Speaker: Prof. Mainak Chaudhuri
Date: 26th September, 2018. (Wednesday)
Time: 5 PM
Venue: KD101

Title: Cache Optimization Needs To Be Technology-aware

Abstract:
Traditional cache design principles abstract away the technology used to build the cache and squarely focus on optimizing reuse probability or hit rate. In the first part of the talk, I will follow this traditional cache design principle and present a technique that dynamically learns reuse probability at run-time and employs the learned probabilities to design high-performance cache management policies. We have applied this generic technique to a number of different scenarios employing SRAM caches such as shared last-level cache of multi-core CPUs, last-level cache of discrete GPUs, and shared last-level cache of CPU-GPU heterogeneous multiprocessor SoCs.

In the second part of the talk, I will attempt to apply similar reuse-driven techniques to optimize DRAM caches and show that such a principle is met with the counter-intuitive trend where performance may degrade with increasing hit rate. I will present a simple theory to understand this phenomenon and briefly outline how this theory can be used to design DRAM caches that perform close to the optimal operating point.