

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ & ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ



Συστήματα Παράλληλης Επεξεργασίας

9ο Εξάμηνο, 2024-2025

Εργαστηριακή Αναφορά

των φοιτητών:

Λάζου Μαρία-Αργυρώ (el20129)

Σπηλιώτης Αθανάσιος (el20175)

Ομάδα: **parlab09**

Περιεχόμενα

1. Conway's Game of Life
 - 1.1 Υλοποίηση
 - 1.2 Αποτελέσματα Μετρήσεων
 - 1.3 Γραφική Απεικόνιση και Παρατηρήσεις
 - 1.4 Bonus
2. Παραλληλοποίηση και βελτιστοποίηση του αλγορίθμου K-means
 - 2.1 Shared Clusters
 - 2.1.1 Υλοποίηση
 - 2.1.2 Εκμετάλλευση του GOMP_CPU_AFFINITY
 - 2.2 Copied Clusters & Reduce
 - 2.2.1 Υλοποίηση
 - 2.2.2 Δοκιμές με μικρότερο dataset
 - 2.2.3 First-touch Policy
 - 2.2.4 Numa-aware initialization
 - 2.3 FLOYD WARSHALL
 - 2.3.1 Recursive
 - 2.3.1.1 Υλοποίηση
 - 2.3.1.2 Αποτελέσματα
 - 2.3.2 Tiled
 - 2.3.2.1 Υλοποίηση
 - 2.3.2.2 Αποτελέσματα
 - 2.4 Αμοιβαίος Αποκλεισμός-Κλειδώματα
 - 2.4.1 Τεχνικές συγχρονισμού
 - 2.4.2 Αποτελέσματα
 - 2.5 Ταυτόχρονες Δομές δεδομένων
 - 2.5.1 Coarse-grain locking
 - 2.5.2 Fine-grain locking
 - 2.5.3 Optimistic synchronization)
 - 2.5.4 Lazy synchronization
 - 2.5.5 Non-blocking synchronization
 - 2.5.6 Αποτελέσματα
 - 2.5.6.1 Αποτελέσματα για 100% reads, 0% updates, 0% inserts
 - 2.5.6.2 Αποτελέσματα για 80% reads, 10% updates, 10% inserts
 - 2.5.6.3 Αποτελέσματα για 20% reads, 40% updates, 40% inserts
 3. Παραλληλοποίηση και βελτιστοποίηση αλγορίθμων σε επεξεργαστές γραφικών
 - 3.1 Naive
 - 3.2 Transpose
 - 3.3 Shared
 - 3.4 Bottleneck Analysis
 - 3.5 Full-offload (All-GPU)
 - 3.6 Delta Reduction (All-GPU)
 4. Παραλληλοποίηση και βελτιστοποίηση αλγορίθμων με χρήση MPI

- 4.1 Άλλη μια παραλληλοποίηση του K-means
- 4.2 Heat transfer
 - 4.2.1 Jacobi
 - 4.2.2 Gauss-Seidel
 - 4.2.3 Red-Black SOR
 - 4.2.3.1 Blocking version (Sendrecv)
 - 4.2.3.1 Non-Blocking version (Isend/Irecv) και ανταλλαγή $N/2$ σημείων
 - 4.2.3.1 Overlapping version
 - 4.2.4 Συνολική Μελέτη επιδόσεων όλων των μεθόδων
 - 4.2.4.1 Σενάριο Σύγκλισης
 - 4.2.4.2 Σενάριο σταθερού αριθμού επαναλήψεων $T = 256$

Conway's GameofLife

Υλοποίηση

Για την παραλληλοποίηση του αλγορίθμου τροποποίησαμε τον κώδικα που δίνεται προσθέτοντας απλώς το #pragma directive στο κύριο loop για τα (i,j) του body:

