

Experimental Verification and Analysis of Dynamic Loop Scheduling in Scientific Applications

Ali Mohammed, Ahmed Eleliemy, Florina M. Ciorba
Department of Mathematics and Computer Science
University of Basel, Switzerland

Franziska Kasielke
Center for Information Services and
High Performance Computing
Technische Universität Dresden, Germany

Ioana Banicescu
Department of Computer Science and Engineering
Mississippi State University, USA

April 30, 2018

Contents

1	Introduction	4
2	Background and Related Work	6
3	Selected Parallel Application and Parallel Computing Systems	9
3.1	The parallel application - PSIA	9
3.2	The parallel computing systems	11
3.3	Performance characterization of PSIA	11
4	Methodology for Experimental Verification and Analysis	13
4.1	Extraction of application and platform characteristics	14
4.2	Employing DLS in native and simulative executions	15
4.3	Simulation of PSIA	15
4.4	Experimental verification and analysis	16
5	Results of Executing PSIA using DLS	17
5.1	Native execution results	17
5.2	Simulative execution results	20
5.3	Discussion	22
6	Conclusion and Future Work	23
7	Reproduction of This Work	24
7.1	Native experiments on miniHPC and Taurus	25
7.2	Simulative experiments using SG-MSG and SG-SD	25

Abstract

Scientific applications are often irregular and characterized by large computationally-intensive parallel loops. Dynamic loop scheduling (DLS) techniques can be used to improve the performance of computationally-intensive scientific applications via load balancing of their execution on high-performance computing (HPC) systems. Identifying the most suitable choices of data distribution strategies, system sizes, and DLS techniques which improve the performance of a given application, requires intensive assessment and a large number of exploratory native experiments (using real applications on real systems), which may not always be feasible or practical due to associated time and costs. In such cases, to avoid the execution of a large volume of exploratory native experiments, simulative experiments which are faster and less costly are more appropriate for studying the performance of applications for the purpose of optimizing it. This motivates the question of *‘How realistic are the simulations of executions of scientific applications using DLS on HPC platforms?’* In the present work, a methodology is devised to answer this question. It involves the experimental verification and analysis of the performance of DLS in scientific applications. The proposed methodology is employed for a computer vision application executing using four DLS techniques on two different HPC platforms, both via native and simulative experiments. Moreover, the evaluation and the analysis of the native and simulative results indicate that the accuracy of the simulative experiments is strongly influenced by the values obtained by the chosen approach used to extract the computational effort of the application (FLOP- or time-based), the choice of application model representation into simulation (data or task parallel), and the choice of HPC subsystem models available in the simulator (multi-core CPUs, memory hierarchy, and network topology). Further insights into the native performance on two HPC platforms one versus the other, the simulated performance using the two SimGrid interfaces, one versus the other, and the native versus the simulated performance for each of the simulated HPC platforms, are also presented and discussed. The minimum and the maximum percent errors achieved between native and simulative experiments are 0.95% and 8.03%, respectively.

Keywords. Dynamic loop scheduling; Self-scheduling; Parallel spin-image; Simulation; Performance; MSG; SimDag.

1 Introduction

In scientific applications, loops are prevalent and represent the primary source of parallelism, most of the application execution time being spent executing loops. Loops of scientific applications typically have a large number of iterations. During the execution of loop iterations, an imbalanced load execution due to problem, algorithmic or systemic characteristics, may result in a performance degradation of the entire application. Static and dynamic loop scheduling (DLS) techniques play an essential role in improving the performance of scientific applications. These techniques balance the assignment and the execution of different loop iterations across the available computing units. Hence, this balancing maximizes the application’s performance. Throughout the present work, a loop iteration refers to a task, both terms being used interchangeably. Identifying the best choices among available DLS techniques for a given application requires intensive assessment and a large number of native experiments¹. This significant amount of experiments may not always be feasible or practical due to their associated time and costs. Simulation mitigates such costs and, therefore, has been shown to be more appropriate to study the performance for the purpose of optimizing it.

The performance of scientific applications is commonly studied natively on real high performance computing (HPC) platforms or using simulators, with simple and straightforward loop scheduling techniques being considered. However, to promote the trustworthiness of the performance insights obtained via simulation, the simulated performance of an application should be studied, analyzed and compared to the native performance of the application. An important source of uncertainty in the performance results obtained via simulation is the degree of trustworthiness in simulation. Attaining a high degree of trustworthiness eliminates this source for the present and further experiments with more complex scheduling techniques and with variable processor and network availabilities. The absence of such analyses motivates the question of ‘*How realistic are the simulations of executions of scientific applications using DLS on HPC platforms?*’

In the present work, a methodology is proposed to experimentally verify and analyze the performance of scientific applications using DLS both, natively and via simulation in an effort to answer the above question. A review

¹Native experimentation is oftentimes also referred to as “direct” or “bare metal” experimentation in the literature, and denotes experiments using real applications on real computing systems

of DLS techniques can be found in the literature [1]. The proposed methodology is built upon three perspectives of comparison of the results of native and simulative experiments: native-to-native, simulative-to-simulative, and simulative-to-native. Through the first perspective, the performance of an application executed on different parallel and distributed platforms using different DLS techniques is examined and compared. This comparison is essential to understand the main characteristics of the application and the application sensitivity to different hardware specifications of the HPC platforms. In the second comparison perspective, the representation of the application and the platform characteristics are used in different simulators which implement DLS techniques. The simulated performance of the application is compared among the different available DLS techniques. Given that the simulators use the same application and platform characteristics, this comparison allows a better assessment of the simulator’s influence on the performance. In the third perspective, the performance of the native experiments is compared against the one of the simulative experiments. Based on the information obtained from the first two comparisons, this comparison is to verify and justify the level of the agreement between the results of native and simulative experiments.

The proposed methodology is applied to an application from computer vision. This application uses the parallel spin-image algorithm (PSIA) [2], an enhanced version of the well-known spin-image algorithm (SIA) [3]. This algorithm has various applications, such as, 3D object recognition, categorization, and face recognition. The usage of the PSIA is considered as an example of scientific applications where the performance of a single large loop dominates the entire execution of the application. Loop scheduling techniques have the potential to enhance the performance of the PSIA [4]. In this work, the performance of the PSIA using different loop scheduling techniques, namely static (STATIC), self-scheduling (SS) [5], fixed size chunking (FSC) [6], guided self-scheduling (GSS) [7], and factoring (FAC) [8], is analyzed, both natively on two HPC platforms (miniHPC and Taurus), and via two SimGrid (SG) [9] simulation interfaces: MetaSimGrid (SG-MSG) and SimDag (SG-SD).

