Coupling Profile and Historical Methods to Predict Execution Time of Parallel Applications

Thierry Monteil
CNRS, LAAS, 7, avenue du Colonel Roche, F-31400 Toulouse, France
Univ. de Toulouse, INSA, LAAS, F-31400 Toulouse, France
monteil@laas.fr

Abstract- This article describes some work in the domain of application execution time prediction, which is always necessary for schedulers. We define a hybrid method of time prediction that is both profile-based and historic-based. This prediction is achieved by combining a program structure analysis with an instance-based learning method. We demonstrate that taking account of an application’s profile improves predictions compared with classical historic-based prediction methods.

Keywords- Performance Prediction; Execution Time; Program Analysis; Historic Model; Parallel Application

I. INTRODUCTION

Much research has been conducted into the prediction of application execution times to determine how to connect this execution time with their launching context (application input, platform performance, etc.). The ultimate aim, therefore, is to estimate the execution time of an application before it starts. In the field of real-time computing, the usefulness of such data is crucial for the proper functioning of the systems, whether critical or not. Indeed, real-time applications are subject to time constraints and should be strictly observed deadlines for hard real-time systems, or at best for soft real-time systems. In all cases, knowledge of the execution time of applications is necessary for real-time schedulers to manage the execution order of applications submitted to them using the WCET (Worst-Case Execution Time) \[1, 2, 3, 4\]. Scheduling mechanisms applied to the fields of clusters and grids also require an estimation of the duration of applications to map the system hardware.

In this article, we focus on regular parallel programs. We propose a hybrid method not to estimate the WCET but to predict execution time depends on specific inputs. This method combines several approaches:

- Analysis of history of past executions
- Statistical analysis of the parameters and input files of the program
- An annotation of source code

II. RELATED WORK

WCET estimation can be done with two main techniques \[8\]:

- The prediction based on a history of past executions (historic-based prediction): this technique is used to predict the time of sequential or parallel applications, in order to schedule them in a cluster or grid computing.
- The prediction based on the profile of applications (profile-based prediction): this type of analysis is commonly used in real time to determine the execution time of an application in the worst case.

A. Dynamic Method for WCET

In this method, the program execution time is measured either on a real system or using a simulator. The application is well executed on the target hardware, and a measure of its execution time is performed \[9\]. Where such execution is impossible, a software simulator can be used to simulate the system hardware.

All methods used to measure the WCET need a set of inputs to run the program and the main difficulty on dynamic methods is to choose a set of inputs in accordance with the execution time duration. To do this, it is possible to use explicit test sets or symbolic test sets.

B. Static Method for WCET

Static analysis analyse the structure of the program from its source code or object code, in order to deduce its WCET \[11\]. It involves three steps:

1. Flow analysis: this phase determines all possible execution paths in the program.
2. The low-level analysis: this allows assessment of the impact of the hardware architecture on the WCET.
3. The calculation of the WCET: this value is determined from the results of the two previous phases.

The flow analysis determines all possible execution paths of a program. For this, the first step is to cut the code of this program into basic blocks.

A basic block is a maximum sequence of instructions with a single entry point and one, and only one exit point in the flood control program. A basic block contains simple instruction which exclude the branch instruction (control structures, function calls, etc) \[10\].

A control flow graph can then be used to display all possible sequences between different basic blocks. In the general, the flow analysis is not a solvable problem \[10\]. Under certain conditions with additional information \[11, 12, 13, 14, 15\], we can bound the number of possible execution paths.
paths and find or improve the WCET.

The phase of low-level analysis estimates the maximum execution time of each block based on a given hardware architecture. This analysis, done from the object code, is mainly dependent on the accuracy of hardware models used. Hardware systems include mechanisms to accelerate the execution time of programs, such as pipelines, branch prediction, multiple execution units or caches. Then, an average:

\[ f \text{ nodes, etc.} \]

The estimation is done using the basic block graph to calculate the worst way. The most common method (Implicit Path Enumeration Technique) transforms the control flow graph into a set of constraints to be respected. This allows the problem to be reduced to a linear optimization of integer variables.

C. Historic Method for WCET

In this approach, the estimated execution time of an application is made according to the execution time of that application obtained in the past. It is considered that the execution time of an application depends on the context in which it is launched: two executions with neighbours context produce relatively close neighbours execution.

