this problem involves *pipelined* execution, which may increase the number of decision variables and complicate the constraints. We solve this infinite problem by reducing it to finding a schedule for a sufficiently-long finite unfolding. Preliminary experiments show that we can treat the periodic case where the number of unfolded tasks is up to 100. Finally, rather than keeping the deadline fixed as a constraint and optimize the platform cost, we reformulate the problem as a bi-criteria optimization problem and compute efficient trade-offs between these two conflicting goals.

The rest of the paper is organized as follows. In Section II we give the basic definitions of execution platforms, task graphs and feasible schedules. In Section III we discuss briefly several approaches for handling mixed constraints in

optimization and describe our encoding of the problem using a Boolean combination of linear constraints. We then present our search procedure along with experimental results obtained using the Yices solver. The extension of the model to treat communication is sketched in Section IV, Section V is devoted to the definition and encoding of a periodic version of the problem and Section VI to the multi-criteria reformulation. All these sections are accompanied with experimental results. Finally in Section VII we mention some related work, discuss the limitations of our models and suggest future research directions.

II. Problem Formulation

A. Execution Platforms

We formalize here the notion of an execution platform, which represents a possible configuration of a multiprocessor system, characterized by the number of active "machines" and their respective speeds. We assume a fixed set of speeds $V = \{v_0, v_1, \ldots, v_m\}$ with $v_0 = 0$ and $v_k < v_{k+1}$, each representing an amount of work (for example, number of instructions) that can be provided per unit of time. We do not bound a priori the number of machines.

Definition 1 (Execution Platform): An execution platform is a non-increasing function $R : \mathbb{N}_+ \to V$ assigning a speed to each machine, with $R(j) = 0$ indicating that machine j is turned off (or does not exist). A platform is finite if $R(j) = 0$ for every $j > j_*$ for some j_*.

It is sometimes convenient to view R as a vector $R = (r_1, \ldots, r_m)$ with r_k being the number of machines working at speed v_k. To compare different platforms we use a *static* cost model that depends only on the membership of machines of various speeds in the platform and not on their actual utilization during execution.

Definition 2 (Platform Cost): The cost associated with a platform $R = (r_1, \ldots, r_m)$ is $C(R) = \sum_{k=1}^{m} c_k \cdot r_k$ where $c_k < c_{k+1}$ are technology-dependent positive constants.

Below we define some useful measures on platforms related to their work capacity.

- Number of active machines: $\mathcal{N}(R) = \sum_{k=1}^{m} r_k$;
- Speed of fastest machine: $\mathcal{F}(R) = R(1)$;
- Work capacity: $\mathcal{S}(R) = \sum_{k=1}^{m} r_k \cdot v_k = \sum_i R(i)$.

The last measure gives an upper bound on the quantity of work that the platform may produce over time when it is fully utilized.

B. Work Specification

The work to be done on a platform is specified by a variant of a *task graph*, where the size of a task is expressed in terms of work rather than duration.

Definition 3 (Task Graph): A task graph is a triple $G = (P, \prec, w)$ where $P = \{p_1, \ldots p_n\}$ is a set of tasks, $\prec$ is a partial order relation on P with a unique[1] minimal element p_1 and a unique maximal element p_n. The function $w : P \to \mathbb{N}$ assigns a quantity of work to each task. When a task p is executed on a machine working in speed v, its execution time is $w(p)/v$.

The following measures give an approximate characterizations of what is needed in terms of time and work capacity in order to execute G.

- The width $\mathcal{W}(G)$ which is the maximal number of $\prec$-incomparable tasks, indicates the maximal parallelism that can be useful;
- The length $\mathcal{L}(G)$ of the longest (critical) path in terms of work, which gives a lower bound on execution time;
- The total amount of work $\mathcal{T}(G) = \sum_i w(p_i)$ which together with the number of machines gives another lower bound on execution time.

A schedule for the pair (G, R) is a function $s : P \to \mathbb{N} \times \mathbb{R}_+$ where $s(p) = (j, t)$ indicates that task p starts executing at time t on machine j. We will sometimes decompose s into s_1 and s_2, the former indicating the machine and the latter, the start time. The duration of task p under schedule s is $d_s(p) = w(p)/R(s_1(p))$. Its execution interval (we assume no preemption) is $[s_2(p), s_2(p) + d_s(p)]$. A schedule is feasible if the execution intervals of tasks satisfy their respective precedence constraints and if they do not violate the resource constraints which means that they are disjoint for all tasks that use the same machine.

Definition 4 (Feasible Schedule): A schedule s is feasible for G on platform R if it satisfies the following conditions:

1) *Precedence*: If $p \prec p'$ then $s_2(p) + d_s(p) \leq s_2(p')$;
2) *Mutual exclusion*: If $s_1(p) = s_1(p')$ then $[s_2(p), s_2(p) + d_s(p)] \cap [s_2(p'), s_2(p') + d_s(p')] = \emptyset$.

The total duration of a schedule is the termination time of the last task $\ell(s) = s_2(p_n) + d_s(p_n)$.

The singular[2] deadline scheduling predicate $\textsc{sds}(G, R, \delta)$ holds if there is a feasible schedule s for G on R such that $\ell(s) \leq \delta$. The problem of finding the cheapest architecture R where this holds is then the constrained optimization problem

$$\min C(R) \ s.t. \ \textsc{sds}(G, R, \delta).$$

The harder part of the problem is to check whether, for a given architecture R, $\textsc{sds}(G, R, \delta)$ is satisfied when δ is close to the duration of the optimal schedule for G on R. Since we will be interested in approaching the cheapest platform we will often have to practically solve the optimal scheduling problem which is NP-hard.

Let us mention some observations that reduce the space of platforms that need to be considered. First, note that if $\textsc{sds}(G, R, \delta)$ is solvable, then there is a solution on a platform R satisfying $\mathcal{N}(R) \leq \mathcal{W}(G)$, because adding processors beyond the potential parallelism in G does not help. Secondly, a feasible solution imposes two lower bounds on the capacity of the platform: 1) the speed of the fastest machine should satisfy $\mathcal{F}(R) \geq \mathcal{L}(G)/\delta$, otherwise there is no way to execute the critical path before the deadline. 2) The total work capacity

[1] This can always be achieved by adding fictitious minimal and maximal tasks with zero work.

