Heap-size analysis for assembly programs
Jean-Yves Marion, Jean-Yves Moyen

To cite this version:
Jean-Yves Marion, Jean-Yves Moyen. Heap-size analysis for assembly programs. 2006. hal-00067838

HAL Id: hal-00067838
https://hal.science/hal-00067838
Preprint submitted on 9 May 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Heap-size analysis for assembly programs

Jean-Yves Marion
INPL-Loria – École Nationale supérieure des Mines de Nancy
Campus scientifique BP239
F-54506 Vandœuvre-lès-Nancy Cédex
Jean-Yves.Marion@loria.fr

Jean-Yves Moyen
LIPN-CNRS – Université Paris 13
99 avenue J.-B. Clément
F-93430 Villetaneuse
jean-Yves.Moyen@lipn.univ-paris13.fr

ABSTRACT
Our objective is to propose methods for resource-aware compilation inspired by the implicit complexity community. We consider a small assembly-like language and we build abstract finite machine models in order to predict a bound on the maximal heap usage of a program. We propose a polynomial time procedure to detect and certify a broad and meaningful class of non-size increasing programs, which run in a constant size heap. We end by discussing about programs running in linear heap space and discussing how to capture logarithmic space computation.

1. INTRODUCTION
1.1 Motivations
The goal of this study is an attempt to predict and control computational resources like space or time, which are used during the execution of a bytecode program. We focus here on heap space management. For this, we have chosen a bytecode language, which is close to a fragment of C. The machine architecture contains registers and a heap memory. The set of instructions consists in arithmetic operations, conditionals and jumps. The memory management is performed by two low level operations new and free dealing with dynamic memory allocations.

We present a data flow analysis of the low-level language sketched by means of “Resource Control Graph”, and we think that this is a generic concept from which several memory management policies could be checked. We establish a criterion which implies that each execution of a program needs only a constant number of dynamic memory allocations. Such programs were dubbed Non Size Increasing (NSI) by Hofmann, who has studied them in the context of functional programming with a linear type discipline [13, 14]. Here, we propose a criterion which is decidable in polynomial time. Besides the interest of this criterion to detect NSI programs, it also illustrates the possibilities and advantages of the concept of Resource Control Graph. Lastly, we describe three other

1Partly supported by the CRISS project.
2Partly supported by the CRISS and NoCost projects.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
Copyright 200X ACM X-XXXX-XX-X/XXX ...$5.00.

applications: (i) computations with a linear amount of space, (ii) LOGSPACE computations, and (iii) termination by lack of memory.

The goal is to develop an efficient static analysis in order to determine whether or not a system is heap safe. Such resource analysis allows to determine how much memory should be allocated and so prevents memory overflow. This is of course crucial for many critical applications, and has strong impact in computer security.

There are several approaches which are trying to solve this problem. The first protection mechanism is by monitoring computations. However, if the monitor is compiled with the program, it could crash unpredictably by memory leak. The second is the testing-based approach, which is complementary to static analysis. Indeed, testing provides a lower bound on the memory while static analysis gives an upper bound. The gap between both bounds is of some value in practical. Lastly, the third approach is type checking done by a bytecode verifier. In an untrusted environment (like embedded systems), the type protection policy (Java or.Net) does not allow dynamic allocation. Actually, the former approach relies on a high-level language, which captures and deals with memory allocation features [5]. Our approach guarantees, and even provides, a proof certificate of upper bound on space computation on a low-level language without disallowing dynamic memory allocations.

There are other motivations, which are more theoretical. Indeed a lot of works have been done the last ten years to provide syntactic characterisations of complexity classes, see [6, 20] among others. Those characterisations are the base bone of recent research on delineating broad classes of programs that run in some amount of time or space, like Hofmann, but also Niggl and Wunderlich [26], Amadio, Coupet-Grimal, Dal Zilio and Jakubiec [4], and Bonfante, Marion, Moyen [7].

Lastly, the method is a first step for us to prove imperative program termination by adapting ideas of Lee, Jones and Ben-Aram [17] or Abel and Altenkirch [1]. The intuition is that a program terminates when there is no more resource to consume.

1.2 Coping with undecidability
All these theoretical frameworks share the common particularity of dealing with behaviours of programs (like time and space complexity) and not only with the inputs/outputs relation which only depends on the computed function.

Indeed, a given function can be computed by several programs with different behaviours (in terms of complexity or other). Classical complexity theory deals with functions and computes extensional complexities. Here, we want to compute intensional or implicit complexity, that is try to understand why a given algorithm is more efficient than another to compute the same function.

The study of extensional complexity quickly reaches the boundary of Rice’s theorem. Any extensional property of programs is
either trivial or undecidable. Intuition and empirical results points out that intensional properties are even harder to decide. Section 2 will formalise this impression.

However, several very successful works do exist for studying both extensional properties (like termination) or intensional ones (like time or space complexity). As these works provide decidable criteria, they must be either incomplete (reject a valid program) or unsound (accept an invalid program). Of course, the choice is usually to ensure soundness: if the program is accepted by the criterion then the property (termination, polynomial bound, ...) is guaranteed. This allows the criterion to be seen as a certificate in a proof carrying code paradigm.

The property that we study here is Non Size Increasingness (NSI) as first studied by Hofmann [13]. A program is NSI if its computation does not require more memory than what is initially allocated (i.e. the total memory usage is the size of the input, plus some constant amount). This is an intensional property and this is an undecidable property.

When studying intensional properties, two different kind of approaches exist. The first one consist of restricting the syntax of programs so that any program written necessarily has the wanted property. This is in the line of the works on primitive recursive functions where the recurrence schemata is restricted to only primitive recursion. This approach gives many satisfactory results, such as the characterisations of PTIME by Cobham [10] or Bellantoni and Cook [6], the works of Leivant and Marion on tiering and predicative analysis [20] or the works of Jones on CONS-free programs [15]. On the logical side, this leads to explicit management of resources in Linear Logic [12].

All these characterisations usually have the very nice property of extensional completeness in the sense that, e.g. every polynomial time computable function can be computed by a bounded recursive function (Cobham). Unfortunately, they’re usually very poor on the intensional completeness, meaning that very few programs fit in the characterisation [11] and programmers have to rewrite their programs in a non-natural way.