The execution time of extended basic blocks is the solution of a system of linear equations. For each execution of the program, we have:

\[ \text{exec}_{\text{time}}(E) = T_{\text{exec}} (E) \]

where:

\[ T_{\text{exec}} (E) = \sum_{i=1}^{n} N_i (E) \cdot T_i (E) \]

The execution time of an extended basic block will be considered constant, i.e. independent of the context of the program. This assumption excludes consideration of the mechanisms present in modern processors, such as caches.

In addition, the execution time of extended basic blocks is independent of the inputs applied to the program, which is due to the lack of branch instruction (including conditional) within extended basic blocks.

The following equation can be used to evaluate the execution time of an application:

\[ T_{\text{exec}} (E) = \sum_{i=1}^{n} N_i (E) \cdot T_i (E) \]

where:

\[ f \text{ set of program functions}, \]

\[ N_i (E) \] the number of executions of the function \( f \), depending on inputs \( E \),

\[ T_b \] the execution time of the function \( f \), depending on inputs \( E \).

The execution time of an application can be expressed as:

\[ T_{\text{exec}} (E) = \sum_{i=1}^{n} N_i^b (E) \cdot T_b^b \]

where:

\[ BB_f^b \] the set of extended basic blocks of the function \( f \),

\[ N_b^b (E) \] the number of executions of basic block \( b \) extended, depending on the inputs,

\[ T_b^b \] the execution time of the extended basic block \( b \),

considered as constant.

The execution time of an application can be easily expressed as:

\[ T_{\text{exec}} (E) = \sum_{i=1}^{n} N_i^b (E) \cdot T_b^b \]

with \( BB \) the set of the extended basic blocks of the program:

\[ BB = \bigcup_{f \in F} BB_f^b \]

The tools gprof (profiler) and gcov (coverage testing) provided by GNU allow us to know respectively \( T_f^b \) and \( N_b^b \). The execution time of extended basic blocks is the solution of a system of linear equations. Each execution of the program for different input values adds an equation to the system of linear equations. For \( X \) executions, we have \( X \) linear equations and \( Card(BB) \) unknowns.

B. Experiments

We use a C code to calculate the power \( p \) of a matrix of dimension \( d \). This program runs on different processor architectures and operating systems.

1) Reproducibility of Execution:

An important aspect that should be checked for consistency of time obtained is the reproducibility of experiments. Indeed, it is essential that two distinct runs of the program for identical inputs produce similar execution times. We chose to run the program 100 times for each set of entries tested. In Figure 1(a) we see the regular increase of execution time depending on the power and dimension of the matrix.
magnitude. On average, the difference is 4.09%. It may be noted, however, that the variation is dependent on the processor architecture.

3) Solving System:

We assume that the system has enough equations to be solved (\( X > Card(BB) \)). The system is potentially overdetermined. Let \( \mathcal{E} \) be the set of input vectors eligible for program. For each execution, there is \( E^c \), with \( c \in [1, X] \), the inputs used in the executions. A system of equations is obtained:

\[
\forall c \in [1, X], \quad T_{App}(E^c) = \sum_{b \in BB} N^B B_b(E^c) \cdot T^B_b
\]

It is expressed thus:

\[
A \cdot z = b
\]

The relations 2 and 3 show the following relations:

\[
A = \begin{bmatrix}
N^B B_1(E^{(1)}) & N^B B_2(E^{(1)}) & \cdots & N^B B_{Card(BB)}(E^{(1)}) \\
N^B B_1(E^{(2)}) & N^B B_2(E^{(2)}) & \cdots & N^B B_{Card(BB)}(E^{(2)}) \\
\vdots & \vdots & \ddots & \vdots \\
N^B B_1(E^{(X)}) & N^B B_2(E^{(X)}) & \cdots & N^B B_{Card(BB)}(E^{(X)})
\end{bmatrix}
\]

\[
z = \begin{bmatrix}
T^B_1 \\
T^B_2 \\
\vdots \\
T^B_{Card(BB)}
\end{bmatrix} \quad b = \begin{bmatrix}
T_{App}(E^{(1)}) \\
T_{App}(E^{(2)}) \\
\vdots \\
T_{App}(E^{(X)})
\end{bmatrix}
\]

The resolution of the system leads directly to results of poor quality due to the ill-conditioned nature of the system to be solved, so we reformulate the problem by introducing an error and moving to an iterative solution. This was validated on pilot matrix multiplication program by running the program and really testing the prediction based on previous executions (Figure 3).

