## Parallelism of Sparse Matrix Operations

Written by Clayton Herbst (22245091)

Focus Questions:

* What operating system did you use in your submission?
* What sparse matrix operations did you use?
* For each matrix operation implemented
  + Description of the parallelism implemented
  + Some informal reasoning about expected run-time and scalability
  + Testing results
  + Comment on the performance observed

##### Background

Matrices are the foundational structure used to represent the data around us. The computer’s ability to naturally store and manipulate this information in contiguous memory locations provides exceptional computation benefits to otherwise complicated calculations. Matrix-like data structures closely represent the relationship between various datasets, often called *vectors*. The term ‘vector’ and matrix have slight difference however are interchangeable for the most part. Both of these structures help physicians & mathematicians represent the complex systems in our world through systems of linear equations and the operations on these linear equations. Matrix computation on computer systems are often used to give good approximations of computationally intensive calculations (MIT News). In this way matrices can be used to store and represent nearly all sets of data. Other examples of applications include complex data structures such as binary trees, which can use matrix representation effectively to represent the relationship between various nodes within a tree (MIT News). Graphics rely heavily on matrices to represent pixels within a file and their respective RGB colours (MIT News). Graphical effects are the result of manipulating these pixel values using operations such as scalar matrix multiplication and matrix vector multiplication to create distortion. An understanding of the many implementations of matrices allows one to gain a greater appreciation for their importance in the modern data driven world. Any improvements gained on the computational efficiency of matrix operations are thus highly valuable to society and the focus of this report.

The apparent computing power of a computer can be increased through the effective use of shared memory between all available cores. Computer architectures traditionally consisted of single core hardware and a kernel managing the memory of various processes operating on this single core. As electronic components drastically reduced in size over time, the apparent computing power drastically increased. This phenomenon is described by Moore’s law. In the early 21st century, electrical components had reached a bottle neck in their ability to decrease in size slowing the growth in computing power. The continued search for computing power encouraged the development of computer systems containing multiple cores, relying on the allocation of processes to certain cores by the kernel and the effective management of memory by the kernel between these processors. My Apple Macbook Pro for instance has 4 processors/cores. The memory sharing between the SSD, cache and EPPROM is managed by the kernel by distributing threads for processors to execute. A thread of execution or just thread is the smallest sequence of programmed instructions that can be managed independently by a scheduler (Wikipedia Thread (Computing)). A single threaded program is a process that is executed as a single thread. Processes can consist of many threads, each thread executing its own sequence of instructions while accessing the shared memory of the process (Lecture 2). The kernel manages the sharing of memory such as to avoid memory overwrites across processes. Modern computer architectures maintain the independent execution environment of processes, however introduce the concept of threads within processes that are able to be executed concurrently and share memory in the processes stack.

>>talk about the process stack and the. Thread stack. And how shared memory works.

Through the implementation of parallel computing using the OpenMP library I have endeavoured to make use of modern computer architecture to gain efficiencies in the calculations of common matrix operations. Parallel computing … . The practical processing times of the parallelised matrix operations will be compared to the practical computation times of a traditional single threaded program. These results will be gathered using the macOS Mojave (v10.14.6) operating system, operating on a single Intel Core i5 processor with 2.3GHz processing speed and four physical cores. The hardware installed provides an L2 cache size of 256KB, L3 cache size of 6MB and 8GB of random-access memory (RAM). All results gathered and inferences made will be highly specific to the architecture used due to the nature of parallel computing and computing phenomena such as cache thrashing.

##### Sparse Matrix Data Structures

Sparse matrices are matrices where most of the elements are zeros. A matrix is generally considered sparse if less than 20 percent of its elements are non-zero. Sparse matrices introduce many complexities and inefficiencies into computer algorithms due to inflated time-complexities and space complexities.

Space complexity is the analysis of the relative amount of space required by a process relative to the size of its inputs. Storing large sparse matrices in their entirety is highly inefficient and often not feasible due to restrictions in hardware capabilities. Time complexity is the analysis of the relative time taken for a program/algorithm to execute relative to its input size. The time complexity of an algorithm and the space complexity required by the algorithm can be reduced by increasing the efficiency of subproblems within the algorithm. For sparse matrices, this can simply be achieved by compressing the sparse matrix into data structures that preserve the location of all elements however store the value of non-zero elements only. Thus, sub-operations can be performed on a reduced number of elements, improving the programs time complexity, while using less memory to store the information.