The present work makes the following contributions: (1) Offers a method for obtaining high confidence in the results obtained both, natively and via simulation; (2) Provides an experimental verification and validation of the use of the different SG interfaces to represent the application loop characteristics for the purpose of developing and testing DLS techniques in simulation;

(3) Evaluates the usefulness of using the floating point operations (FLOP) count vs. time-based measurements to represent the application characteristics in simulation; and (4) The results of the verification and the analysis strongly indicate that the absence of modern CPU and memory models in SG may adversely influence the close agreement between native and simulative experimental results, and thus, the relevance of the insights reported regarding the application performance in simulation.

The rest of this work is organized as follows. In Section 2, a review is presented of the selected DLS techniques, the SG simulation toolkit, as well as of the relevant research efforts regarding the use of simulation in performance studies of loop scheduling techniques. The description of the selected parallel application and HPC systems is given in Section 3. The proposed methodology for experimental verification and analysis of DLS is introduced and discussed in Section 4. The experimental results of executing PSIA using DLS, both natively and via simulation are presented and discussed in Section 5. The conclusions and the potential future work are outlined in Section 6.

2 Background and Related Work

This section consists of three parts. The first two parts discuss essential concepts concerning the loop scheduling techniques and the SG simulation toolkit. The last part includes a review of the relevant research efforts on the use of simulation in studying the performance of loop scheduling techniques.

Loop scheduling. There are two main categories of loop scheduling techniques: static and dynamic. The essential difference between static and dynamic loop scheduling is the time when the scheduling decisions are taken. Static scheduling techniques, such as block, cyclic, and block-cyclic [10], divide and assign the loop iterations (or tasks) across the processing elements before the loop executes. The tasks division and assignment do not change during execution. In this work, block scheduling is considered and is denoted as **STATIC**.

Dynamic loop scheduling (DLS) techniques divide and self-schedule the loop iterations during the execution of the loop. As a result, DLS techniques balance the execution of the loop iterations at the cost of increased overhead compared to the static techniques. DLS techniques consider independent

loop iterations (or independent tasks) of applications. For dependent tasks,, several loop transformations, such as loop peeling, loop fission, loop fusion, and loop unrolling can be used to eliminate loop dependencies [11]. DLS techniques can be categorized as *non-adaptive* and *adaptive*. During the application execution, the non-adaptive techniques calculate the chunk sizes based on certain parameters that can be obtained prior to the application execution. The adaptive DLS techniques exploit during execution the latest information on the state of both the application and the system to predict the next sizes of the chunks of the iterations to be executed. In highly irregular environments, the adaptive DLS techniques balance the execution of the loop iterations significantly better than the non-adaptive techniques. However, adaptive techniques may result in significant scheduling overhead compared to the non-adaptive techniques and are, therefore, recommended in cases characterized by highly imbalanced execution.

This work considers the non-adaptive DLS techniques while the adaptive techniques are planned as future work. In particular, SS [5], FSC [6], GSS [7], and FAC [8] are used herein. SS [5] is one of the simplest DLS techniques. When a processing element becomes free and available, it retrieves a single loop iteration from a central work queue. In general, SS can achieve a high load balancing between all processing elements. However, this advantage is at the cost of higher execution overhead compared to other DLS techniques. FSC [6] avoids the large overhead of single loop iterations being retrieved at a time by grouping iterations into chunks at each scheduling round. In FSC, the chunk size is fixed and plays a critical role in determining the performance of this technique. FSC needs profiling to obtain certain information such as the mean of iterations' assignment overheads and the standard deviation of loop iteration execution times. GSS [7] and FAC [8] are improvements to SS in terms of decreased scheduling overhead at decreased load balancing. GSS divides the total number of loop iterations into decreasing-sized chunks. Upon a work request, the remaining loop iterations are divided by the total number of processing elements. FAC improves GSS by scheduling the loop iterations in batches of equal-sized chunks. The initial chunk size of GSS is usually larger than the size of the initial chunk using FAC. If more time-consuming loop iterations are at the beginning of the loop, FAC balances the execution better than GSS. The chunk calculation in FAC is based on probabilistic analysis to balance the load among the processes, depending on the prior knowledge of the mean and the standard deviation of the loop iterations execution times. Due to the fact that loop characteristics are not

known apriori and typical loop characteristics that can cover many probability distributions, a practical implementation of FAC was suggested [8] that assigns half of the remaining work in a batch. This work considers this practical implementation. Compared to STATIC and SS, GSS and FAC provide better trade-offs between load balancing and scheduling overhead.

SimGrid simulation toolkit. SG [9] is a well-known toolkit based on event-based simulation. It supports the development of parallel and distributed applications in homogeneous/heterogeneous parallel and distributed environments. SG has been selected for the current work due to its reliability and its active support in the community. SG has three main interfaces: SG-MSG, SG-SD, and SG-SMPI. SG-MSG is used for the simulation of independent tasks. SG-SD supports simulation of tasks that have dependencies and represented as directed acyclic graphs (DAGs). SG-SMPI is designed to simulate applications written using the message passing interface (MPI).

Related work on DLS in simulation. Two interfaces of SG, SG-MSG and SG-SD, were used to implement various DLS techniques. For instance, eight DLS techniques were implemented using the SG-MSG interface in the literature [12]: five non-adaptive, SS, FSC [6], GSS [7], FAC [8], and weighted factoring (WF) [13], and three adaptive techniques, adaptive factoring (AF) [14], adaptive weighted factoring (AWF-B and AWF-C) [15]. The weak scalability of these DLS techniques was assessed in the presence of certain load imbalance sources (algorithmic and systemic). The robustness of the same DLS techniques implemented using SG-MSG was also studied [16]. Moreover, the resilience of these DLS techniques on a heterogeneous computing system was studied using the SG-MSG interface [17]. Another research effort used the SG-MSG interface to reproduce certain experiments of DLS techniques [18]. Therein, a successful reproduction of the past DLS experiments was presented. The results were compared to the experiments from the past to verify the implementation of the DLS techniques. A similar approach of verifying the implementation of certain DLS techniques via reproduction was proposed using the SG-SD interface [19]. The present work aims to assess the usefulness of these two SG interfaces for achieving realistic simulations of scientific applications scheduled using the DLS implemented in SG.

Related work on performance prediction. A method was introduced for predicting the performance of dynamic load balancing techniques of geophysics applications using SG [20]. In several geophysics applications, the level of the over-decomposition of the input domain and the chosen load balancing technique are key factors to achieve the desired performance. How-

ever, the evaluation process of the possible combinations of these two elements is not feasible in many cases due to its time and resource consuming process. Therefore, the SMPI interface of the SimGrid [9] simulation toolkit is extended to support the Charm++’s adaptive message passing interface (AMPI) [21]. This extension is referred to SAMPI. The main idea is to port any MPI application to the AMPI library, then for once, the AMPI application is executed to obtain a time-independent trace [22]. The SAMPI replayed the acquired trace using different load balancing techniques. The time needed for this replay is small compared to executing the real application. The accuracy of this replay is subject to the accuracy of the hybrid flow-level network models of SimGrid [23]. This method relies on porting applications to the AMPI library which may require certain developing efforts. AMPI uses the concept of processor virtualization that is argued in [24] to have no overhead. However, there is a broad range of scientific applications that directly implement DLS techniques to balance the execution load. The methodology presented in the present work addresses the same concerns as Tesser et al. [20] given that the evaluation of different load balancing techniques is time- and resource-consuming.

