Exploring Data Locality for GPU Unified Virtual Memory

ABSTRACT

To facilitate GPU programming and accelerate applications with large working set, recent GPUs support unified CPU/GPU memory addressing and on-demand page migration. However, our investigation shows that current page migration mechanism is not aware of data locality, resulting in poor performance and redundant data movement. To improve the performance and energy efficiency of GPU unified memory, we propose a framework to achieve intelligent unified memory management using both compile-time and runtime information. Compiler collects information such as data access pattern and reuse frequency and passes them to runtime library, and then runtime makes data placement decisions based on both current hardware status and compiler analysis results. The proposed scheme can benefit both directive based GPU programming models such as OpenMP and native GPU programming languages such as CUDA. Our preliminary evaluation shows that our framework can significantly improve GPU performance and reduce redundant data movement between CPU and GPU for benchmarks with large working set.

KEYWORDS

GPU, unified virtual memory, compiler optimization

ACM Reference format:
. 2017. Exploring Data Locality for GPU Unified Virtual Memory. In *Proceedings of ACM Conference, Washington, DC, USA, July 2017 (Conference'17),* 1 pages.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In this paper, we focus on discreted GPUs.

REFERENCES

Conference'17, July 2017, Washington, DC, USA
2017. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn