Squeezing Energy Savings Out Of Similar Data and Computation in GPGPUs

Daniel Wong
University of California, Riverside

ABSTRACT

General Purpose Graphical Processing Units (GPGPUs) consume a significant amount of power. A large portion of computation and data values stored in hardware exhibit value similarity [1], that is, values that are the same, or similar and differs only in the least significant bits. GPGPUs execute groups of 32 threads in lockstep, called a warp. We identify a form of value similarity called *intra-warp operand value similarity*, where all input operands within the warp are value similar. We demonstrate that intra-warp operand value similarity can tradeoff accuracy for energy. By dynamically identifying value similarity in hardware, and executing a single representative thread on behalf of all the active threads in a warp, we can save execution unit energy. This generates a single representative value that is an approximation of the precise values. This representative value can then be stored compactly in the register file as a value similar scalar, reducing the read and write energy when dealing with approximate data. We show that these techniques can reduce execution unit energy by 37%, register file energy by 28%, and improve overall GPGPU energy efficiency by 26% with less than 5.5% quality degradation. This work targets GPGPU applications where output errors can be tolerated, such as image processing applications. This work enables a new dimension in energy savings tradeoff for GPGPUs.

BODY

Significant energy savings (with $<5.5\%$ output quality loss) can be achieved by merging computation and storage of similar data.

REFERENCES

[1] D. Wong, N. Kim, and M. Annavaram. Approximating warps with intra-warp operand value similarity. In *Proceedings of the 2016 International Symposium on High Performance Computer Architecture (HPCA)*, HPCA '16, 2016.

Volume 4 of Tiny Transactions on Computer Science