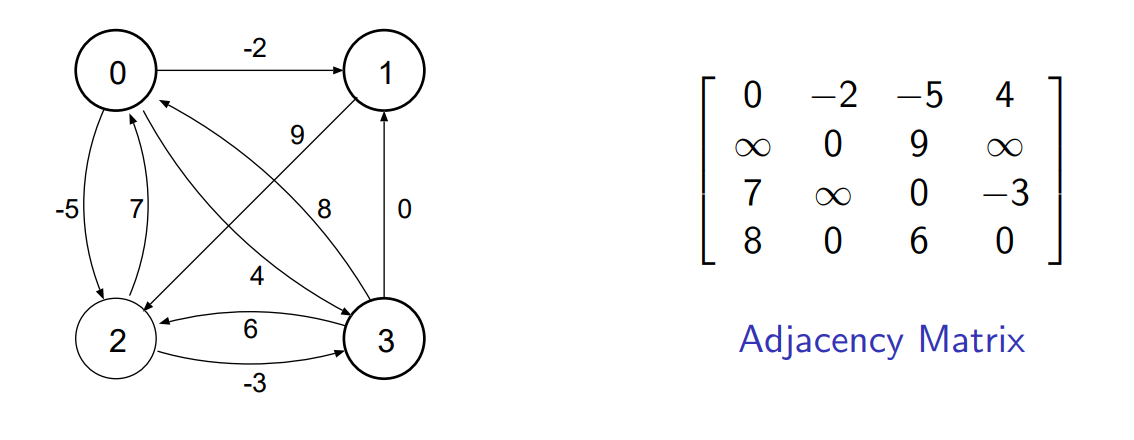
**ALL-PAIRS SHORTEST PATH USING FLOYD-WARSHALL ALGORITHM**

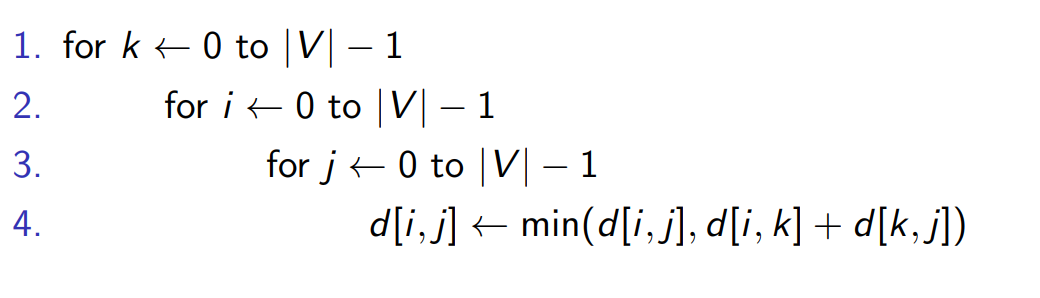
Alexander Rohl – 22233158

Farruh Mavlonov – 22252282

**Shortest Paths**

Given a weighted (positive only) directed graph, determine the shortest path between any two nodes in the graph. 

**Floyd-Warshall Algorithm**

Recursive solution based on *intermediate* vertices with the following sequential sudo code:

Sequential complexity:

**Parallel algorithms using MPI**

The most common way to write a parallel program is to use a sequential language and a subroutine library. For our project, we will be using C to write the bodies of processes and calling the MPI library for process creation, process management, communication and synchronization.

­­

**Program Architecture and OS**

For this project, we are running on the cluster **\_\_\_\_\_\_\_\_\_cluster details?\_\_\_\_\_\_\_\_\_\_\_\_**

We compile our code in terminal by running:

mpicc main\_with\_matrix.c -o mysolution.exe

We measure program speed using wall-clock time with *MPI\_Wtime().* The regular c clock function takes the sum of all runtimes for all threads and therefore doesn’t return an accurate reading of speed up via threading.

Our code consists of the files implemented on the right in *main.c*. *parse\_input* is responsible for reading the binary matrix input and *helper\_functions* consists of any subfunctions required by main.

We execute our code in terminal by running:

 mpirun -n 20 --hostfile hostfile ./mysolution.exe examples/512.in

**Sequential Approach**

*Time analysis of the sequential algorithm*

We first parse the matrix sequentially. Since there are N^2 elements to be parsed, we expect this to grow O(N^2). Over 5 runs for each matrix size, we notice the following time complexity.

Figure 1: Sequential Parsing Time Complexity Plot

The plot above shows an approximate quadratic fit. However, we can confirm this by considering a log-log plot of the data. By doing so, we see below a gradient of 1.8352, which is reasonably close to 2 and hence implies a quadratic relationship.

Figure 2: Sequential Parsing Time Complexity Log-Log Plot

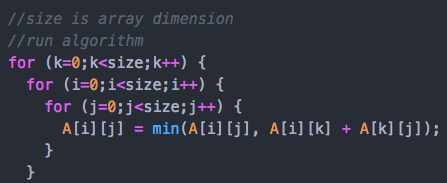
We now test the sequential Floyd-Warshall algorithm. Since we have 3 nested for loops each having ‘size’ number of iterations, we expect the time complexity to grow O(N^3).