Game_Of_Life.c

```
1  ****
2  ***** Conway's game of life ****
3  ****
4
5  Usage: ./exec ArraySize TimeSteps
6
7  Compile with -DOUTPUT to print output in output.gif
8  (You will need ImageMagick for that - Install with
9  sudo apt-get install imagemagick)
10 WARNING: Do not print output for large array sizes!
11 or multiple time steps!
12 ****
13
14
15 #include <stdio.h>
16 #include <stdlib.h>
17 #include <sys/time.h>
18
19 #define FINALIZE \
20 convert -delay 20 `ls -1 out*.pgm | sort -V` output.gif\n\
21 rm *pgm\n\
22 "
23
24 int ** allocate_array(int N);
25 void free_array(int ** array, int N);
26 void init_random(int ** array1, int ** array2, int N);
27 void print_to_pgm( int ** array, int N, int t );
28
29 int main (int argc, char * argv[]) {
30     int N;           //array dimensions
31     int T;           //time steps
32     int ** current, ** previous; //arrays - one for current timestep, one for previous timestep
33     int ** swap;      //array pointer
34     int t, i, j, nbrs; //helper variables
35
36     double time;      //variables for timing
37     struct timeval ts,tf;
38
39     /*Read input arguments*/
40     if ( argc != 3 ) {
41         fprintf(stderr, "Usage: ./exec ArraySize TimeSteps\n");
42         exit(-1);
43     }
44     else {
45         N = atoi(argv[1]);
46         T = atoi(argv[2]);
47     }
48
49     /*Allocate and initialize matrices*/
50     current = allocate_array(N);      //allocate array for current time step
51     previous = allocate_array(N);    //allocate array for previous time step
52
53     init_random(previous, current, N); //initialize previous array with pattern
54
55     #ifdef OUTPUT
56     print_to_pgm(previous, N, 0);
57     #endif
58
59     /*Game of Life*/
60
61     gettimeofday(&ts,NULL);
62     for ( t = 0 ; t < T ; t++ ) {
63         #pragma omp parallel for shared(current, previous) private (nbrs, i, j)
64         for ( i = 1 ; i < N-1 ; i++ ) {
```

```

65     for ( j = 1 ; j < N-1 ; j++ ) {
66         nbrs = previous[i+1][j+1] + previous[i+1][j] + previous[i+1][j-1] \
67             + previous[i][j-1] + previous[i][j+1] \
68             + previous[i-1][j-1] + previous[i-1][j] + previous[i-1][j+1];
69         if ( nbrs == 3 || ( previous[i][j]+nbrs ==3 ) )
70             current[i][j]=1;
71         else
72             current[i][j]=0;
73     }
74 }
75
76 #ifdef OUTPUT
77 print_to_pgm(current, N, t+1);
78#endif
79 //Swap current array with previous array
80 swap=current;
81 current=previous;
82 previous=swap;
83 }
84 gettimeofday(&tf,NULL);
85 time=(tf.tv_sec-ts.tv_sec)+(tf.tv_usec-ts.tv_usec)*0.000001;
86
87 free_array(current, N);
88 free_array(previous, N);
89 printf("GameOfLife: Size %d Steps %d Time %lf\n", N, T, time);
90 #ifdef OUTPUT
91 system(FINALIZE);
92#endif
93 }
94
95 int ** allocate_array(int N) {
96     int ** array;
97     int i,j;
98     array = malloc(N * sizeof(int*));
99     for ( i = 0; i < N ; i++ )
100        array[i] = malloc( N * sizeof(int));
101        for ( i = 0; i < N ; i++ )
102            for ( j = 0; j < N ; j++ )
103                array[i][j] = 0;
104    return array;
105 }
106
107 void free_array(int ** array, int N) {
108     int i;
109     for ( i = 0 ; i < N ; i++ )
110         free(array[i]);
111     free(array);
112 }
113
114 void init_random(int ** array1, int ** array2, int N) {
115     int i,pos,x,y;
116
117     for ( i = 0 ; i < (N * N)/10 ; i++ ) {
118         pos = rand() % ((N-2)*(N-2));
119         array1[pos%(N-2)+1][pos/(N-2)+1] = 1;
120         array2[pos%(N-2)+1][pos/(N-2)+1] = 1;
121
122     }
123 }
124
125 void print_to_pgm(int ** array, int N, int t) {
126     int i,j;
127     char * s = malloc(30*sizeof(char));
128     sprintf(s,"out%d.pgm",t);
129     FILE * f = fopen(s,"wb");
130     fprintf(f, "P5\n%d %d 1\n", N,N);
131     for ( i = 0; i < N ; i++ )
132         for ( j = 0; j < N ; j++)
133             if ( array[i][j]==1 )
134                 fputc(1,f);
135             else
136                 fputc(0,f);
137     fclose(f);
138 }
```

```
139     free(s);  
140 }
```

Για την μεταγλώττιση και εκτέλεση στον scirouter χρησιμοποίησαμε το ακόλουθα scripts :

```
#!/bin/bash  
## Give the Job a descriptive name  
#PBS -N make_gameoflife  
## Output and error files  
#PBS -o make_gameoflife.out  
#PBS -e make_gameoflife.err  
## How many machines should we get?  
#PBS -l nodes=1:ppn=1  
## Start  
## Run make in the src folder (modify properly)  
module load openmpi/1.8.3  
cd /home/parallel/parlab09/a1  
make
```

```
#!/bin/bash  
## Give the Job a descriptive name  
#PBS -N run_gameoflife  
## Output and error files  
#PBS -o omp_gameoflife_all.out  
#PBS -e omp_gameoflife_all.err  
## Limit memory, runtime etc.  
#PBS -l walltime=01:00:00  
##Number of nodes aka threads  
#PBS -l nodes=1:ppn=8  
module load openmpi/1.8.3  
cd /home/parallel/parlab09/a1  
for threads in 1 2 4 6 8  
do  
export OMP_NUM_THREADS=$threads  
echo "Running with OMP_NUM_THREADS=$OMP_NUM_THREADS"  
.omp_gameoflife 64 1000  
.omp_gameoflife 1024 1000  
.omp_gameoflife 4096 1000  
echo "Finished run with OMP_NUM_THREADS=$OMP_NUM_THREADS"  
echo "-----"  
done
```

