IPM

پژوهشگاه دانش‌های بنیادی

پژوهشکده علوم کامپیوتر

سخنرانی علمی

Designing a Scalable Memory System for GPUs

Negar Akbarzadeh, Sharif University of Technology

Abstract

Graphics Processing Units (GPUs) are widely used for modern applications with huge data sizes. However, the performance benefit of GPUs is limited by their memory capacity and bandwidth. This talk begins with an overview of the memory system of GPUs, followed by a discussion about the bottlenecks that prevents further scaling of the memory system and reaching the maximum performance. We then briefly review the current approaches to remedy capacity and bandwidth bottlenecks. However, these works mainly focus on one bottleneck or do not provide a scalable solution that fits future requirements. We explore the possibility of true-3D stacking of PCM memory layers on top of a powerful GPU. Such a 3D structure provides two main benefits. First, PCM is a suitable candidate for architecting a true-3D GPU memory architecture with lower power consumption and no refresh overhead, significantly increasing the GPU memory bandwidth. Second, PCMs' higher density and lower power consumption enable high-capacity memories through 1) integrating more cells in each 3D layer, and 2) increasing the number of layers. However, PCM faces some challenges such as longer write latency, higher write energy, and lower endurance, that need to be addressed carefully. Hence, the talk continues with the proposal of architectural solutions to address these challenges. Our experimental results show considerable performance improvement while maintaining the area and power budgets. The proposed architecture also provides substantial memory space, enabling the execution of applications with huge datasets.

Biography

Negar Akbarzadeh received her B.Sc. and M.Sc. in Electrical Engineering from Shahid-Beheshti University in 2014 and 2016, respectively. She is currently a Ph.D. candidate in Computer Engineering at the Sharif University of Technology. Her research interests include computer architecture, memory systems, NoCs, and GPUs. Currently, she is actively seeking a postdoctoral research position within the Department of Computer Science at the Institute for Research in Fundamental Sciences (IPM).

زمان : چهارشنبه ۱۴۰۲/۱۱/۲۵ – ساعت ۱۵:۰۰

ارائه به صورت مجازی انجام خواهد شد.

https://vmeeting2.ipm.ir/b/com-hh1-n07-vil

*****شرکت برای عموم علاقه‌مندان آزاد است*****