Question – 5

Assignment - 3

Find a published paper from an ACM or IEEE conference that discusses a novel sparse matrix format that was not covered in class. Discuss why the proposed format is superior to the CSR or CSC format. Make sure cite your sources.

Answer:

CSR Delta Unit (CSR-DU) : It is a compressed format that targets the index arrays of CSR using delta encoding. It compresses column indices and row pointers in CSR by breaking the matrix into units and applying delta encoding within each unit, using flags to store the delta values. Each unit has a control byte that flags when a new row begins within that unit, since a single unit may contain the end of one row and start of the next. The delta values are stored in a compact byte-oriented manner: if the differences are small, they take fewer bytes. By adjusting unit sizes, the format balances compression vs. loop overhead.

Advantages over CSR and CSC format:

- In CSR and CSC formats, the indices for rows and columns are explicitly recorded for non-zero elements. CSR-DU improves this by employing delta encoding for the column indices. Delta encoding preserves the variations between successive indices instead of the indices themselves. This frequently leads to reduced numerical values, which can be represented with fewer bits.
- The CSR format is typically effective for row-wise traversal and operations such as matrix-vector multiplication, while the CSC format is favored for column-wise operations. CSR-DU preserves the benefits of CSR for row-wise operations while improving performance by the utilization of delta units, which can diminish the computational burden of decoding column indices during processes such as matrix multiplication, particularly when non-zero entries are aggregated.
- The delta units in CSR-DU can enhance cache utilization. Smaller deltas can yield smaller values that are more conducive to cache efficiency, especially during processes requiring many traversals of the matrix data. This can enhance the overall efficiency of computations utilizing sparse matrices.
- Sparse matrices in scientific computing, machine learning, and graph algorithms frequently display patterns characterized by the non-uniform distribution of non-zero entries. CSR-DU can more effectively adapt to such patterns than CSR or CSC, especially when clusters of non-zero data exist, as delta encoding compresses these patterns efficiently.
- Delta encoding incorporates a phase of calculating deltas during compression and cumulative sums during decompression; the entire procedure can be improved for enhanced speed compared to conventional approaches, provided the deltas are minimal and the operations performed on them are efficient. This is especially beneficial in situations when matrices are often accessed and infrequently altered.

Source:

Optimizing Sparse Matrix-Vector Multiplication Using Index and Value Compression by Kornilios Kourtis: https://kkourt.io/papers/cf08-spmv-kkourt-pr.pdf#:~:text=CSR,compression%20ratioinnermost%20loops%20without%20branches

"Optimizing sparse matrix-vector multiplication using index and value compression," by K. Kourtis, G. I. Goumas, and N. Koziris:

https://courses.e-ce.uth.gr/CE432/voh0hmata/bibliographic%20project/papers2/SC2013%20-%20Tang%20et%20al%20-%20Accelerating%20sparse%20matrix-vector%20multiplication%20on%20GPUs%20using%20bit-representation-optimized%20schemes.pdf

Compressed Sparse Row 2 (CSR2): It is a newer variant of CSR that uses a two-dimensional tiling with padding to align data for wide SIMD vectors. CSR2 maintains the identical three arrays as CSR (row pointers, column indices, values) but restructures the nonzero elements so that each tile (sub-matrix block) contains a constant number of elements aligned with the SIMD width. This may entail inserting zero-valued values as padding to ensure that each row segment corresponds to the vector length. By judiciously selecting this parameter, CSR2 minimizes the additional space compared to CSR while facilitating efficient vector loads. The format retains a logical CSR structure; however, the "values" and "col_index" arrays are organized to prevent misaligned or strided memory access during SpMV operations utilizing vector instructions.

Advantages over CSR and CSC format:

- CSR2 is engineered to enhance cache utilization by reorganizing non-zero elements and indices to augment cache hits during matrix computations. This is accomplished by structuring data into blocks that conform to the cache size, hence minimizing cache misses and accelerating operations such as matrix-vector multiplication.
- In CSR2, the matrix is partitioned into smaller, fixed-size pieces that can be processed autonomously. This is especially advantageous for parallel processing and vectorized processes, facilitating more efficient utilization of multi-core processors and SIMD (Single Instruction, Multiple Data) instructions.
- The block structure of CSR2 promotes cache performance and increases the capacity for parallel processing. Various blocks can be concurrently processed without dependency conflicts, presenting a notable advantage over conventional CSR and CSC formats, where parallel operations may be more difficult due to irregular data distribution.
- CSR2 permits adaptability in block dimensions, which can be adjusted based on particular hardware attributes (such as cache size) or the inherent qualities of the matrix. This adjustment can result in significant performance enhancements by adjusting the equilibrium between overhead and computational efficiency.
- CSR2 minimizes the overhead linked to row pointers required in CSR for each row by storing non-zero elements in blocks. This may result in decreased memory consumption and expedited access times, especially in matrices with several rows.

- CSR2 can be particularly effective for matrices with certain sparsity patterns, such as block-diagonal matrices or matrices where non-zero elements are clustered. The block format can exploit these patterns more effectively than CSR or CSC.

Source:

https://dl.acm.org/doi/pdf/10.1145/3168818

https://www.mdpi.com/2076-3417/12/19/9812#:~:text=CVR%20The%20CVR%20format%20is,format%2C%20with%20two%20SIMD%20lanes