

Methodology for predicting the energy consumption of SPMD application on virtualized environments *

Javier Ballardini*, Ronal Muresano⁺, Remo Suppi⁺, Dolores Rexachs⁺ and Emilio Luque⁺

* Computer Engineering Department, National University of Comahue, Argentina

+ Computer Architecture and Operating System Department, University Autònoma of Barcelona (SPAIN)

* javier.balladini@fi.uncoma.edu.ar

+ rmuresano@caos.uab.es, {remo.suppi, dolores.rexachs, emilio.luque}@uab.es

Abstract—Over the last decade, the computing clusters have been updated in order to satisfy the increasing demand of greater computational power for running applications. However, this increasing is transformed in more system energy consumption, which results in financial, environmental and in some cases with social consequences. Hence, the ideal is to achieve an scenario that allows the system administrator to find a trade-off between time and energy-efficiency for parallel algorithms on virtualized environments. The main objective of this work is based on developing an analytical model to predict the energy consumption and energy delay product (EDP) for SPMD applications on virtual environments. The SPMD applications selected are designed through a message passing interface (MPI) library with high communication volumes, which can generate imbalance issues that affect seriously the execution time and also the energy-efficiency. Our method is composed by four phases (characterization, tile distribution model, mapping and scheduling). This method has been validated using scientific applications and we observe that the minimum Energy and EDP values are located close to the values calculated with our analytical model with an error rate between 4% and 9%.

Keywords: Performance, Energy, EDP, Prediction, Virtualization

1. Introduction

The cloud platforms are become increasingly popular together with the virtualization technology, which often used in cloud and they offer several advantages specially in efficiently managing of resources [1]. However, when these environments are used for executing parallel applications, we have to consider a set of challenges that have to be analyzed in order to improve the application efficiency (we are considering the term efficiency in two directions: the computing resources usage and the energy required for some computation). However, a large-scale computing infrastructure consumes enormous amount of electrical power which

results in financial, environmental and in some cases with social consequences [2].

The cloud computing systems have been updated in order to satisfy the increasing demands of greater computational power for running parallel applications. However, this increasing is transformed in more system energy consumption. For this reason, we have to deal with one of the most important challenges, use cloud computing systems (normally with virtualized instances) for running HPC (High Performance Computing) applications, whose resource requirements are very different from the original target applications (business and web) for which the cloud was designed [3]. HPC applications typically require low latency and high bandwidth inter-processor communication to achieve best performance [4]. These two factors affect seriously the performance especially for tightly coupled applications such as Single Program Multiple Data (SPMD).

The parallel processes of SPMD applications have to exchange information between them, and these can be located in different instances of the virtual machine, where the network is the bottleneck resource to be managed [5]. These instances can be located in different cores of the same node or other nodes of the virtualized environment. In this sense, communications are performed using diverse communication paths, which are included in the hierarchical communication architecture of the virtualized environments. So, communications exchange is one parameter to be considered in order to improve performance and efficiency in both computation resource usage and energy consumption.

Hence, the ideal target is to achieve an scenario that allows the system administrator to find a trade-off between time and energy-efficiency for SPMD applications and virtualized environments. In this sense, in a previous work [6], we have presented a method to manage the CPU inefficiency by properly selecting the number of cores to be used and the problem size needed in order to find the maximum speedup, while the efficiency is maintained over a defined threshold, for SPMD applications on a hierarchical communication architecture. However, this work does not consider the energy consumption and the use of virtual machines (and their effects on performance and energy consumption).

The energy efficiency of computing systems depends not

*This research has been supported by the MICINN Spain under contract TIN2007-64974, the MINECO (MICINN) Spain under contract TIN2011-24384

only on the hardware but also on the used CPU clock frequencies, the application type and its implementation in a specific programming model between other factors [7]. So, it is needed to consider the energy efficiency for each application that is executed on certain hardware. Thus, the main objective of this work is focused on developing an analytical model to predict the energy consumption and the energy delay product (EDP) of SPMD applications on virtual environments. The EDP is a metric capable of coupling both energy consumption and performance [8]. The novel contribution of this work is to determine the ideal number of processing element and frequency, in which the SPMD application has to be executed in order to find the minimum energy or EDP for the different frequencies of the parallel machine.

