Transforming Source Code to Mathematical Relations for Performance Evaluation

Habib Izdikhah

Department of Computer Science, Faculty of Mathematical Sciences, University of Tabriz, Tabriz, Iran

Abstract — Assessing software quality attributes (such as performance, reliability, and security) from source code is of the utmost importance. The performance of a software system can be improved by its parallel and distributed execution. The aim of the parallel and distributed execution is to speed up by providing the maximum possible concurrency in executing the distributed segments. It is a well known fact that distributing a program cannot be always cause speeding up the execution of it; in some cases, this distribution can have negative effects on the running time of the program. Therefore, before distributing a source code, it should be specified whether its distribution could cause maximum possible concurrency or not. The existing methods and tools cannot achieve this aim from the source code. In this paper, we propose a mathematical relationship for object oriented programs that statically analyze the program by verifying the type of synchronous and asynchronous calls inside the source code. Then, we model the invocations of the software methods by Discrete Time Markov Chains (DTMC). Using the properties of DTMC and the proposed mathematical relationship, we will determine whether or not the source code can be distributed on homogeneous processors. The experimental results showed that we can specify whether the program is distributable or not, before deploying it on the distributed systems.

Keywords: Distributed Software Systems, Source Code, Speedup, Discrete Time Markov Chains

(Received: 18.05.2015; Revised: 21.07.2015; Published: 21.09.2015)

1 Introduction

The need for high speed computation in large-scale scientific applications for analyzing complex scientific problems is very high, so that the common computers would not be able to satisfy it. Therefore, nowadays, using the distributed systems and processing power of numerous processors or cores to reach the favorable speed is known as a fact [1]. Yet, as a fact, creating a large scale distributed program is always more difficult than creating a non-distributed program with the same functionality, as the creation of a distributed system can change into a tedious and error-prone task.

Since the computational programs have many computations, so its execution requires more time. Therefore, if a program does not have the ability to distribute, there will be a lot of waste time. The most important time of a distributed program is invocation or communication time of their methods. These calls spend the most execution time. Certainly, by distributing a program, if two classes of it can be distributed on two different machines, the invocations between those classes will turn into the remote calls. As reference [2] specifies, in some cases, the program distribution can have negative effects on the running time of the program. When there are many calls between two methods, the network traffic increases and as a result, efficiency of the distributed program will be lower than the initial sequential program. So, regarding that constructing the distributed program from the source code is complex and time consuming, it is better to predict whether the source code is distributable or not, before distributing a program on the machines. None of the existing methods and tools can achieve this goal from source code.

1.1 The Problem and the Claim

The overall problem addressed in this paper is to specify whether the source code has the potential for parallelization on homogeneous processors; i.e., in case of distribution, whether it brings the maximum concurrency compared to the sequential mode. We claim that it is possible to provide a solution to the mentioned problem by doing the following tasks:

(1) Modeling software's method invocations by Markov chains as (described in section III) as:
   - Markov chains nodes represent methods and edges between nodes represent calls between methods,
• weight of the edges in Markov chains, shows
  the number of calls between the methods.
(2) Determine the maximum potential of distributability of each method (described in
section III)
(3) Determine the expected performance of the
source code from obtained Markov chain (de-
scribed in section III)
(4) compute the speedup. Speedup is defined as the
execution time of a sequential program divided
by the execution time of a parallel program that
computes the same result. In particular,
\[ \text{Speedup} = \frac{T_s}{T_p} \]
where \(T_s\) the sequential time
and \(T_p\) is the expected performance.

1.2 The Paper Outline

The other sections of the paper are organized as fol-
ows: A literature review on the researches conducted by
others is discussed in section II. In section III, we propose
a mathematical relation of time estimation by which the
potential for distribution of the source code can be spec-
ified. Case study is discussed in sections IV. At the end,
section V deals with conclusions and future works.