So, the motto of this first family of methods can be described as leaving the proof burden to the programmer rather than to the analyser. If you can write a program with the given syntax (which, in some cases, can be a real challenge), then certain properties are guaranteed. The other family of methods will go in the other way. Let the programmer write whatever he wants but the analysis is not guaranteed to work.

Since syntax is not hampered in these methods, decidability is generally achieved by loosening the semantics during analysis. That is, one will consider more that all the executions a program can have. A trivial example of this idea would be “a program without loop uniformly terminates”. The reason we consider loops as bad is because we assume it is always possible to go through the loop infinitely many time. That is, the control of the loop is completely forgotten by this “analysis”.

A more serious example of this kind of characterisation is the Size Change Principle [17]. The set FLOW that is build during the analysis contains all “well-formed call sequences”. Every execution of the program can be mapped to a well-formed call sequence but several (most) of the call sequences do not correspond to any execution of the program. Then, properties (termination) of call sequences in FLOW are necessarily shared by all execution of the program.

However, the methods sometimes fail – which is normal since it’s a decidable method for partly solving an undecidable problem – because FLOW does contain well formed call sequences which correspond to no execution of the program but nonetheless do not have the wanted property.

This second kind of methods can thus be described as not meddling with the programmer and let the whole proof burden lay on the analysis.

Now, what is the situation on the Non Size Increasingness side? All the existing works, either by Hofmann [13] or Aehlig and Schwichtenberg [2] or Aspinall and Compagnoni [5] belongs to the first family of methods. In each case, resource management has to be done by the programmer via the diamond type, meaning that one has to explicitly reuse pointers that are no more needed (via the linear typing).

This work is intended to belong to the second family of methods. This means that programs do not have to reuse pointers but the analysis will discover if a reuse is possible. This has a nice consequence in the way that programmers can now freely allocate and de-allocate memory (with the diamonds, one cannot allocate and everything de-allocated is gone for good). So our analysis do not prevent malloc.

Of course, in order to have decidability back, we have to loosen semantics as mentioned above. In our case, similarly to the Size Change Principle, this is done by “rendering tests non deterministic”, that is considering that any branch of a if can be taken without looking at the actual result of the test.

And of course, there are several programs that are Non Size Increasing but won’t be analysed correctly. The set of falsely rejected programs may appear big at first, however, it is not bigger than in Hofmann’s framework since we can capture any algorithm that is writable in this model. We will discuss this further at the end of Section 5.2.

2. SOME DECIDABILITY ISSUES

In this section, we hint that intensional properties are more undecidable than extensional ones by proving this result for polynomial time computation and NSI.

Rice’s theorem [28] implies that any extensional property of Turing Machines (or programs) is either trivial or undecidable. An extensional property is one that depends only of the inputs and the output of the machines, i.e. that depends only of the function computed by the machine. A direct application of this theorem leads to the following result:

**Proposition 1. The set of programs p whose output size is bounded by their input size \(|p(x)| \leq |x| + c^k\) is not recursive.**

However, the Non Size Increasingness property that we want to study is stronger. Indeed, there exists programs who need tremendous amount of memory to compute but only output a small result. Typically, any program solving a decision problem will only output either “yes” or “no”, i.e. bounded size information, but may require arbitrary large amount of memory to compute this answer.

The NSI property that we study is actually an intensional property of programs. It depends not only on the inputs and outputs but also on the way the computation is performed. Intensional properties are really properties of programs and not properties of functions. Indeed, a given function can be computed by several programs (e.g. Fibonacci’s numbers can be computed with a naive exponential time algorithm or by a dynamic programming, linear time algorithm). Extensional properties depending only on the inputs/outputs of the programs are shared by all programs computing the same function. Intensional properties are not. For example, the intensional property of “computing in polynomial time” is not shared by all the program computing Fibonacci’s numbers.

Intuitively, intensional properties seem even harder to decide that
extensional ones. This can be formalised a bit by the following theorem. Here, we consider that a machine is NSI if and only if it never requires more space than a constant value plus the initial space (needed to store the input).

**Theorem 1.** Let $p$ be a program. The question "is $p$ NSI?" is undecidable even if we know that $p$ uniformly terminates.

Here, uniform termination means that the program terminates for all inputs.

**Proof.** Let $q$ be a program and consider the program $p$ that works as follows:
- $p$ answer 1 if its input is 0.
- On input $x \neq 0$, $p$ simulates $q(0)$ for $x$ steps.
  - If $q(0)$ halts within $x$ steps, then $p$ answers 0.
  - Else, $p$ answers $2^x$ (or any other large value depending on $x$).

Obviously, $p$ uniformly terminates. Is it NSI?

If $q(0)$ terminates, then it does so in $n$ steps. The space used to compute $p$ is bounded by $2^n + S_q(0)$, where $S_q(0)$ is the space needed to compute $q(0)$, i.e. a constant value and so $p$ is NSI. If $q(0)$ does not terminates, then $p$ compute the exponential and is certainly not NSI.

So, $p$ is NSI if and only if $q(0)$ terminates. Since the halting problem is not decidable, so is belonging to NSI.

Uniform termination of programs, in itself, is a non semi-recursive property. So even with an oracle powerful enough to solve (some) non semi-recursive problems, the computational property of being NSI is still undecidable! Intensional properties are, indeed, much harder than extensional ones.

Notice that this proof can be easily adapted to show the undecidability of any complexity class (of programs). It is sufficient to change the function computed by $p$ if $q(0)$ does not terminates.

**Remark 1.** A similar proof, for the undecidability of running in polynomial time, based on Hilbert’s tenth problem[23] ($p(x + 1) = 0$ if some polynomial $P$ has a root $x$, $2 \times p(x)$ otherwise) was presented as the Geocal ICC workshop in Feb. 2006 by Terui. This work has been done independently from a similar result presented by Marion in March 2000 at a seminar in ENS Lyon. This is kind of a folklore result but is nonetheless worth mentioning because lot of confusion is done on the subject.