Our method starts with a characterization phase in which the application and the environment are evaluated in order to obtain some parameters, which are later introduced in the analytical model of the second phase. The tile distribution model phase predict the number of processing elements, supertile size, application execution time, energy consumption and EDP. Mapping phase assign tiles to a set of processing elements according to the values obtained through the analytical model. Finally, the scheduling phase manages the overlapping strategy between computation and communication in order to avoid inefficiency.

This paper is structured as follows. Section 2 presents the impact of virtual environment on SPMD applications. Section 3 exposes the method for predicting the energy consumption. Section 4 illustrates the experimental validation. Finally, section 5 draw the main conclusions.

2. SPMD applications over virtualized environments

The SPMD applications used have to accomplish the following characteristics: static, where the communication pattern is known prior to the execution of the algorithm, local, where applications do not have collective communications, grid application, and regular, that is, that communications are repeated for several iterations. In this sense, there are some benchmarks that have these characteristics, for example the NAS parallel benchmarks in the CG and BT algorithms [9], and some real applications such as: heat transfer simulation, Laplace equation, applications focus on fluid dynamics, application of finite differences, etc.

When these SPMD applications are executed on a hierarchical communication architecture, they are strongly affected by the latency and bandwidth of different communication links [10]. This problem exacerbated when the applications are executed on virtualized environments because the communications need to go through other protocol stacks that penalize the latency and bandwidth [11]. So, the analysis of the communication delay, for these applications and

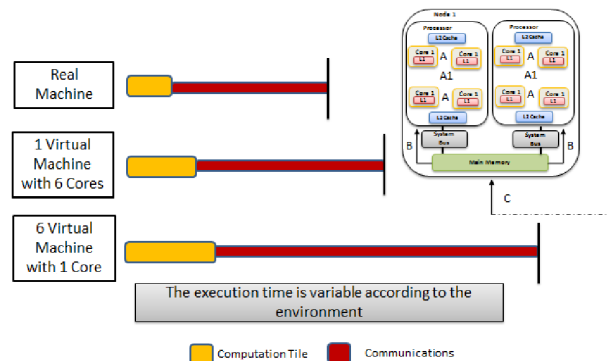


Fig. 1: SPMD application on different environments

execution environment, is a critical issue.

An example of the communication delays can be evidenced in figure 1, where the computation and communication are affected due to virtualization. However, these idle times allows us to establish strategies in order to organize how SPMD tiles could be distributed on different environment configuration with the aim of managing these communications inefficiency. These variations are a limiting factor to improve application performance, due to the latency of the slower link, which determines when iteration has been completed (Fig. 1).

To manage this communication issues, the tiles comprising the problem of the analyzed SPMD application are grouped in a number of SuperTile (ST). Each ST will be assigned to one processing element, and each processing element will only process one ST (in each iteration). The problem of finding the optimal ST size is formulated as an analytical problem, in which the ratio between computation and communication of the tile has to be founded with the objective of searching the relationship between efficiency and speedup [10]. The improvement in the execution time can allow us to minimize one of the influencing factor in energy (time). Then, the ST has been defined as a group of tiles in the form of a grid of $K \times K$ tiles, which have to be assigned to each core with the aim of maintaining an ideal

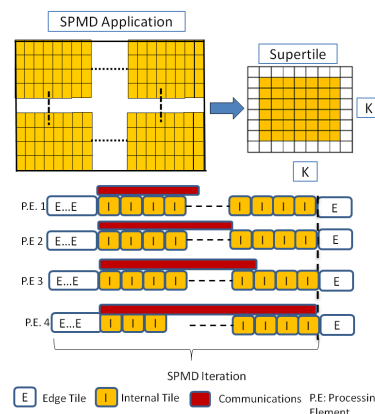


Fig. 2: Supertile creation for improving the efficiency

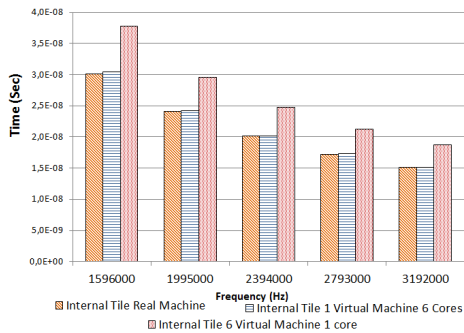


Fig. 3: Impact of CPU frequency changes in the execution time of tiles

relationship between efficiency and speedup.