[2] To distinguish it from the periodic problem defined in Section V.

should satisfy $\mathcal{S}(R) \geq \mathcal{T}(G)/\delta$, otherwise even if we manage to keep the machines busy all the time we cannot finish the work before the deadline.

III. Constrained Optimization Formulation

A. Background

The problem of optimizing a linear function subject to linear constraints, also known as linear programming [37], is one of the most studied optimization problems. When the set of feasible solutions is *convex*, that is, it is defined as a *conjunction* of linear inequalities, the problem is easy: there are polynomial algorithms and, even better, there is a simple worst-case exponential algorithm (simplex) that works well in practice. However, all these nice facts are not of much help in the case of scheduling under resource constraints. The mutual exclusion constraint, an instance of which appears in the problem formulation for every pair of unrelated tasks executing on the same machine, is of the form $[x, x'] \cap [y, y'] = \emptyset$, that is, a *disjunction* $(x' \leq y) \vee (y' \leq x)$ where each disjunct represents a distinct way to solve the potential resource conflict between the two tasks. As a result, the set of feasible solutions is decomposed into an exponential number of disjoint convex sets, a fact which renders the nature of the problem more combinatorial than numerical. Consequently, large scheduling problems do not benefit from progress in relaxation-based methods for mixed integer-linear programming (MILP).

Techniques that are typically applied to scheduling problems [19] are those originating from the field known as constraint logic programming (CLP) [35]. These techniques are based on heuristic search (guessing variable valuations), constraint propagation (deducing the consequences of guessed assignments and reducing the domain of the remaining variables) and backtracking (when a contradiction is found). A great leap in performance has been achieved during the last decade for search-based methods for *the* generic discrete constraint satisfaction problem, the satisfiability of Boolean formulae given in CNF form (SAT). Modern SAT solvers [40] based on improvements of the DPLL procedures [26] can now solve problems comprising of hundreds of thousands of variables and clauses and are used extensively to solve design and verification problems in hardware and software.

Recently, efforts have been made to leverage this success to solve satisfiability problems for Boolean combinations of predicates, such as numerical inequalities, belonging to various "theories" (in the logical sense), hence the name *satisfiability modulo theories* (SMT) [32], [16]. SMT solvers combines techniques developed in the SAT context (search mechanism, unit resolution, non-chronological backtracking, learning, and more) with a theory-specific solver that checks the consistency of truth assignments to theory predicates and infers additional assignments. The relevant theory for standard scheduling problems is the theory of *difference constraints*, a sub theory of the theory of *linear inequalities*, but in order to cover costs and speeds we use the latter theory. To this end we use the powerful solver Yices [30] which excels on SMT problems that involves linear constraints. In the sequel we describe the problem encoding.

B. Problem Encoding

Solutions to the problem are assignments to decision variables $\{u_j\}$, $\{e_i\}$ and $\{x_i\}$ where variable u_j ranging over V, indicates the speed of machine j, integer variables e_i indicates the machine on which task p_i executes and variable x_i is its start time. Constants $\{v_k\}$ indicates the possible speeds, $\{w_i\}$ stand for the work in tasks and $\{c_k\}$ is the cost contributed by a machine running at speed v_k. We use auxiliary (derived) variables $\{d_i\}$ for the durations of tasks based on the speed of the machine on which they execute and C_j for the cost of machine j in a given configuration. The following constraints define the problem.

- The speed of a machine determines its cost:
$$\bigwedge_j \bigwedge_k (u_j = v_k \Rightarrow C_j = c_k)$$
- Every task runs on a machine with a positive speed and this defines its duration:
$$\bigwedge_i \bigwedge_j ((e_i = j) \Rightarrow (u_j > 0 \wedge d_i = w_i/u_j))$$
- Precedence: $\bigwedge_i \bigwedge_{i' : p_i \prec p_{i'}} x_i + d_i \leq x_{i'}$
- Mutual exclusion:
$$\bigwedge_i \bigwedge_{i' \neq i} ((e_i = e_{i'}) \Rightarrow ((x_i + d_i \leq x_{i'}) \vee (x_{i'} + d_{i'} \leq x_i))$$
- Deadline: $x_n + d_n \leq \delta$
- Total architecture cost: $C = \sum_j C_j$.

We use additional constraints that do not change the satisfiability of the problem but may reduce the space of the feasible solutions. They include the above mentioned lower bounds on architecture size and a "symmetry breaking" constraint which orders the machines according to speeds, avoiding searching among equivalent solutions that can be transformed into each other by permuting the indices of the machines.

C. Implementation and Experimental Results

We have implemented a prototype tool that takes a task graph as an input, performs preliminary preprocessing to compute width, critical path and quantity of work and then generates the constraint satisfaction problem in the solver input language. The solver has no built-in optimization capabilities and we can pose only queries of the form $\textsc{sds}(G, R, \delta) \wedge C(R) \leq \mathbf{c}$ for some cost $\mathbf{c}$. We use $\psi(\mathbf{c})$ as a shorthand for this query. The quest for a cheap architecture is realized as a sequence of calls to the solver with various values of $\mathbf{c}$. Performing a search with an NP-hard problem in the inner loop is very tricky since, the closer we get to the optimum, the computation time becomes huge, both for finding a satisfying solution and for proving unsatisfiability (from our experience, the procedure may sometimes get stuck for hours near the optimum while it takes few seconds for slightly larger or

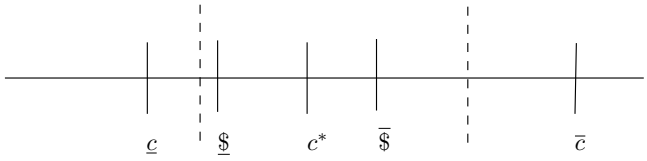

Fig. 1. Searching for the optimum. The dashed line indicate the interval toward which the algorithm will converge, the best estimation of the optimum for time budget θ.

smaller costs). We have implemented the following search algorithm. We fix a time budget θ beyond which we do not wait for an answer (currently 5 minutes on a modest laptop). The outcome of the query $\psi(\mathbf{c})$ can be either

$$\begin{array}{rcl} (1, c) & : & \text{There is a solution with cost } c \leq \mathbf{c} \\ 0 & : & \text{There is no solution} \\ \$ & : & \text{The computation is too costly} \end{array}$$