2 Related Work and Background

The complicated computational applications cannot be
executed in an acceptable time on the computation ma-
chine, so they should be divided into small tasks. We can
use distributed or multiprocessor systems for executing
of these tasks. Nowadays, most distributed and mul-
tiprocessor tools use scheduling methods for distribution.
The aim of scheduling is execution of a program on several
processors such that the time of execution of the whole
program will be minimal, considering the time of tasks
and communication time between the processors [3]. The
scheduling methods can be divided into two groups; in-
cluding those which can assurance the quality of service,
and those which cannot. The former scheduling systems
are preferred to the latter ones. CONDOR [4], SGE [5],
PBS [6] and LSF [7] can be referred to as some of the
most popular and widely used scheduling systems. These
scheduling systems do not guarantee the service quality.
These tools perform the scheduling only at the job level
and not at applications’. Unlike the above systems, there
are some which observe the service quality in schedul-
ing. Such systems observe Job Characteristics, Planning
in Scheduling, Rescheduling and Scheduling Optimization
in their scheduling. AppleS [8], GrADS [9] and Nimrod/G
[10] are among the most famous systems of this kind.
Moreover, none of the aforementioned schedulers can pre-
dict whether an offered program has the potential to be-
come parallelized, or whether speedup can be achieved in

2.1 Overview of Discrete Time Markov Chains

In this section, we discuss Discrete Time Markov
Chains (DTMCs), which we use to model the source
code’s invocations [13]. A DTMC is described by its
states and transition probabilities between the states;
where we indicate the transition probabilities between
the states as one-step transition probability matrix. The
one-step transition probability is the probability that the
process, when in state \(i\) at time \(n\), will next transition to
state \(j\) at time \(n+1\). We write:

\[ P_{ij}(n) = P(X_{n+1} = j \mid X_n = i) \]  

Note that all the elements in a row of \(P\) add up to 1 and
each of the \(P_{ij}\)’s lie in the range \([0, 1]\). For our purpose,
we use absorbing DTMC. One DTMC is called absor-
ing if at least one state has no outgoing transition. Each
DTMC with several final states can be converted into an
absorbing DTMC. It is performed by adding a final state
to DTMC. Next, a transition is drawn to the added ab-
sorbing state from all the final states available in DTMC.
We can partition the transition probability matrix of an
absorbing DTMC as:

\[ P = \begin{bmatrix} 1 & 0 \\ C & Q \end{bmatrix} \]

If the DTMC has \(n\) states with \(m\) absorbing states, \(Q\)
would be a \((n-m) \times (n-m)\) sub-stochastic matrix (with
at least one row sum \(< 1\) describing the probabilities of
transition only between transient states, \(C\) being a \(m \times m\)
identity matrix, \(0\) would be an \(n \times (n \times m)\) matrix of zeros,
and \(C\) would be an \((n-m) \times m\) matrix describing the
probabilities of transition between transient states and
absorbing state. The \((i,j)\)-th entry of \(Q^k\) denotes the
probability of arriving to state \(s_j\) after exactly \(k\) steps,
starting from state $s_i$. Hence the inverse matrix $(I - Q)^{-1}$ exists. This is called the fundamental matrix $F$:

$$F = (I - Q)^{-1} = I + Q + Q^2 + Q^3 + \cdots + \sum_{i=0}^{\infty} Q^i.$$ (3)

Let $X_{i,j}$ represent the number of visits to state $j$ starting from the state $i$ before process is absorbed. It can be shown that the expected number of visits to state $j$ with starting from state $i$ (i.e., $E[X_{i,j}]$), before entering an absorbing state is given by the $(i,j)$-th entry of the fundamental matrix $F$ [14, 15]. So

$$E[X_{i,j}] = m_{i,j},$$ (4)

$m_{i,j}$ is the $(i,j)$-th entry of the fundamental matrix $F$. The variance of the expected number of visits could also be computed using the fundamental matrix. Let $\sigma_{i,j}$ denote the variance of the number of visits to the state $j$ starting from state $i$. Define $F_D = [m_{i,j}]$ such that:

$$md_{i,j} = \begin{cases} m_{i,j} & \text{if } i = j \\ 0 & \text{otherwise}. \end{cases}$$ (5)

