2016-1 Computer Architecture

**Final Project**

1. **Objective:** Design and simulate a NAND flash memory based main memory system for data intensive processing such as in-memory databases.
2. **Description:** By exploiting NAND flash memory as a component of main memory, this system can provide cost-effectiveness such as cost-per-bit and power consumption. However, as shown in Table 1, since NAND flash has long access time compared to current DRAM based main memory, such high access latencies of NAND flash induce performance degradation in the overall system. To overcome this problem, one approach is to use a small amount of DRAM space as a sort of caching buffer between NAND flash based main memory and last level cache. This approach can hide long access latencies of flash memory and Figure 1 shows this kind of operation.

|  |
| --- |
| **Table 1. Characteristics of HDD,DRAM, and NAND Flash** |
| |  |  |  |  | | --- | --- | --- | --- | | Device Type | SRAM | DRAM | NAND Flash | | Read Access Time | 1 ns  (per cache line) | 10 ns  (per cache line) | 25 us  (per page-4KB) | | Write Access Time | 1 ns  (per cache line) | 10 ns  (per cache line) | 200 us  (per page-4KB) | |

Table 2. Access Latencies

|  |  |  |  |
| --- | --- | --- | --- |
|  | Access Latency |  | Access Latency |
| L1 (at once) | 1 ns | Main memory  (at once) | 200 ns |
| L2 (at once) | 20 ns |  |  |

|  |
| --- |
|  |
| Fig. 1 DRAM cache buffer |

1. **Based on the above mentioned**, you should design a NAND flash memory based main memory system for in-memory database computing using the proposed basis approach and/or any other new methods. And design your best configuration and management method, for example any prefetch/eviction scheme, an overall managing flow, best prefetch size (unit), and etc.

(1) Basic cache structure - L1 instruction: 32KB, L1 data: 32KB, L2 unified: 512KB (HW4 2-level cache)

(2) Total capacity - 200MB (DRAM - 40MB, NAND Flash- 160MB) – 1GB

If possible, find best ratio of DRAM and NAND Flash memory space for cost-effectiveness.

(3) NAND Flash prefetch unit- 1KB, 4KB, 8KB, 16KB, etc..

(4) Use simulation parameters as in Table 1 and 2.

- Total Access Time = Device read/write access time + cache/main memory access latency

(5) Assume that all data are stored in NAND Flash and DRAM devices and do not consider secondary storage.

1. **Option:** Flash memory based main memory for conventional systems, PIM (processor in memory) architectural model for Deep learning applications.
2. **Submission**: Report & simulation source code
3. **Presentation & Due date**: June 10 (Fri), 13:00-15:00pm