3 Selected Parallel Application and Parallel Computing Systems

In this section, the application of interest and the computing systems under test are introduced.

3.1 The parallel application - PSIA

The application considered in this work is an application from the computer vision domain, namely, the parallel spin-image algorithm (PSIA) [2]. The SIA is a computationally-intensive application. The core computation of the SIA is the generation of the 2D spin-images. The PSIA exploits the fact that spin-images generations are independent of each other. The size of a single spin-image is small (200 bytes) and fits in the lower level (L1) cache. Therefore, the memory subsystem has an impact on the application performance.

A pseudocode of the PSIA [4] is described in Algorithm 1. According to Algorithm 1, lines 10 and 13, the amount of computations to generate

Algorithm 1: Spin-image calculation algorithm [4]

```
1 adCalculateSpinImages (W, B, S, OP, M, spinImages, start, end)
  Inputs : W: image width, B: bin size, S: support angle,
           OP: list of oriented points, M: number of oriented points,
           spinImages: list of spin-images to be filled
2 for imageCounter = start → end do
3   P = OP[imageCounter]
4   tempSpinImage[W, W]
5   init(tempSpinImage)
6   for j = 0 → M do
7     X = OP[j]
8     npi = getNormal(P)
9     npj = getNormal(X)
10    if  $\text{acos}(np_i \cdot np_j) \leq S$  then
11       $k = \left\lceil \frac{W/2 - np_i \cdot (X - P)}{B} \right\rceil$ 
12       $l = \left\lceil \frac{\sqrt{\|X - P\|^2 - (np_i \cdot (X - P))^2}}{B} \right\rceil$ 
13      if  $0 \leq k < W$  and  $0 \leq l < W$  then
14        | tempSpinImage[k, l]++
15      end
16    end
17  end
18  add(spinImages, tempSpinImage)
19 end
```

spin-images is data-dependent and not identical over all the spin-images generated from the same object. This introduces an algorithmic source of load imbalance among the parallel processes generating the spin-images. The performance of the PSIA has been previously enhanced by using non-adaptive DLS techniques to balance the load between the parallel processes [4]. Using DLS improved the performance of the PSIA by a factor of 1.2 and 2 for homogeneous and heterogeneous computing systems. The number of spin-images generated by each process is governed by the `start` and `end` variables in

Algorithm 1, line 2. These variables represent the lower and upper bound of the indices of the images generated by a process, and the difference between them represents the chunk size calculated by the selected DLS technique.

3.2 The parallel computing systems

The miniHPC system The miniHPC is a high performance computing cluster of the Department of Mathematics and Computer Science at University of Basel, Switzerland. It consists of 26 compute nodes, a login node, and a storage node. The miniHPC cluster has a theoretical peak performance of 30 *TFLOP/s*. For the experimental studies in this work, 22 dual-socket nodes are used. Each node has two Intel Broadwell CPUs. The four remaining compute nodes have standalone Intel Xeon Phi processors. The software and hardware characteristics of the *Broadwell partition* of the miniHPC system are listed in Table 1.

The Taurus system Taurus is a Bull HPC system at the Technische Universität Dresden, Germany. It comprises 2,085 nodes with a total theoretical peak performance of 2,087 *TFLOP/s*. For the experimental studies in this work, 22 dual-socket Intel Broadwell nodes are used. The software and hardware characteristics of the *Broadwell partition* of Taurus are listed in Table 1.

3.3 Performance characterization of PSIA

To test the application performance on both HPC platforms from Table 1, the application is executed to generate 400,000 spin-images from the Ramesses object [25] using 352 processes on 22 compute nodes. The application is configured and bounded to use 16 cores on each node, to leave the rest of the cores for the operating system and other system-related processes. Each MPI process is pinned to a single core among the cores of the two available processor sockets (16 processes on two sockets, 8 per socket) to uniformly scatter the processes among the two non-uniform memory access (NUMA) domains and to avoid memory contention on a single NUMA domain. Each execution is repeated 20 times to obtain representative results. The *coefficient of variation* (c.o.v.) of the processes finishing times of the parallel loop

Table 1: The characteristics of the HPC systems

Parameter	Broadwell partition of miniHPC	Broadwell partition of Taurus
Operating system	CentOS Linux release 7.2.1511	Red Hat Enterprise Linux Server release 6.9
Job scheduler	Slurm v. 17.02.7	Slurm v. 16.05.7
MPI	Intel MPI v. 2017 update 1	Intel MPI v. 2017 update 1
File system	NFS4	NFS4
Number of nodes	22	32
Processor	Intel Xeon E5-2640 v4	Intel Xeon E5-2680 v4
Number of sockets	2	2
Cores per socket	10	14
Hyper-threading	enabled	disabled
Operating frequency	2.4 – 3.4 <i>GHz</i>	2.4 <i>GHz</i>
Peak performance per core	38.4 – 54.4 <i>GFLOP/s</i>	38.4 <i>GFLOP/s</i>
L1 cache	32 <i>KB</i> per core	32 <i>KB</i> per core
L2 cache	256 <i>KB</i> per core	256 <i>KB</i> per core
L3 cache	25 <i>MB</i> per socket	35 <i>MB</i> per socket
RAM	64 <i>GB</i> per node	64 <i>GB</i> per node
Topology	non-blocking fat tree	non-blocking fat tree
Interconnection	Intel Omni-Path	Infiniband FDR
Bandwidth	100 <i>Gbit/s</i>	54.4 <i>Gbit/s</i>
Latency	100 <i>ns</i>	700 <i>ns</i>

is used as a measure of the load imbalance [8]. The c.o.v. is defined as

$$c.o.v. = \frac{\sigma}{\mu}, \quad (1)$$

where σ is the standard deviation of the processes' parallel loop finishing times and μ is their average. The c.o.v. is unitless and is bounded by

$$0 \leq c.o.v. \leq \sqrt{(P - 1)}, \quad (2)$$

where P is the number of processes [26]. To quantify load imbalance in PSIA, the finishing times of the processes are measured and the c.o.v. of these times is calculated.

The results of executing PSIA with the two load balancing extremes, STATIC and SS, are shown in Table 2. The application execution time is denoted by T_{par} and the parallel loop execution time is denoted by T_{par}^{loop} . The results show that using SS achieved a balanced load execution with a c.o.v. of 0.003 compared to 0.022 with STATIC on miniHPC. Even though the c.o.v. values achieved by SS and STATIC are relatively small, the c.o.v. value achieved by SS is one order of magnitude lower than that of STATIC, which indicates an improved load balance. Similarly, the load balancing using SS on Taurus achieved a better performance than with STATIC.