Typical sparse matrix data structures used to preserve the contents of sparse matrices are the Coordinate (COO) format, Compressed Row Storage (CRS) format and the Compressed Column Storage (CCS) format. CRS and CCS are very similar containing a subtle difference in how the order in which the non-zero elements are stored. It is thus assumed the characteristics of CRS and CCS with respect to storage and time complexities are very similar.

The Coordinate (COO) format sparse matrix data structure is the simplest compressed representation of sparse matrices. This data structure consists of three arrays the size of the number of non-zero elements. The non-zero element is stored in one array and the other two store the coordinates of the non-zero element. The space complexity of the COO data structure is , where n is the number of non-zero elements in the matrix. The coordinate format is advantageous when constructing sparse matrices, performing item-wise operations and allows for fast conversion into other sparse matrix data structures. However, improvements can be made on the storage complexity of the data structure.

The Compressed Row Storage (CRS) format and Compressed Column Storage (CCS) format provide increased non-zero element storage density. The space complexity required by the CRS data structure is , where *n* is the number of non-zero elements and *m* are the number of rows in the sparse matrix. This is achieved by storing all non-zero elements in a contiguous memory locations (), array referencing the number of non-zero elements in each row ( and an array referencing the column each non-zero element belongs to (). The trade-off of the reduced space-complexity is increased algorithm complexity due to the extra addressing step required for every scalar operation since the coordinates of the non-zero elements are not explicitly stored (Dongarra 1995). The time taken to construct the CCS data structure is also when reading from a dense matrix file format since elements cannot be read in a contiguous manner, rather need to jump ahead to fetch column elements and store elements in an erratic fashion. This data access pattern requires an interim COO data structure storage step and then reading into CCS data structure.

All three storage structures were implemented in my matrix operation command line tool in order to maximise the efficient and simple element-wise access provided by COO and make use of the unique data access patterns of CRS and CCS data structure to perform row-wise/column wise operations.

>>> Insert information about advantages and disadvantages of parallelising certain data structures.

##### Parallel Sparse Matrix Scalar Multiplication

Scalar multiplication involves the traversal of each non-zero element of a matrix once. The time complexity of this algorithm is thus , where n is the number of non-zero elements. Linear order complexity is optima. Further efficiencies can be gained by taking advantage of modern computer architectures and shared memory systems using a multithreaded approach.

I have implemented the COO sparse matrix data structure for two main reasons. Firstly, the COO sparse matrix data structure is more time efficient when converting from a dense sparse matrix format to the COO sparse matrix format and vice versa. This is due to the linear traversal of all entries and the storage of these elements in their order of discovery. The simplistic storage structure of COO results in fast conversions to and from dense sparse matrices as well as fast element wise access. Secondly the COO sparse matrix data structure efficiently preserves the location of each non-zero element and hence any zero elements as well. Data structures such as CRS and CCS rely on additional lookup step to determine the location of an element. The computational intensity of element wise scalar multiplication cannot justify the conversion costs or time cost of additional address lookups steps that are required of other data structures.

The lack of computational intensity results in very fast and efficient execution of the single threaded sparse matrix scalar multiplication program. The following graph demonstrates the execution times achieved with various sparse matrix input files sizes.

>> graph on synchronous results.

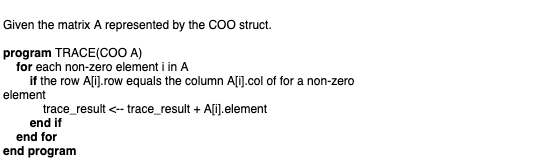
The possibility of gaining further speedup through multithreading was investigated. The OpenMP library *pragma omp parallel for* directive was used in to effectively distribute the outstanding workload amongst the available threads. The *pragma omp for* directive in particular provides useful functionality in managing the even distribution of tasks to outstanding threads; thus, limiting uneven work distribution and any implicit barriers that may be experienced as a result. The following diagram illustrates the results captured in comparison to the single threaded program.