The above result can be improved. Indeed, the set of programs that run in polynomial time, or which are NSI, is $\Sigma_2$-complete. Recall, that a typical $\Sigma_2$-complete set is the set of partial computable functions.

In order to establish the fact that NSI programs is a $\Sigma_2$-complete set, we take a class $C$ of computable functions which contains all constant functions. Assume also that there is a computable set of function codes $C$ which enumerates all functions in $C$. The set of linear functions $\{x + b \mid b \in \mathbb{N}\}$ satisfies the above hypothesis. Another example is the set of polynomials or the set of affine functions $a \times x + b$.

Next, let $\llbracket p \rrbracket$ be the function computed by the program $p$ with respect to an acceptable enumeration of programs. We refer to Rogers’ textbook [16] for background. Say that $T_p(x)$ is the number of steps to execute the program $p$ on input $x$ with respect to some universal (Turing) machine.

Now, define the set of programs whose runtime is uniformly bounded by functions in $C$:

$$A_T = \{ p \mid \exists e \in C \quad \forall x, T_p(x) < [e](|x|) \}$$

**Theorem 2.** The set $A_T$ is $\Sigma_2$-complete.

**Proof.** It is clear that the statement which defines $A_T$ is a $\Sigma_2$ statement.

Let $B$ be any $\Sigma_2$ set defined as follows:

$$B = \{ q \mid \exists y \forall x, R(q, y, x) \}$$

$R$ is a computable predicate

We prove that $B$ is reducible to $A_T$.

For this, we construct a binary predicate $Q$ as follows. $Q(q, t)$ tests during $t$ steps whether there is a $y$ with respect to some canonical ordering such that $\forall x, R(q, y, x)$ holds. If it holds, $Q(q, t)$ also holds. Here, the predicate $Q$ is computable with a complete $\Pi_1$ set as oracle.

Using the s-m-n theorem, there is a program $q'$ such that $\llbracket q' \rrbracket(t) = Q(q, t)$.

1. Suppose that $q \in B$. We know that there is an $y$ such that $\forall x, R(q, y, x)$. It follows that the witness $y$ will be found by $Q$ after $t$ steps. Since the constant function $\lambda x.t$ is in $C$, we conclude that $q'$ is in $A_T$.

2. Conversely, suppose that $q' \in A_T$. This means that $Q(q', t)$ holds for some $t$, which yields an $y$ verifying $\forall x, R(q, y, x)$

Notice that we may change time by space in the above proof, which leads to the following consequence.

**Corollary 1.** Let

$$A_S = \{ p \mid \exists e \in C \quad \forall x, S_p(x) < [e](|x|) \}$$

where $S_p(x)$ is the space use by $p$ on $x$. The set $A_S$ is $\Sigma_2$-complete.

Depending on the choice of the set of functions $C$, this proves the $\Sigma_2$-completeness of the following sets:

- Non-Size Increasing programs (for linear functions and $A_S$).
- PTIME (for polynomials functions and $A_T$).
- LOGSPACE, PSPACE, . . . any classical complexity class.

## 3. PROGRAMS

### 3.1 Syntax

**Definition 1.** A program is defined by the following grammar:

- **Programs**
  - $p ::= \text{lbl1 : i; ... lbln : in;}$
  - $i ::= r_1, \ldots, r_l := \text{op(r'_{i_1}, \ldots, r'_{i_k})}$
  - $r ::= \text{r'} \mid \text{r := r'}$
  - $r ::= \text{newn \mid freer}$
  - $\text{jmp~lbl} \mid \text{jr~lbl0 lbl1}$
  - $\text{end}$

- **Labels**
  - $L \ni \text{lbl}$ finite set of labels

- **Registers**
  - $R \ni \text{r}$ finite set of registers

- **Operators**
  - $O \ni \text{op}$ finite set of operators

Each operator has a fixed arity $k$ and co-arity $l$ and $n$ is an integer constant. The syntax of a program induces a function $\text{next} : L \rightarrow L$ such that $\text{next(lbl1)} = \text{lbl}_{i+1}$ and a mapping $\text{instr} : L \rightarrow I$ such that $\text{instr(lblk)} = i_k$. The above result can be improved. Indeed, the set of programs that run in polynomial time, or which are NSI, is $\Sigma_2$-complete. Recall, that a typical $\Sigma_2$-complete set is the set of partial computable functions.

In order to establish the fact that NSI programs is a $\Sigma_2$-complete set, we take a class $C$ of computable functions which contains all constant functions. Assume also that there is a computable set of function codes $C$ which enumerates all functions in $C$. The set of linear functions $\{x + b \mid b \in \mathbb{N}\}$ satisfies the above hypothesis. Another example is the set of polynomials or the set of affine functions $a \times x + b$.

Next, let $\llbracket p \rrbracket$ be the function computed by the program $p$ with respect to an acceptable enumeration of programs. We refer to Rogers’ textbook [16] for background. Say that $T_p(x)$ is the number of steps to execute the program $p$ on input $x$ with respect to some universal (Turing) machine.
Machines that we consider are close to the RAM-Machines. There is a memory device, similar to heaps on computers, and a finite number of registers. The heap consists in an unbounded number of memory cells. Each cell stores an integer. Each register also stores an integer whose value may be the address of a heap-cell.

Roughly speaking, \( *r \) denotes the value of the memory cell whose address is stored in \( r \). The instruction \( jmp \ lbl \) gives the control to label \( lbl \), \( jz \ r \ lbl \ l1' \) gives the control to either \( l1' \) (if \( r \) contains 0) or \( l1' \). The instruction \( new \ n \) allocates \( n \) consecutive cells of memory in the heap and returns the address of the first one and \( free \) frees the memory cell in the heap at the given address.

Memory allocation is done by block because we may require two given to cells have consecutive addresses (e.g. to represent lists) but freeing memory is done cell by cell (unlike, e.g. in \( \mathbb{C} \)) to avoid remembering the size of each allocated block.