In other words, $F_D$ represents a diagonal matrix with the diagonal entries the same as that of $F$. If we define $F_2 = [m_{i,j}^2]$, we have:

$$\sigma^2 = F(2F_D - I) - F_2.$$ (6)

Hence:

$$\text{Var}[X_{i,j}] = \sigma_{i,j}^2.$$ (7)

3 Predicting Performance Of A Source Code

In this section we describe our approach for modeling a software system that method invocations are represented by an absorbing DTMC, such that DTMC states represent the software methods, and the transitions between states represent transfer of control from one method to another. We assume that the system consists of $n$ methods, and has a single initial state denoted by 1, and a single absorbing or exit state denoted by $n$. Consider Fig. 1. Numbers on edges indicate the probability of movement from one method to another method. In this paper the probability to go from method $x$ to method $y$ is computed as [number of method call from $x$ to $y$ / total number of out method call of $x$ (i.e. fan out)]. The method invocations of the source code are given by the one-step transition probability matrix $P$.

![Figure 1. Modelling method invocations for a sample program with DTMC](http://ai.annales.umcs.pl)

Equation (8) shows the one-step transition probability matrix $P$ for Figure 1.

$$P = \begin{pmatrix} 0 & 0.5 & 0.5 & 0 & 0 & 0 & 0 \\ 0 & 0.25 & 0 & 0.25 & 0 & 0.25 & 0 \\ 0 & 0 & 0 & 0 & 0.5 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0.5 & 0 & 0 & 0.5 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$ (8)

Let $PD_i$ denotes the potential of distributability of method $i$ that indicated by node $i$ in the DTMC. During a single execution, the performance of the software, denoted by the random variable $P$ is given by:

$$P = \prod_{i=1}^{n} PD_{X_{i,i}}.$$ (9)

where $X_{i,i}$ denotes the number of visits to the transient state $i$ starting from the state $1$. Therefore, the expected performance of a software system is as follows:

$$E[P] = E\left[\prod_{i=1}^{n} PD_{X_{i,i}}\right] = \prod_{i=1}^{n} E\left[PD_{X_{i,i}}\right].$$ (10)

Thus to obtain the expected performance of the source code, we need to obtain $E[PD_{X_{i,i}}]$, which is the expected potential of distributability of method $i$ for a single run of the software. Using the Taylor series expansion, $E[PD_{X_{i,i}}]$ in relation 10 can be written as relation 11.

$$E\left[PD_{X_{i,i}}\right] = PD_{i}^{E[X_{1,i}]} + \frac{1}{2}(PD_{i}^{E[X_{1,i}]})(\log PD_{i})^{2}\text{Var}PD_{i}.$$ (11)

Let $E[X_{1,i}] = m_{1,i}$ and $\text{Var}[X_{i,j}] = \sigma_{i,j}^2$, relation (11) may be written as:
(12) \[ E \left[ PD_i^{X_{1,i}} \right] = PD_i^{m_{1,i}} + \frac{1}{2} (PD_i^{m_{1,i}})(\log PD_i)^2 \sigma^2_{1,i}, \]

\( m_{1,i} \) is the expected number of visits to state \( i \) and \( \sigma^2_{1,i} \) is the variance of the number of visits to state \( i \). \( m_{1,i} \) and \( \sigma^2_{1,i} \) can be obtained from DTMC analysis. Relation (10) can thus be written as:

(13) \[ E[P] = \prod_{i}^{n} (PD_i^{m_{i,i}} + \frac{1}{2} (PD_i^{m_{i,i}})(\log PD_i)^2 \sigma^2_{1,i}) . \]

3.1 Computing Potential Of Distributability Of Method \( i \)

In this section, we are going to determine Potential of Distributability (PD) of each method to determine overall performance (i.e., \( P \)) of a program. For this aim, we determine \( PD_i \) to measure the values of different distributions for method \( i \). Invocation (or call) between methods are two types of asynchronies and sequential. If by distributing a program, two methods of the program distribute in two different machines, calls between those methods will turn into asynchronies; and in sequential call, two methods of the program are placed on the same machine. Considering of communication time, our method considers two asynchronies and sequential mode for each call; to determine which mode (sequential or parallel) can reach a maximum speed up.