IV. CORRELATION BETWEEN EXECUTION TIME AND INPUTS

A. Estimation of the Number of Execution of the Extended Basic Blocks

Now, the next step is to estimate the number of executions \( N^B B_b(E) \) of each extended basic block based on the input of the program contained in a database of previous versions (also called experiments) and the instance of the application, which is needed in order to predict the execution time. The model defined below is used to select, from the knowledge base, a set of experiments similar to the query, and then combine them in order to estimate the
number of executions of basic blocks of a program for a given input vector.

Each experiment is associated with a vector input \( E^{(i)} \), with \( i \in [1; N] \):

\[
E^{(i)} = \left\{ E^{(1)}_v \right\}_{v \in [1; N^V]} = \begin{bmatrix} E^{(1)}_1 \\ E^{(1)}_2 \\ \vdots \\ E^{(1)}_{N^V} \\ E^{(2)}_1 \\ E^{(2)}_2 \\ \vdots \\ E^{(2)}_{N^V} \end{bmatrix} \quad \text{with} \quad E^{(i)} \in \mathbb{R}
\]

where \( N^V \) is the number of values that form the inputs of the program. To measure the distance of two execution contexts, the Euclidean distance is used \([35]\). Let two input vectors \( E^{(1)} \) and \( E^{(2)} \) of the program, be expressed as follows:

\[
\forall E^{(1)} \in \mathbb{R}, \quad \forall E^{(2)} \in \mathbb{R}, \quad D( E^{(1)}, E^{(2)}) = \sqrt{\sum_{v=1}^{N^V} d( E^{(1)}_v, E^{(2)}_v)^2}
\]

where \( d(E^{(1)}_v, E^{(2)}_v) \) is the distance between two input values defined using a heterogeneous distance \( d(E^{(1)}_v, E^{(2)}_v) \):

\[
\begin{cases} 
1 & \text{if } E^{(1)}_v \text{ or } E^{(2)}_v \text{ is unknown}, \\
0 & \text{if the value } E^{(1)}_v = E^{(2)}_v, \\
1 & \text{in the other case}. \\
\end{cases}
\]

The estimated number of occurrences of extended basic blocks lying in the path of execution for a vector input is achieved through a weighted average of the values contained in the knowledge base. The weighting will be based on the distance between the experiences of the application, in order to promote the influence of the closest experience of the application \([7, 33]\). If \( E^* \) is the vector entries for the query, the number of executions of each extended basic blocks \( b \) in the program can be estimated as follows: \( \forall b \in \mathbb{B} \)

\[
N^{BB}(E^*) = \frac{\sum_{b \in \mathbb{B}} K(D(E^*, E^*)) \cdot N^{BB}(E^*),}{\sum_{b \in \mathbb{B}} K(D(E^*, E^*))}
\]

Where \( \mathbb{B} \) is a set of input vectors corresponding to the selected experiences in the knowledge base. The weighting function chosen is the Gaussian function:

\[
K(d) = e^{-d^2}
\]

\( k \) is a constant for varying the width of the Gaussian. This constant can be adapted to the density of experience in the region containing the query.

To validate the approach, a particle-filtering program parallelized with MPI was used \([20]\). It is built on a master / slave model. Figure 4 shows a comparison between the actual number of executions of extended basic blocks of the program for a set of inputs and that estimated by the prediction model defined. The two curves have the same shape. Thus, the prediction model is able to identify blocks in which the number of executions is varying, depending on the inputs, as well as to estimate the effects of input values of these variations. The mean relative error, found on the prediction of all blocks, amounted to 8.8% for the set of inputs considered.

\[
\text{Fig. 4 comparing the real and predicted number of executions of extended basic blocks}
\]

B. Influence of the Base of Knowledge

Next, we want to determine the influence that the knowledge base can have on the estimation accuracy, focusing more precisely on the influence of the number experimentations it contains. The experiments described also allow us to study the relationship between the coefficient \( k \) and the degree of filling of the knowledge base. For this, the above experiment is repeated. Several sets of estimates will be made, using for each 500 requests. Each set of estimates is performed several times, varying the number of experiments in the database of knowledge on the one hand, and the coefficient \( k \) to adjust the width of the Gaussian function weighting of other part.