At every stage of the search we maintain 4 variables: $\underline{c}$ is the maximal value for which the query is not satisfiable, $\underline{\$}$ is the minimal value for which the answer is \$, $\overline{\$}$ is the maximal value for which the answer is \$, and $\overline{c}$ is the minimal solution found (see Figure 1). Assuming that computation time grows monotonically as one approaches the optimum from both sides, we are sure not to get answers if we ask $\psi(\mathbf{c})$ with $\mathbf{c} \in [\underline{\$}, \overline{\$}]$. So we ask further queries in the intervals $[\underline{c}, \underline{\$}]$ and $[\overline{\$}, \overline{c}]$ and each outcome reduces one of these intervals by finding a larger value of $\underline{c}$, a smaller value of $\underline{\$}$, a larger value of $\underline{\$}$ or a smaller value of $\overline{c}$. Whenever we stop we have a solution $\overline{c}$ whose distance from the real optimum is bounded by $\overline{c} - \underline{c}$. This scheme allows us to benefit from the advantages of binary search (logarithmic number of calls) with a bounded computation time.

We did experiments with this algorithm on a family of platforms with 3 available speeds $\{1, 2, 3\}$. The costs associated with the speeds are, respectively, 1, 8 and 27, reflecting the approximate cubic dependence of energy on speed. We have experimented with numerous graphs generated by TGFF tool [28] and could easily find solutions for problems with 40-50 tasks. Figure 2 illustrates the influence of the deadline constraints on the platform and the schedule for a 10-task problem of width 4. With deadline 100 the problem can be scheduled on the platform $R = (0, 0, 3)$, that is, 3 slow processors, while when the deadline is reduced to 99, the more expensive platform $R = (0, 1, 1)$ is needed.

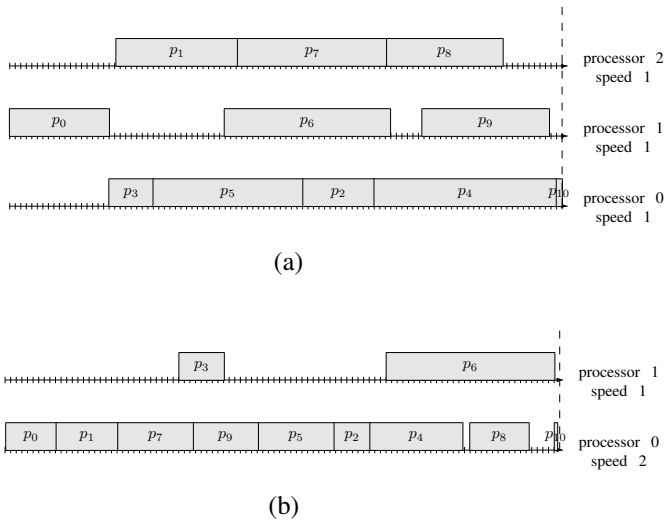

Fig. 2. The effect of deadline: (a) a schedule on a cheap architecture with deadline 100; (b) a more expensive architecture needed for deadline 99.

IV. Adding Communication

The model described in the preceding section neglects communication costs. While this may be appropriate for some traditional applications of program parallelization, it is less so for modern streaming applications that have to pass large amounts of data among tasks and there is a significant variation in communication time depending on the *distance* between the processors on which two communicating tasks execute. To this end we extend the models as follows.

On the application side we annotate any edge $p \prec p'$ in the precedence graph with a number indicating the *quantity* of data that p transmits to p'. We assume that the data items sent by a task to each of its successors are disjoint. Furthermore, we assume that a task waits for the arrival of *all* its data before execution and that it transmits them *after* termination. On the architecture side we assume a network topology defined by a strongly-connected *connectivity graph*, a subgraph of the full directed graph whose nodes are the processors existing in the configuration (those with positive speed). We assume all these channels have the same speed normalized to 1, hence transmitting data of quantity Q on such a channel will occupy it for Q time. The contribution of an existing channel to the (static) cost of the architecture is a positive constant. Clearly, the model can be extended to admit channels with different speeds and hence different costs as we did with processors. We assume a *routing function* that maps any pair (m, m') of distinct processors to a loop-free path on the connectivity graph leading from m to m'. Hence when a task p and its successor p' are mapped to m and m' respectively, the termination of p should be followed by the execution of additional communication tasks that have to be scheduled successively on the channels on the path from m to m' before p' may be executed.[3] We assume the communication time between two tasks running on the same machine to be negligible. We spare the reader from the formal definitions of the extended task-data graph, the connectivity graph and the routing function and just illustrate them in Figure 3.

We have coded this problem and ran experiments with TGFF-generated graphs to find cheap and deadline-satisfying configurations of an architecture with up to 8 processors, $\{0, 1, \ldots, 7\}$ equipped with a *Spidergon* network topology developed in ST [24] which is used in the XSTREAM archi-

[3]Here too, the model can be refined to have task-dependent routing.

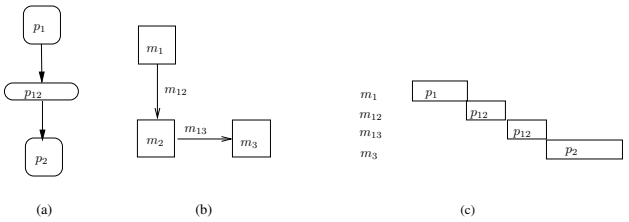

Fig. 3. (a) a part of a task-data graph with data transmission from p_1 to p_2 denoted by p_{12}; (b) a part of an architecture where communication from m_1 to m_3 is routed via channels m_{12} and m_{23}; (c) a part of a schedule, including communication, where p_1 executes on m_1 and p_2 on m_3.

tecture. A spidergon network has links of the form $(i, i + 1 \ mod \ 8)$, $(i, i - 1 \ mod \ 8)$ and $(i, i + 4 \ mod \ 8)$ for every i and there is a path of length at most 2 between any pair of processors. In this setting we could solve easily problems with 15-20 tasks. Figures 4 shows the effect of the deadline on the architecture for a task-data graph with 16 tasks. For a deadline of 25 the task graph can be scheduled on an architecture with 3 machines while for deadline 24, 4 machines are needed. Note that the schedule for the first case is very tight and is unlikely to be found by heuristics such as list scheduling.

