Abstract

Sparse Basic Linear Algebra Subroutines (Sparse BLAS) is an important library. Sparse BLAS includes three levels of subroutines. Level 1, Level2 and Level 3 Sparse BLAS routines. Level 1 Sparse BLAS routines do computations over sparse vector and spare/dense vector. Level 2 deals with sparse matrix and vector operations. Level 3 deals with sparse matrix and dense matrix operations. The computations of these Sparse BLAS routines on General Purpose Processors (GPPs) not only suffer from less utilization of hardware resources but also takes more compute time than the workload due to poor data locality of sparse vector/matrix storage formats.

In the literature, tremendous efforts have been put into software to improve these Sparse BLAS routines performance on GPPs. GPPs best suit for applications with high data locality, whereas Sparse BLAS routines operate on applications with less data locality hence, GPPs performance is poor. Various Custom Function Units (Hardware Accelerators) are proposed in the literature and are proved to be efficient than soft wares which tried to accelerate Sparse BLAS subroutines. Though existing hardware accelerators improved the Sparse BLAS performance compared to software Sparse BLAS routines, there is still lot of scope to improve these accelerators.

This thesis describes both the existing software and hardware software co-designs (HW/SW co-design) and identifies the limitations of these existing solutions. We propose a new sparse data representation called Sawtooth Compressed Row Storage (SCRS) and corresponding SpMV and SpMM algorithms. SCRS based SpMV and SpMM are performing better than existing software solutions. Even though SCRS based SpMV and SpMM algorithms perform better than existing solutions, they still could not reach theoretical peak performance.

The knowledge gained from the study of limitations of these existing solutions including the proposed SCRS based SpMV and SpMM is used to propose new HW/SW co-designs. Software accelerators are limited by the hardware properties of GPPs, and GPUs itself, hence, we propose HW/SW co-designs to accelerate few basic Sparse BLAS operations (SpVV and SpMV). Our proposed Parallel Sparse BLAS HW/SW co-design achieves near theoretical peak performance with reasonable hardware resources.