**Example 1.** The following program reverses a list. A list is here represented by a set of positions, each position consists in two consecutive memory cells. The first cell contains the value of the element in the list and the second contains the address of the next position (or 0 for the end of the list). For the sake of clarity, only labels which play a special role (e.g. destination of jumps) have been mentioned in this example.

```
begin : next := 0; jz r0 end loop;
loop : val := *r0; tmp := r0 + 1; free r0;
    r0 := *tmp; free tmp;
    r1 := new 2; r1 := val; r1 + +; *r1 := next; r1 := - -;
next := r1; jz r0 end loop;
end : end;
```

Of course, the use of low-level data structure may imply tedious pointer management. However, we could introduce new operators acting both on registers and on the heap that can be seen as macros.

**Definition 2.** A program is a sequence of configurations that can be formally defined as:

- \( l := \text{cons}(a, l') \) is equivalent to \( l := \text{new} 2; *l := a; l + +; l := l'; l := - -; \)
- \( a, l' := \text{destr}(l) \) is equivalent to \( a := *l; \text{tmp} := l + 1; \text{free} l'; l := \text{tmp}; \text{free} \text{tmp} \)

Using these, the program thus becomes:

```
begin : next := 0; l1 := jz r0 end loop;
loop : val, r0 := destr(r0);
l1 := cons(val, next);
l1 := r1; l1 := jz r0 end loop;
end : end;
```

We illustrate an execution of the program in Figure 1.

Notice that if one wants to write the reverse program within Hofmann’s or Aspinall and Compagnoni’s formalism, one cannot use \( \text{new} \) and \( \text{free} \) and has to explicitly keep the values of pointer (i.e. the diamonds) and tells when they have to be reused.

### 3.2 Semantics

The domain of the computation is \( V = \mathbb{N} \), and we interpret each operator \( \circ \) by a function \( [\circ] \) of the same arity an co-arity over \( V \).

An environment is a mapping \( \sigma : \mathcal{R} \rightarrow V \), which associates to each register a value. We note \( \{r_i \rightarrow v\} \) the environment that maps \( r_i \) to \( v \) and \( r_j \) to \( r_j' \) (\( i \neq j \)). When no ambiguity can arise, we just write \( r \) instead of \( \{r\} \).

A heap is a mapping \( \mu : \mathbb{N} \rightarrow V \cup \{\bot\} \), which returns a value if a heap address is allocated. Otherwise, it returns \( \bot \) when a heap address is unallocated. We note \( \{n \rightarrow v\} \mu \) to say that the heap is updated in such a way that \( \{n \rightarrow v\} \mu(n) = v \) and \( \{n \rightarrow v\} \mu(m) = \{m\} (m \neq n) \).

Memory allocation is performed by some external mechanism similar to a call system. In our framework, we have a predicate \( \text{memfree}(k, n) \) which tests whether the memory cells between \( k \) and \( k + n - 1 \) are currently unallocated:

\[
\text{memfree}(k, n) = \text{and} (\mu(k) = \bot) \land \ldots \land (\mu(k + n - 1) = \bot)
\]

During computation, three kind of errors can happen. Namely, (i) accessing an unallocated memory cell (MemError), (ii) lacking free memory to perform a new instruction (MemFull) or (iii) freeing an already unallocated memory cell (FreeError).

The MemError rule means that there is a mechanism (either internal or external via system calls performed when memory is read) able to detect dangling pointers. This, of course, is a rather costly operation at runtime and quite hard to perform on static code. On the other hand, memory leaks are not detected.

Both these issues can be discarded if we consider that our programs are obtained by compilation of a higher level language with efficient memory management and garbage collection. In the following, we will assume that programs are safe in the sense that they never cause one of the error rules to be triggered.

**Definition 3.** Programs are executed using a straightforward small-step semantics whose rules are described in Figure 2. The relation \( p \vdash \theta_1 \rightarrow \theta_2 \) means that the new configuration is \( \theta_2 \) after executing the instruction \( \theta = \text{instr}(\text{IP}) \) where \( \text{IP} \) is the instruction pointer of \( \theta_1 \).

An execution of a program \( p \) is a sequence of configurations computed step by step. We write \( p \vdash \theta_0 \rightarrow \theta_1 \rightarrow \ldots \rightarrow \theta_n \) if \( p \vdash \theta_0 \rightarrow \theta_1 \rightarrow \ldots \rightarrow \theta_n \) for each \( i = 0, n - 1 \).

Programs compute functions from heaps to heaps as soon as certain conditions on the initial environment (with respect to the initial heap) are respected. The program of Example 1 reverses a list assuming that the heap is really the encoding of a list and that \( r_0 \) initially contains the address of the first cell of the list.

### 3.3 Measuring heap Usage

In order to control space resources, we consider the number of new heap cells allocated during any execution of a program. We do not consider the number of cells of the input. So, only the “workspace” is counted like in read-only model of computation.

If we want to control the real space usage, we have to take values into account. Indeed, working with unbounded values means that one needs logarithmic space to actually store them. To constantly
keep this into mind, we speak of heap complexity rather than the usual space complexity.

Of course, in any real computer, the set of values is bounded (e.g., it is the set long int of 32 bits integers). In this case, the real space usage is directly proportional to the heap usage, heap and space complexity are the same up to a multiplicative constant.

Definition 4. Let $\theta = (T, P, \mu, \sigma)$ be a configuration. The heap size $|\theta|$ of $\theta$ is $\text{card}\{n, \mu(n) \neq \bot\}$, that is the number of non-$\bot$ elements in $\mu$.

The heap measure is a function such that for any program $p$, and any configuration $\theta_0$,

$$\text{heap}_p(\theta_0) = \begin{cases} \max_{\theta_1, \ldots, \theta_n} |\theta_1| - |\theta_0| & \text{if } p \vdash \theta_0 \xrightarrow{t_1} \theta_1 \ldots \xrightarrow{t_n} \theta_n \\
\bot & \text{and } \theta_n \text{ final} \\
\perp & \text{otherwise} \end{cases}$$

In other words, the heap usage $\text{heap}_p(\theta_0)$ is the maximal heap size of an encountered configuration when the computation terminates.

Given a total function $f : N \rightarrow N$, we define the set of heap-bounded programs:

$$\text{Heap}(f) = \{ p \mid \text{heap}_p(\theta) \leq f(|\theta|) \forall \theta, \text{heap}_p(\theta) \neq \bot \}$$

Definition 5. $\text{LinHeap}$ is the class of programs running with a linear heap. That is $\text{LinHeap} = \bigcup \text{Heap}(f)$ where the union ranges over all $f : x \mapsto \beta \times x + \alpha$.

