A Survey of Graphical Processing Units

Inji Kim and Kavish Ranawella and Andrew Davis

***Abstract*—**

I. Introduction

APPENDIX

Notes that won't be included in the final draft:

A. Topics

List of project topics and the papers that correspond to those topics

1) Memory Models: [2] [6] [17] [1] [14] [13] [4] [7] [8] [12]

2) Architecture: [2] [18] [5] [16] [11] [19] [3]

3) GPU simulators: [10] [9] [2] [5] [15] [12]

REFERENCES

[1] P. B, N. Jawalkar, and A. Basu, "Designing virtual memory system of mcm gpus," in *Proceedings of the 55th Annual IEEE/ACM International Symposium on Microarchitecture*, ser. MICRO '22. IEEE Press, 2023, p. 404–422. [Online]. Available: https://doi.org/10.1109/MICRO56248.2022.00036

[2] A. Bakhoda, G. L. Yuan, W. W. L. Fung, H. Wong, and T. M. Aamodt, "Analyzing cuda workloads using a detailed gpu simulator," in *2009 IEEE International Symposium on Performance Analysis of Systems and Software*, 2009, pp. 163–174.

[3] W. W. L. Fung and T. M. Aamodt, "Thread block compaction for efficient simt control flow," in *2011 IEEE 17th International Symposium on High Performance Computer Architecture*, 2011, pp. 25–36.

[4] W. W. L. Fung, I. Singh, A. Brownsword, and T. M. Aamodt, "Hardware transactional memory for gpu architectures," in *Proceedings of the 44th Annual IEEE/ACM International Symposium on Microarchitecture*, ser. MICRO-44. New York, NY, USA: Association for Computing Machinery, 2011, p. 296–307. [Online]. Available: https://doi.org/10.1145/2155620.2155655

[5] A. B. Hayes, F. Hua, J. Huang, Y. Chen, and E. Z. Zhang, "Decoding cuda binary," in *Proceedings of the 2019 IEEE/ACM International Symposium on Code Generation and Optimization*, ser. CGO 2019. IEEE Press, 2019, p. 229–241.

[6] M. A. Ibrahim, O. Kayiran, Y. Eckert, G. H. Loh, and A. Jog, "Analyzing and leveraging shared l1 caches in gpus," in *Proceedings of the ACM International Conference on Parallel Architectures and Compilation Techniques*, ser. PACT '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 161–173. [Online]. Available: https://doi.org/10.1145/3410463.3414623

[7] A. Jog, O. Kayiran, N. Chidambaram Nachiappan, A. K. Mishra, M. T. Kandemir, O. Mutlu, R. Iyer, and C. R. Das, "Owl: cooperative thread array aware scheduling techniques for improving gpgpu performance," in *Proceedings of the Eighteenth International Conference on Architectural Support for Programming Languages and Operating Systems*, ser. ASPLOS '13. New York, NY, USA: Association for Computing Machinery, 2013, p. 395–406. [Online]. Available: https://doi.org/10.1145/2451116.2451158

[8] G. Kadam, D. Zhang, and A. Jog, "Rcoal: Mitigating gpu timing attack via subwarp-based randomized coalescing techniques," *2018 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, pp. 156–167, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:4565978

[9] V. Kandiah, S. Peverelle, M. Khairy, J. Pan, A. Manjunath, T. G. Rogers, T. M. Aamodt, and N. Hardavellas, "Accelwattch: A power modeling framework for modern gpus," in *MICRO-54: 54th Annual IEEE/ACM International Symposium on Microarchitecture*, ser. MICRO '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 738–753. [Online]. Available: https://doi.org/10.1145/3466752.3480063

[10] M. Khairy, Z. Shen, T. M. Aamodt, and T. G. Rogers, "Accel-sim: An extensible simulation framework for validated gpu modeling," in *2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA)*, 2020, pp. 473–486.

[11] I. Laguna and G. Gopalakrishnan, "Finding inputs that trigger floating-point exceptions in gpus via bayesian optimization," in *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, ser. SC '22. IEEE Press, 2022.

[12] H. Liu, S. Pai, and A. Jog, "Why gpus are slow at executing nfas and how to make them faster," in *Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems*, ser. ASPLOS '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 251–265. [Online]. Available: https://doi.org/10.1145/3373376.3378471

[13] Y. Liu, X. Zhao, M. Jahre, Z. Wang, X. Wang, Y. Luo, and L. Eeckhout, "Get out of the valley: Power-efficient address mapping for gpus," in *2018 ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA)*, 2018, pp. 166–179.

[14] X. Mei and X. Chu, "Dissecting gpu memory hierarchy through microbenchmarking," *CoRR*, vol. abs/1509.02308, 2015. [Online]. Available: http://arxiv.org/abs/1509.02308

[15] M. A. Raihan, N. Goli, and T. M. Aamodt, "Modeling deep learning accelerator enabled gpus," *CoRR*, vol. abs/1811.08309, 2018. [Online]. Available: http://arxiv.org/abs/1811.08309

[16] I. Sañudo, N. Capodieci, J. Martinez, A. Marongiu, and M. Bertogna, "Dissecting the cuda scheduling hierarchy: a performance and predictability perspective," 04 2020.

[17] I. Singh, A. Shriraman, W. W. L. Fung, M. O'Connor, and T. M. Aamodt, "Cache coherence for gpu architectures," in *2013 IEEE 19th International Symposium on High Performance Computer Architecture (HPCA)*, 2013, pp. 578–590.

[18] D. Yan, W. Wang, and X. Chu, "An llvm-based open-source compiler for nvidia gpus," in *Proceedings of the 27th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*, ser. PPoPP '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 448–449. [Online]. Available: https://doi.org/10.1145/3503221.3508428

[19] M. Zhu, T. Zhang, Z. Gu, and Y. Xie, "Sparse tensor core: Algorithm and hardware co-design for vector-wise sparse neural networks on modern gpus," in *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture*, ser. MICRO '52. New York, NY, USA: Association for Computing Machinery, 2019, p. 359–371. [Online]. Available: https://doi.org/10.1145/3352460.3358269