Table 2: The performance of the PSIA using STATIC and SS scheduling. PSIA is configured to run with 352 processes to generate 400,000 spin-images. The median of 20 repetitions of each experiment is reported.

HPC system	miniHPC		Taurus	
Loop scheduling	STATIC	SS	STATIC	SS
T_{par} (s)	113.935	110.111	181.516	177.890
T_{par}^{loop} (s)	109.061	106.078	174.702	172.357
C.o.v.	0.022	0.003	0.030	0.005

4 Methodology for Experimental Verification and Analysis

In this section, the proposed approach to analyze the application performance on different platforms is presented. In addition, this section describes how to extract the application and the platform characteristics and how to represent them in simulation. The experimental verification methodology of the simulative execution of the application is described next.

4.1 Extraction of application and platform characteristics

Characterizing the behavior of a parallel application on an HPC system can be challenging as it involves the representation of two major components that contribute to its performance: (1) The application representation; and (2) The HPC system. Using the approach introduced in an earlier work [27], the representation of the computing system can be verified via a separation of the application representation by using the SG-SMPI interface. The SG-SMPI interface simulates the execution of native message passing interface (MPI) codes on a simulated computing platform. Both the native and simulative executions using SG-SMPI share the application’s native code. The difference between the native execution and the simulative SG-SMPI-based execution is the computing system component. The representation of the computing system can be verified by comparing the native and SG-SMPI simulative performance results. The SG-SMPI simulation produces a special type of text-based execution trace called time independent trace (TiT) [22]. The TiT contains a trace of the application execution as a series of computation and communication events, with their amounts specified in FLOP and bytes, respectively. The TiT is used to understand the application flow and to represent the application in the SG-MSG/SG-SD interfaces. The same computing system representation used earlier in the SG-SMPI simulation is used for the SG-MSG/SG-SD simulations. The performance results of the SG-MSG/SG-SD simulations are compared to the native execution results to verify the application representation.

The amount of work contained in each iteration of the loop is measured using PAPI [28]. The FLOP count obtained with PAPI is used to represent the amount of work in each iteration in SG-MSG/SG-SD. The core speed needs to be estimated to obtain more accurate simulation results, due to the fact that the application does not execute at the theoretical peak performance. The core speed is calculated by measuring the loop execution time in a sequential run to avoid any parallelization or communication overhead. The sum of the total number of FLOP in all iterations is divided by the measured loop execution time to estimate the core processing speed. This core speed is used in the SG `platform file` to represent the computing system core speed in processing the application loop iterations.

4.2 Employing DLS in native and simulative executions

In a recent work [4], DLS has been used to balance the load of PSIA application executing on homogeneous and heterogeneous computing platforms. The performance of the PSIA was studied using strong and weak scalability on more than 300 heterogeneous cores. Using DLS enhanced the performance of PSIA by a factor of 1.2 and 2 compared to STATIC, for homogeneous and heterogeneous platforms, respectively.

4.3 Simulation of PSIA

MSG simulation. The SG-MSG module implements a master-worker execution model. Two-sided communication is used for information exchange between master and workers. This characteristic fits perfectly the demands of studying scheduling algorithms using a central entity for coordination of the work distribution. However, in this work an approach that coordinates the distribution of work via common state information held in memory is investigated. Instead of using a central entity and two-sided communication, the access to the state information needed for scheduling decisions is achieved via one-sided communication. Therefore, the master in the SG-MSG representation of the PSIA application represents this state information. Whenever a worker is available or becomes idle, it makes a request to the master before computing the next chunk size. This request represents the remote memory access.

SimDag simulation. In SG-SD, the applications are represented as directed acyclic graph (DAG) of tasks. Dependencies can be added between tasks to represent execution precedence. Tasks can be computation tasks or communication tasks. To represent the PSIA application, each loop iteration is represented as a computation task. The amount of work in a computational task is equal to the FLOP counted by PAPI for the corresponding loop iteration. The FLOP count of all iterations is read from a file to create computational tasks in simulation that represent loop iterations. Whenever a process is available, the scheduler calculates a chunk size and allocates it to this process. A computation and a communication tasks are created at each scheduling step to represent the scheduling overhead in calculating a chunk size and the communication with the requesting process. The

amount of work contained in the tasks denoting the computation scheduling overhead is acquired with PAPI to count the FLOP in the functions that calculate and assign chunks of work in the native code. The amount of communication in the tasks denoting the scheduling overhead is equal to four bytes, which represents the communication of one integer (chunk size). The DLS implementation in SG was verified in previous work [19].

SimGrid platform files. The characteristics of the computing platform are provided to the simulator using an XML file, called the `platform file`. Each computing node is represented a host in the `platform file`. The number of cores of a host is equal to the number of cores in a node of the computing platform (20 cores per node in miniHPC and 28 cores per node in Taurus). The core speed calculation is described in Section 4.1 for both platforms. The core speeds are found to be 0.705 GFLOP/s and 0.439 GFLOP/s for miniHPC and Taurus, respectively. Both platforms use the non-blocking fat tree topology with different parameters to describe the number of fat tree levels, nodes, and links. The link bandwidth and latency are also specified in the `platform files` of the two systems. The SG-based calibration procedure [29] is used to calibrate the representations of both platforms to better adjust the network bandwidth and latency in their respective `platform files`.

4.4 Experimental verification and analysis

Three perspectives of comparisons are taken in this work to analyze the performance of the PSIA in native and simulative executions as depicted in Figure 1. Through the first perspective, the native performance of the PSIA on both HPC platforms is compared. This comparison allows the analysis of how the native platform characteristics influence the performance of the application for the DLS techniques considered. In the second perspective, the performance of the simulative executions from SG-MSG and SG-SD is compared to evaluate how the application representation (using data or task parallelism) can affect the simulative results. Through the third perspective, the native and simulative results are compared to answer: *‘How realistic are the simulations of executions of scientific applications using DLS on HPC platforms?’*. The first and the second comparisons are prerequisites for the third comparison, to understand the application characteristics (first) and to evaluate the effect of the application representation on the simulative results

(second).

To compare which interface better predicts the performance of the application of interest, the performance of the native and the simulated application from the SG-MSG simulation and the SG-SD simulation is compared. The percent error $\%E$ between native performance (T_{nat}) and simulative performance (T_{sim}) is calculated as

$$\%E = \left(1 - \frac{T_{sim}}{T_{nat}}\right) \times 100. \quad (3)$$

The percent errors between native and simulative performance are compared to answer the following additional questions: (1) *Which interface simulates the application performance with a reduced $\%E$ for the same computing system?* (2) *Which interface simulates the application performance with a minimum change in the $\%E$ in predicting the performance on the two machines (miniHPC and Taurus)?* The answers to these questions are essential to understand the accuracy of the simulation performance predictions on different systems. These results will guide future decisions regarding which interface to use for a given application and computing system and on the expected accuracy of the simulation results.

