ABSTRACT

The state-of-the-art computing systems are based on traditional Von-Neumann architecture and are suffering from limited throughput and energy inefficiency due to well known Von-Neumann bottleneck. With data intensive applications such as Artificial intelligence, neural computing etc., this problem of limited throughput and energy efficiency is more pronouncing. A possible approach to break the Von-Neumann bottleneck is to enable compute in memory.

For the past several decades, 6T SRAM has found a prominent place as conventional memory for caches, and is an obvious choice for in-memory computing (IMC) as well because of its less area requirement. With minimally sized 6T SRAM, there is a problem of compute-disturb i.e. flipping of data bits stored in two activated bit-cells for computation. Many techniques have been proposed to avoid compute-disturb problem. However, these techniques either degrade performance or incur architecture-level overheads.

In this work, we have analyzed different existing techniques for performing in-memory computing (IMC) in 6T SRAM and present two different 6T SRAM cells which are robust, have the capability to avoid compute-disturbs and require only 13-14% additional area compared to conventional 6T SRAM cell. We have implemented 2KB IMC array in CMOS 28nm technology and performed logical operations like AND, OR, NAND and NOR and arithmetic operations like addition. We also ensured that the proposed cell designs do not affect the other memory operations and for that we used negative bit-line technique to improve write noise margin. The holistic *post-facto* analysis of write and hold failures is one of the important contributions of this work. Also, our proposed IMC is implemented as conventional 6T SRAM with minimal mod-

ification of the peripheral circuits, thus allowing ease of migration from existing memory to IMC. The proposed bit-cells based IMC architectures are compared with other existing works by taking the same peripherals for all the works. It is observed that the proposed IMC architecture delivers 5.04 TOPS/Watt for 8-bit addition, supports a wide supply range (0.6-0.9)V, is fast (3.704 GHz), and exhibits compute-disturb free operation at all process corners.