One important parameter that partly determines the shape of optimal solutions for task-data graph scheduling problems is the computation/communication to ratio [33]. Computation and communication are antagonist in nature: while computation calls for parallelism, communication is reduced by sequential execution of two communicating tasks on the same processor. This is illustrated in the schedules of Figure 5 for two task-data graphs which are identical except for the fact that all quantities of communication in second are doubled and this increases the amount of communication to the point that parallel execution becomes infeasible and a sequential solution on a faster machine is the only possibility.

V. Periodic Scheduling

In this section we extend the problem to deal with a *stream* of instances of G that arrive periodically.

Definition 5 (Periodic Scheduling Problem): Let ω (arrival period) and δ be positive integers. The periodic deadline scheduling problem $\text{PDS}(G, R, \omega, \delta)$ refers to scheduling an infinite stream $G[0], G[1], \ldots$ of instances of G such that each $G[h]$ becomes available for execution at $h\omega$ and has to be completed within a (relative) deadline of δ, that is, not later than $h\omega + \delta$.

Let us make some simple observations concerning the relation between SDS and PDS for fixed G and R.

1) $\text{PDS}(\omega, \delta) \Rightarrow \text{SDS}(\delta)$. This is trivial because if you cannot schedule one instance of G in isolation you cannot schedule it when it may need to share resources with tasks of other instances.
2) If $\delta \leq \omega$ then $\text{PDS}(\omega, \delta) \Leftrightarrow \text{SDS}(\delta)$ because in this case each instance should terminate *before* the arrival of the next instance and hence in any interval $[h\omega, (h + 1)\omega]$ one has to solve one instance of $\text{SDS}(\delta)$. Thus we consider from now on that $\omega < \delta$.

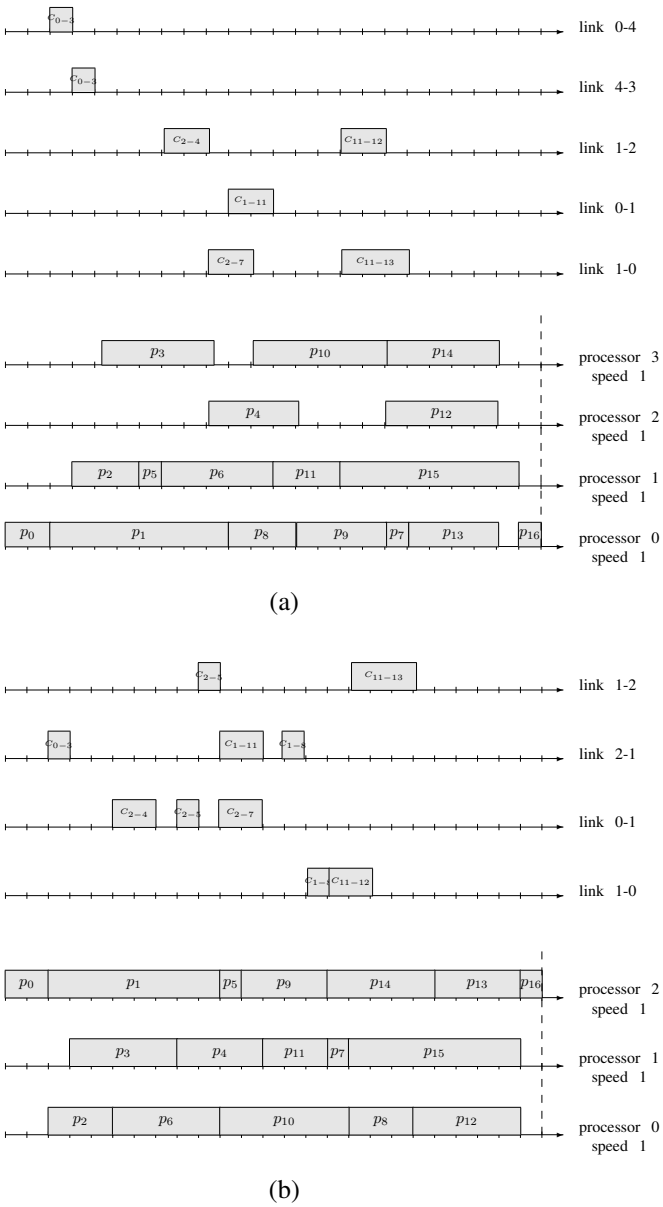

Fig. 4. Scheduling with communication: (a) $\delta = 24$; (b) $\delta = 25$.

3) Problem $\text{PDS}(\omega, \delta)$ is solvable only if $W(G) \leq \omega \mathcal{S}(R)$. The quantity of work demanded by an instance should not exceed the amount of work that the platform can supply within a period. Otherwise, backlog will be accumulated indefinitely and no finite deadline can be met.
4) When $\text{SDS}(\omega)$ is not solvable, $\text{PDS}(\omega, \delta)$ can only be solved via *pipelined* execution, that is, executing tasks belonging to successive instances simultaneously.

We first encode the problem using infinitely many copies of the task related decision variables where $x_i[h]$ and $e_i[h]$, denotes, respectively, the start time of instance h of task p_i and the machine on which it executes, which together with the speed of that machine determines its duration $d_i[h]$. The major constraints that need to be added or modified are:

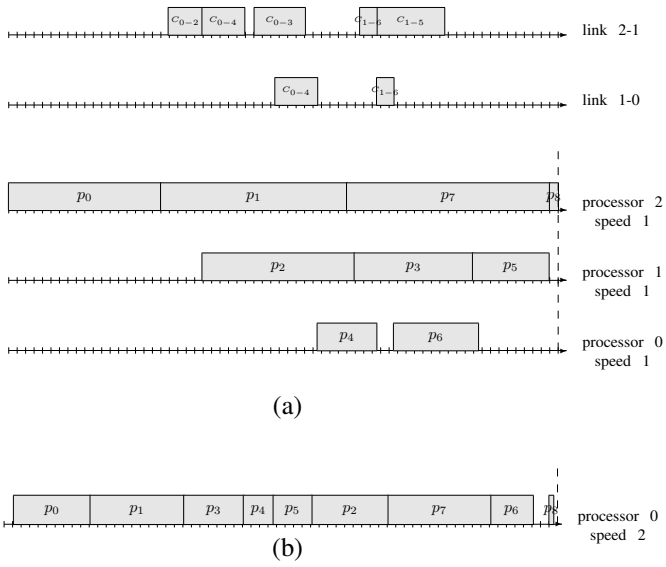

Fig. 5. The effect of changing the computation to communication ratio: (a) a schedule for a task graph G with parallelism; (b) A sequential schedule for G' constructed from G by doubling the amount of all communications.