$\text{NSI}$ is the class of programs running with at most a constant increase in heap size, that is $\text{NSI} = \bigcup \text{Heap}(f)$ where the union ranges over all $f : x \mapsto \alpha$.

4. RESOURCE PETRI NET

4.1 Petri Nets in a nutshell

We briefly describes Petri nets and reader may consult the survey of Murata [25] for more details. A Petri net $N$ is a quadruplet $(S, T, A, \omega)$ where $S$ is a set of places, $T$ is a set of transitions and $A$ is a set of arcs such that $(S, T, A)$ is a bipartite oriented graph. That is, $S$ and $T$ are disjoint and each arrow is either from $S$ to $T$ or from $T$ to $S$. An arrow between $s$ and $t$ ($s,t$) has a weight $\omega(s,t)$ ($\omega(t,s)$). As usual, places are graphically represented by circles while transitions are represented by squares.

A marking $M$ assigns to each place $s$ a natural number $M(s)$ which indicates the number of tokens in it. The pre-set of a transition $t$ is $\cdot t = \{ s \in S \mid (s, t) \in A \}$. The post-set of a transition $t$ is $t^* = \{ s \in S \mid (t, s) \in A \}$.

A transition $t$ is enabled at marking $M$ if and only if for any place $s$ in $t^*, M(s) - \omega(s,t) \geq 0$. It can then be fired thus reaching marking $M'$:

$$M'(s) = M(s) - \omega(s,t) + \omega(t,s)$$

We write $N \vdash M_{n} \xrightarrow{t_{1}} \ldots \xrightarrow{t_{m}} M_{n}$ is a run of the Petri net.

4.2 Control Flow Petri Net

Definition 6. Let $p$ be a program, its control flow Petri net (CFPN) $\text{cfpn}(p)$ is defined as follows:

- The set of places is exactly the set $L$ of labels of $p$.
- For each instruction that is neither $\text{zj}$ nor end, there is a transition $\text{instr}(\text{lbl})$ bearing the same name and for each $\text{zj}$ or $\text{lbl}$ instruction, there are two transitions $r = 0$ and $r \neq 0$.
- For each label $\text{lbl}$ such that $\text{instr}(\text{lbl}) \notin \{ \text{zj}, \text{jmp}, \text{end} \}$, there is one arrow from $\text{lbl}$ to $\text{instr}(\text{lbl})$ and one from $\text{instr}(\text{lbl})$ to $\text{next}(\text{lbl})$.

Figure 1: An example run of the reverse program.
\[
\text{instr}(\text{IP}) = r_1, \ldots, r_k := \text{op}(r'_1, \ldots, r'_l) \quad v_j = \sigma(r'_j) \quad v_1, \ldots, v_l = [\text{op}](v'_1, \ldots, v'_k)
\]

\[
p \vdash (\text{IP}, \mu, \sigma) \xrightarrow{\text{instr}(\text{IP})}\ (\text{next}(\text{IP}), \mu, \{r_i \leftarrow v_i\}_{i \in 1 \ldots k})
\]

\[
\text{instr}(\text{IP}) = r := *r' \quad \text{or} \quad \text{instr}(\text{IP}) = *r := r\quad \mu(r') = \bot
\]

\[
p \vdash (\text{IP}, \mu, \sigma) \xrightarrow{\text{instr}(\text{IP})}\ (\text{next}(\text{IP}), \mu, \{r \leftarrow v\}_{\sigma})
\]

\[
\text{instr}(\text{IP}) = r := *r' \quad \mu(r') = v
\]

\[
p \vdash (\text{IP}, \mu, \sigma) \xrightarrow{\text{instr}(\text{IP})}\ (\text{next}(\text{IP}), \mu, \{r \leftarrow v\}_{\sigma})
\]

\[
\text{instr}(\text{IP}) = *r := r' \quad \mu(r) \in V
\]

\[
p \vdash (\text{IP}, \mu, \sigma) \xrightarrow{\text{instr}(\text{IP})}\ (\text{next}(\text{IP}), \{\sigma(r) \leftarrow \sigma(r')\}_{\mu}, \sigma)
\]

\[
\text{instr}(\text{IP}) = r := \text{new } n \quad \forall k \in \mathbb{N}, \neg \text{memfree}(k, n)
\]

\[
p \vdash (\text{IP}, \mu, \sigma) \xrightarrow{\text{instr}(\text{IP})}\ (\text{next}(\text{IP}), \{k \leftarrow 0, \ldots, k + n - 1 \leftarrow 0\}_{\mu}, \{r \leftarrow k\}_{\sigma})
\]

\[
\text{instr}(\text{IP}) = \text{free } r \quad \mu(r) = \bot
\]

\[
p \vdash (\text{IP}, \mu, \sigma) \xrightarrow{\text{instr}(\text{IP})}\ (\text{next}(\text{IP}), \{\sigma(r) \leftarrow \bot\}_{\mu}, \{r \leftarrow 0\}_{\sigma})
\]

\[
\text{instr}(\text{IP}) = \text{free } r \quad \mu(r) \neq \bot
\]

\[
p \vdash (\text{IP}, \mu, \sigma) \xrightarrow{\text{instr}(\text{IP})}\ (\text{next}(\text{IP}), \{\sigma(r) \leftarrow \bot\}_{\mu}, \{r \leftarrow 0\}_{\sigma})
\]

\[
\text{instr}(\text{IP}) = \text{jmp } \text{lbllbl}
\]

\[
p \vdash (\text{IP}, \mu, \sigma) \xrightarrow{\text{instr}(\text{IP})}\ (\text{lbllbl}, \mu, \sigma)
\]

\[
\text{instr}(\text{IP}) = \text{jz } r \text{ } \text{lbllbl' } \sigma(r) = 0
\]

\[
p \vdash (\text{IP}, \mu, \sigma) \xrightarrow{\text{instr}(\text{IP})}\ (\text{lbllbl}, \mu, \sigma)
\]

\[
\text{instr}(\text{IP}) = \text{jz } r \text{ } \text{lbllbl' } \sigma(r) \neq 0
\]

\[
p \vdash (\text{IP}, \mu, \sigma) \xrightarrow{\text{instr}(\text{IP})}\ (\text{lbllbl'}, \mu, \sigma)
\]

\[\text{Figure 2: Small-steps semantics}\]
The number of tokens in the place $\text{Free}$ represents a heap cell. At the beginning of a program execution, the place $\text{Free}$ contains the number of free cells. When a program requests new cells, then we need to have as many tokens in the corresponding marking $M^0$ of $\text{Free}$ as the current size of the heap (the number of non-$\bot$ cells). The place $\text{Free}$ should contain as many tokens as there is free memory (this depends on the machine used, not only on the program).

