Grid'5000 energy-aware experiments with DVFS

Tom Guérout, Georges da Costa, Thierry Monteil, Mihai Alexandru

▶ **To cite this version:**

Tom Guérout, Georges da Costa, Thierry Monteil, Mihai Alexandru. Grid'5000 energy-aware experiments with DVFS. grid'5000 school, Dec 2012, Nantes, France. 7p. hal-01228321

HAL Id: hal-01228321
https://hal.science/hal-01228321

Submitted on 12 Nov 2015

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

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Grid'5000 energy-aware experiments with DVFS

Tom Guérout[123], Georges Da Costa[123], Thierry Monteil[13], and Mihai Alexandru[13]

[1] CNRS; LAAS; 7 avenue du Colonel Roche, F-31077 Toulouse, France,
[2] IRIT; 118 route de Narbonne, 31000 Toulouse, France,
[3] Université de Toulouse; UPS, INSA, INP, ISAE; LAAS; F-31077 Toulouse, France
tguerout@laas.fr

Abstract. In recent years, much research has been conducted in the area of energy efficiency in distributed systems. To analyze, understand and improve their behaviour, simulators provide useful tools, to achieve energy-aware simulation like DVFS (Dynamic Voltage and Frequency Scaling). This paper presents current work on Grid'5000 to deploy a specific distributed electromagnetic application called TLM (Transmission Line Matrix), using DVFS and power measurements. The aim is to launch different set of experiments using different DVFS configurations, and then compare simulations and real experiments results.

Keywords: power consumption, dvfs, simulations

1 Introduction

Nowadays, the increasing use of datacenters makes the analysis of energy consumption more and more important. Some metrics to evaluate their efficiency are known (PUE, ERE) [1] and much research are ongoing to find new ways to reduce the energy. This leads to improvement and/or creation of new algorithms that require extensive testing phases to validate their effectiveness.

To compare real energy-aware experiments, some simulators allow reducing the consumption of machines/cpu, while others give the possibility to measure the total energy consumption of a simulation, but it seems that none of them provide all tools required to achieve a complete energy-aware simulator. For these experiments, the simulator CloudSim has been improved with DVFS support to be able to run simulations with similar behaviour as Grid'5000 experiments.

After the DVFS implementation into CloudSim is validated with a simple test application, the next step is to compare Grid'5000 experiments of a real distributed application, the TLM, using DVFS and power measurements with CloudSim simulation results.

This paper details specifically a TLM experiment. It is organized as follows. First section 2.1 explains how to find the optimal frequency. Section 2.2 shows how experiments have been conducted in term of energy and how to configure DVFS. Results of these experiments are analyzed in Section 2.3, and the current work on hard disk throughput is explained in section 3. Finally, section 4 concludes the paper.

2 Deployment of TLM Electromagnetic application

TLM implements a numerical method for the electromagnetic simulation which fills the environment of the electromagnetic field propagation with a network of transmission lines. This propagating field model inside a given medium becomes possible thanks to the equivalence that exists between the electric and magnetic fields and voltages and currents in a transmission line network. A detailed presentation of TLM method, also in a three-dimensional approach, can be found in [2].

The parallel resolution process of this method on a grid is based on a division of the structure along the three axes. Volumes obtained are assimilated to tasks executed on each machine, CPU, or CPU core, sending data over MPI library (Message Passing Interface).

The aim of these experiments is to compare the power consumption using different DVFS configurations (i.e different frequencies). After referencing power consumption of Reims nodes at 0% and 100% of use, the next step is to determine the theoretical optimal fixed frequency.

2.1 CPLEX Solver

IBM CPLEX solver was used to compute the optimal fixed frequency, depending on TLM input parameters (size of the structure and number of iterations), network bandwidth and latency and power values of machines. The problem has been written in JAVA and described as follows:

H : Current node used

T_{net} and T_{cpu} : Network and CPU time

$F = \{f_1, f_2, ..., f_{n-1}, f_n\}$: Available frequencies on node H, with $f_{n-1} < f_n$.

$l^H_{cpu}(f_i)$ et $d^H_{cpu}(f_i)$: Power of node H at 0% and 100% of CPU load.

z^f_i : Binary variable, equal to 1 if f_i is used, 0 otherwise.

The prediction model of CPU time and network TLM presented in [3] was used:

$T_{cpu} = C_1 + n_l n_x n_y n_z C_2$, $T_{net} = \left(L + \frac{n_x n_y}{D}\right) * 4 n_l$, with n_x, n_y,n_z dimensions of TLM environment, n_l the number of iterations, C_1 et C_2 constants estimated by past experiments, L et D are latency and network bandwidth.