The ST is composed by two type of tiles: internal and edge tiles. This is done with the objective of creating an overlapping strategy that minimize the communication effects in the execution time, an example can be evidenced in figure 2, where the ST is composed of a set of tiles that hide all the communication effects by overlapping computation and communication. However, the computation time of the ST can present alterations at different CPU frequencies. In this sense, the figure 3 shows evidence of how tiles have a huge variation when we modify the CPU frequency in all scenarios with real and virtualized machine. Thus, the tile computation is another variable that we have to consider inside the analytical model.

3. Methodology for predicting energy consumption

This methodology is focused on managing the different communication latencies and bandwidths with the objective of finding a trade-offs between time and energy-efficiency for SPMD applications on virtualized environments. This process is realized through four phases: a characterization, a tile distribution model, a mapping strategy and a scheduling policy. These phases allow us to handle the latencies and the imbalances created due to the different communication paths. Also, these phases permit us to predict the execution time, energy consumption and the ideal processing elements used to execute the application with minimum EDP on a real or virtualized machine. The method works by managing the hierarchical communication architecture of both real and virtualized environments.

To begin the analysis we have to consider that energy depends on two main factors: power and time. For this reason, our method analyzes diverse characteristics of the application and environment in both power and time. Then, a set of variables are collected to our analytical model in order to obtain the prediction for both execution time and energy consumption in the tile distribution phase. Next, the mapping phase allocates the set of tiles (ST) among the cores, which are calculated with the model defined in the tile distribution

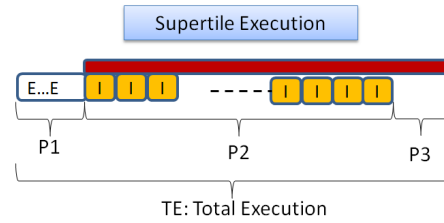


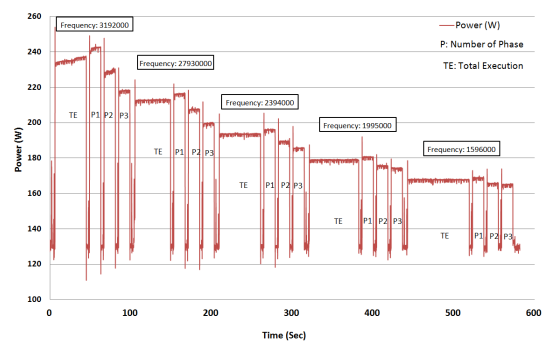
Fig. 4: Power Analysis of ST execution

phase. Finally, the scheduling phase has two functions, one of them is to assign tile priorities and the other is to control the overlapping process. Later, once the methodology is shown, we evaluate the obtained performance results.

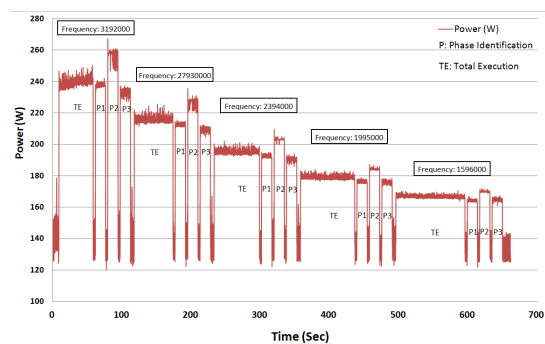
3.1 Characterization Phase

The objective of this phase is to gather the necessary parameters of both SPMD application and environment. These characterization parameters are classified in two groups: power and application analysis.

Power Analysis: To characterize the power, we have to divide the SPMD application in phases with the aim of obtaining a precise measure of the events that are included in the execution.



