Graphics processor (GPU) architectures
Preeti Ranjan Panda
-June 30th 2011, 16:15, E23, OH14-

Graphics processor (GPU) architectures have evolved rapidly in recent years with increasing performance demanded by 3D graphics applications such as games. However, challenges exist in integrating complex GPUs into mobile devices because of power and energy constraints, motivating the need for energy efficiency in GPUs. While a significant amount of power optimization research effort has concentrated on the CPU system, GPU power efficiency is a relatively new and important area because the power consumed by GPUs is similar in magnitude to CPU power. Power and energy efficiency can be introduced into GPUs at many different levels: (i) Hardware component level - queue structures, caches, filter arithmetic units, interconnection networks, processor cores, etc., can be optimized for power. (ii) Algorithm level – the deep and complex graphics processing computation pipeline can be modified to be energy aware. Shader programs written by the user can be transformed to be energy aware. (iii) System level - co-ordination at the level of task allocation, voltage and frequency scaling, etc., requires knowledge and control of several different GPU system components.

We outline two strategies for applying energy optimizations at different levels of granularity in a GPU. (1) Texture Filter Memory is an energy-efficient an augmentation of the standard GPU texture cache hierarchy. Instead of a regular data cache hierarchy, we employ a small first level register based structure that is optimized for the relatively predictable memory access stream in the texture filtering computation. Power is saved by avoiding the expensive tag lookup and comparisons present in regular caches. Further, the texture filter memory is a very small structure, whose access energy is much smaller than a data cache of similar performance. (2) Dynamic Voltage and Frequency Scaling, an established energy management technique, can be applied in GPUs by first predicting the workload in a given frame, and, where sufficient slack

exists, lowering the voltage and frequency levels so as to save energy while still completing the work within the frame rendering deadline. We apply DVFS in a tiled graphics renderer, where the workload prediction and voltage/frequency adjustment is performed at a tile-level of granularity, which creates opportunities for on-the-fly correction of prediction inaccuracies, ensuring high frame rates while still delivering low power.

Preeti Ranjan Panda received his B. Tech. degree in Computer Science and Engineering from the Indian Institute of Technology Madras, and his M. S. and Ph.D. degrees from the University of California at Irvine. He is currently a Professor in the Department of Computer Science and Engineering at IIT Delhi. He has previously worked at Texas Instruments and Synopsys. His research interests are: Embedded Systems Design, CAD/VLSI, Post-silicon Debug/Validation, Memory Architectures and Optimizations, and Low Power Design. He is the author of two books: 'Memory issues in Embedded Systems-on-chip: Optimizations and Exploration' and 'Power-efficient System Design'. Prof. Panda has served on the program committees and chaired sessions at several conferences in the areas of Embedded Systems and CAD/VLSI, including DAC, ICCAD, DATE, CODES+ISSS, etc. He is a present/past member of the editorial boards of IEEE TCAD, ACM TODAES, and IJPP.