- The whole task graph has to be executed between its arrival and its relative deadline:

$$\bigwedge_{h\in\mathbb{N}} x_1[h] \geq h\omega \ \wedge \ x_n[h] + d_n[h] \leq h + \delta$$

- Precedence:

$$\bigwedge_{h\in\mathbb{N}} \bigwedge_{i} \bigwedge_{i':p_i\prec p_{i'}} x_i[h] + d_i[h] \leq x_{i'}[h]$$

- Mutual exclusion: execution intervals of two *distinct* task instances that run on the same machine do not intersect.

$$\bigwedge_{j} \bigwedge_{i,i'} \bigwedge_{h,h'} (i \neq i' \vee h \neq h') \wedge e_i[h] = e_{i'}[h'] \Rightarrow$$
$$(x_i[h] + d_i[h] < x_{i'}[h']) \vee (x_{i'}[h'] + d_{i'}[h'] < x_i[h])$$

Note that the first two constraints treat every instance *separately* while the third is machine centered and may involve several instances due to pipelining.

While any satisfying assignment to the infinitely-many decision variables is a solution to the problem, we are of course interested in solutions that can be expressed with a finite (and small) number of variables, solutions which are *periodic* in some sense or another, that is, there is an integer constant β such that for every h, the solution for instance $h + \beta$ is the solution for instance h shifted in time by $\beta\omega$.

Definition 6 (Periodic Schedules):

- A schedule is β-machine-periodic if for every h and i, $e_i[h + \beta] = e_i[h]$;
- A schedule is β-time-periodic if for every h and i, $x_i[h+\beta] = x_i[h] + \beta\omega$;
- A schedule is (β_1, β_2)-periodic if it is β_1-machine-periodic and β_2-time-periodic.

We say that β-periodic solutions are *dominant* for a class of problems if the existence of a solution implies the existence of a β-periodic solution. It is not hard to see that there is some β for which β-periodic solutions are dominant: the deadlines are bounded, all the constants are integers (or can be normalized into integers), hence we can focus on solutions with integer start times. Combining with the fact that overtaking (executing an instance of a task before an older instance of *the same* task) can be avoided, we end up with a finite-state system where any infinite behavior admits a cycle which can be repeated indefinitely. However, the upper bound on dominant periodicity derived via this argument is too big to be useful.

It is straightforward to build counter-examples to dominance of β-machine-periodic schedules for any β by admitting a task of duration $d > \beta\omega$ and letting $\delta = d$. Each instance of this task has to be scheduled immediately upon arrival and since each of the preceding β instances will occupy a machine, the new instance will need a different machine. However, from a practical standpoint, for reasons such as code size which are not captured in the current model, it might be preferable to execute all instances of the same task on the same processor and restrict the solutions to be 1-machine-periodic. For time periodicity there are no such constraints unless one wants to use a very primitive runtime environment.

We encode the restriction to (β_1, β_2)-periodic schedules as an additional constraint.

- Schedule periodicity:

$$\bigwedge_{h\in\mathbb{N}} \bigwedge_{i} e_i[h + \beta_1] = e_i[h] \wedge x_i[h + \beta_2] = x_i[h] + \beta_2\omega$$

We denote the infinite formula combining feasibility and (β_1, β_2)-periodicity by $\Phi(\beta_1, \beta_2)$ and show that it is equi-satisfiable with a finite formula $\Phi(\beta_1, \beta_2, \gamma)$ in which h ranges over the finite set $\{0, 1, \ldots \gamma - 1\}$. In other words, we show that it is sufficient to find a feasible schedule to the finite problem involving the *first* γ instances of G.

Claim 1 (Finite Prefix):
Problems $\Phi(\beta_1, \beta_2)$ and $\Phi(\beta_1, \beta_2, \gamma)$ are equivalent when $\gamma \geq \beta + \lceil \frac{\delta}{\omega} \rceil - 1$. where $\beta = lcm(\beta_1, \beta_2)$.

Proof: Intuitively, the prefix should be sufficiently *long* to demonstrate repeatability of a segment where β instances are scheduled, and sufficiently *late* to demonstrate steady-state behavior where resource constraints are satisfied by the maximal number of instances whose tasks may co-exist simultaneously without violating the deadline.

Suppose we scheduled $\gamma = \beta + \lceil \frac{\delta}{\omega} \rceil - 1$ instances successfully. Since $\gamma \geq \beta$ we can extract a pattern that can be repeated to obtain an infinite schedule that clearly satisfies β-periodicity and precedence constraints. We now prove that this periodic schedule satisfies resource constraints. Suppose on the contrary that instance h of a task i is in conflict with instance h' of a task i', $h \leq h'$ and let $h = (k\beta + k')$, $k' \in \{0 \ldots \beta - 1\}$. The deadline constraint, together with the fact that task i and i' overlap in time, limits the set of possible values for h' to be of the form

$$h' = h + \Delta = k\beta + k' + \Delta$$

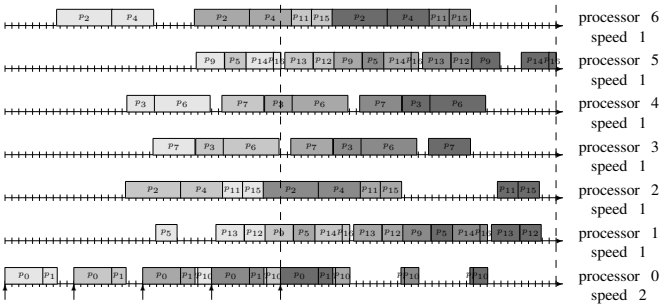

Fig. 6. A $(2,1)$-periodic schedule for a task graph with 16 tasks.