Αποτελέσματα Μετρήσεων:

```
Running with OMP_NUM_THREADS=1  
GameOfLife: Size 64 Steps 1000 Time 0.023112  
GameOfLife: Size 1024 Steps 1000 Time 10.965944  
GameOfLife: Size 4096 Steps 1000 Time 175.900314  
Finished run with OMP_NUM_THREADS=1
```

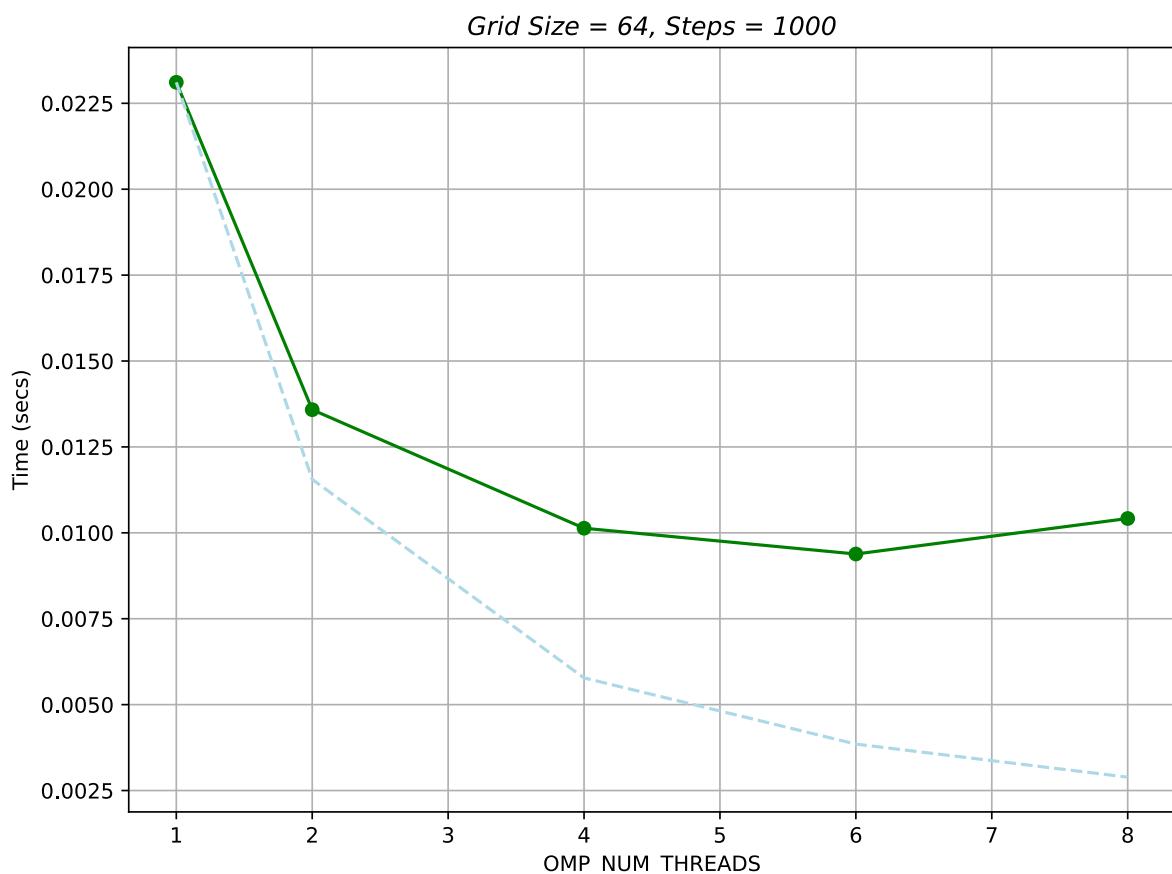
```
-----  
Running with OMP_NUM_THREADS=2  
GameOfLife: Size 64 Steps 1000 Time 0.013583  
GameOfLife: Size 1024 Steps 1000 Time 5.458949  
GameOfLife: Size 4096 Steps 1000 Time 88.263665  
Finished run with OMP_NUM_THREADS=2
```

```
-----  
Running with OMP_NUM_THREADS=4  
GameOfLife: Size 64 Steps 1000 Time 0.010134  
GameOfLife: Size 1024 Steps 1000 Time 2.723798  
GameOfLife: Size 4096 Steps 1000 Time 45.901567  
Finished run with OMP_NUM_THREADS=4  
-----
```