5 Results of Executing PSIA using DLS

The results of native executions of the PSIA on the two considered HPC systems, miniHPC and Taurus, are presented in this section. The simulative executions results of the PSIA using the two simulation interfaces, SG-MSG and SG-SD, are also illustrated and compared to the native executions results. A discussion on the results of the comparisons between the performance on the two HPC systems in native executions, the performance obtained from simulative executions using SG-MSG and SG-SD, and the percent error between the native and simulative executions is also included.

5.1 Native execution results

The results of the native execution of the PSIA on miniHPC and Taurus are depicted in Figure 2. The results in Figure 2c and Figure 2f show that using SS, GSS, and FAC achieved a balanced load execution on both systems. The FSC failed to achieve a balanced execution on Taurus compared to the other

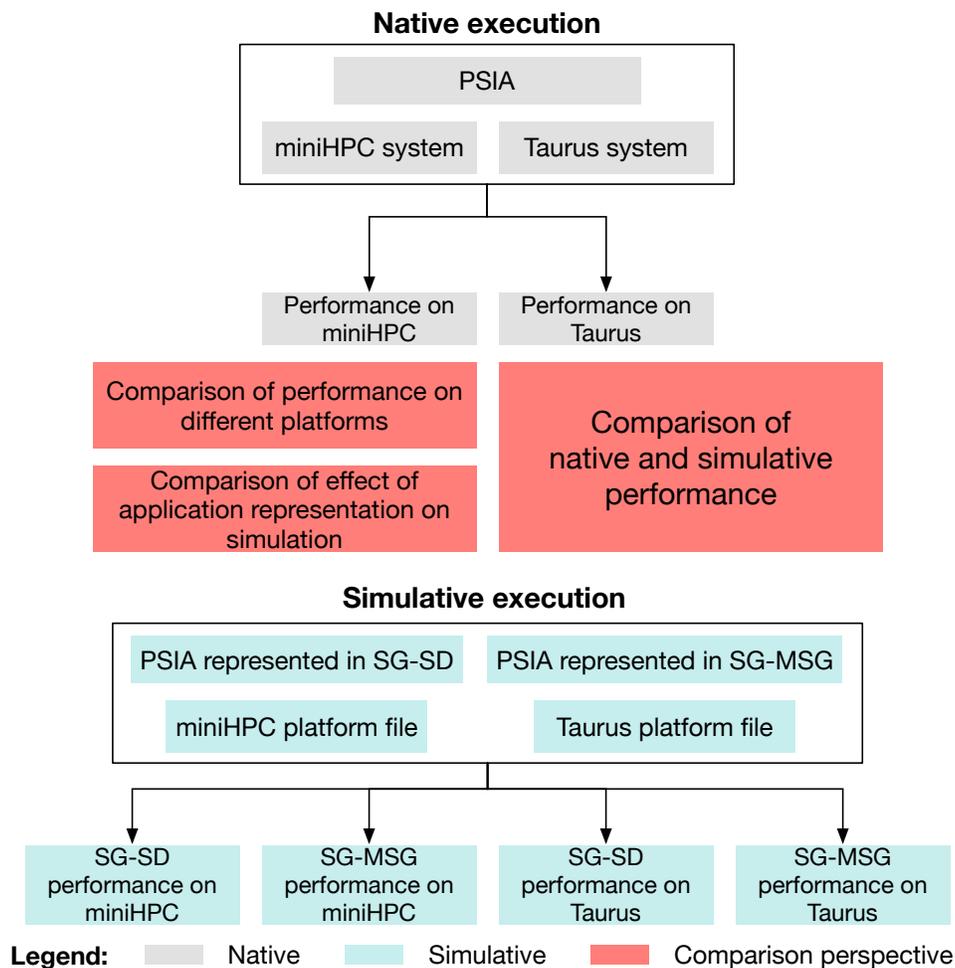
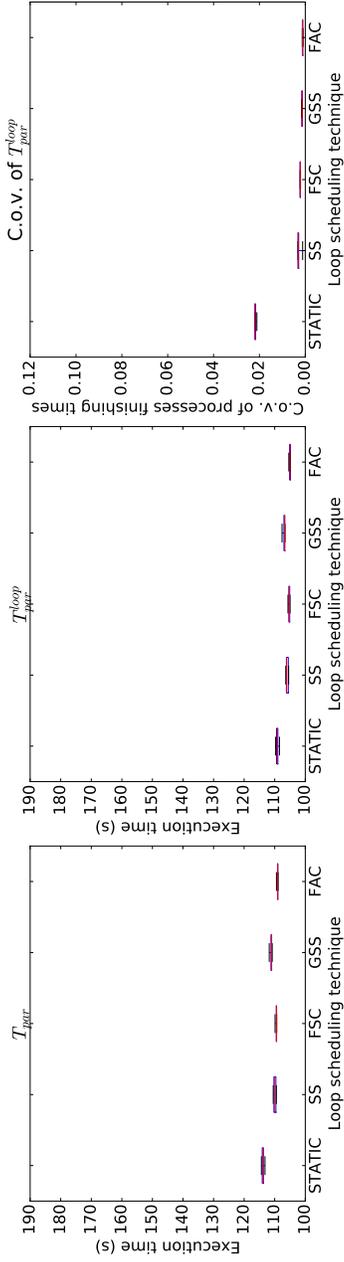
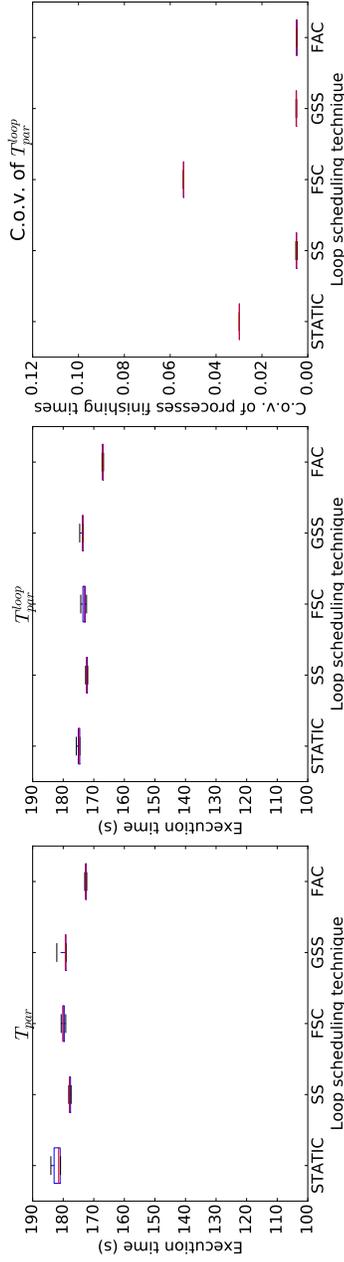


Figure 1: Experimental verification and analysis methodology.