To estimate the speed-up, the execution time of all instructions should be estimated. The execution time of all instructions, except the nested calls, can be computed by the existing methods [16-17]. The existing methods cannot be applied easily to calculate the execution time of nested calls because the execution time of a caller method is depending on the fact that the calls inside it are carried out in a sequential or asynchronous manner. For example, consider Listing 1. In the Listing 1, in the time \( t_1 \), the current method (caller method) will continue to work in a non-stop manner until reaching the use point of the results of a callee method. We call these points' synchronization points [18] and is shown by \( S \). So, one method continues to work after calling a method from a remote locations (other distributed segments) and waits for a call response only when requires that response. As shown in Listing 1, the level of concurrency in executing the caller and the callee methods depends on the time interval between the call point and use point of the call results. The problem is the estimation of this interval time. As shown in Listing 1, there may be other calls between the call point and use point and the execution of these calls can be either synchronous or asynchronous.

**Listing 1. Several nested calls**

```plaintext
Method m ( ) {
    Some statements // t0
    Call R
    Some statements // t1
    Use R // S
    Some statements // t2
}
Method R ( ) {
    Some statements // t3
    Call P
    Some statements // t4
    Use P // S
    Some statements // t5
}
Method P ( ) {
    Some statements // t6
}
```

3.1.1 Estimated execution time for sequential mode

In Listing 1, considering methods \( m \), \( R \) and \( P \), if all of them executed sequentially (or synchronously), the estimated execution time will be calculated as follows:

(14) \[ PD_{sequential} = t_0 + t_3 + t_6 + t_4 + t_5 + t_1 + t_2. \]

We can write above relation for Listing 1 in the recursive form and expand it for the nested call with any depth.

(15) \[ PD_{sequential} = t_0 + PD_{sequential}^R + t_1 + t_2. \]

(16) \[ PD_{sequential}^R = t_3 + PD_{sequential}^{p} + t_4 + t_5. \]

(17) \[ PD_{sequential}^p = t_6. \]

Generally, for the sequential call, estimated execution time relation, is as relation:

(18) \[ PD_{sequential} = \sum t_i + PD_{sequential}^R. \]

3.1.2 Estimated execution time for asynchronous mode

Now we calculate the estimated execution time when methods are executed parallel (or asynchronously). See again Listing 1. If methods \( m \), \( R \) and \( P \) are executed
asynchronously, the estimated execution time will be calculated as follows:

\[
P_{D_{async}}^{m} = t_0 + t_1 + I_{init} + t_2.
\]

(19)

\[
max(P_{D_{R}}^{async} - t_1 + C_t + I_{init}, 0) + t_2.
\]

(20)

\[
P_{D_{async}}^{R} = t_3 + t_4 + T_{init} + t_5.
\]

(21)

\[
max(P_{D_{async}}^{P} - t_4 + C_t + I_{init}, 0) + t_5.
\]

\[
P_{D_{async}}^{P} = t_6.
\]

C_t is the communication time and I_{init} shows the preparation time for doing remote call. Generally, the estimated time relation for the parallel (or asynchronous) is calculated as follows:

\[
P_{D_{async}}^{m} = \sum t_i + \sum I_{init} +
\]

\[
max(P_{D_{async}}^{P} - t_i + C_t + I_{init}, 0).
\]

(22)

3.1.3 Determining the Potential of Distribution

Considering the relations (18) and (22), the general mathematical form of a PD relation is written as follows:

\[
P_{D_{m}} = \sum t_i + \sum a_i \times P_{D_{I_i}} + \sum (1 - a_i) \times (I_{init} +
\]

\[
max((P_{D_{I_i}} + C_t) - t_i + I_{init}, 0)).
\]

(23)

In the above relation, depending on the call to be synchronous or asynchronous, the value of a_i is considered as 1 and 0, respectively. The goal is to determine a_i, so that this minimizes P_{D_{m}}. In the relation (23), the communication time is C_t and t_i is the estimated time between the call point of I_i and the synchronization point of S_i (use point).