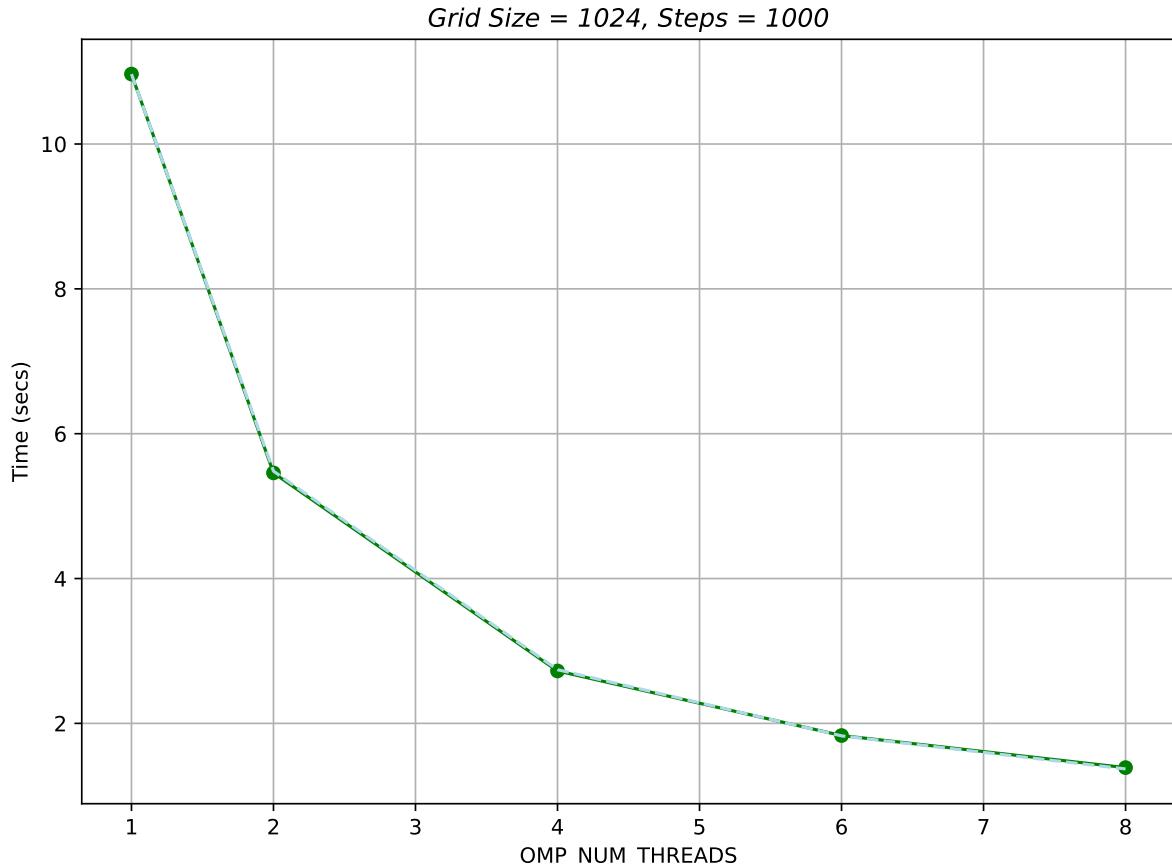
```
Running with OMP_NUM_THREADS=6  
GameOfLife: Size 64 Steps 1000 Time 0.009383  
GameOfLife: Size 1024 Steps 1000 Time 1.832227  
GameOfLife: Size 4096 Steps 1000 Time 43.661123  
Finished run with OMP_NUM_THREADS=6  
-----
```

```
Running with OMP_NUM_THREADS=8  
GameOfLife: Size 64 Steps 1000 Time 0.010417  
GameOfLife: Size 1024 Steps 1000 Time 1.389175  
GameOfLife: Size 4096 Steps 1000 Time 43.186379  
Finished run with OMP_NUM_THREADS=8  
-----
```