Definition 7. Let $p$ be a program. The resource Petri net $\text{rpn}(p)$ of $p$ is obtained by adding to $\text{cfpn}(p)$ two places $\text{Free}$ and $\text{Heap}$ and four kind of arrows

- An arrow from the $\text{Free}$ place to each new $n$ transition of weight $n$.
- An arrow from each new $n$ transition to $\text{Heap}$ place of weight $n$.
- An arrow from each free transition to the $\text{Free}$ place of weight 1.
- An arrow from the $\text{Heap}$ place to each free transition of weight 1.

Inversely, when a free $r$ instruction is executed, then one token is moved from the $\text{Heap}$ place to the $\text{Free}$ one. Thus we recycle discarded heap cells. So, a RPN is a simple model which allows to simulate executions of programs and to check memory usage at the same time. The $\text{Heap}$ place should contain as many tokens as the current size of the heap (the number of non-$\bot$ cells). The $\text{Free}$ place should contain as many tokens as there is free memory (this depends on the machine used, not only on the program).

Proposition 2. Let $p$ be a program. If $p \models \theta_0 \xrightarrow{t_1} \ldots \xrightarrow{t_n} \theta_n$ is an execution of $p$, then $\text{cfpn}(p) \models M^{\theta_0} \xrightarrow{t_1} \ldots \xrightarrow{t_n} M^{\theta_n}$ is a run of $\text{cfpn}(p)$.

4.3 Resource Petri Net

We now turn to represent the heap memory mechanism by a Resource Petri Net (RPN). In order to define a RPN, we add to CFPN construction described above two new places $\text{Heap}$ and $\text{Free}$. As previously, each program execution is mapped on a RPN run which, unlike the previous CFPN model, is resource-aware. For this, a token represents a heap cell. At the beginning of a program execution, the place $\text{Free}$ contains the number of free cells. When a program requests $n$ new cells, then we need to have $n$ tokens in the place $\text{Free}$. In this case, we put $n$ tokens in the place $\text{Heap}$. Otherwise, there is not enough free memory, and the run of the RPN stalls. The number of tokens in the place $\text{Heap}$ is the number of cells which have been allocated during a program execution.
of this edge is the absolute value of $\omega(\text{instr})$ and the direction of it is given by the sign of the weight. Figure 4 illustrates this by showing the RPN of the reverse program.

Tokens in the $\text{Heap}$ and $\text{Free}$ places correspond to Hofmann’s diamonds ($\diamond$), that is a fixed amount of space in the heap. Each token represents a diamond, either used by (one of the variable of) the program if in the $\text{Heap}$ place or available if in the $\text{Free}$ place. Memory allocation (such as $\text{cons}$), needs diamonds in order to be performed while memory de-allocation (such as $\text{destr}$) releases diamonds to be used later.

**Lemma 1.** In each run of a RPN, the total number of tokens in the $\text{Free}$ and $\text{Heap}$ places is constant.

To each configuration $\rho = (\text{IP}, \mu, \sigma)$ of $p$, we associate a marking $M^\rho$ of $\text{rpm}(p)$:

\[ M^\rho(\text{Heap}) = |\rho| \quad M^\rho(\text{Free}) = i \]
\[ M^\rho(\text{IP}) = 1 \quad M^\rho(1b1) = 0 \quad \text{otherwise} \]

**Lemma 2.** Let $p$ be a program, $\rho$ be a configuration such that $p \vdash \rho \Rightarrow \rho'$ and $i$ be an integer such that $i + \omega(t) \geq 0$. Then, $\text{rpm}(p) \vdash M^\rho \rightarrow M^{\rho'} + M^{\rho + \omega(t)}$.

**Proof.** Consider that $t = \text{new} n$. We can remove $n$ tokens from $\text{Free}$ because $i + \omega(t) = i - n \geq 0$. So, we can fire the $\text{new} n$ transitions.

Next, consider $t = \text{free} r$. The place $\text{Heap}$ contains $|\rho|$ tokens and by hypothesis $p \vdash \rho \Rightarrow \rho'$, so we know that $|\rho| > 0^2$. Therefore, we can remove one token from the place $\text{Heap}$ and fire the transition. \qed

**Proposition 3.** Let $p$ be a program and let $N$ be its RPN. Assume that $p \vdash \theta_0 \xrightarrow{t_1} \ldots \xrightarrow{t_n} \theta_n$. Let $i_0$ be a natural number such that $\forall k \leq n$, $i_0 + |\theta_k| - |\theta_0| \geq 0$. Then, there is a run $N \vdash M^{\theta_0}_0 \xrightarrow{t_1} \ldots \xrightarrow{t_n} M^{\theta_n}_{i_n}$ which satisfies $i_k = i_0 + |\theta_k| - |\theta_0|$ for each $k = 0, \ldots, n$.

### 5. Characterization of Heap Complexity

The condition $|\rho| - \omega(t) \geq 0$ is not necessarily here.

#### 5.1 Heap complexity and resource Petri nets

**Theorem 3.** Let $f : \mathbb{N} \rightarrow \mathbb{N}$ be a total function. Let $p$ be a program.

If $p \in \text{Heap}(f(x))$ if and only if for each initial configuration $\theta_0$ and for each execution $p \vdash \theta_0 \xrightarrow{t_1} \ldots \xrightarrow{t_n} \theta_n$, we have $\text{rpm}(p) \vdash M^\rho_{f(\theta_0)} \xrightarrow{t_1} \ldots \xrightarrow{t_n} M^\rho_{f(\theta_n)}$ for some $i_1, \ldots, i_n$.