Figure 5 represents the overall relative error found for each set of queries based on the number of experiments in the present knowledge base and the coefficient \( k \). This curve highlights a steady increase in the observed error, which may be 0 if the parameters tested are well adjusted, but can also exceed 200% if this is not the case. It thus appears that the accuracy of the model depends strongly on the choice set of these parameters.

\[
\text{Fig. 5 Relative error in each set of 500 queries with the number of experiments in the database and the coefficient } k
\]

To investigate more thoroughly the influence of the degree of filling of the knowledge base and that of the coefficient \( k \), we make two cuts (Figure 6) of the three-dimensional view of Figure 5.
Error predictions based on the coefficient $k$ (500 experiments found in the knowledge base)

Error predictions based on the number of experiments in the database of knowledge ($k=0.02$)

For a knowledge base with enough experience, a too high value of the coefficient $k$ implies a significant error in predictions (Figure 6(a)), due to the disruption of the prediction by experimental noise. In practice values of $k$ between 0.01 and 0.1 offer good results (cf. Figure 7).

In addition, the Curve 6(b) shows that the accuracy of the prediction of behaviour is essentially dependent on the knowledge base. A large number of experiments provide increased accuracy, while the probability of finding one or more experiments similar to the query is higher.

However, the lack of experience can be compensated, to some extent, by increasing the coefficient $k$, in order to expand the number of experiments included in the prediction process. Indeed, Figure 8 shows that for low values of $k$, the curve has a shape of a hollow: the error decreases first, before increasing when the value $k$ increases.

The previous formulation assumes that all input values of the program have a similar influence on the number of executions of each extended basic block of the program. This hypothesis is strong. One way to better take into account the impact of each entry is to add a factor accounting for this: $w_{v,b}$, in calculating the distance to the input $v$ and the block $b$:

$$D_k(E^{(1)},E^{(2)}) = \frac{\sum_{v \in B} w_{v,b} \cdot d(E^{(1)},E^{(2)})^2}{\sum_{v \in B} w_{v,b}}$$

The choice of values: $w_{v,b}$ depends entirely on the program structure. Thus, the person most able to make such a choice is undoubtedly the application developer.

The annotations are a flexible and easy way to allow the programmer to learn these values. They must meet the following characteristics:

- They allow the user to give the dependence of the number of execution blocks depending on the inputs of the program.
They require no knowledge of the mathematical model for predicting the behaviour of applications,

They can be inserted directly into the source code of the program,

They do not block compilation or execution in the proper functioning of the application.

We have chosen as part of C, the use of directives #pragma that are ignored by the compiler. Similar concepts exist in most of programming languages. The following syntax is used:

```
#pragma etp annotation (parameters)
```

where:

- etp means execution time prediction. This keyword characterizes the set of annotations that we create, so they are immediately identifiable by the tool in place to interpret them.
- annotation designates the type of annotations. There are four different types of annotations, described below.
- (parameters) is a list of optional parameters, separated by commas and enclosed in parentheses.

Four types of annotations are defined:

- **Annotations type “inputs”:** This type of annotation defines the inputs considered in the prediction of execution time. Such annotations can be inserted at the beginning of the source code alongside the traditional guidelines #define.

- **Annotations type “begin dependency”:** This type of annotation starts a sequence of instructions whose execution depends on number of entries in the program. The list of these entries, as defined by Directive “input”, is specified parameters.

- **Annotations type “begin independency”:** This type of annotation starts a sequence of instructions on which the number of executions depend in any input of the program.

- **Annotations type “end”:** This type of annotations closes any sequence of instructions, thus following the annotation types “begin dependency” or “begin independency”.

The annotation allows also to improve the sensibility of communication part. Communication block can be enclosed with the specific parameters that characterize the amount of data exchanged. In MPI communication most of the time, arrays are exchanged. So the time of communication could be connected to the amount of data send or received.

For example, the following can be added to the program of particle filtering, in order to define the three inputs considered:

```
#pragma etp inputs (particles, time_intervals, slave_tasks)
#pragma etp begin dependency (particles)
for (int i=0 ; i<particles ; i++) {
    instructions_1;
    #pragma etp begin dependency (time_intervals)
    for (int j = 0 ; j < time_intervals ; j++) {
        instructions_2;
    }
    #pragma etp end
    instructions_3;
}
#pragma etp end
```

Thus, the instruction blocks 1 and 3 depend only on the number of particles, while the statement block 2 depends on both the number of particles and the number of time intervals.