with $\Delta \in \{0 \dots \lceil \frac{\delta}{\omega} \rceil - 1\}$. Because of β-periodicity we can conclude that task i and i' also experience a conflict in the respective instances k' and $k' + \Delta$. Since $k' < \beta$ and $\Delta \leq \lceil \frac{\delta}{\omega} \rceil - 1$ we have

$$k' + \Delta < \beta + \left\lceil \frac{\delta}{\omega} \right\rceil - 1$$

which contradicts our assumption. ∎

Hence the problem can be reduced to $\Phi(\beta_1, \beta_2, \gamma)$ with γ copies of the task-related decision variables. Note however that in fact there are only β "free" copies of these variables and the values of the other variables are tightly implied by those. The unfolding is required only to add additional resource constraints, not decision variables. Table I shows performance results on several graphs with different values of δ/ω, β_1 and β_2. In general we can treat problems where the number of unfolded tasks is around 100. Figure 6 shows a $(2,1)$-periodic schedule for a task graph 1 of Table I with 17 tasks and $\delta/\omega = 4$ which requires 5 unfoldings. Although there are 85 tasks the problem is solved quickly in 3 minutes as there are only 34 decision variables. Figure 7 shows an example where making β_2 larger improves the solution. We consider a task-graph of 5 tasks, $\omega = 7$ and $\delta = 12$ with machine periodicity $\beta_1 = 2$. When we impose 1-time-periodicity we get a solution with 5 machines and cost 45, while when we move to 2-time-periodicity we can do with 3 machines and cost 43.

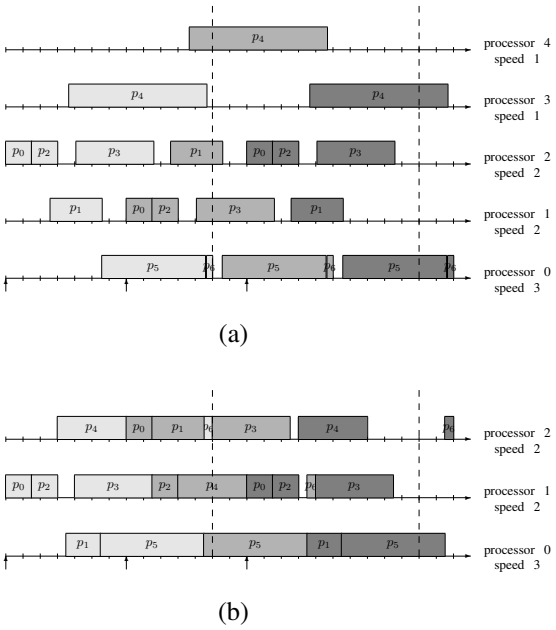

Fig. 7. The effect of time periodicity: (a) A $(2,1)$-periodic solution; (b) A cheaper $(2,2)$-periodic schedule. Tasks p_1 and p_3 are not executed 1-periodically.

VI. Bi-Criteria Optimization

We presented a method for optimizing the cost of a tunable multi-processor platform when scheduling an application under a strict deadline constraint. However rather than keeping the deadline fixed and optimizing the cost, we can adapt the search algorithm to provide a good approximation of the trade-off curve between cost and deadline, a valuable information in the process of design-space exploration. To this end the problem is viewed as a bi-objective optimization problem involving two conflicting criteria to be minimized: schedule duration and platform cost. In that case we do not seek a unique optimum but rather a set of optimal trade-offs, also known as *Pareto* solutions [1], characterized by the fact that their cost cannot be improved in one dimension without being worsened in another. In our context a solution to the scheduling problem would be Pareto optimal if one cannot decrease the platform cost without increasing the duration of the schedule. The purpose of a multi-criteria optimization algorithms is generally to provide a good approximation of the Pareto front (the set of all Pareto solutions) to help the designer in choosing among the trade-offs, a choice that may vary at different contexts, for example, depending on the battery level. For this reason, multi-objective optimization problems have been studied since the early days of modern optimization using diverse techniques, depending on the nature of the underlying optimization problems (linear, nonlinear, combinatorial) [11], [9], [10], [8]. In particular, there is a huge effort in developing efficient meta-heuristics such as evolutionary algorithms [2], [3] and local search [6] to handle complex engineering problems.

Recently we proposed an alternative methodology which approximates the Pareto front of a multi-criteria optimization problem using queries to a constraint solver [7]. This method can be viewed as multi-dimensional generalization of the binary search method described in Section III and it has the same appealing property: it provides a bound on the quality of the obtained solution. In the context of multi-criteria optimization the algorithm returns a set of Pareto solutions which is provably an ϵ-approximation of the Pareto front, that is, a set of points whose Hausdorff distance from the actual Pareto front is smaller than ε (see [14] for a discussion on quality measures for multi-criteria optimization).

We have implemented an algorithm solving the bi-objective formulation of the singular deadline scheduling problem using

	tasks	work	cp	ω	δ	$\frac{\delta}{\omega}$	β_2	β_1	platform/cost	time
0	10	78	49	8	24	3	1	2	(1,2,5)/48	1’
							2	2	(1,2,4)/47	2’
							1	3	(1,2,4)/47	2’
					32	4	1	2	(0,3,5)/29	1’
					32	4	2	2	(0,3,4)/28	4’
1	17	77	48	10	30	3	1	2	(0,2,4)/20	4’
							2	2	(0,2,4)/20	5’
					40	4	1	2	(0,1,6)/14	3’
2	21	136	65	20	40	2	1	1	(0,2,3)/19	3’
					60	3			(0,1,5)/13	5’
								2	(0,1,5)/13	6’
3	29	199	89	25	50	2	1	1	(1,2,2)/45	4’
								2	(1,2,2)/45	8’
4	35	187	63	40	80	2	1	1	(0,0,5)/5	6’
				30	60	2			(0,1,5)/13	5’
								2	?	$\perp$
5	40	210	56	35	70	2	1	1	(0,4,4)/(34,36)	11’
6	45	230	45	70	140	2	1	1	(0,0,4)/4	18’

TABLE I

RESULTS FOR THE PERIODIC DEADLINE SCHEDULING PROBLEM. THE COLUMNS STAND FOR: NUMBER OF TASKS IN G, THE QUANTITY OF WORK, THE LENGTH OF THE CRITICAL PATH, THE GRAPH INPUT PERIOD, THE DEADLINE, THE MAXIMUM PIPELINING DEGREE, THE TIME AND MACHINE PERIODICITIES, THE PLATFORM FOUND AND ITS COST AND EXECUTION TIME. PLATFORMS ARE WRITTEN AS $(R(1), R(2), \ldots)$. A PAIR (c_1, c_2) STANDS FOR A SOLUTION WITH COST c_2 WHOSE OPTIMALITY HAS NOT BEEN PROVED, BUT FOR WHICH THE STRICT LOWER BOUND c_1 WAS FOUND. THE SYMBOL $\perp$ MEANS THAT NO SOLUTION WAS EVER FOUND, I.E. A TIMEOUT OCCURRED AT THE FIRST CALL TO THE SOLVER NEAR THE UPPER BOUND.