The fact that the run exists means that $f(|\theta_0|)$ is a sufficiently large number of tokens to put in $\text{Free}$ to avoid any deadlock.

**Proof.** If for each initial configuration $\theta_0$ and each execution $p \vdash \theta_0 \xrightarrow{t_1} \ldots \xrightarrow{t_n} \theta_n$, there is a corresponding run

\[ \text{rpm}(p) \vdash M^\rho_{f(\theta_0)} \xrightarrow{t_1} \ldots \xrightarrow{t_n} M^\rho_{f(\theta_n)} \]

then for any $k$,

\[ |\theta_k| = M^\rho_{f(\theta_0)}(\text{Heap}) \leq M^\rho_{f(\theta_0)}(\text{Heap}) + M^\rho_{f(\theta_0)}(\text{Free}) = |\theta_0| + f(|\theta_0|) \]

Conversely, deadlocks in the run may come from three sources: the label places, the $\text{Heap}$ place or the $\text{Free}$ place. The label places cannot cause deadlock due to Proposition 2. The $\text{Heap}$ place may only cause deadlock if a $\text{free}$ transition should be fired but the $\text{Heap}$ place is empty. However, this would correspond to executing a $\text{free}$ instruction on an empty heap, thus triggering the (FreeError) rule.

The $\text{Free}$ place can cause deadlock if a $\text{new} n$ transition has to be fired but $\text{Free}$ contains less than $n$ tokens. However, Prop. 3 claims that if we start from marking $M^\rho_{\theta_0}$ such that for all $1 \leq k \leq n$, $i_0 + |\theta_0| - |\theta_k| \geq 0$ there will be no deadlock. Since $p \in \text{Heap}(f)$, $|\theta_k| \leq f(|\theta_0|) + |\theta_0|$ and thus $i_0 + |\theta_0| - |\theta_k| \geq i_0 + f(|\theta_0|)$. So by starting from $M^\rho_{f(\theta_0)}$, there will be no deadlock. \qed

Notice that the (implicit) quantifier on the Petri net runs is existential and not universal. To each execution of the program corresponds a run, but some runs correspond to no execution of the program (because the Petri net does not mimic the semantics tightly, especially tests). This means that unless one can decide which run corresponds to an execution, analysis of the Petri net will not capture every programs (because in some cases, the “invalid” runs do not correspond to an execution). This is a necessarily lose if we want our criterion to be decidable, as mentioned in the discussion at the end of Section 1.2.

**Corollary 2.** $p \in \text{NSI}$ if and only if there exists a constant $\alpha$ such that for each initial configuration $\theta_0$ and each execution $p \vdash \theta_0 \xrightarrow{t_1} \ldots \xrightarrow{t_n} \theta_n$, we have $\text{cfn}(p) \vdash M^\rho_{\theta_0} \xrightarrow{t_1} \ldots \xrightarrow{t_n} M^\rho_{\theta_n}$ where $i_k = |\theta_0| + \alpha - |\theta_k|$. $\alpha \neq 0$

$p \in \text{LINHEAP}$ if and only if there exists $\alpha$ and $\beta$ such that for each initial configuration $\theta_0$ and each execution $p \vdash \theta_0 \xrightarrow{t_1} \ldots \xrightarrow{t_n} \theta_n$, we have $\text{cfn}(p) \vdash M^\rho_{\theta_0} \xrightarrow{t_1} \ldots \xrightarrow{t_n} M^\rho_{\theta_n}$ where $i_k = (\beta + 1)|\theta_0| + \alpha - |\theta_k|$.

#### 5.2 Detection of non-size increasing programs

**Theorem 4.** There is a polynomial time procedure to detect if a program is $\text{NSI}$.

**Proof.** Let $p$ be a program. We construct a directed graph $G$ from $\text{rpm}(p)$ that we name the resource control graph of $p$ as follows. Nodes are label place. There is an edge from $1b1$ to $1b1'$. 

---

Figure 5: Control Flow Graph of the reverse program. Edges are labelled after the instruction they represent.
Figure 6: Resource Control Graph of the reverse program. Edges are labelled after the instruction they represent.

if there is a transition from \( l_1 \) to \( l_1' \) in \( rpm(p) \). The weight assigned to an edge is the weight of the corresponding transition, that is the weight \( \omega(\text{instr}(l_1)) \). In other words, the resource control graph is the control flow graph where each arrow as the same weight as the corresponding instruction.

If the resource control graph contains no cycle of strictly negative weight, then the program is in NSI. This criterion can be decided in polynomial time via Bellman-Ford's algorithm.

Indeed, any path in the resource control graph correspond to a run in the RPN. The weight of the path being exactly the number of token removed from (if negative) the \( \text{Free} \) place at the end of the run. If there is no cycle of negative weight, then the minimum weight of all paths, \( \alpha \), is well defined. So, no run can remove more than \( \alpha \) token from \( \text{Free} \). Hence, each execution corresponds to a run that can be performed with \( \max(-\alpha, 0) \) token initially in \( \text{Free} \) and \( p \) is NSI.

The above demonstration provides also the number of cells which are necessary to perform any computation. Therefore, we may be able to set up a proof procedure, like in proof carrying code paradigm, in polynomial time to calculate and certify that a program uses a fixed amount of heap.

Hofmann has shown that Non Size Increasing programs can be compiled into \( C \) without using \texttt{malloc} [13]. That is, in our case, we can write an equivalent program without the \texttt{new} instruction, or, more precisely, starting with a single \texttt{new} \( \alpha \) instruction during the initialization process. Amadio already has a \texttt{PTIME} bound to detect NSI programs in terms of quasi-interpretation [3]. And linear space usage can also be characterized by quasi-interpretations [8].

Lastly, notice that a cycle of positive weight may cause damages by releasing potentially non-allocated heap cells. We do not control this kind of troubles here.

Example 2. The resource graph of the reverse program is given in Figure 6. Since there is no cycle of negative weight, so the program is NSI. Moreover, the weight of any path starting from begin is either 0 or +2, so choosing \( \alpha = 0 \) is enough, meaning that the program doesn’t use more memory that initially allocated.