DLS techniques. This may be due to a suboptimal estimation of the scheduling overhead, h , and the standard deviation of the loop iterations execution times, σ , that are needed by the FSC to properly calculate the chunk sizes. Even though SS, GSS, and FAC achieved a balanced load execution and very small values of c.o.v., FAC outperformed all other scheduling techniques considered in this work. This is due to the better load balancing of FAC and its lower scheduling overhead compared to SS and GSS.



(a) Application execution time (b) Parallel loop execution time (c) C.o.v. of processes parallel loop finishing times (miniHPC)



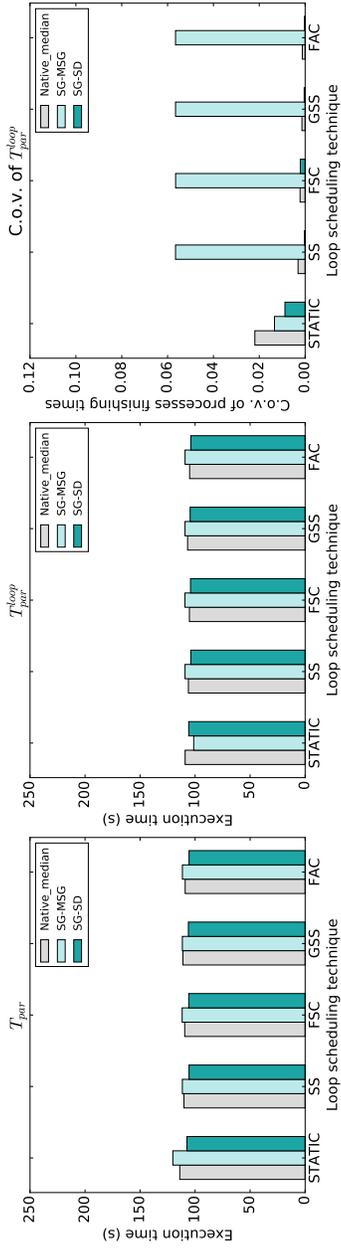
(d) Application execution time (e) Parallel loop execution time (f) C.o.v. of processes parallel loop finishing times (Taurus)

Figure 2: *Native* performance of the PSIA using DLS on miniHPC and Taurus. PSIA is configured to generate 400,000 spin-images from the Ramesses object [25]. The PSIA is executed using 352 processes, on 22 compute nodes, and 16 processes per node. The processes are distributed and binded to processor cores from the two processor sockets, 8 processes per socket. Experiments are repeated 20 times. The whiskers represent the maximum and the minimum values. The box represents the first and the third quartiles of the results data. The red line represents the median value.

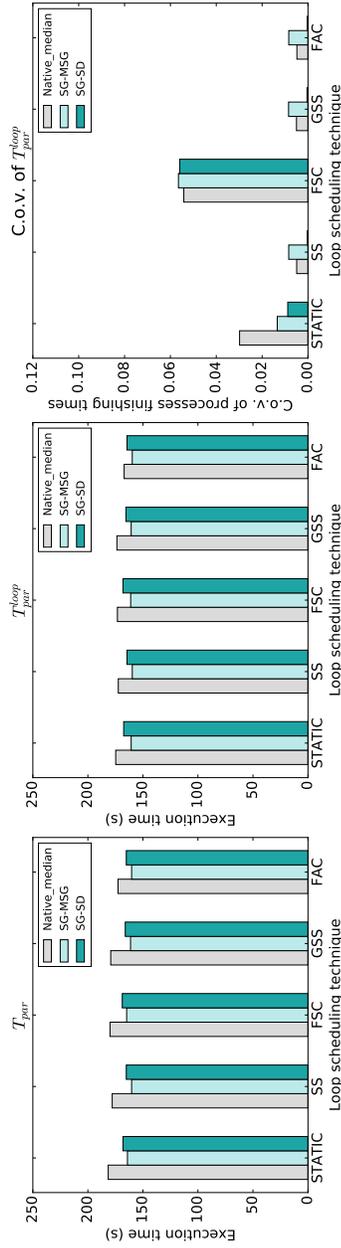
5.2 Simulative execution results

The results of the simulative executions with SG-MSG and SG-SD compared to the median of the native execution on both, miniHPC and Taurus, are shown in Figure 3. The results show that, in general, the simulative execution tends to underestimate the execution times. Both simulations using SG-MSG and SG-SD captured the large value of the c.o.v. in the case of FSC on Taurus, as can be seen in Figure 3f. The SG-MSG simulation tends to overestimate the c.o.v. values, especially on miniHPC, whereas SG-SD simulation underestimates the c.o.v. values in most cases. The results of both simulation interfaces (SG-MSG and SG-SD) tend to underestimate the execution time on Taurus for scheduling techniques that incur high overhead, such as, SS, as the scheduling overhead is not fully captured by both simulations. For example, SG-MSG only accounts for the messages to send the chunk size whereas SG-SD considers the FLOP count of the chunk calculation and the messages to send the chunk size. The SG-MSG simulation tends to overestimate the execution time on miniHPC in most cases, contrary to the SG-SD simulation, which always underestimates the execution time on miniHPC and on Taurus.

For the execution on miniHPC, both the SG-MSG and the SG-SD simulations correctly predict that STATIC results in the worst performance and that FAC outperforms all other techniques, similar to the native execution. For the execution on Taurus, both simulators correctly predict that FAC outperforms other loop scheduling techniques. Both simulations incorrectly predict that the worst performance occurs with FSC, instead of STATIC as the native execution.



(a) Application execution time (b) Parallel loop execution time (c) C.o.v. of processes parallel loop finishing times (miniHPC)



(d) Application execution time (e) Parallel loop execution time (f) C.o.v. of processes parallel loop finishing times (Taurus)

Figure 3: *Simulative* performance of the PSIA using DLS obtained using SG-MSG and SG-SD. Simulative execution results are compared with the median of the *native* executions results.

Table 3: The percent error $\%E$ between native and simulative parallel loop execution times of the PSIA on miniHPC and Taurus. The $\%E$ is calculated as Section 3.