a variant of the technique proposed in [7][4]. The program returns a set of duration/platform cost trade-offs and the associated platform-configurations and schedules. The periodic version which was not modeled in this context, is left as future work.In our experiments we were capable of approximating the Pareto front of up to 25-tasks graphs on an 8-spidergon architecture with reasonably small error (less than 5% for each objective) in a total computational time of a few minutes. Interestingly there was not a big overhead involved in moving to a bi-ojective version of the problem. This may be due to the *learning* capability of modern satisfiability solvers which allow them to keep information about the problem through successive calls. Figure 8 shows as an example the approximation obtained for a 20-tasks tgff-generated graph on the 8-spidergon architecture. As one can see, the unsatisfiability information gives a good guarantee on the quality of the approximation returned.

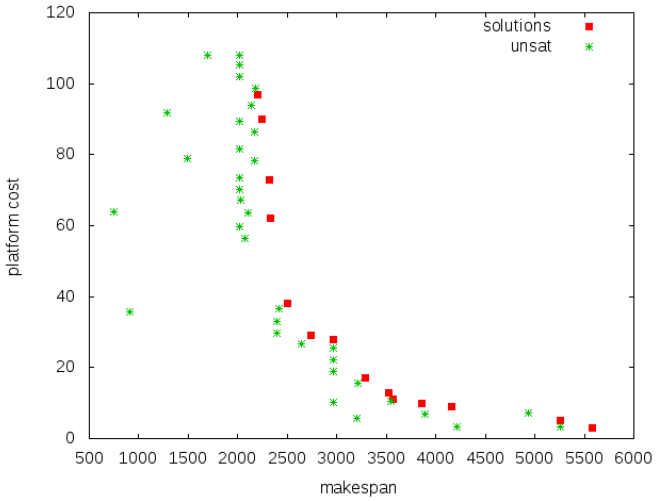

Fig. 8. A 2% approximation of the Pareto front for a 20-tasks tgff-generated task-graph.

VII. DISCUSSION

We have formulated several design questions, involving cost optimization, scheduling, communication and pipelining, inspired by real problems. We have demonstrated how modern solvers for constraint satisfaction problems can handle decent-size instances of these problems. Of course, in order to be used in practice, we will need more refined models that will reflect more closely the reality of applications and architectures. We list below some related work and then discuss limitations of the current models which are subject to ongoing and future work.

The problem of mapping and scheduling of real-time applications on multiprocessor architecture has been subject of numerous publications. We restrict ourselves to few of those that deal with closely-related problems. In [34], a list scheduling technique is developed for design space exploration, with mapping and scheduling of communicating tasks. It also takes communication resource constraints into account but it is

[4]We use the algorithm in combination with the solver Z3[5] which provides a C API and enables to save the context between different calls.

purely heuristic. An adaptation of the list scheduling heuristic has been proposed in [38] in the context of DSP processors. In [39], [20] methods based on ILP/CP decomposition are used to find accurate solutions to mapping/scheduling problems. They take more realistic constraints into account but do not explore pipelining as we do. Concerning the periodic version of the problem, let us remark that we have not found a similar problem formulation in the scheduling literature. In real-time systems [18], [36] one often deals with *independent* tasks that arrive periodically, possibly with distinct periods and deadlines, but not with structured "jobs" consisting of partially-ordered sets of tasks. On the other hand, problems associated with cyclic task graphs as in program loop parallelization [29] or manufacturing, are typically different since no *external* arrival constraints are imposed and a new instance is ready for execution once the previous instance has terminated. Let us mention also the recent work [27] which introduces a more general model where a dynamic scheduler has to cope with a stream of requests, chosen non-deterministically by a request generator from a finite sets of task graphs.

The treatment of communication in the present paper suffers from two related shortcomings. First, the model we use may not be the most faithful model for certain situations where the execution of a task is strongly *interleaved* with communication. In such situations, refining the granularity of tasks to a level where the assumption of their separation holds is not realistic due to the huge number of decision variables. We are exploring alternative formulations which do not attempt to fix a precise schedule for these communications but rather attempt to minimize and balance the load on the channels. The disadvantage of such approaches is that in the absence of a well-defined schedule, the evaluation of the mapping solution is left to stochastic simulation. Our communication model is useful for situations were predictability is important such as hard real-time systems [22], for scheduling large volume data transfers using a DMA and also as a first approximation of feasibility.

Another simplifying aspect of our model is its being *static* in several senses. First, we keep the configuration fixed and do not apply dynamic voltage scaling. The applications that we have in mind admit a more or less regular pattern of activity and the overhead in changing configuration may be too high to implement at the granularity of tasks and should be delegated to higher levels. Secondly we use a static cost model which does not distinguish between periods where a processor is executing instructions and periods where it idles. We believe that this is a good first approximation because the energy consumption of a processor that operates in a given frequency is significantly larger than that of a processor operating in lower frequency or a processor which is turned off. Refining the model in this direction will increase the number of linear constraints. Finally we assumed full information and determinism in the durations of executions and communications and in arrival times. Variability in these parameters may benefit from more adaptive scheduling strategies [15], [27] but the computational and observational overhead in implementing such strategies may turn out to be higher than their gains.

The current model does not capture other factors that affect the feasibility and quality of solutions, most notably those related to memory constraints due to code size and buffers [17]. These aspects can be incorporated into the model while remaining in the domain of linear inequalities. It also restricts itself to what is called task and pipelined parallelism, not touching the issue of data parallelism which is very important for streaming applications. Finally, constraint-based methods could also be used to automate the process of splitting and merging data which is often done manually today.

Acknowledgements: This work was done while the first author was an employee of STMicroelectronics. We thank Bruno Jego and Gilbert Richard for explaining us the XSTREAM architecture, Aldric Degorre for discussions on periodic scheduling and Scott Cotton for his advice on SAT and SMT.

REFERENCES

[1] V. Pareto. Manuel d'économie politique. *Bull. Amer. Math. Soc.*, 18:462–474, 1912.