(a) Real Machine



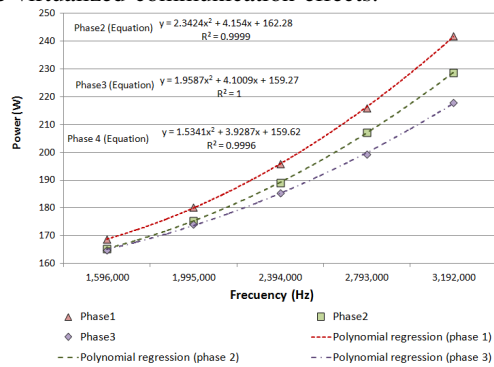
(b) One virtual machine per core

Fig. 5: Power Analysis

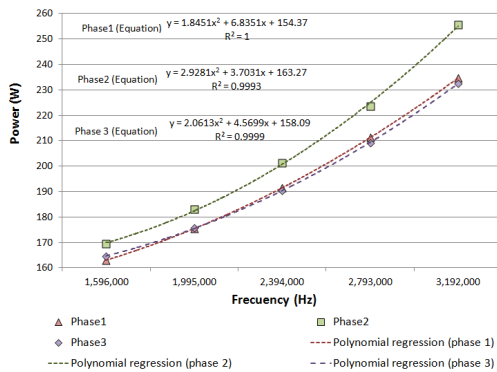
An example of this division is shown in figure 4, where the application has been divided in three parts. The first one represent the edge computation, follow the phase 2 include the overlapping area where is executed the internal

computation and the communications, and the last phase 3 is responsible of measure when communication is longer than internal computation. The knowledge of per phase power rather than an average application power allows us to improve the model accuracy.

Two examples of this characterization are illustrated in figures 5(a) and 5(b), where has been analyzed a heat transfer application for both real and virtualized environment. As can be evidenced, the phases have variations depending on the environment used. For example, the phase 2 of the virtualized environment has an power increment of around 10% over the other phases (Fig. 5(b)) while in real environment (Fig. 5(a)) this effect does not occur. This could be motivated by the virtualized communication effects.



(a) Real Machine



(b) One virtual machine per core

Fig. 6: Power Regression Analysis

Also, we can observe how is the increment in power (and execution time) when we increase the CPU frequencies. Power variations can modify the energy consumption for a specific scenario. So, we have to find the power equation for an specific application in function of frequency. Then, once the phases are characterized, we have to apply a regression analysis with the objective of finding the equation that represents the power in function of CPU frequency.

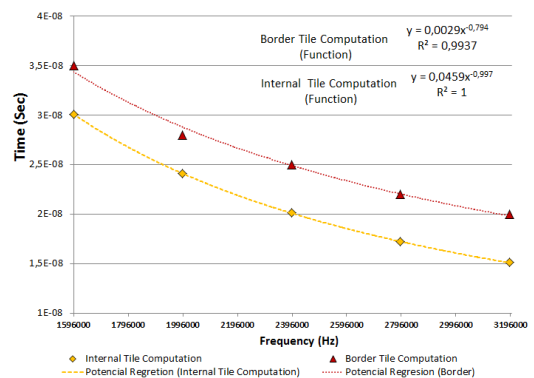
The figures 6(a) and 6(b) illustrate the power behavior for the different frequencies. Also, we apply an polynomial regression and we obtain a polynomial of degree 2 where the error is less than 0,01% for the worst case. All these equation

obtain using the regression will be used in the model for predicting the energy consumption.

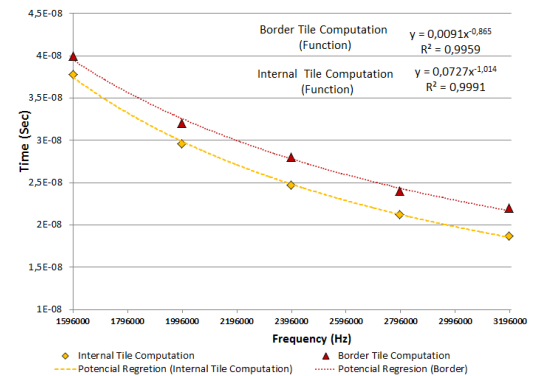
Application analysis: The main idea of this analysis is to find a nearest relationship between the machine and the SPMD application. The parameters determined allow us to establish the communication and computational ratio time of a tile inside of the hierarchical communication architecture of real and virtualized environment. This relationship will be defined as $\lambda(p)(w)$, where p determine the link where the communication of one tile to another neighboring tile has been performed and w describes the direction of the communication processes (e.g. up, right, left or down in a four communications pattern). This ratio is calculated with equation 1, where $CommT(p)(w)$ determines the time of communicating a tile for a specific p link and the $Cptint(freq)$ is the value of computing one tile on a processing element using a determined CPU frequency.

$$\lambda(p)(w) = CommT(p)(w)/Cptint(freq) \quad (1)$$

However, as was observed in figure 3, the tile computation time is affected due to CPU frequencies. So, if we decrease the CPU frequency, the tiles can need more time to be computed and this affect the value of the ratio communication-computation of all scenarios (real and virtualized).



