Northeastern University

Model-based Concurrent Kernel Execution on GPU

Leiming Yu, Xun Gong, Fanny Nina-Paravecino,
Qianqian Fang, Norm Rubin and David Kaeli

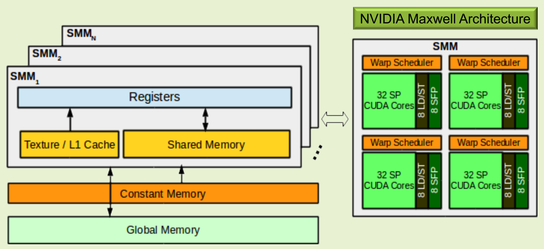
NIH NIGMS

NVIDIA

Abstract

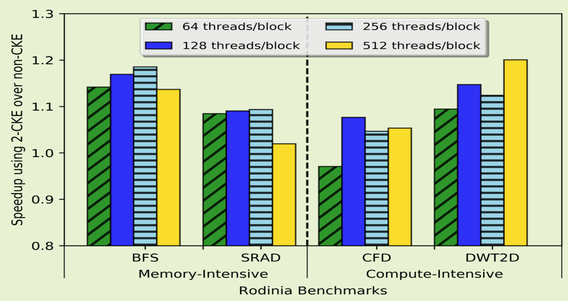
- Modern GPUs have become a mainstream accelerator to speed up scientific computing, large-scale machine learning algorithms and big data analysis.
- GPUs support concurrent kernel execution to achieve a high device utilization.
- Maximizing the overall performance using concurrent kernels involves significant programming effort.
- Model-based analysis can accelerate design space exploration.
- Our proposed model can help tune GPU kernels and predict concurrent kernel performance accurately.

GPU Background



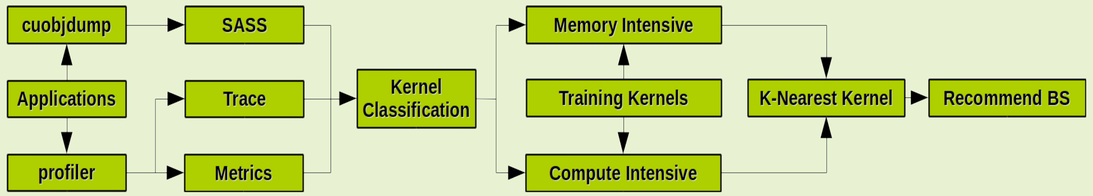
- More computing units and higher memory bandwidth.
- Memory hierarchy is specialized for various optimizations

Motivation



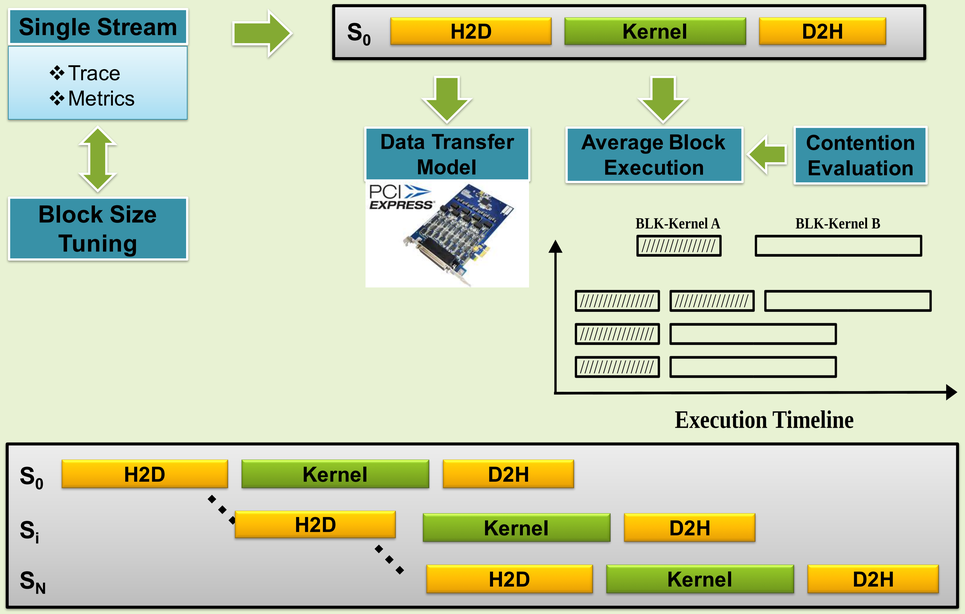
- Select the best block size for your application.
- Find the best stream number for the overall performance
- Schedule kernels more efficiently, less contention.

Block Size Tuning



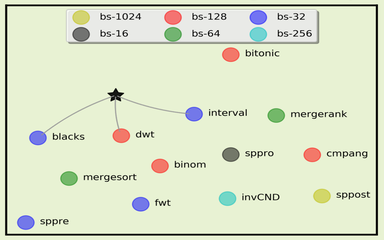
- Run GPU kernels to obtain performance profiles and capture configuration information
- SASS instruction cycles are benchmarked
- Categorize GPU kernels
- Apply t-SNE to select top3 nearest kernels
- Recommend the best block size from the majority

Model-based Concurrent Kernel Execution

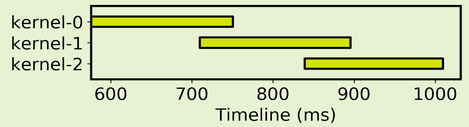


Results

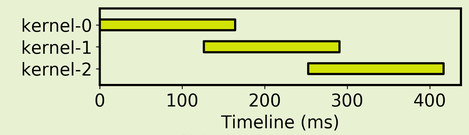
Monte Carlo Photon Migration (MCX[1,2])



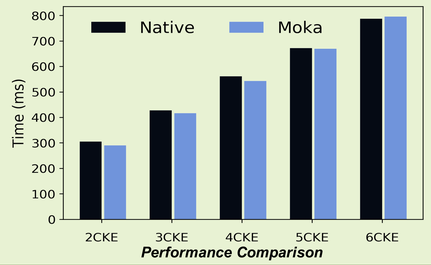
Block Size Tuning



Native Runtime on GTX 950



Moka Predicted Runtime



Performance Comparison

Conclusion

- Scaling the workload size can affect the accuracy of the block size prediction.
- Kernel configuration and resource requirements result in contention.
- Resource contention factors have strong impact on the prediction accuracy.

Future Work

- Predict performance on heterogeneous architecture environments.
- Optimize workload scheduling based on the proposed performance model.

References

1. Q. Fang and D. A. Boas. "Monte Carlo Simulation of Photon Migration in 3D Turbid Media Accelerated by GPUs." Optics Express 17.22 (2009): 20178-20190.
2. F. Nina-Paravecino, L. Yu, Q. Fang and D. Kaeli, "MCX-Accelerated Monte Carlo Simulation of Photon Migration in 3D Turbid Media for Single and Multiple GPUs." NVIDIA GPU Technology Conference, April, 2016.

Acknowledgement

- This project is supported in part by the NIH/NIGMS under grant: R01-GM114365.
- We would like to acknowledge NVIDIA for their support for this work through the NVIDIA Research Center program.