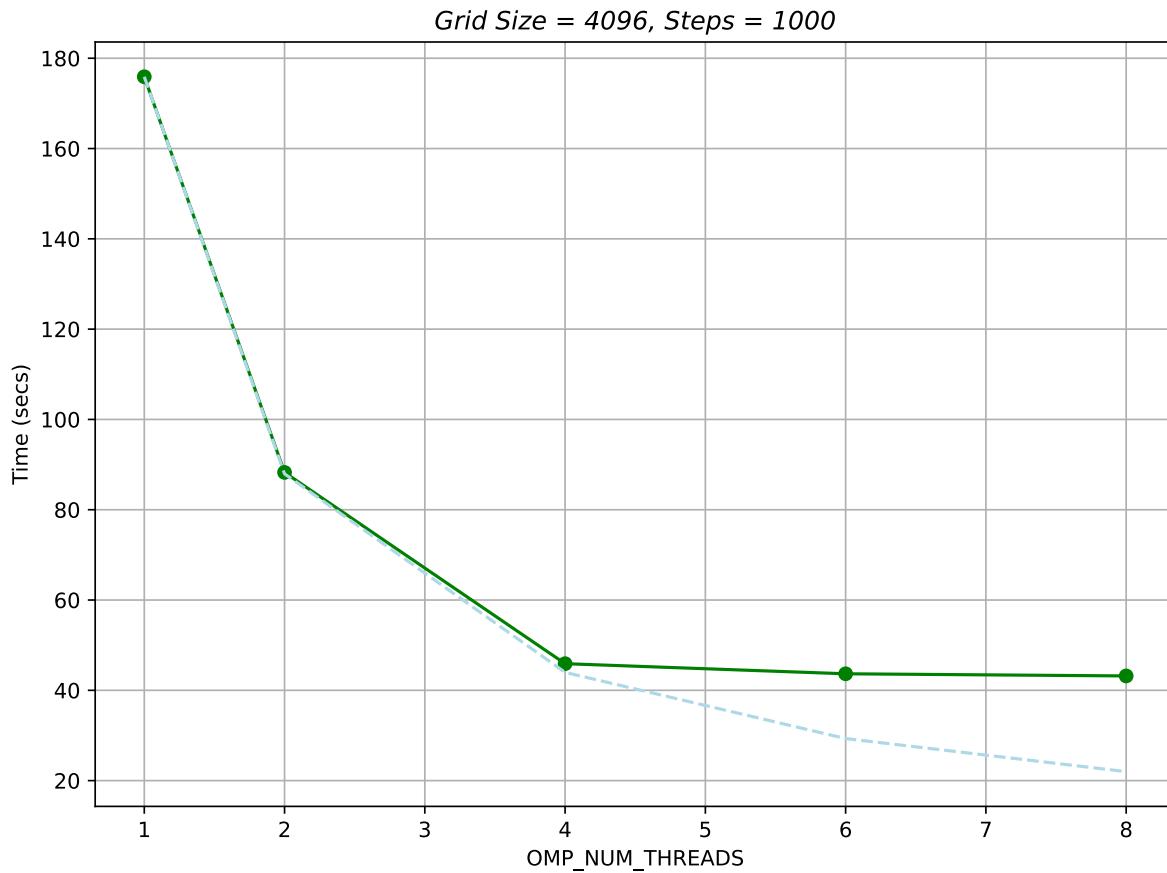
Γραφική Απεικόνιση και Παρατηρήσεις



Παρατηρούμε ότι για μικρό μέγεθος grid (με συνολική απαίτηση μνήμης $4*64*64\text{bytes} = 16\text{KB}$), δεν υπάρχει ομοιόμορφη κλιμάκωση της επίδοσης με αύξηση των νημάτων από 4 και πάνω. Bottleneck κόστους θα θεωρήσουμε την ανάγκη συγχρονισμού των threads και το overhead της δημιουργίας τους συγκρυτικά με τον φόρτο εργασίας που τους ανατίθεται (granularity).



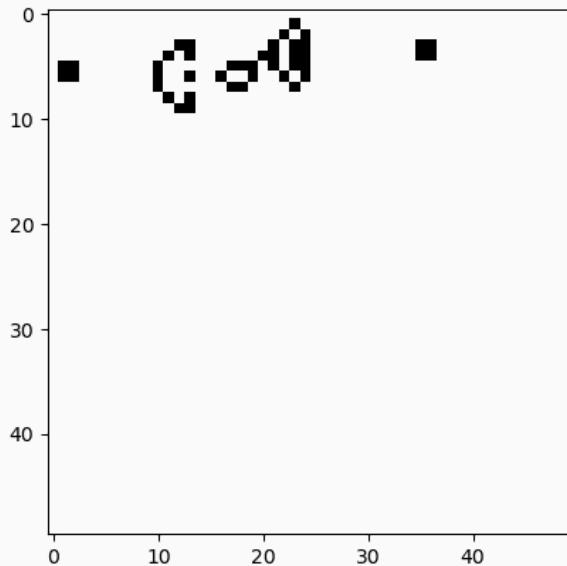
Για μέγεθος grid με συνολική απαίτηση μνήμης $4*1024*1024 \text{ bytes} = 4\text{MB}$, η επίδοση βελτίωνεται ομοιόμορφα και ανάλογα με το μέγεθος των νημάτων . Εικάζουμε, λοιπόν, πως η cache χωράει ολόκληρο το grid ώστε το κάθε νήμα να μην επιβαρύνει την μνήμη με loads των αντίστοιχων rows, ο φόρτος εργασίας είναι ισομοιρασμένος στους workers και το κόστος επικοινωνίας αμελητέο. Συνεπώς, προκύπτει perfect scaling.



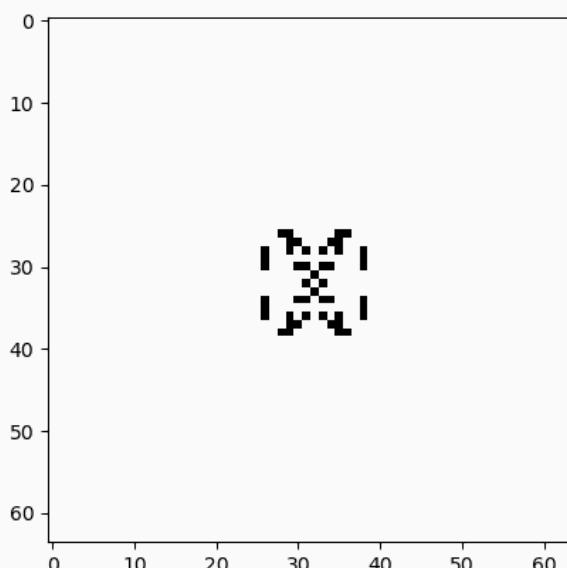
Για μεγάλο grid (με συνολική απαίτηση μνήμης $4 * 4096 * 4096$ bytes = 64MB), η κλιμάκωση παύει να υφίσταται για περισσότερα από 4 νήματα. Bottleneck κόστους εδώ θεωρούμε το memory bandwidth. Επειδή ολόκληρο το grid δεν χωράει στην cache, δημιουργούνται misses όταν ξεχωριστά νήματα προσπαθούν να διαβάσουν ξεχωριστές γραμμές του previous. Σε κάθε memory request αδειάζουν χρήσιμα data για άλλα νήματα, φέρνοντας τις δικές τους γραμμές και στο μεταξύ οι υπολογισμοί stall-άρουν.