Loop scheduling technique	SG-MSG simulation of		SG-SD simulation of	
	miniHPC	Taurus	miniHPC	Taurus
STATIC	7.14%	8.03%	2.99%	4.14%
SS	-2.89%	7.37%	2.05%	4.58%
FSC	-3.83%	6.96%	0.95%	2.91%
GSS	-2.21%	7.46%	2.16%	4.72%
FAC	-3.93%	4.42%	1.04%	1.55%

Comparing the two simulative execution results obtained with SG-MSG against results obtained with SG-SD from the perspective of the percent errors calculated between the native and simulative parallel loop execution times in Table 3, one can see that SG-MSG simulation tends to overestimate the execution time on miniHPC and underestimate the execution time on Taurus. Inspecting the percent errors in the case of the SG-SD execution, one can see that it always underestimates the execution time on both HPC systems. The median of the $\%E$ of the SG-MSG simulations is 2.89% compared to 2.05% with the SG-SD simulation for the miniHPC execution. For the execution on Taurus, the medians of the $\%E$ are 7.37% and 4.14% for SG-MSG and SG-SD, respectively. These median $\%E$ values are considered small, and the larger values of the $\%E$ can be decreased in the future through a better representation of the scheduling overhead.

5.3 Discussion

The native and simulative experiments performed in this work and the analysis of their results have revealed certain key aspects.

First, the application representation and the platform representation in simulation *can not be decoupled*. One needs to take into consideration how the application characteristics specified in the simulation interact with the represented computing platform. For example, the processor core speed is measured as described in Section 4.1 to achieve more accurate simulation results.

Second, representing the computational effort in an application using FLOP count was used as it represents the amount of work regardless of the platform computing speed as opposed to time measurements. Also, FLOP count can be accurately measured even at the fine-grained loop iterations. Measuring the execution times of loop iterations to represent the computational effort in the loop iterations was also not very successful. The time-based measurements could not capture the dynamic behavior of the application and is affected by the measurement process. The results of simulations performed using time-based measurements to represent the application can be found online [30]. The time measurement is used at the gross grain of the loop execution time to estimate the core speed as described in Section 4.1.

Third, the comparison between native and simulative execution results confirms that a close agreement thereof is limited by the absence of modern CPU and memory models. SimGrid uses a simple CPU model, where the computation time is equal to the computational effort in a loop iteration divided by the core speed [9]. The application and platform characteristics need to be aligned with the simulator and the subsystems models it offers. For example, to accurately predict the performance of a memory-bound application, a simulator that offers a precise memory model, in addition to other subsystems models, is required. Other simulators may provide more complex models and more accurate results, however, they may not be adequate for the purpose of studying scheduling techniques. Finally, the choice of the application model representation in simulation (using data or task parallelism) may affect the simulation results.

6 Conclusion and Future Work

In this work, a methodology is devised to answer the question of how realistic are the simulations of executions of scientific applications using DLS on HPC platforms, and involves the experimental verification and the performance analysis of DLS in scientific applications. The answer to this question helps to eliminate the uncertainty regarding the performance results obtained via simulation. An approach is proposed to analyze the performance of an application on different platforms. This work described how to extract the application and the platform characteristics and how to represent them in simulation. Furthermore, the experimental verification methodology of the simulative execution of the application is explained. The proposed method-

ology is employed for a computer vision application executing using four DLS techniques on two different HPC platforms, both via native and simulative (using two SimGrid interfaces) experiments. The evaluation and the analysis of the native and simulative results indicate that the accuracy of the simulative experiments is strongly influenced by the values obtained by the chosen approach used to extract the computational effort of the application (FLOP- or time-based), the choice of application model representation into simulation (data or task parallel), and the choice of HPC subsystem models available in the simulator (multi-core CPUs, memory hierarchy, and network topology). The minimum percent error achieved between native and simulative experiments was 0.95%, while the maximum was 8.03%. Further work remains for arriving at an even closer agreement between the native and simulative results via more precise representation of the application and system characteristics. The study of the effect of memory system on the performance of scientific application would lead to a more accurate application representation, thus, closer simulative results to the native results. Furthermore, the presented methodology can be employed for other computationally-intensive, scientific applications using adaptive DLS techniques, with the goal to improve their performance on real HPC systems by selecting via simulation the best suited DLS.

Acknowledgment

This work was in part supported by the Swiss National Science Foundation in the context of the “Multi-level Scheduling in Large Scale High Performance Computers” (MLS) grant number 169123 and the USA National Science Foundation under grant number NSF CGI-1034897.

7 Reproduction of This Work

Reproducibility of the execution of scientific applications on parallel and distributed computing systems is of a growing interest, underlying the trustworthiness of the experiments and the conclusions derived from experiments. In the following subsections, it is described how the source codes were compiled and how experiments were executed.

7.1 Native experiments on miniHPC and Taurus

The native PSIA code was compiled with the Intel MPI compiler version 2017 update 1 with `-O0` compiler optimization level flag. For each of the loop scheduling techniques STATIC, SS, FSC, GSS, and FAC, 20 runs were performed. Each run of the application was executed in a single job. Slurm was used for job scheduling. The application was executed using 352 tasks (processes) on 22 compute nodes, with 16 processes per node. The Slurm `exclusive` flag was used to prevent the scheduler from assigning the same nodes to other jobs simultaneously and avoid interference.

The Intel MPI `I_MPI_ASYNC_PROGRESS` flag was set to speed up the execution of the one-sided communications. The “Tag Matching Interface (TMI)” MPI fabrics library was used on the miniHPC as it provides improved performance on the Intel Omni-Path interconnection fabric, whereas the “Direct Access Programming Library (DAPL)” fabrics library was used on the Taurus. The `srun`-command was used to launch the application processes on the 22 nodes. Processes were pinned to cores from the two processor sockets using scatter strategy to balance the load among the two sockets.

7.2 Simulative experiments using SG-MSG and SG-SD

The SimGrid simulation framework version 3.16 was compiled using Intel compiler 2017 update 1. The SG-MSG and SG-SD codes were compiled using the Intel C compiler (`icc`) with `-g -Wall` compilation flags.

References

- [1] I. Banicescu and R. L. Cariño, “Addressing the stochastic nature of scientific computations via dynamic loop scheduling,” *Electronic Transactions on Numerical Analysis (ETNA)*, vol. 21, pp. 66–80, 2005.
- [2] A. Eleliemy, M. Fayze, R. Mehmood, I. Katib, and N. Aljohani, “Load-balancing on Parallel Heterogeneous Architectures: Spin-image Algorithm on CPU and MIC,” in *Proceedings of the 9th EUROSIM Congress on Modelling and Simulation*, September 2016, pp. 623–628.

