IPM

"سخنرانی‌های علمی"

پژوهشگاه دانشهای بنیادی
پژوهشکده علوم کامپیوتر

# Highly Concurrent Latency-Tolerant Register Files for GPUs

سید محمد صدرالساداتی

پژوهشکده علوم کامپیوتر، پژوهشگاه دانش‌های بنیادی

## Abstract

Graphics Processing Units (GPUs) employ large register files to accommodate all active threads and accelerate context switching. Unfortunately, register files are a scalability bottleneck for future GPUs due to long access latency, high power consumption, and large silicon area provisioning. Prior work proposes hierarchical register file to reduce the register file power consumption by caching registers in a smaller register file cache. Unfortunately, this approach does not improve register access latency due to the low hit rate in the register file cache.

In this work, we propose the Latency-Tolerant Register File (LTRF) architecture to achieve low latency in a two-level hierarchical structure while keeping power consumption low. We observe that compile-time interval analysis enables us to divide GPU program execution into intervals with an accurate estimate of a warp's aggregate register working-set within each interval. The key idea of LTRF is to prefetch the estimated register working-set from the main register file to the register file cache under software control, at the beginning of each interval, and overlap the prefetch latency with the execution of other warps. We observe that register bank conflicts while prefetching the registers could greatly reduce the effectiveness of LTRF. Therefore, we devise a compile-time register renumbering technique to reduce the likelihood of register bank conflicts. Our experimental results show that LTRF enables high-capacity yet long-latency main GPU register files, paving the way for various optimizations. As an example optimization, we implement the main register file with emerging high-density high-latency memory technologies, enabling 8× larger capacity and improving overall GPU performance by 34%.

## Biography

Mohammad Sadrosadati is a postdoctoral researcher at Institute for Research in Fundamental Sciences (IPM). He received his PhD in computer engineering from Sharif University of Technology in 2019. His research interests include Heterogeneous Computing, Processing-in-Memory, Memory Systems, and Interconnection Networks.

**زمان : چهارشنبه ۱۹ آذرماه ۱۳۹۹- ساعت ۱۵:۰۰**

**ارائه به صورت مجازی انجام خواهد شد.**

***** شرکت برای عموم علاقه مندان آزاد است*****