(a) Real Machine



(b) One virtual machine per core

Fig. 7: Tile computation characterization

Hence similarly to the power analysis, the internal and edge tile computation time depend of CPU frequencies (C_{ptint} and C_{ptedge} respectively). Thus, we have to apply a regression analysis in order to find the equation in function of the frequencies for both internal and edge tile behavior. Figures 7(a) and 7(b) show the behavior of tiles computation for both real and virtualized environments. In this case, the potential regressions present the best fit to the tiles computation time (for each CPU frequency), whose equations are shown inside the figures.

3.2 Tile Distribution Model Phase

The analytical model for predicting the energy delay product (EDP) starts using the equation 2 with the objective of determining the ideal scenario which finds the trade-off between execution time $Time(freq)$ and energy consumption $E(freq)$ under a frequency $freq$.

$$EDP = Time(freq) * E(freq) \quad (2)$$

The equation 3 defines $Time(freq)$ that represents the execution time of a SPMD application using the overlapping strategy for a specific frequency $freq$. This equation first calculate the edge tile computation time $EdgCpT$ and then we add the maximum value between internal tile computation time $IntCpT$ and edge tile communication time $CommT$. This process obtain the time used to compute each iteration of the SPMD algorithm, and these values are added to get the execution time for all iterations ite . A first approach of this model can be found in [10].

$$Time(freq) = \sum_{i=1}^{ite} (EdgCpT + \text{Max} \left(\begin{matrix} IntCpT \\ CommT \end{matrix} \right)) \quad (3)$$

$$EdgCpT = (K^n - (K - 2)^n) * C_{ptedge}(freq) \quad (4)$$

$$IntCpT = (K - 2)^n * C_{ptint}(freq) \quad (5)$$

$$CommT = K^{(n-1)} * \text{Max}(CommT_{(p)(w)}) \quad (6)$$

From the foregoing, the next step is to find the value of K that determines the ideal size of ST with K^n tiles, where n is the application dimension (e.g 1, 2, 3, etc.). K is defined by considering the overlapping strategy between internal computation and edge communication such that CPU efficiency will be maintained over a threshold $effic$. The equation 7 shows how both values, internal computation time and edge communication time, can be equalized with the aim of finding the value of K . Using the equation 1 we can equalize the equation 7 in function of $C_{ptint}(freq)$. Having both internal computation time and edge communication time in function of $C_{ptint}(freq)$, the next step is to find the value of K by replacing all the values in equation 7. Depending on the dimension of the SPMD application, we can obtain a quadratic equation, cubic equation, etc.

$$K(freq)^{(n-1)} * \text{max}(\lambda_{(p)(w)} * C_{ptint}(freq)) = ((K(freq) - 2)^n / effic) * C_{ptint}(freq) \quad (7)$$

At this point we have calculated the ideal value of $K(freq)$ that allow us to obtain the minimum execution time $Time(freq)$ while the CPU efficiency is maintained over the threshold $effic$. The next step is to predict the energy consumption of whole system $E(freq)$, defined in equation 8. $E(freq)$ is the sum of the energy consumption produced by the execution of each iteration $iter$ of the application. The energy consumption of an iteration is calculated from the energy consumption produced by a core when executes (part of) the application $EC(freq)$, multiplied by the number of cores $Ncores(freq)$ used to the execution.

$$E(freq) = \sum_{i=1}^{ite} (EC(freq) * Ncores(freq)) \quad (8)$$

Equation 9 represents a simple manner to calculate the ideal number of cores $Ncores(freq)$, where the problem size (M^n) is divided by the size of the ideal ST (K^n).