Of course, the above method does not capture all non size increasing programs. For example, if in a loop like (written in a \( C \)-like syntax for readability):

\[
\text{for}(i = 1, i \leq r, i++) \{ r' := \text{new} 1; \}
\]

the method fails. However, we can roughly distinguish three cases for this kind of loops.

- \( r \) is an immediate value, that is the loop actually is, say,

\[
\text{for}(i = 1, i \leq 10, i++) \{ r' := \text{new} 1; \}
\]

In this case, the allocation can be performed before the loop:

\[
r := \text{new} 10; \text{for}(i = 1, i \leq 10, i++) \{ r' := r + i; \}
\]

This allows to have the same effect but put the allocation out of the loop.

- \( r \) is a value stored in a register which happen to be a constant. In this case, constant propagation allows to detect at compile time that \( r \) can be replaced by an immediate value.

- \( r \) is the value of a register which depends on the inputs. In this case, the program is probably not NSI because it perform a number of allocations dependant on the inputs.

Both the first two transformations can be performed automatically at compile time, just before the analysis is performed. Moreover, construction of the CFG and RCG can help find potentially harmful loops (those with a positive weight) and thus help the transformation of these loops. Of course, programs where constant propagation fails or programs with loops of the third kind will not be detected as NSI. Hopefully, there are few of them among the commonly used algorithms...

It is also important to notice that any algorithm that fits into Hofmann’s framework can be detected as NSI by our method. Indeed, Hofmann provides a way to compile NSI programs into \texttt{malloc}-free programs and our method obviously detect \texttt{malloc}-free programs as NSI (since they do not allocate memory, the RCG has no edge of negative weight and cannot have cycle of negative weight).

6. OTHER CHARACTERIZATIONS

6.1 Linear Heap Space

The detection of linear space computation is closely related to the one of non size increasing computation, and the previous method can be modified to handle linear heap run. The idea is that each input cell gives new cells when it is deallocated, instead of one in the case of a non-size increasing memory allocation. On the other hand, cells which are allocated during an execution only count for one. To perform this analysis, we consider two kinds of heaps: an \texttt{input heap} that cannot be further re-allocated ; and a \texttt{work heap}, initially empty, that can be allocated. Then, programs have two \texttt{free} instructions: \texttt{free} to free input cells and \texttt{free}_w to free work cells\(^3\). The \texttt{new} only allocates memory on the work heap \( \mu_w \).

Now, we extend RPN in order to take into account both new free instructions by assigning the following weights to them:

\[
\omega(\text{free}, r) = \beta \quad \omega(\text{free}_w r) = 1
\]

Theorem 5. There is a polynomial time procedure to detect if a program is linear heap computable.

Proof. The proof is similar to the demonstration of Theorem 4. We need to find \( \beta \) such that the resource control graph has no cycles of strictly negative weight. Notice that this will also give us the values \( \beta \) and \( \alpha \) and a precise bound on the space usage. \( \square \)

\(^3\)We cope with the difficulty to detect which free instruction releases memory from the input or work heap. A way is to restrict the use of input register like in [15].
But this value might well be unrelated to the size of the heap (that is the number of allocated cells) just because the allocated cells might be non-contiguous.

Indeed, pointers (that is registers containing addresses in the heap) has access to these maps and in addition to allocating some real compute NSI-programs in logarithmic space as soon as the values might have a value as big as the address of any allocated heap-cell.

6.3 Lack of Resources

Since it is not possible to free more memory that what is allocated, no execution can free an unbounded number of cells. So, any potentially infinite execution whose corresponding run in the RPN (resp. path in the RCG) removes infinitely many tokens from Heap (resp. has weight $+\infty$) will stop once the heap is empty and does not correspond to a real infinite execution.

Similarly, on any real computer, the available memory is finite and bounded. So, any potentially infinite execution whose corresponding run in the RPN (resp. path in the RCG) removes infinitely many tokens from Free (resp. has weight $-\infty$) will stop once all the memory has been allocated and does not correspond to any real infinite execution. This leads to a simple criterion to detect termination by lack of resources.

**Proposition 4.** Let $p$ be a program and $G$ be its RCG. If all the infinite paths of $G$ have infinite weight ($\leq \infty$), then all executions of the program terminate by lack of resources.

This approach is somewhat similar to the Size Change Principle (SCP) [17] in the sense that we check if a given finite resource (in our case memory) is exhausted during computation. However, SCP is much more efficient that our simple criterion since it takes into account the values of the variables and not only the global amount of memory. Typically, this lack of resources criterion is unable to detect termination of the reverse program while SCP detects it by checking the length of the list.

7. CONCLUSION

The main result that we have presented is a (incomplete) criterion to determine if a low-level program runs within a fixed amount of heap memory, and to calculate an upper bound on this heap size. Moreover, this criterion is polynomial time computable. One of the main advantages of our approach compared with the other ones that we have cited in the text, is that we are dealing with a low-level programming language without type annotations. Aspinall and Compagnoni [5] derived a similar characterisation for assembly language but still by using the type discipline originally introduced by Hofmann [14]. We believe that the representation in term of Resources Petri Nets shed some light on Hofmann’s diamond-concept. In particular, we do not need any more the linearity conditions introduced in these works. Moreover, we do not forbid dynamic heap allocation.

More generally, our method shows that it is quite easy to control the size of any buffer. In the same way that we control the size of the heap via the `new` and `free` instructions, we could similarly control depth of a data stack (via `push` and `pop`), of a control stack (via `call` and `return`), or the value of a counter (via increment and decrement). We may also deal with several kinds of tokens, which represent different amount of memory. A challenging direction is to compare our (more) theoretical method with the exciting approaches of [27, 9]. Another interesting direction would be to incorporate our methods into the Hoare or Dynamic logic paradigm. For this, we could follow the recent works of Leivant [18, 19].

Finally, we think that our method can be extended in order to control the size of each variable of a program independently. We aim at having a criterion for termination similar to the Size Change Principle. We have started to do so on simple basis in [24, 21] and a bit more elaborated in [22]. We intend to go further in this direction.

8. REFERENCES


