ABSTRACT:
We have entered the multi-core era! In a multi-core processor, each core has its own private resources, such as pipeline, registers, and L1 caches. The aggregated computing resources allow a multi-core processor to offer greater computing capability with less power consumption than that of a single-core processor by executing multiple threads concurrently. However, the multiple cores share the last level cache (LLC) and a narrow data path to the memory. Thus, concurrent running threads can contend for these resources in an out-of-control fashion to severely degrade performance. Keeping the working sets of the running threads in LLC is the pivotal point to reduce the contention by effectively reducing long-latency memory accesses and lowering the usage of the shared data path. The operating system, being a key resource manager, can play a critical role for this purpose.

I will introduce a set of effective enhancements in OS scheduling and buffer cache management in order to maintain the working sets of running threads in LLC. The main system component is a shared cache aware scheduling framework for multi-core processors by extending the LLC hardware design with a set of low overhead profiling units to measure working set sizes and to collect other locality information of running threads. With rich knowledge of locality related information, the OS scheduler can effectively make thread assignments such that the working sets of running threads can be held within the capacity of LLC. A supporting system component is a buffer cache design for multi-core systems, which avoids the working sets of the running threads being polluted by burst file accesses. Experimental performance results will be presented to show the effectiveness of the proposed methods. I will present my future research at the end of the talk.