For example, to obtain PD for Listing 1, we need to combine the estimated times for the asynchronous (relation 22) and sequential execution (relations 15-17) as follows:

\[
P_{D_{m}} = t_0 + a_1 \times P_{D_{R}} + t_1 +
\]

\[
(1 - a_1) \times (I_{init} + max(P_{D_{R}} - t_1 + C_t + I_{init}, 0)) + t_2,
\]

(24)

\[
P_{D_{R}} = t_3 + a_2 \times P_{D_{P}} + t_4 +
\]

\[
(1 - a_2) \times (I_{init} + max(P_{D_{P}} - t_4 + C_t + I_{init}, 0)) + t_5,
\]

\[
GTE_p = t_6.
\]

In relation 24, the aim is to determine a_1 and a_2 in a way to minimize P_{D_{m}}, P_{D_{R}} and P_{D_{P}}.

Listing 2. A sample program code

Class A {
    
    Public void m () {
        // some statements T1
        B b=new B();
        int r1 = b. m();
        print (r1); //S1
        C c=new C();
        int r2= c. n();
        D d=new D();
        int r3= d.p();
        // some statements T2
        if (r2==1) {...} //S2
        // some statements T3
        F f=new F();
        int r4= f.g();
        If((r1>r2 & & r1>r3 & & r1>r4)
        {...} // S3 and S4
        // some statements T4
    }
}

} // class

Class B extends A{
    
    static int m() { // some statements T5
        }
}

} // Class

Class C extends A{
    
    static int n() { // some statements T6
        }
}

} // Class

Class D {
    
    int p() {
        D d=new D();
        int r=d.p();
        Print (r); //S5
        F f=new F();
        int r1= f.g();
        If(r>r1) {...} // S6
    }
}

} //Class

Class F {
    
    // some statements T7
}

} //Class

Considering the program code in the Listing 2, P_{D_{A,m}} can be written as (25).
\[
PD(A.m) = T_1 + a_1 \cdot PD(B.m) + (1 - a_1) \cdot T(S_1) + a_2 \cdot PD(C.n) + a_3 \cdot PD(D.p) + T_2 + (1 - a_2) \cdot T(S_2) + T_3 + a_4 \cdot PD(F.g) + (1 - a_3) \cdot T(S_3) + (1 - a_4) \cdot T(S_4) + T_4,
\]

\[
PD(B.m) = T_5, PD(C.n) = T_6, PD(D.p) = a_5 \cdot PD(D.p) + (1 - a_5) \cdot T(S_5) + a_6 \cdot PD(F.g) + (1 - a_6) \cdot T(S_6),
\]

\[
PD(F.g) = T_7,
\]

\[
T(S_1) = \max(PD(B.m) + 2t_{c_1}, 0),
\]

\[
T(S_2) = \max(T_2 + a_3 \cdot PD(D.p) + 2t_{l}, 0),
\]

\[
T(S_3) = \max((PD(F.g) + 2t_{c_1}) + T_3 + ((1 - a_2) \cdot T(S_2) + T_2), 0),
\]

\[
T(S_4) = \max((PD(F.g) + 2t_{c_1}), 0),
\]

\[
T(S_5) = \max((PD(D.p) + 2t_{c_1}), 0),
\]

\[
T(S_6) = \max((PD(F.g) + 2t_{c_1}), 0).
\]

<table>
<thead>
<tr>
<th>Sequential Time (seconds)</th>
<th>Expected Distributed Time (seconds)</th>
<th>Speed-up</th>
</tr>
</thead>
<tbody>
<tr>
<td>380</td>
<td>261</td>
<td>1.455</td>
</tr>
</tbody>
</table>