Bonus

Δύο ενδιαφέρουσες ειδικές αρχικοποιήσεις του ταμπλό είναι το pulse και το gosper glider gun, για τις οποίες η εξέλιξη των γενιών σε μορφή κινούμενης εικόνας φαίνεται με μορφή gif παρακάτω:



glider_gun animation



pulse animation

KMEANS

1) Shared Clusters

Υλοποίηση

Για την παραλληλοποίηση της συγκεκριμένης έκδοσης χρησιμοποιήσαμε το parallel for directive του omp και για την αποφυγή race conditions τα omp atomic directives. Αυτά εμφανίζονται όταν περισσότερα από 1 νήματα προσπαθούν να ανανεώσουν τιμές στους shared πίνακες newClusters και newClusterSize σε indexes, τα οποία δεν είναι μοναδικά για το καθένα, καθώς και στην shared μεταβλητή delta. Για αυτήν, προσφέρεται η χρήση reduction και εδώ μπορεί να αγνοηθεί εντελώς, αφού η σύγκλιση του αλγορίθμου καθορίζεται από τον πολύ μικρό αριθμό των επαναλήψεων(10). Ωστόσο, χρησιμοποιούμε atomic για ορθότητα της τιμής του και για παρατήρηση με βάση το μεγαλύτερο δυνατό overhead.

```
omp_naive_kmeans.c

1 #include <stdio.h>
2 #include <stdlib.h>
3 #include "kmeans.h"
4 /*
5  * TODO: include openmp header file
6  */
7
8 // square of Euclid distance between two multi-dimensional points
9 inline static double euclid_dist_2(int      numdims, /* no. dimensions */
10                                double * coord1,   /* [numdims] */
11                                double * coord2)   /* [numdims] */
12 {
13     int i;
14     double ans = 0.0;
15
16     for(i=0; i<numdims; i++)
17         ans += (coord1[i]-coord2[i]) * (coord1[i]-coord2[i]);
18
19     return ans;
20 }
21
22 inline static int find_nearest_cluster(int      numClusters, /* no. clusters */
23                                       int      numCoords, /* no. coordinates */
24                                       double * object,    /* [numCoords] */
25                                       double * clusters) /* [numClusters][numCoords] */
26 {
27     int index, i;
28     double dist, min_dist;
29
30     // find the cluster id that has min distance to object
31     index = 0;
32     min_dist = euclid_dist_2(numCoords, object, clusters);
33
34     for(i=1; i<numClusters; i++) {
35         dist = euclid_dist_2(numCoords, object, &clusters[i*numCoords]);
36         // no need square root
37         if (dist < min_dist) { // find the min and its array index
38             min_dist = dist;
39             index   = i;
40         }
41     }
42     return index;
43 }
44
45 void kmeans(double * objects,           /* in: [numObjs][numCoords] */
46             int      numCoords,        /* no. coordinates */
47             int      numObjs,          /* no. objects */
48             int      numClusters,      /* no. clusters */
49             double   threshold,        /* minimum fraction of objects that change membership */
50             long    loop_threshold,    /* maximum number of iterations */
51             int     * membership,      /* out: [numObjs] */
```

```

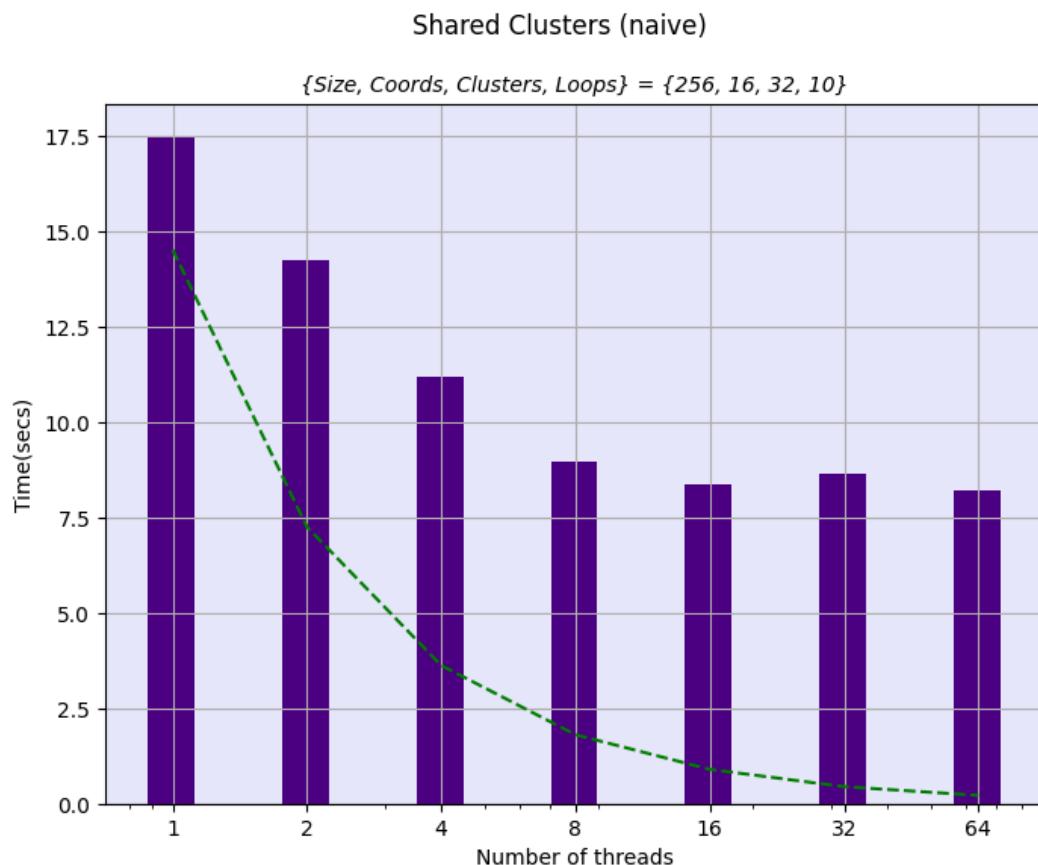
52     double * clusters)          /* out: [numClusters][numCoords] */
53 {
54     int i, j;
55     int index, loop=0;
56     double timing = 0;
57
58     double delta;           // fraction of objects whose clusters change in each loop