The same experiments as above were performed and the prediction error with or without annotation is compared in Figure 9, which shows the significant improvement in prediction using annotations.

**VI. FULL MODEL OF PREDICTION**

With profiled applications, the execution time of each basic block of the program is determined based on the iterative resolution of the equations system formed by the data from experiments in the present knowledge base. This phase of learning and estimation can be done in the background and stop when the times of extended basic blocks are stable (Figure 10). Then, when the execution of an application is submitted to a scheduler a pre-processing task are carried out:

- Query the database and get the time of basic blocks
- Estimate the characteristics of a selected set of applications according to their proximity to the application in the database
- Deduce the prediction of execution time that can be used by the scheduler to refine the mapping.

The pre-processing time needed to predict execution time depends mainly on the size of the database, the number of inputs and the number of extended basic blocks. In our experiments, this time (a few seconds) is neglected compared to a submission time of the grid scheduler.
VII. FULL MODEL OF PREDICTION

This paper proposes a method to predict performance of regular applications running in parallel. The difficult problem of communication time is hidden within the extended basic block and annotations principle. This is possible only if the network is not overloaded and the synchronisation between the different tasks of the parallel application stays similar.

The hybrid approach for predicting execution time is based on two prediction methods:

- A method using a history of past performances and based on a learning-based capabilities,
- A method based on the profile of applications, as it can be applied in determining the WCET of real-time applications.

This approach requires two types of program information:

- Temporal information, that is to say, the execution time of each extended basic block. This information can be obtained from solving a system of linear equations obtained from a set of executions of the program. This system is poorly conditioned, so we have adapted a classical numerical method to solve this system of equations by successive iterations.
- Behavioural information, that is to say the number of executions of each basic block. We have shown how the estimate can be done from a history of past executions, by implementing a learning approach based on the proceedings. In addition, we also defined a system of annotations of the program source code to improve this estimate.

The complete model for prediction using the hybrid approach has been tested and shows satisfactory results for the different set of parallel applications. Indeed, the hybrid approach allows us to take into account the structure of the program. The result is a more detailed prediction as shown on Figure 9 since it is possible to identify portions of the program that have a large execution time and link them directly to a set of inputs of the program. This method has been also used to predict the execution time of electromagnetic simulation to improve the utilisation of data center with autonomous policies. In this case, the MPI application follows another scheme of communication based on Simple Program Multiple Data architecture and communication with its neighbours only.

Further improvements of this method are possible:

- Instrumentation of object code would no longer be estimated by calculating the execution time of basic blocks, but by measuring them directly. The method would then gain accuracy.
- An analysis of the data flow of the program could be used to infer some annotations automatically. So even if all the dependencies of basic blocks with the entries could not be deducted automatically, the task of the programmer would still be simplified.
- It may also be worth considering creating analytical models of parallel applications to improve the prediction. These models describe the structure of applications. A model corresponding to a family of applications, such as master / slave applications will allow the consideration of common features with known effects on the execution time.
- The comparison with other methods coming from the real-time domain is difficult because they are more interested to have very good WCET in sequential program than estimation in parallel application. Nevertheless, benchmark of high number of parallel applications will allow improving the method.

REFERENCES


Thierry Monteil is associate professor in computer science since 1998 at INSA Toulouse and researcher at LAAS-CNRS. He received the Engineering degrees in Computer Science and applied mathematics from ENSEEIHT in 1992. He had a Doctorate in parallel computing in 1996 and a HDR degree in 2010.

He works on parallel computing middleware (LANDA parallel environment), Grid resources management (AROMA project), computer and network modeling, load balancing with prediction models, autonomous policies to improve performance on distributed applications, parallelization of large electromagnetic simulation, autonomic middleware (FrameSelf project) and machine-2-machine system. He has managed a SUN microsystems center of excellence in the field of grid and cluster for network applications and a Cisco academy. Since 2011, he coordinates the industrial SOP project funded by ANR that creates hybrid cloud for personal service over ADSL network under energy and quality of service constraints. He is author of more than 50 regular and invited papers in conferences and journals.