Table 1. Distributed execution times, sequential execution times and speed-up for Listing 2

The aim of PD relations in (25) is to determine \(a_1, a_2, a_3, a_4\) and \(a_5\) in a way to minimize \(PD(A.m), PD(B.m), PD(C.n), PD(D.p)\) and \(PD(F.g)\). We use the Dantzig’s simplex algorithm [20] to determine the binary values of \(a_1\) for synchronous call the value of \(a_1\) is considered as 1 and for asynchronous calls, the value of \(a_1\) is considered as 0). Simplex method is a popular algorithm for linear programming. Then, after determining PD for methods \(m, n, p\) and \(g\), we make DTMC for the program of Listing 2 and then we compute the potential of distributability (using relation 24) for each method and then of course we will determine expected performance (relation 12). Also, the sequential execution time of the program is calculated as well. Finally, the speedup is calculated by dividing the sequential time to the expected performance. For relations (25), the communication overhead is considered as 1 second and \(T_1, T_2, T_3, T_4\) and \(T_5\) (execution time of non-call statements) are considered as 40, 35, 45, 50 and 20 seconds. Table 1 shows the expected distributed potential (using relation 13), sequential, and speed-up execution times for Listing 2. Since speed-up is bigger than one, this indicates that the program is capable of parallel execution; i.e., the parallel execution of the program is faster than the sequential execution of the program.

4 Evaluation Result

In this section, we evaluate the performance of the proposed method. We want to determine when the speed-up achieved by our method is greater than one; the actual execution will speed up. For achieve this goal, we use jDistributor [2] tool. jDistributor is a tool for automatic distribution of the sequential program on the homogeneous distributed systems using the Java Symphony middleware [19]. The algorithm used in the jDistributor is a hierarchical clustering method and its goal is to find an appropriate clustering for distribution. We use the well-known travelling salesman problem (TSP) for evaluating of the proposed method. We compute \(PD_{sequence}\) and \(PD_{async}\) from source code. We then predict from \(PD_{sequence}\) relation, the estimated time of the parallel and sequential execution for different graph nodes and then calculate speed-up by them. Afterwards, we distribute the TSP on the network including three computers by use of the jDistributor tool and then we calculate the parallel and sequential execution time. The results are shown in Table 2.

5 Conclusion

In this paper, we introduced a new approach to specify whether the source code is distributable or not, before the distribution. For achieve this goal, by considering asynchronous and sequential calls, a mathematical relationship was proposed to measure different distributions values from the same program code. Then, we model the software’s method invocations by Discrete Time Markov Chains (DTMC). DTMC and its properties and proposed mathematical relationship can determine whether or not the source code distribution capabilities on homogeneous processors.
Table 2. Comparison of estimated execution time using PD relation with its actual execution time

<table>
<thead>
<tr>
<th>Graph No.</th>
<th>Estimated Execution Time</th>
<th>Actual Execution Time (Using Distributor tool [2])</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Nodes</td>
<td>Edges</td>
</tr>
<tr>
<td>20</td>
<td>30</td>
<td>405</td>
</tr>
<tr>
<td>30</td>
<td>50</td>
<td>801</td>
</tr>
<tr>
<td>60</td>
<td>100</td>
<td>2230</td>
</tr>
<tr>
<td>80</td>
<td>180</td>
<td>7569</td>
</tr>
<tr>
<td>100</td>
<td>310</td>
<td>19341</td>
</tr>
<tr>
<td>130</td>
<td>1420</td>
<td>35987</td>
</tr>
<tr>
<td>170</td>
<td>686</td>
<td>59811</td>
</tr>
</tbody>
</table>

5.1 Future Work

We plan to extend and improve this work as follows: Our aim is to propose an algorithm, which attempts to improve the speed-up as much as possible in the distribution environments by reordering instructions at the compilation time. Therefore, it attempts will be made to increase distance between the caller points to its use point using the techniques known as instructions scheduling, for increase concurrent time of caller and callee methods as much as possible.

References