Performance Prediction of Stencil Applications on Accelerator Architectures

Víctor Martínez[1], **Philippe Navaux**[1]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

{victor.martinez,navaux}@inf.ufrgs.br

Resumo. *As aplicações stencil são comuns na solução de problemas relacionados com as equações diferenciais parciais, por exemplo nas simulações de geofísica. Consistem em um padrão que vai calculando a mesma operação em múltiples dados. Este trabalho apresenta um modelo para a predição do tempo de execução das aplicações stencils, baseado em Machine Learning. Os resultados mostram que é possível treinar o modelo e conseguir uma alta precisão.*

1. Introduction

The performance of HPC applications depends on many factors: architecture, code optimization, compiler and runtime frameworks. For example, [Dupros et al. 2015, de la Cruz and Araya-Polo 2015] presents cache-efficient algorithms for stencil computations on HPC architectures. On the other hand, Machine Learning (ML) is a comprehensive methodology for optimization. Recently, ML algorithms have been used on HPC systems. In [Martínez et al. 2017], the authors introduce a ML model to predict the performance of stencil computations on multicore architectures.

Stencil computations consist in using the neighboring points to evaluate a current point. The algorithm then moves to the next point applying the same computation pattern until the entire domain has been traversed. In this works, we study three well-known stencil kernels: the 7-point Jacobi for heat transfer and the seismic wave propagation explained in [Martínez et al. 2017], and the isotropic acoustic wave propagation explained in [Martínez et al. 2018]

2. Machine Learning Methodology

2.1. Testbed

We used Intel Xeon Phi (*Knights Landing*) to carry out the experiments. The detailed configuration are shown in Table 1. Based on this platform, Table 2 details all the configurations available. As it can be noted, a brute force approach would be unfeasible due to the large number of simulations required, because some of these executions can take many hours (or days).

2.2. Prediction Model

The proposed ML model is based on Support Vector Machines (SVM) and was built on top of three consecutive layers. The input layer contains the configuration values from the input vector presented in table 2. The hidden layer contains two SVMs that take values from the input vector to simulate the hardware counters, measured by PAPI library: L2 total cache misses and total cycles. Finally, the output layer contains one SVM to obtain the execution time value. SVMs were implemented in R language.

Table 1. Testbed

Processor	Intel Xeon Phi 7520
Clock(GHz)	1.40
Cores	68
Sockets	1
Threads	272
L2 cache size (MB)	32

Table 2. Configuration Domain

Optimization	Parameters	Total configurations
Number of threads	1	272
Scheduling policy	1	3
Chunk size	1	512
Total	**3**	**417,792**

2.3. Training and Validation

We created a training set by randomly selecting a subset from the configuration set presented in Table 2. Then, for each experiment we measured the hardware counters and execution time. A random testing set was used to calculate new execution time values. Table 3 presents the total number of experiments that were performed to obtain the training and validation sets. After that, we measured the accuracy of the model using the coefficient of determination (R-square). R-square ranges from zero to one, equal to one indicates a perfect fit of data prediction. As it can be noted in Table 4, the R-square is close to 99%, then we get a highly accurate regression.

Table 3. Experiment sets

	Jacobi	**Seismic**	**Isotropic**
Training set	103	163	159
Testing set	929	1469	1436
Total	1032	1632	1595

Table 4. Prediction Accuracy

	R-square
Jacobi	0.9949
Seismic	0.9993
Isotropic	0.9610

3. Conclusion

In this paper, we introduced a predictive performance modeling strategy for stencil applications on accelerator architectures. We showed that performance can be predicted with a high accuracy (96-99%). Our model is not restricted to accelerators platforms and can also be implemented into architectures with the available hardware counters to obtain the cache-related measurements. Future work is oriented to unsupervised algorithms to avoid training stage.

References

de la Cruz, R. and Araya-Polo, M. (2015). *Modeling Stencil Computations on Modern HPC Architectures*, pages 149–171. Springer International Publishing, Cham.

Dupros, F., Boulahya, F., Aochi, H., and Thierry, P. (2015). Communication-avoiding seismic numerical kernels on multicore processors. In *International Conference on High Performance Computing and Communications (HPCC)*, pages 330–335.

Martínez, V., Dupros, F., Castro, M., and Navaux, P. (2017). Performance improvement of stencil computations for multi-core architectures based on machine learning. *Procedia Computer Science*, 108:305 – 314. International Conference on Computational Science, {ICCS} 2017, 12-14 June 2017, Zurich, Switzerland.

Martínez, V., Serpa, M., Dupros, F., Padoin, E. L., and Navaux, P. (2018). *Performance Prediction of Acoustic Wave Numerical Kernel on Intel Xeon Phi Processor*, pages 101–110. Springer International Publishing, Cham.