59     int * newClusterSize; // [numClusters]: no. objects assigned in each new cluster
60     double * newClusters; // [numClusters][numCoords]
61     int nthreads;         // no. threads
62
63     nthreads = omp_get_max_threads();
64     printf("OpenMP Kmeans - Naive\t(number of threads: %d)\n", nthreads);
65
66     // initialize membership
67     for (i=0; i<numObjs; i++)
68         membership[i] = -1;
69
70     // initialize newClusterSize and newClusters to all 0
71     newClusterSize = (typeof(newClusterSize)) calloc(numClusters, sizeof(*newClusterSize));
72     newClusters = (typeof(newClusters)) calloc(numClusters * numCoords, sizeof(*newClusters));
73
74     timing = wtime();
75
76     do {
77         // before each loop, set cluster data to 0
78         for (i=0; i<numClusters; i++) {
79             for (j=0; j<numCoords; j++)
80                 newClusters[i*numCoords + j] = 0.0;
81             newClusterSize[i] = 0;
82         }
83
84         delta = 0.0;
85
86         /*
87          * TODO: Detect parallelizable region and use appropriate OpenMP pragmas
88         */
89         #pragma omp parallel for private(i, j, index) shared(newClusters, newClusterSize,
90         membership) schedule(static)
91         for (i=0; i<numObjs; i++) {
92             // find the array index of nearest cluster center
93             index = find_nearest_cluster(numClusters, numCoords, &objects[i*numCoords], clusters);
94
95             // if membership changes, increase delta by 1
96             if (membership[i] != index)
97                 #pragma omp atomic
98                 delta += 1.0;
99
100            // assign the membership to object i
101            membership[i] = index;
102
103            // update new cluster centers : sum of objects located within
104            /*
105             * TODO: protect update on shared "newClusterSize" array
106             */
107            #pragma omp atomic
108            newClusterSize[index]++;
109            for (j=0; j<numCoords; j++)
110                /*
111                 * TODO: protect update on shared "newClusters" array
112                 */
113                #pragma omp atomic
114                newClusters[index*numCoords + j] += objects[i*numCoords + j];
115
116            // average the sum and replace old cluster centers with newClusters
117            // #pragma omp parallel for private(i,j)
118            for (i=0; i<numClusters; i++) {
119                if (newClusterSize[i] > 0) {
120                    for (j=0; j<numCoords; j++)
121                        clusters[i*numCoords + j] = newClusters[i*numCoords + j] / newClusterSize[i];
122                }
123            }
124        }
125

```

```

126 // Get fraction of objects whose membership changed during this loop. This is used as a
127 convergence criterion.
128     delta /= numObjs;
129
130     loop++;
131     printf("\r\ntcompleted loop %d", loop);
132     fflush(stdout);
133 } while (delta > threshold && loop < loop_threshold);
134 timing = wtime() - timing;
135     printf("\n          nloops = %3d    (total = %7.4fs)  (per loop = %7.4fs)\n", loop, timing,
136 timing/loop);
137
138     free(newClusters);
139     free(newClusterSize);
140 }
```