[2] M. Mitchell. *An introduction to genetic algorithms*. The MIT press, 1998.

[3] K. Deb. *Multi-objective optimization using evolutionary algorithms*. Wiley, 2001.

[4] M.W. Moskewicz., C.F. Madigan, Y. Zhao, L. Zhang, and S. Malik. Chaff: Engineering an efficient SAT solver. *IEEE Design Automation Conference, 2001. Proceedings*, 530–535, 2001.

[5] L. De Moura and N. Bjorner. Z3: An efficient SMT solver. In *TACAS*, pages 337–340, 2008.

[6] L. Paquete and T. Stützle. Stochastic local search algorithms for multiobjective combinatorial optimization: A review. Technical Report TR/IRIDIA/2006-001, IRIDIA, 2006.

[7] J. Legriel, C. Le Guernic, S. Cotton, and O. Maler. Approximating the Pareto front of multi-criteria optimization problems. In *TACAS*, pages 69–83, 2010.

[8] M. Ehrgott. *Multicriteria optimization*. Springer Verlag, 2005.

[9] M. Ehrgott and X. Gandibleux. A survey and annotated bibliography of multiobjective combinatorial optimization. *OR Spectrum*, 22(4):425–460, 2000.

[10] J. Figueira, S. Greco, and M. Ehrgott. *Multiple criteria decision analysis: state of the art surveys*. Springer Verlag, 2005.

[11] R.E. Steuer. *Multiple criteria optimization: Theory, computation, and application*. John Wiley & Sons, 1986.

[12] E. Zitzler and L. Thiele. Multiobjective evolutionary algorithms: A comparative case study and the strength Pareto approach. *IEEE transactions on Evolutionary Computation*, 3(4):257–271, 1999.

[13] X. Fan, W.D. Weber and L.A. Barroso. Power provisioning for a warehouse-sized computer. *ACM Proceedings of the 34th annual international symposium on Computer architecture*, 13–23, 2007.

[14] E. Zitzler, L. Thiele, M. Laumanns, C.M. Fonseca, and V.G. da Fonseca. Performance assessment of multiobjective optimizers: An analysis and review. *IEEE Transactions on Evolutionary Computation*, 7(2):117–132, 2003.

[15] Y. Abdeddaim, E. Asarin and O. Maler Scheduling with timed automata , *Theoretical Computer Science* 354, 272-300, 2006.

[16] C. Barrett, R. Sebastiani, S.A. Seshia and C. Tinelli, Satisfiability modulo theories, *Handbook of satisfiability*, IOS Press, 2009

[17] S.S Battacharyya, P.K Murthy and E.A. Lee, *Software synthesis from dataflow graphs*, Kluwer, 1996.

[18] G. Bottazzo. *Hard Real-Time Computing Systems: Predictable Schedulding Algorithms and Applications*, Springer, 2005.

[19] Ph. Baptiste , C. Le Pape, W. Nuijten, *Constraint-based Scheduling: Applying Constraint Programming to Scheduling*, Springer, 2001.

[20] L. Benini, D. Bertozzi , A. Guerri and M. Milano, *Allocation and scheduling for MPSoCs via decomposition and no-good generation*, *CP'05*, 107-121, 2005.

[21] L. Benini and G. De Micheli, Networks on chips: A new SoC paradigm, *Computer* 35, 70-78, 2002.

[22] P. Caspi, A. Curic, A. Maignan, C. Sofronis, S. Tripakis and P. Niebert, From Simulink to SCADE/Lustre to TTA: a layered approach for distributed embedded applications, *LCTES'03*, 2003.

[23] E.G. Coffman, Computer and job-shop scheduling theory, Wiley, New York, 1976.

[24] M. Coppola, R. Locatelli, G. Maruccia, L. Pieralisi and A. Scandurra, Spidergon: A novel on-chip communication network, *System-on-Chip*, 2004.

[25] S. Cotton, *A study of some problems in satisfiability solving*, PhD Thesis, University of Grenoble, June 2009.

[26] M. Davis, G. Longemann, and D. Loveland. A machine program for theorem proving, *CACM* 5, 394-397, 1962.

[27] A. Degorre and O. Maler, On scheduling policies for streams of structured jobs, *FORMATS'08*, 141-154, LNCS 5215, 2008.

[28] R.P. Dick, D.L. Rhodes and W. Wolf, TGFF: Task graphs for free, *CODES/CASHE'98*, 97-101, 1998.

[29] A. Darte, Y. Robert, and F. Vivien. *Scheduling and automatic parallelization*. Birkhauser Boston, 2000.

[30] B. Dutertre, L.M. de Moura: A fast linear-arithmetic solver for DPLL(T), *CAV'06*, 81-94 LNCS 4144, 2006

[31] H. El-Rewini, T.G. Lewis and H.H. Ali *Task scheduling in parallel and distributed systems*, Prentice-Hall, 1994

[32] H. Ganzinger, G. Hagen, R. Nieuwenhuis, A. Oliveras, and C. Tinelli, DPLL(T) fast decision procedures, *CAV04*, 175-188, 2004.

[33] M.I. Gordon, W. Thies, and S. Amarasinghe, Exploiting coarse-grained task, data, and pipeline parallelism in stream programs, *ASPLOS*, 2006.

[34] J. Hu and R. Marculescu, Energy-aware communication and task scheduling for network-on-chip architectures under real-time constraints, *DATE'04*, 234- 239, 2004.

[35] J. Jaffar and M.J. Maher, Constraint logic programming: a survey, *J. of Logic Programming* 19/20, 503-581, 1994

[36] J.W.S Liu, *Real-Time Systems*, Prentice Hall, 2000.

[37] A. Schrijver *Theory of linear and integer programming*, Wiley, 1998.

[38] G. Sih and E.A. Lee, List scheduling modifications to account for interprocessor communication within interconnection-constrained heterogeneous processor networks, *Int. Conf. on Parallel Processing*, 1990.

[39] M. Ruggiero, A. Guerri, D. Bertozzi, F. Poletti and M. Milano, Communication-aware allocation and scheduling framework for stream-oriented multi-processor systems-on-chip, *DATE'06*, 2006.

[40] L. Zhang and S. Malik: The quest for efficient boolean satisfiability solvers, *CAV'02* 17-36, LNCS 2404, 2002.