The expression of the energy E to be minimized is:

$$\begin{aligned} E = \quad & T_{net} * \left[l^H_{cpu}(f_1) * z^f_a + ... + l^H_{cpu}(f_n) * z^f_n \right] \\ & + (T_{cpu} * f_n) \left[\frac{d^H_{cpu}(f_1) * z^f_a}{f_1} + ... + \frac{d^H_{cpu}(f_n) * z^f_n}{f_n} \right] \end{aligned}$$

with constraint : $\sum_{i=1}^{n} z^f_i = 1$, which allows to use only one frequency.

The energy equation E takes into account available host frequencies and power consumption values. Reims Grid'5000 values are in Table 1.

With these available frequencies and power consumption values at *Idle* and *Full* state on all 24 cores of a node, CPLEX solver is able to find the optimal

Table 1. Frequencies available and power consumption values (at *idle* and *full* state) for each frequency on Grid'5000 Reims site

Reims site characteristics						
Available Frequencies (GHz)		0.8	1.0	1.2	1.5	1.7
Power Consumption (W)	***Idle***	140	146	153	159	167
	Full	228	238	249	260	272

frequency that minimize E for different ratio values $\frac{T_{cpu}}{T_{net}}$. For example, one instance of this problem with the ratio $\frac{T_{cpu}}{T_{net}} = 1$ looks like :

```
Minimize
 obj: 52.0406573169004 z_a + 45.8853564547206 z_b + 42.1450984558719 z_c
      + 37.8042459957691 z_d + 36.5833333333333 z_e
Subject To
 c1: z_a + z_b + z_c + z_d + z_e  = 1
Bounds
 0 <= z_a <= 1
 0 <= z_b <= 1
 0 <= z_c <= 1
 0 <= z_d <= 1
 0 <= z_e <= 1
Binaries
 z_a  z_b  z_c  z_d  z_e
End
```

Once all energy consumption values computed (for each frequencies) and the optimal frequencies for each ratio $\frac{T_{cpu}}{T_{net}}$ found by the solver, it's interesting to merge these two results in one figure (shown in figure 1). The TLM configuration used for these experiments give a ratio of about 0.16, so the theoretical optimal frequency that will be use is 1.5GHz.

2.2 Experiments on Grid'5000

These experiments were conducted on the Grid'5000 Reims site, whose configuration is: *HP ProLiant DL165 G7 CPU: AMD Opteron 6164 HE 1.7GHz, 12 MB L3 cache, 44 nodes with 2 CPUs to 12 cores per CPU (1056 cores).* TLM was deployed on 2 nodes used at 100 %, indeed using 48 cores, with one task TLM on each core.

Input parameters are: TLM parameters : n_x=172 n_y=90, n_z=12 and n_l=26000. Throughput and latency measured (with a simple MPI application) in Reims are : D=700Mbit/s and L=4*10^{-5}s.

Deployment of TLM on Grid'5000 Reims site has been done in a Kadeploy environment. Once the image has been deployed on all nodes, the first step is to configure the DVFS. In this experiment, 3 modes has been used :

– Maximum Frequency (1.7GHz) in Performance mode.

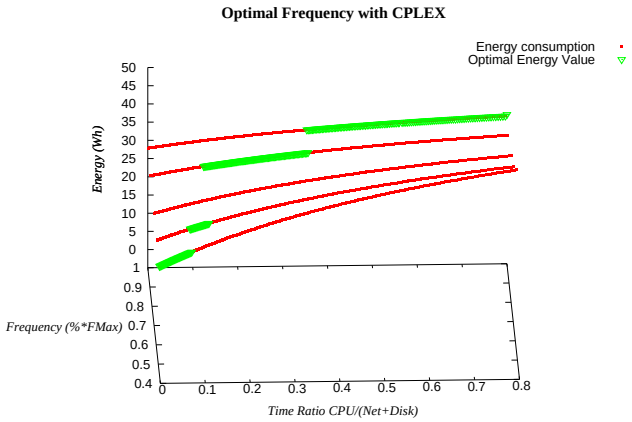

Fig. 1. Optimal frequencies (in green) for each ratio values

– Optimal Frequency (1.5GHz) determined by CPLEX solver in UserSpace mode.
– OnDemand mode, which allow the kernel to dynamically change the frequency of the CPU regarding its current load.

Then, another bash script to manage the power consumption is launched. First, it has to wait for the start of the experiment and then save power consumption of all deployed nodes into files. Each second the script runs the command line to get the power consumption. At the end of the experiment, all output files are sent to the frontend.

2.3 Results

Table 2 show all detailed experiments results in term of duration and energy consumption values on each node.