- [3] A. E. Johnson and M. Hebert, “Using spin images for efficient object recognition in cluttered 3D scenes,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 5, pp. 433–449, 1999.
- [4] A. Eleliemy, A. Mohammed, and F. M. Ciorba, “Efficient Generation of Parallel Spin-images Using Dynamic Loop Scheduling,” in *Proceedings of the 19th IEEE International Conference for High Performance Computing and Communications Workshops (HPCCS 2017)*, December 2017, p. 8.
- [5] T. Peiyi and Y. Pen-Chung, “Processor Self-Scheduling for Multiple-Nested Parallel Loops,” in *Proceedings of the International Conference on Parallel Processing*, August 1986, pp. 528–535.
- [6] C. P. Kruskal and A. Weiss, “Allocating Independent Subtasks on Parallel Processors,” *IEEE Transactions on Software Engineering*, vol. SE-11, no. 10, pp. 1001–1016, 1985.
- [7] C. D. Polychronopoulos and D. J. Kuck, “Guided Self-Scheduling: A Practical Scheduling Scheme for Parallel Supercomputers,” *IEEE Transactions on Computers*, vol. 100, no. 12, pp. 1425–1439, 1987.
- [8] S. Flynn Hummel, E. Schonberg, and L. E. Flynn, “Factoring: A method for scheduling parallel loops,” *Communications of the ACM*, vol. 35, no. 8, pp. 90–101, 1992.
- [9] H. Casanova, A. Giersch, A. Legrand, M. Quinson, and F. Suter, “Versatile, Scalable, and Accurate Simulation of Distributed Applications and Platforms,” *Journal of Parallel and Distributed Computing*, vol. 74, no. 10, pp. 2899–2917, 2014.
- [10] H. Li, S. Tandri, M. Stumm, and K. C. Sevcik, “Locality and Loop Scheduling on NUMA Multiprocessors,” in *Proceedings of the International Conference on Parallel Processing*, August 1993, pp. 140–147.
- [11] D. F. Bacon, S. L. Graham, and O. J. Sharp, “Compiler Transformations for High-performance Computing,” *ACM Computing Surveys*, vol. 26, no. 4, pp. 345–420, 1994.

- [12] M. Balasubramanian, N. Sukhija, F. M. Ciorba, I. Banicescu, and S. Srivastava, “Towards the Scalability of Dynamic Loop Scheduling Techniques via Discrete Event Simulation,” in *Proceedings of the International Parallel and Distributed Processing Symposium Workshops*, May 2012, pp. 1343–1351.
- [13] S. Flynn Hummel, J. Schmidt, R. Uma, and J. Wein, “Load-sharing in Heterogeneous Systems via Weighted Factoring,” in *Proceedings of the Annual ACM Symposium on Parallel Algorithms and Architectures*. ACM, June, 1996, pp. 318–328.
- [14] I. Banicescu and Z. Liu, “Adaptive Factoring: A Dynamic Scheduling Method Tuned to the Rate of Weight Changes,” in *Proceedings of the High Performance Computing Symposium*, April 2000, pp. 122–129.
- [15] R. L. Cariño and I. Banicescu, “Dynamic Load Balancing With Adaptive Factoring Methods in Scientific Applications,” *Journal of Supercomputing*, vol. 44, no. 1, pp. 41–63, 2008.
- [16] N. Sukhija, I. Banicescu, S. Srivastava, and F. M. Ciorba, “Evaluating the Flexibility of Dynamic Loop Scheduling on Heterogeneous Systems in the Presence of Fluctuating Load Using SimGrid,” in *Proceedings of the International Parallel and Distributed Processing Symposium Workshops*, May 2013, pp. 1429–1438.
- [17] N. Sukhija, I. Banicescu, and F. M. Ciorba, “Investigating the Resilience of Dynamic Loop Scheduling in Heterogeneous Computing Systems,” in *Proceedings of the International Symposium on Parallel and Distributed Computing*, June 2015, pp. 194–203.
- [18] F. Hoffeins, F. M. Ciorba, and I. Banicescu, “Examining the Reproducibility of Using Dynamic Loop Scheduling Techniques in Scientific Applications,” in *International Parallel and Distributed Processing Symposium Workshops*, May 2017, pp. 1579–1587.
- [19] A. Mohammed, A. Eleliemy, and F. M. Ciorba, “Towards the Reproduction of Selected Dynamic Loop Scheduling Experiments Using SimGrid-SimDag,” Poster at IEEE International Conference on High Performance Computing and Communications (HPCC), 2017.

- [20] R. K. Tesser, L. M. Schnorr, A. Legrand, F. Dupros, and P. O. A. Navaux, “Using Simulation to Evaluate and Tune the Performance of Dynamic Load Balancing of an Over-decomposed Geophysics Application,” in *Proceedings of the International Conference on Parallel and Distributing Computing*, August 2017, pp. 192–205.
- [21] C. Huang, O. Lawlor, and L. V. Kale, “Adaptive MPI,” in *Proceedings of the International Workshop on Languages and Compilers for Parallel Computing*, October 2003, pp. 306–322.
- [22] F. Desprez, G. S. Markomanolis, and F. Suter, “Improving the accuracy and efficiency of time-independent trace replay,” in *Proceedings of the International High Performance Computing, Networking, Storage and Analysis*, November 2012, pp. 446–455.
- [23] P. Bedaride, A. Degomme, S. Genaud, A. Legrand, G. S. Markomanolis, M. Quinson, M. Stillwell, F. Suter, and B. Videau, “Toward Better Simulation of MPI Applications on Ethernet/TCP Networks,” in *Proceedings of the International Workshop on Performance Modeling, Benchmarking and Simulation of High Performance Computer Systems*, November 2013, pp. 158–181.
- [24] C. Huang, G. Zheng, L. Kalé, and S. Kumar, “Performance Evaluation of Adaptive MPI,” in *Proceedings of the ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*, March 2006, pp. 12–21.
- [25] K. Wang, G. Lavoué, F. Denis, A. Baskurt, and X. He, “A benchmark for 3D mesh watermarking,” in *Proceedings of the 9th IEEE International Conference on Shape Modeling and Applications*, 2010, pp. 231–235.
- [26] J.-Y. Le Boudec, *Performance evaluation of computer and communication systems*. EPFL Press, 2010.
- [27] A. Mohammed, A. Eleliemy, and F. M. Ciorba, “A Methodology for Bridging the Native and Simulated Execution of Parallel Applications,” Poster at ACM/IEEE International Conference for High Performance Computing, Networking, Storage, and Analysis, 2017.

- [28] S. Browne, J. Dongarra, N. Garner, G. Ho, and P. Mucci, “A portable programming interface for performance evaluation on modern processors,” *International Journal of High Performance Computing Applications*, vol. 14, no. 3, pp. 189–204, 2000.
- [29] SimGrid, “SimGrid Calibrations documentation,” <http://simgrid.gforge.inria.fr/contrib/smpi-calibration-doc/>, 2014, [Online; accessed 17 April 2018].
- [30] A. Mohammed, A. Eleliemy, F. M. Ciorba, F. Kasielke, and I. Banicescu, “Experimental Verification and Analysis of Dynamic Loop Scheduling in Scientific Applications,” <https://drive.switch.ch/index.php/s/xZD0gPPkTMWxZit>, February 2018, [Online; accessed 26 February 2018].