# Adaptive Partitioning Strategies for Loop Parallelism in Heterogeneous Architectures

Angeles Navarro*, Antonio Vilches*, Rafael Asenjo*, Francisco Corbera*
*Universidad de Málaga, Andalucía Tech, Dept. of Computer Architecture, Spain.
{angeles,vilches,asenjo,corbera}@ac.uma.es

***Abstract*—This paper explores the possibility of efficiently using multicores in conjunction with multiple GPU accelerators under a parallel task programming paradigm. In particular, we address the challenge of extending a `parallel_for` template to allow its exploitation on heterogeneous systems. The extension is based on a two-stages pipeline engine which is responsible for partitioning and scheduling the chunks into the computational resources. Under this engine, we propose a dynamic scheduling strategy coupled with an adaptive partitioning heuristic that resizes chunks to prevent underutilization and load unbalance of CPUs and GPUs. In this paper we introduce the adaptive partitioning heuristic which is derived from an analytical model that minimizes the load unbalance while maximizes the throughput in the system. Using two benchmarks we evaluate the overhead introduced by our template extensions finding that it is negligible. We also evaluate the efficiency of our adaptive partitioning strategies and compared them with related work.**

## References

[1] J. Reinders, *Intel Threading Building Blocks: Multi-core parallelism for C++ programming*. O'Reilly, 2007.

[2] C. Augonnet, J. Clet-Ortega, S. Thibault, and R. Namyst, "Data-aware task scheduling on multi-accelerator based platforms," in *Parallel and Distributed Systems (ICPADS), 2010 IEEE 16th International Conference on*, Dec. 2010, pp. 291–298.

[3] J. Bueno, J. Planas, A. Duran, R. Badia, X. Martorell, E. Ayguade, and J. Labarta, "Productive programming of GPU clusters with OmpSs," in *Parallel Distributed Processing Symposium (IPDPS), IEEE 26th Intl.*, May 2012, pp. 557–568.

[4] J. Lima, T. Gautier, N. Maillard, and V. Danjean, "Exploiting concurrent GPU operations for efficient work stealing on multi-GPUs," in *Computer Architecture and High Perf. Comp. (SBAC-PAD), IEEE 24th Intl. Symp. on*, Oct. 2012.

[5] T. Ibaraki and H. Katoh, *Resource Allocation Problems: Algorithmic Approaches*. Cambridge, Mass.: MIT Press, 1988.

[6] M. E. Belviranli, L. N. Bhuyan, and R. Gupta, "A dynamic self-scheduling scheme for heterogeneous multiprocessor architectures," *ACM Trans. Archit. Code Optim.*, vol. 9, no. 4, pp. 57:1–57:20, Jan. 2013. [Online]. Available: http://0-doi.acm.org.jabega.uma.es/10.1145/2400682.2400716

[7] D. C. Rudolph and C. D. Polychronopoulos, "An efficient message-passing scheduler based on guided self scheduling," in *3rd Intl. Conf. on Supercomputing*, ser. ICS '89. New York, NY, USA: ACM, 1989, pp. 50–61. [Online]. Available: http://0-doi.acm.org.jabega.uma.es/10.1145/318789.318796

[8] M. Kulkarni, M. Burtscher, C. Cascaval, and K. Pingali, "Lonestar: A suite of parallel irregular programs," in *IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS'09)*, 2009.

[9] C. Augonnet, S. Thibault, R. Namyst, and P.-A. Wacrenier, "StarPU: A unified platform for task scheduling on heterogeneous multicore architectures," *Concurrency and Computation: Practice and Experience*, no. 23, pp. 187–198, February 2011.

[10] *CUDA Toolkit 5.0 Performance Report*, NVidia, Jan. 2013, https://developer.nvidia.com/nvidia-gpu-programming-guide.

[11] S. Russel, "Levering GPGPU and OpenCL technologies for natural user interaces," You i Labs inc., Tech. Rep., 2012.

[12] A. Hart, "The OpenACC programming model," Cray Exascale Research Initiative Europe, Tech. Rep., 2012.

[13] C.-K. Luk, S. Hong, and H. Kim, "Qilin: Exploiting parallelism on heterogeneous multiprocessors with adaptive mapping," in *Microarchitecture, 2009. MICRO-42. 42nd IEEE/ACM Intl. Symp. on*, Dec. 2009, pp. 45–55.

[14] S. Venkatasubramanian and R. W. Vuduc, "Tuned and wildly asynchronous stencil kernels for hybrid CPU/GPU systems," in *23rd International Conference on Supercomputing (ICS'09)*, Jun. 2009, pp. 244–255.

[15] V. Ravi and G. Agrawal, "A dynamic scheduling framework for emerging heterogeneous systems," in *High Performance Computing (HiPC), 2011 18th International Conference on*, Dec. 2011, pp. 1–10.

This work has been supported by Andalucía Tech, Campus Internacional de Excelencia, and by the following Spanish projects: TIN2010-16144 from Ministerio de Ciencia e Innovación, and P08-TIC-3500 and P11-TIC-08144 from Junta de Andalucía.