$$Ncores(freq) = M^n / K^n \quad (9)$$

The energy consumption by core can be calculated using the equation 11, considering the average power of each phase (1 to 3 in our case of study) and the time that the application spent in the phases. As the average power of each phase was obtained (in the characterization phase) for the entire computing node, we define $Pw(i)(freq)$ in equation 10 (i identifies the phase) to calculate the power demanded by only one core. So, this new equation divide the power of the entire node between the number of cores $CoresByNode$.

$$Pw(i)(freq) = Phase(i)(freq) / CoresByNode \quad (10)$$

$$EC(freq) = Pw(1)(freq) * EdgCpT + \text{if}(IntCpT \leq CommT)$$

$$Pw(2) * IntCpT + Pw(3) * (CommT - IntCpT) \text{ else } Pw(2) * CommT + Pw(1) * (IntCpT - CommT) \quad (11)$$

The energy consumption by core (eq. 11) considers all the power phases analyzed in the characterization phase, where the first step is evaluate the energy consumed by the edge computation (phase 1) and then next step is analyze the overlapping strategy between internal computation (eq. 4) and edge communication (eq. 6). However, the overlapping can present two scenarios. The first is when the internal computation is lower than or equal to edge communication. In this case, we add to the edge computation the energy consumption of the part where the communication is overlapped with computation of internal tiles (phase 2), and then we add the energy consumption of the remaining communication that occur without internal computation (phase 3).

The second scenario is when the internal computation is longer than edge communication. In this case, the energy consumption correspond to the part where computation of internal tiles overlaps with edge communication (phase 2) and the part with computation of internal tiles and without communication (equivalent to phase 1).

3.3 Mapping phase

The main purpose of this phase is to apply a distribution of STs in cores. The ST assignments are made applying a core affinity which allocates the set of tiles according to the policy of minimizing the communications delays. This core affinity permits us to identify where the processes have to be allocated and how STs are assigned to each core. However, the ST assignments should maintain the initial considered allocation used in the characterization phase.

This phase is divided in three key points. The first point performs a logical processes distribution of the MPI processes. The second function is to apply the core affinity, and the last one is the division and distribution of the STs. The mapping has to divide the tiles in order to create the ST considering the value of K obtained by the analytical model. It is important to understand that an incorrect distribution of the tiles can generate different application behaviors.

3.4 Scheduling phase

The main function of the scheduling phase is to assign a execution priority assignment to each tile with the aim of applying the overlapping strategy. The scheduling establishes the highest priority to the computation of tiles which have communications through slower paths, and slower priority to internal tile computation. This phase performs an overlapping strategy, which is the main key of our method.

4. Experimental Validation

To validate our method we have used a DELL node with a Intel Xeon Processor W3670 3.2 GHz, 24 GB of main memory and 12 MB of cache memory. This machine has 13 frequencies available from 3.19 Ghz to 1.59 GHZ. We use the KVM virtualization environment and the MPI library Open MPI version 1.6.4. The scenarios defined to test our method are: (A) real machine in which we execute the application without using any virtualization, (B) one virtual machine that uses all computing cores (e.g. X-large instance in Amazon EC2 or BonFIRE cloud) and (C) a virtual machine per core (e.g. an small instance in Amazon EC2 or BonFIRE cloud). Furthermore, we have tested with different SPMD application that accomplish the characteristics defined before of (regular, local and static). Specifically for this work, we have evaluated a heat transfer simulation with the aim of showing the efficacy of our method for both time and energy prediction.

The first step is to characterize the application in order to obtain the ratio computation–communication (eq. 1),

where the regression analysis is used for the tile computation characterization as was observed in figures 7(a) and 7(b). Similarly the power is characterized using regression analysis. Part of this characterization is illustrated in table 1, where we can observe the characterization values for a frequency of 3,19 GHz. This process was done for all frequencies and scenarios. Then, we proceed to apply the analytical model. Table 2 summarize the analytical results obtained using our model for a defined problem size of 500x500 tiles.

