3D DRAM Based Application Specific Hardware Accelerator for SpMV

Fazle Sadi, Larry Pileggi and Franz Franchetti
Department of Electrical and Computer Engineering
Carnegie Mellon University, Pittsburgh, PA, USA
Email: {fsadi,pileggi,franzf}@andrew.cmu.edu

ABSTRACT

For numerous scientific applications Sparse Matrix-Vector multiplication (SpMV) is one of the most important kernels. Unfortunately, due to its very low ratio of computation to memory access SpMV is inherently a memory bound problem. On the other hand, the main memory bandwidth of commercial off-the-shelf (COTS) architectures is insufficient for available computation resources on these platforms, well known as the *memory wall* problem. As a result, COTS architectures are unsuitable for SpMV. Furthermore, SpMV requires random access into a memory space which is far too big for cache. Hence, it becomes difficult to utilize the main memory bandwidth which is already scarce.

With the emergence of 3D IC technology, stacked DRAM such as High Bandwidth Memory (HBM) can be deemed as a solution to the memory wall problem. Nonetheless, full utilization of this high bandwidth in an energy efficient manner remains as a challenge for COTS architectures. On the other hand, advanced circuit level techniques such as Logic in Memory (LiM) based hardware accelerators are recently being used to provide low power solution to big data problems. Moreover, interposer technology has become available to provide high speed channel between compute core and memory.

Given these advancements in hardware, we propose an algorithm for large SpMV problems which is specially optimized to fully exploit the underlying micro-architecture and overall system capabilities. This algorithm is implemented in two steps. The key feature of the first step is that it converts all the main memory random access into streaming access. This reduces the overall data transfer volume significantly and ensures full utilization of the memory bandwidth. On top of that, we propose a meta-data compression technique, namely Variable Length Delta Index (VLDI), to decrease the data transfer volume even further. VLDI is particularly effective for sparse matrices where meta-data to payload ratio is high, e.g. sparse bit matrices.

After the first step, the SpMV problem effectively converts into a big multi-way merge problem. For the second step, we demonstrate a novel fast multi-way merge algorithm, namely Propagated Active Path (path) merge. This algorithm helps to consume the all the bandwidth provided by HBM and maintain high throughput. PAP merge is independent of the problem size. Hence, the problem size can be easily scaled without sacrificing performance and efficiency. Furthermore, the hardware implementation of PAP merge is enhanced to provide wide output interface. This increases the throughput of the computation core linearly which is especially helpful for systems with multiple HBM.

The overall hardware system for the proposed SpMV algorithm is also demonstrated. It constitutes of application specific LiM based accelerator core, eDRAM (embedded DRAM) scratchpad, 3D stacked DRAM and interposer platform. To test the effectiveness of the proposed SpMV algorithm on the accelerator system, we investigate the performance and energy efficiency of various COTS architectures using finely tuned standard libraries such as Intels Math Kernel Library (MKL). Our experimental results show that the proposed algorithm, along with the data compression and fast multi-way merge technique, implemented on the application specific hardware can achieve at least two orders of magnitude improvement in performance and energy efficiency over the available COTS architectures.

978-1-5090-3525-0/16/$31.00 ©2016 IEEE