Results analysis Simulator results show the expected logic: theoretical optimum frequency given by the solver obtains a power consumption value lower than the maximum frequency, despite a longer execution time and OnDemand mode is more efficient with a gain of -10.5% relative to the optimum frequency. Network latency and throughput are hidden in the simulator, while real experiments are strongly influenced by these changes. In analysing the results of two different modes that have been executed on the same node, or two experiments running at same frequency on both same nodes, we can note that this is only the communication time between the tasks which has varied. This variation disrupts the overall result, while the mode change only DVFS impacts the computational time and does not affect the time spent in communication. The standard deviation value indicate high dispersion in all measures, due to to the variation of network performance on the environment used. For calculating energy PDUs

Table 2. Detailed results of experiment on Grid'5000 Reims site

DVFS mode	Duration (s)		Energy (Wh)				Nodes
	Total	% Error	Node 1	Node 2	Total	% Error	
Performance (1.7GHz)	7788	-0.15	391	372	763	-6.4	25-26
	7953	1.96	332	424	756	-7.2	1-10
	8529	9.34	398	427	825	1.22	5-7
	8627	10.6	457	443	900	10.4	27-29
	8681	11.3	453	465	918	12.6	25-26
Optimal (1.5GHz)	8030	-0.37	393	400	793	-2.57	27-28
	8071	0.14	400	416	816	0.24	25-26
	8686	7.76	433	439	872	7.12	25-26
OnDemand	8084	3.64	401	394	795	7.43	2-20
	8088	3.69	384	405	789	6.62	2-21
	8170	4.74	431	407	838	13.24	2-20
	8248	5.74	400	430	830	12.16	5-7
	8324	6.71	406	440	846	14.32	3-28
	8401	7.71	421	440	861	16.35	3-28
	8603	10.29	439	475	914	23.51	25-26

available on the website of Reims were used once per second, but their power measures are updated only each 3 seconds which also affects the accuracy of the total energy consumed value.

Table 3. Comparison between Grid'5000 experiments and CloudSim (bolds values)

DVFS Mode	Duration (s)			Energy (Wh)		
	Average	Deviation	% Error	Average	Deviation	% Error
Performance (1.7GHz)	**8320**			**833**		
	8315	414	-0.06	832	75	-0.12
Optimal (1.5GHz)	**8580**			**831**		
	8262	367	-3.7	832	30	0.12
OnDemand	**8320**			**743**		
	8320	367	0	827	40	11.31

3 Next step of Grid'5000 experiments

The TLM application need to write a lot of informations into several files. To avoid overloading the RAM, the application can write a part of these data during each iteration. The first step is to estimate the rate of hard drive in Reims to determined the time needed to write these data at each iterations. To do these measures, a C language application has been implemented. This application writes files from 20Mo to 180Mo , ten times for each size, and compute the average and the standard deviation to estimate the rate. Table 4 show results of the first measures :

Table 4. Hard Disk throughput measures on Reims site

File size (Mo)	Average (s)	Standard Deviation (s)	Throughput (Mo/s)
20	0.038072	0.0031	525.32
40	0.038072	0.0030	517.37
60	0.115149	0.0030	521.06
80	0.154235	0.0044	518.69
100	0.193702	0.0038	516.26
120	0.239641	0.0069	500.75
140	0.283617	0.0091	493.6
160	0.341733	0.0877	468.20
180	0.379805	0.0484	473.93

With these values, the aim is to determine how much information can be written in each iteration of TLM, without too much slow down the application. To do that, others experiments will be done, CPU time and Network time will be estimate with different configuration of TLM input parameters, and the file size to write will be determine to have an efficient $\frac{\text{CPU}}{\text{Network+I/O}}$ time ratio.

4 Conclusion

These experiments highlights the difficulty to obtain reliable power consumption values on a real platform, here due to the large standard deviation between experiments. One reason is that power measurements are not enough precise but also because the communication time varies a lot depending on experiments and resources used. To show the efficiency of DVFS, next TLM experiments will have a shorter communication time and use also I/O more constant.

References

1. Patterson, M., Tschudi, B., Vangeet, O., Cooley, J., Azevedo, D.: ERE: A metric for measuring the benefit of reuse energy from a data center. Technical Report White Paper 29, The Green Grid (2010)
2. Hoeffer, W.: The transmission-line matrix method–theory and applications. Microwave Theory and Techniques, IEEE Transactions on **33**(10) (oct 1985) 882–893
3. Alexandru, M., Monteil, T., Lorenz, P., Coccetti, F., Aubert, H.: Efficient large electromagnetic problem solving by hybrid tlm and modal approach on grid computing. In: International Microwave Symposium, Montral, Canada, 17-22 june 2012. (2012)