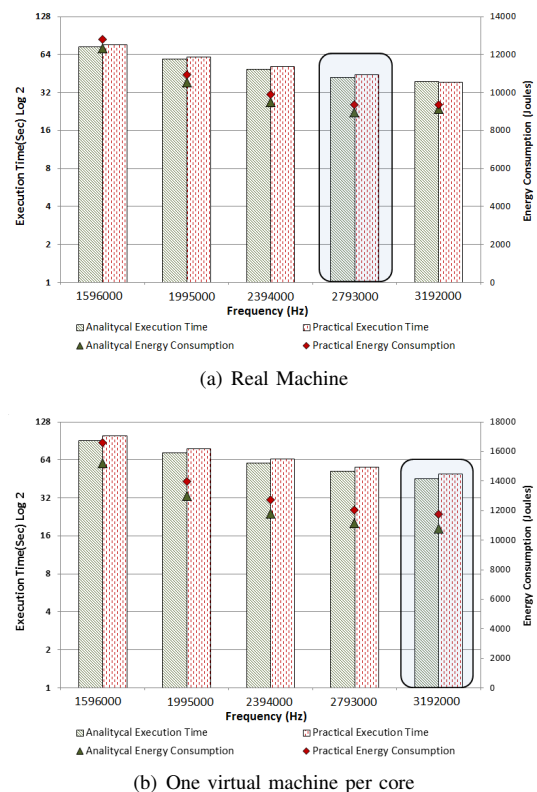


Fig. 8: Energy and time prediction

These results show that lower energy consumption is located for the scenarios (A) and (B) using a frequency of 2.79 GHz. However, for scenario (C) the lower energy consumption is at 3,19 GHz. These results are also evidenced in figures 8(a) and 8(b), where we compare practical and analytical results. The error rate is around 4% for first (A) scenario (fig. 8(a)) and 9% for the worst case using one virtual machine per core (fig. 8(b)). The results evidence that using the same program with the same workload, the frequency must be changed depending on the scenario executed (real or virtualized). Also, the results show the effectiveness of the prediction for different available frequencies.

Table 1: Characterization at a frequency of 3,19 GHz

Scenario	Cptint	CommT	ratio	Power Av
A	$1.5E - 8Sec$	$3.18E - 6Sec$	211	235.20 W
B	$2.05E - 8Sec$	$3.6E - 6Sec$	243	235.29 W
C	$2.22E - 8Sec$	$6.75E - 6Sec$	304	238.03 W

Table 2: Analytical values for different frequencies

Scenario	Freq	Time (Sec)	Energy (Joules)	EDP
A	3.19Ghz	38.8	9145.9	3.56E+5
A	2.79Ghz	42.1	8954.3	3.77E+5
A	1.59Ghz	73.5	12338.3	9.08E+5
B	3.19Ghz	42.8	10086.0	4.32E+5
B	2.79Ghz	44.8	9488.3	4.33E+5
B	1.59Ghz	74.1	12460.6	9.24E+5
C	3.19Ghz	45.2	10775.7	4.88E+5
C	2.79Ghz	51.8	11139.3	5.77E+5
C	1.59Ghz	91.4	15217.9	1.39E+6

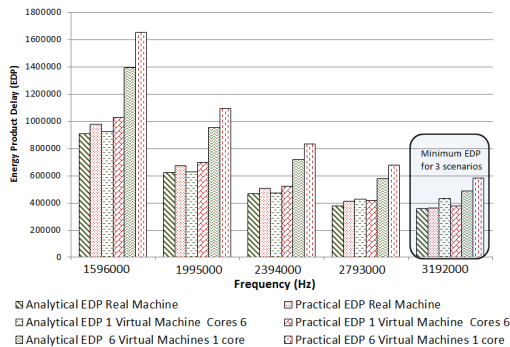


Fig. 9: EDP analysis

To analyze the EDP metric, we have to consider that time has more influence than energy (eq. 2). In this sense, we have evaluated the three scenarios and the results are illustrated in figure 9. As can be shown in figure 9, the minimum EDP value for the three scenarios and available frequencies are located in the highest frequency of 3.19 GHz. These values can vary depending on the machine architecture and the values obtained in the characterization phase.

Finally, figure 10 shows the overhead added by the virtualization. As can be detailed, the overhead added depends on the environment configuration. For example, when we set up a virtual machine using a set of cores, the overhead is lower than 2%, and when we use one virtual machine per core the overhead is around 30% for all available frequencies.