>> graph of asynchronous results.

>>Discussion of finding in the asynchronous results.

##### Parallel Sparse Matrix Trace Calculation

The sum of all elements across the diagonal of a matrix is defined to be the trace of that matrix (Trace (Linear Algebra). The computation of a matrices trace is useful in many applications involving linear algebra and is only strongly defined for square matrices. The algorithm I have implemented to compute the trace of a given matrix traverses through all non-zero elements of the given matrix summing all elements with equivalent column and row coordinates. The resulting time complexity is , where n is the number of non-zero elements. The following pseudo code illustrates the algorithm.



As the pseudo code above suggests above, the COO sparse matrix data structure was implemented due to similar reasonings provided for sparse matrix scalar multiplication. Other sparse matrix data structures can be implemented however the added complexity in computation is unjustified for the computation required in the trace calculation. The data access pattern for elements within the COO data structure is the best fit to satisfy the algorithms sole purpose of traversing the given matrices non-zero elements and quickly determining their location.

The *pragma omp parallel for* directive was used to efficiently distribute tasks to outstanding threads. In order to ensure no race conditions between the shared memory of the trace sum between unique threads the OpenMP clause, *reductions(+:var)* was used. Due to the recurring nature of data sharing patterns such as summing values between parallel processes such as in the pseudo code provided above, the OpenMP library provides the functionality of managing any potential race conditions between threads through the *reductions(+:var)* clause. This functionality reduces the need to declare a critical section using the OpenMP directive *omp critical* or the need to perform thread specific operations within the containing for loop. The following graph represents the results achieved.

>> Insert graphical comparison between sync and async variables

>> Analysis of results captured above.

##### Parallel Sparse Matrix Addition

Matrix addition involves the element wise addition of the given matrices. Due to the nature of element wise operations, both matrices specified need to have equivalent dimensions in order to traverse all the elements. The resulting time complexity will be relative to the number of rows and columns the program has to traverse through. Thus, the time complexity expressed using Big O notation is , where n is the product of the matrix’s rows and columns (i.e the total number of matrix elements). The time complexity clearly remains linear to the size of the input matrix, the optimal time complexity of a sequential algorithm needing to traverse all elements.

>> Explain choice of data structure

>> Explain parallel code and use of openmp

>> results graph

>>Description of results.

##### Parallel Sparse Matrix Transposition

The transpose of a matrix is a matrix that has been flipped over its diagonal (Wikipedia Transpose). This operation results in the rows of the transposed matrix reflecting the columns of the original matrix. The operation of swapping matrix coordinates requires the element wise traversal of the original matrix. Considering sparse matrices and how they are represented in sparse matrix data structures, only the coordinates of all non-zero elements need to be traversed and have the element wise swapping operation performed on them. The time complexity of finding the transposed matrix is thus , where n is the number of non-zero elements represented in the sparse matrix data structure.

The COO sparse matrix data structure seems initially to be the most natural of data structure to represent the given input matrix. The explicit storage of each non-zero elements coordinate metadata is naively the most appropriate data structure for the operation. The algorithm would consist of a single loop traversing through all non-zero elements, swapping the stored column and row index values for each element. The pseudo code is given below.

A picture containing animal

Description automatically generated

**Figure 3.2: Transpose operation pseudo code**

Parallelising the algorithm described about would involve similar techniques to those implemented in the scalar multiplication operation and trace calculation. The for loop containing the swapping operation can be optimised by distributing the workload across outstanding threads. Since the operation of every inner block of code is independent from the other the algorithm is suited to incorporate multithreading techniques. The *#pragma omp parallel for* directive can hence be applied. Special consideration must be taken for the cache usage as to avoid any race conditions as this would be the most critical bottleneck of the operation. In order to avoid this, pre-processing may need to be performed in order to better prepare how the coordinate index values are stored in contiguous memory addresses.