Figure 3: Floyd-Warshall Algorithm Time Complexity Plot

Unsurprisingly, the tests closely follow a cubic trend. To confirm this, we consider a log-log plot. By doing so, we see below a gradient of 2.8633 which is reasonably close to 3 and confirms a grow rate of O(N^3).

Figure 4: Floyd-Warshall Algorithm Time Complexity Log-Log Plot

***Naïve Parallel Approach***

This approach is based on partitioning the matrix among processors and sending a copy of the entire matrix after each iteration to each of the arrays. Refer to the sudo code below, note that in the actual implementation of this, in **main\_with\_matrix.c**, the code is modified to consider cases where the number of processors does not evenly divide the number of elements in the matrix. To account for this, we give the left-over elements to the root process.

np <- number of processes

***--- MASTER PROCESS ---***

read matrix size

size <- matrix size

num\_local\_elements <- local number of elements = size/np

lo <- remaining elements after splitting = size - num\*np

read matrix

convert zeros to infinity

convert diagonals to zeroes

***----------------------***

MPI\_Bcast size

MPI\_Bcast num\_local\_elements

MPI\_Bcast lo

global\_index = np\*pid

SCATTER matrix to sub\_array

MPI\_Bcast matrix

for iteration (k) in range(size):

for local\_index in range(num\_local\_elements):

i (global row) <- (global\_index + local\_index)/size

j (global col) <- (global\_index + local\_index)%size

sub\_array[local\_index] = min(sub\_array[local\_index], matrix[i\*size+k] + matrix[k\*size+j])

GATHER sub\_array into matrix

MPI\_Bcast matrix

print matrix

Running our code on the cluster we see the following times:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | | **MPI OLD (with 1 ppn on 4 nodes)** | | | | |
|  | | MATRIX SIZE | | | | |
|  | 32 | 512 | 1024 | 2048 | 4096 |
| 1 | | 0.001078 | 1.920015 | 13.93651 | 188.09429 |  |
| 2 | | 0.001123 | 1.928698 | 13.74084 | 185.030222 |  |
| 3 | | 0.00119 | 1.928318 | 13.743335 | 190.333041 |  |
| 4 | | 0.001121 | 1.912096 | 13.533887 | 183.23194 |  |
| 5 | | 0.001059 | 1.909035 | 13.471012 | 186.294095 |  |
| **Average** | | 0.0011142 | 1.9196324 | 13.6851168 | 186.5967176 | 0 |

Note in the previous table that we had difficulties running the algorithm for matrices of size 4096. Assuming a time complexity growth rate of O(N^3), we can estimate the time this computation would take to be approximately 186.6\*2^3=1493 seconds. Since this is well beyond the cluster time-limit of 600 seconds, we were unable to test this.

Comparing this algorithm to the sequential we see:

|  |  |  |  |
| --- | --- | --- | --- |
| Speedup OLD for 4 processors | | | |
| **Number of vertices** | **Sequential time (s)** | **Parallel time (s)** | **Speedup** |
| 32 | 0.00048 | 0.00111 | 0.4265 |
| 512 | 0.97625 | 1.91963 | 0.5086 |
| 1024 | 8.12674 | 13.68512 | 0.5938 |
| 2048 | 72.69271 | 186.59672 | 0.3896 |
| 4096 | 524.43165 |  |  |

***Better Parallel Approach***

Why the method described above is “naïve” is that the entire matrix is copied and queried for each process. However, we notice that in each iteration (*k*) only the *k*th row and *k*th column is actually required. Hence, are more optimal algorithm should only share this information with each process rather than the full matrix.

In order to preserve the functionality of being able to effectively run the algorithm with the number of processors not necessarily dividing the total number of matrix elements, we consider the following method:

1. Broadcast kth row and kth column.
2. Collectively build the next (k+1)th row and (k+1)th column while the kth iteration is being computed.
3. Repeat.

Sudo Code: (items crossed out are from the naïve approach)

**~~MPI\_Bcast matrix~~**

MPI\_Bcast k\_row <- 0row

MPI\_Bcast k\_col <- 0col

for iteration (k) in range(size):

for local\_index in range(num\_local\_elements):

i (global row) <- (global\_index + local\_index)/size

j (global col) <- (global\_index + local\_index)%size

sub\_array[local\_index] = min(sub\_array[local\_index], **~~matrix[i\*size+k]~~** k\_col[i] + **~~matrix[k\*size+j]~~** k\_row[j])

if i==k+1:

(k+1)row element found

if j==k+1:

(k+1)col element found

GATHER sub\_array into matrix

**~~MPI\_Bcast matrix~~**

MPI\_Bcast k\_row <- (k+1)row

MPI\_Bcast k\_col <- (k+1)col

We see the performance of this algorithm on the cluster as follows:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **MPI ALGORITHM TIMING (with 2048 vertices)** | | | | |
| NUMBER OF NODES | 1 | 2 | 4 | 4 | 4 |
| NUMBER OF PROCESSES | 1 | 2 | 4 | 8 | 16 |
| 1 | 158.751591 | 80.36369 | 40.630325 | 21.450497 | 12.993871 |
| 2 | 159.708206 | 80.388465 | 40.348857 | 21.464516 | 12.95157 |
| 3 | 158.738431 | 80.275108 | 40.351639 | 21.47298 | 13.008058 |
| 4 | 158.64898 | 80.453915 | 40.299123 | 21.454593 | 13.23492 |
| 5 | 158.582564 | 80.29967 | 40.301254 | 21.452171 | 12.893623 |
| **Average time** | 158.8859544 | 80.3561696 | 40.3862396 | 21.4589514 | 13.0164084 |
|  |  |  |  |  |  |
|  | **MPI ALGORITHM TIMING (with 1 ppn on 4 nodes)** | | | | |
| MATRIX SIZE | 32 | 512 | 1024 | 2048 | 4096 |
| 1 | 0.000793 | 0.767922 | 5.392001 | 40.630325 | 321.570263 |
| 2 | 0.000856 | 0.769462 | 5.384408 | 40.348857 | 320.174032 |
| 3 | 0.000892 | 0.770977 | 5.372458 | 40.351639 | 334.600912 |
| 4 | 0.000878 | 0.770193 | 5.384641 | 40.299123 | 333.497945 |
| 5 | 0.000882 | 0.770851 | 5.380796 | 40.301254 | 326.014467 |
| **Average time** | 0.0008602 | 0.769881 | 5.3828608 | 40.3862396 | 327.1715238 |

More interestingly, we can plot this and compare against the previous implementations.

In the above plots, the log-log plot tells the story. By studying the gradients we see the ‘Fast MPI’ algorithm clearly surpassed the previous two algorithms. Furthermore, although at the 4096 node scale the ‘Slow MPI’ algorithm was impossible to run, the log-log plot suggests a marginally superior time complexity for larger graphs.

We also compare speedup with respect to the number of processors implemented:

The faster algorithm follows a clear power trend as we increase the number of processors at our disposal. This is reasonable as an efficient parallel program should increase the code speed *n*-fold for *n* processors. However, note the contrast to the slower algorithm which from the data we have does not follow a clear trend. This is consistent with our intuition that this algorithm is not an effective parallelisation.

**Farruh.c parallel code – when I get it working**

* Let be the time to compute an iteration
* Sequential execution time:
* Computational complexity:
  + Innermost loop:
  + Middle loop: at most
  + Outer loop:
  + Overall complexity:
* Communication complexity
  + No communication in innermost loop
  + No communication in middle loop
  + Broadcast in outer loop:
  + Number of broadcasts: (one per outer loop iteration)
  + Overall complexity:
* Broadcast time:
  + Each broadcast has steps
  + is the message latency
  + is the bandwidth
  + Each broadcast sends bytes
* Overall parallel complexity:
* Expected parralel execution time:
* Deadlock: process waiting for a condition that will never become true
  + 2 processes both receive before send
  + Send tag doesn’t match receive tag
  + Process sends message to wrong destination process
* Strategy
  + Agglomerate tasks to minimise communication
  + Create one combined task per MPI process
* Row-wise block striped decomposition
  + Broadcast within rows eliminated
  + Reading matrix from file simpler because elements in C are stored in row major order
* A master process reads all the rows and send the data to the appropriate process

**Performance Analysis of MPI Collective Operations**

Previous studies of application usage show that the performance of collective communications are critical for high performance computing (HPC) and are often overlooked when compared to the point-to-point performance. It is essential for MPI implementations to provide high-performance collective operations. The optimal implementation of a collective for a given system depends on many factors, including physical topology of the system, number of processes involved and message sizes. There are many parallel communicational models that predict performance of any given collective operation based on standardized system parameters.

**Speedup Analysis**

|  |  |  |  |
| --- | --- | --- | --- |
| **Number of processes** | **Time (s)** | **Speedup** | **Efficiency (%)** |
| 1 | 158.8860 | 0.4575 | 45.75 |
| 2 | 80.3562 | 0.9046 | 45.23 |
| 4 | 40.3862 | 1.7999 | 45.00 |
| 8 | 21.4590 | 3.3875 | 42.34 |
| 16 | 13.0164 | 5.5847 | 34.90 |

Table 1 - Speedup for graphs of 2048 vertices vs 1, 2, 4, 8, 16 processors (sequential execution time 72.6927)

|  |  |  |  |
| --- | --- | --- | --- |
| **Number of vertices** | **Sequential time (s)** | **Parallel time (s)** | **Speedup** |
| 32 | 0.00048 | 0.00086 | 0.5524 |
| 512 | 0.97625 | 0.76988 | 1.2681 |
| 1024 | 8.12674 | 5.38286 | 1.5097 |
| 2048 | 72.69271 | 40.38624 | 1.7999 |
| 4096 | 524.43165 | 327.17152 | 1.6029 |

Table 2 - Speedup for 4 processors vs graphs of 32, 512, 1024, 2048, 4096