Architecting Emerging Memory Technologies for Energy-efficient Computing in Modern Processors

Mr Wang Jianxing

12 November 2015, Thursday, 02:00 PM to 03:30 PM

Executive Classroom, COM2-04-02

Dr Wong Weng Fai, Associate Professor, School of Computing

The concern about energy consumption in current technology node is becoming more serious. As we are already in deep sub-micron era, the relentless process scaling makes existing energy reduction techniques less efficient. It is predicted that we might soon face the situation where most of the silicon area is forced to be dark and can not be powered on simultaneously.

As alternative solutions, a number of emerging memory technologies are being actively studied. They allow us to design more energy efficient architectures and potentially solve the dark silicon problem from the ground. However, they also incur certain weaknesses such as high write energy and limited endurance. Previous studies showed that an effective way to mitigate these negative impacts is by optimizing the memory architecture.

Therefore, this thesis will focus on architecture designs for deploying emerging memories at various levels in the memory hierarchy. First of all, we exploit both the storage and computing capabilities of a particular emerging memory: the memristor, in the design of a neural branch predictor. Experiments showed that our design not only increases the accuracy, but also consumes far less energy than traditional schemes. Secondly, we present a L1 cache architecture that utilizes both the conventional SRAM and the emerging STT-RAM technology. Our scheme mitigated the performance impacts from the expensive write operations of STT-RAM and achieved energy savings with enhanced reliability. Thirdly, we propose a dynamic block reconfiguration mechanism in the design of multi level cell (MLC) STT-RAM last-level caches. Our design leveraged the latency/capacity trade-off in MLC STT-RAM to accelerate some portions of the cache with less energy consumption. Lastly, the use of STT-RAM in GPGPU has also been investigated. A hybrid L2 architecture employing a large STT-RAM part and a small SRAM augment is demonstrated. Our design monitors the read/write behavior of cache blocks in STT-RAM part and attempts to offload writes to SRAM augment, reducing the write penalties. Simulation shows our design
achieved minor performance improvement over both pure SRAM and STT-RAM with significant energy reduction.