In my implementation I have however used the CRS sparse matrix representation to store the non-zero elements of the given sparse matrix. This decision has been made in order to better from the improved storage complexity of the CRS data structure as well as take advantage of the element non-zero element access pattern of the data structure. The contiguous row-wise storage of the matrices non-zero elements allows for a niche conversion to be made between the compressed row storage data structure and the compressed column storage data structure through a single system call. The operation assumes sufficient addressable memory is available for the array of non-zero elements to be copied to the CCS data structure as well as the metadata surrounding the non-zero values. The time complexity of the algorithm remains , where n is the number of non-zero values. *Memcpy* is the system call used to copy the values of the contiguous memory locations within the CRS struct. System calls provide optimised execution times as the source code’s reliance on compiler optimisation is reduced, thus the actual time performance of the algorithm is expected to be significantly improved relative to the number of non-zero elements in the sparse matrix. Due to this system optimisation, I have avoided implementing multithreading techniques over the system call.

>> results

>> discussion of results

##### Parallel Sparse Matrix Vector Multiplication

Matrix vector multiplication or the ‘*matrix dot-product’* is a critical calculation in fields such as physics to determine the magnitude of mechanical work by finding the dot product of the force vector and the displacement vector. Matrix multiplication is also commonly used in graphics to apply distortions to graphical elements. The operation involves taking two equal-length arrays of numbers, finding the product of each element with identical indexing and summing all values in the resultant array. The result for a single operation as described above is a single number (Wikipedia Dot Product). Mathematically this can be described using two vectors *a* and *b* as:

The algorithm for the dot-product of two matrices traverses each row of the first matrix. Let represent the row vector at row index *r*. Let represent the column vector at column index *c* in the second input matrix. For each row in the first matrix, calculate the dot-product of each column in the second matrix. The scalar result is stored in the resultant matrix at the coordinates . Considering the computation of such an operation on sparse matrices, the memory access patterns of both the CRS and CCS data structures are closely aligned with the row to column comparisons made between input matrix one and two respectively. The resultant matrix can be stored in a CRS sparse matrix data structure to allow for synchronous storage of computed vector dot-products. The space complexity of such an algorithm relies on three data structures consisting of . Time complexity analysis results in a derived time complexity of , where:

* represents the number of rows in the first matrix
* represents the number of rows in the second matrix
* represents the number of columns in the second matrix

Evidently this is the most computationally intensive operation considered within this report. The fundamental structure of the algorithm involves three nested for-loops with only the inner most block containing execution dependencies. Within my implementation, the use of the CRS sparse matrix data structure results in a sequential execution dependency forcing the use of blocking operations. This is due to the CRS structures use of element ordering to implicitly determine the coordinates of the non-zero element. If the resultant matrix was instead represented by a COO data structure, where the coordinates of a non-zero element can be explicitly stated. The revised non-blocking functionality of such an algorithm introduces the concepts of partially and fully distributed memory sharing during sparse matrix vector multiplication.

The partially distributed memory sharing technique for matrix vector multiplication involves the partitioning of the first matrix row wise allowing allocated threads to yield submatrices rather than scalar quantities (Albert-Jan et al). This technique would involve creating tasks to be executed by the team of threads the size of the external for-loop block. Each submatrix created is thus stored in unique chunks of memory. A major factor to consider other than data races is the load balance between threads. In order to ensure efficient parallelisation, each thread in the team should perform similar amounts of *‘work’*. The OpenMP library directive *#pragma omp for* manages the efficient load distribution between threads in the team.

The fully distributed memory sharing construct for matrix vector introduces added complexity as the referential locality of the rows and columns is exploited (Albert-Jan et al). Modern supercomputer systems implement fully distributed algorithms that are additionally parallelised over multiple nodes. Alternative partitioning algorithms have been derived by Yzelman and Bisseling such as the partitioning in the doubly separated block diagonal (SBD) form (Albert-Jan et al). These algorithms however are designed to be implemented in high performance distributed systems in addition to local memory sharing, rather than sole local multithreading techniques. As a result I have attempted to implement partially distributed memory algorithm.

>> Insert info about my solution

>> Insert results comparison

>> Insert analysis of results.

>> resultant matrix execution order dependency provides problems when implementing multithreading

##### Appendices:

##### References:

Hardesty L., 2013. *Explained: Matrices,* MIT News. http://news.mit.edu/2013/explained-

matrices-1206. [17 Sep 2019].

Dongarra J., 1995. *Compressed Row Storage. http://netlib.org/linalg/html\_templates*

*/node91.html. [22 Sep 2019].*