5. Conclusion

This paper has presented a novel methodology based on characterization, tile distribution model, mapping and scheduling. These phases allow us to find through an analytical model the optimal size of the SuperTile (group of tiles assigned to each cores) and the number of processing elements needed in order to find the minimum energy delay product. Also, our model allows us to predict the energy consumption and the minimum execution time for an SPMD application. This model is focused on managing the hierarchical communication architecture in order to hide the communication effects as was evidenced.

Experimental evaluation makes clear that to achieve the best scenario for reducing the energy consumption in SPMD applications, we have to manage properly the inefficiencies generated by communications. Thus, our method evaluates the environment through the characterization phase in order

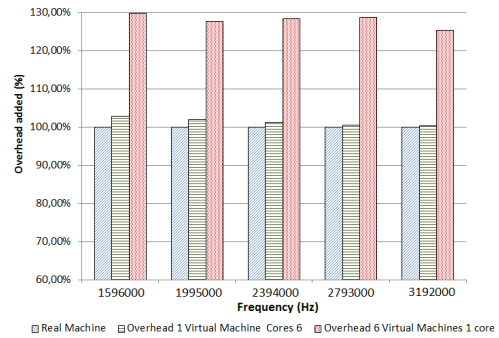


Fig. 10: Overhead of the virtualization

to apply with real values of the architecture. The model can predict with a good level of accuracy the energy consumption and execution time as was shown by the experimental results (less than 9% for a virtualized environment and lower than 4% for real machine). The mapping distribute the set of tiles for each core according to the different communication delays present in the machine architecture, and the scheduling allows us to perform the overlapping method. Finally, finding a trade-off between execution time and energy consumption allow us to improve the manner of administering the virtualized environments.

References

- [1] Q. Huang, F. Gao, R. Wang, and Z. Qi, "Power consumption of virtual machine live migration in clouds," in *Third Int Conf on Comm and Mobile Computing (CMC)*, 2011, pp. 122–125.
- [2] A. Beloglazov and R. Buyya, "Energy efficient resource management in virtualized cloud data centers," in *10th IEEE/ACM Int Conf on Cluster, Cloud and Grid Computing (CCGrid)*, 2010, pp. 826–831.
- [3] J. Ekanayake and G. Fox, "High performance parallel computing with clouds and cloud technologies," *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering*, vol. 34, pp. 20–38, 2010.
- [4] G. Mercier and J. Clet-Ortega, "Towards an efficient process placement policy for mpi applications in multicore environments," vol. 5759, pp. 104–115, 2009.
- [5] A. Iosup, S. Ostermann, N. Yigitbasi, R. Prodan, T. Fahringer, and D. Epema, "Performance analysis of cloud computing services for many-tasks scientific computing," *IEEE Trans. Parallel Distrib. Syst.*, vol. 22, no. 6, pp. 931–945, June 2011.
- [6] R. Muresano, D. Rexachs, and E. Luque, "Methodology for efficient execution of spmd applications on multicore environments," *10th IEEE/ACM Int Conf on Cluster, Cloud and Grid Comp, CCGrid 2010, Australia*, pp. 185–195, 2010.
- [7] J. Balladini, R. Suppi, D. Rexachs, and E. Luque, "Impact of parallel programming models and cpus clock frequency on energy consumption of hpc systems," pp. 16–21, 2011.
- [8] R. Gonzalez and M. Horowitz, "Energy dissipation in general purpose microprocessors," *Solid-State Circuits, IEEE Journal of*, vol. 31, no. 9, pp. 1277–1284, 1996.
- [9] R. F. V. der Wijngaart Ana Haoqiang Jin, "Nas parallel benchmarks, multi-zone versions," NASA Advanced Supercomputing (NAS) Division, Tech. Rep., 2003.
- [10] R. Muresano, D. Rexachs, and E. Luque, "Combining scalability and efficiency for spmd applications on multicore clusters," *The 2011 International Conference on Parallel and Distributed Processing Techniques and Applications, Las Vegas, USA*, pp. 43–49, 2011.
- [11] F. Schatz, S. Koschnicke, N. Paulsen, C. Starke, and M. Schimmler, "Mpi performance analysis of amazon ec2 cloud services for high performance computing," *Advances in Computing and Communications*, vol. 190, pp. 371–381, 2011.