Απεικονίζουμε παρακάτω τα αποτελέσματα των δοκιμών στον sandman για τις διάφορες τιμές της environmental variable OMP_NUM_THREADS:



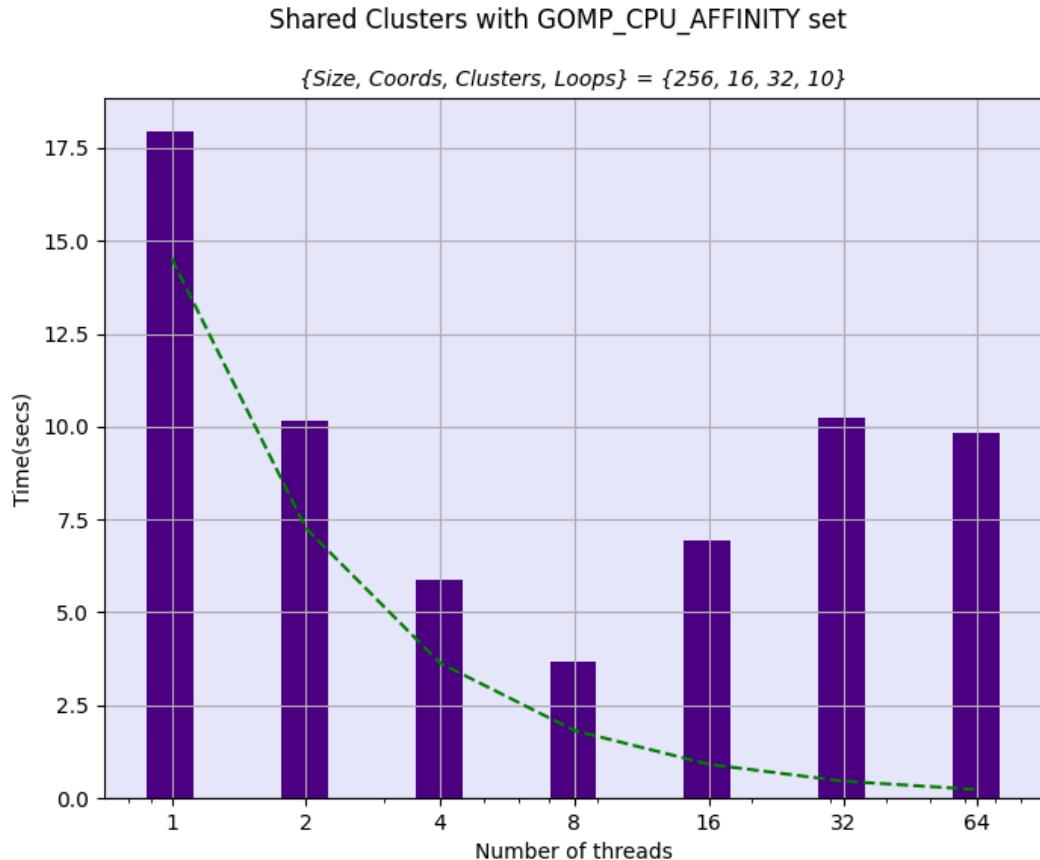
Παρατηρούμε πως ο αλγόριθμος δεν κλιμακώνει καθόλου καλά από 8 και πάνω νήματα εξαιτείας της σειριποίησης των εγγραφών, ολοένα και περισσότερων νημάτων που επιβάλλει η omp atomic, και της αυξανόμενης συμφόρησης στο bus κατά την απόκτηση του lock.

Εκμετάλλευση του GOMP_CPU_AFFINITY

Με την χρήση του environmental variable GOMP_CPU_AFFINITY και στατικό scheduling κάνουμε ριν νήματα σε πυρήνες(εφόσον δεν υπάρχει ανάγκη για περίπλοκη δυναμική δρομολόγηση). Έτσι, δεν σπαταλάται καθόλου χρόνος σε flash πυρήνων και αχρείαστη μεταφορά δεδομένων από πυρήνα σε άλλον.

Για την υλοποίηση τροποποίησαμε κατάλληλα το script υποβολής στον sandman και προσθέσαμε την παράμετρο **schedule (static)** στο parallel for.

Αποτελέσματα



Παρατηρούμε σημαντική βελτίωση στην κλιμάκωση μέχρι 8 νήματα, όμως μετά σταματάει να κλιμακώνει ο αλγόριθμος λόγω της δομής που έχει ο sandman. Για 16 νήματα και πάνω δεν μπορούμε να τα κάνουμε ριν στο ίδιο cluster, οπότε δεν μοιράζονται τα νήματα την ίδια L3 cache και υπάρχει συνεχής μεταφορά δεδομένων των shared πινάκων και bus invalidations λόγω του cache coherence protocol. Ακόμη τα L3 misses κοστίζουν ξεχωριστά για κάθε cluster. Εαν αξιοποιήσουμε το hyperthreading και κάνουμε ριν τα threads 8-15 στους cores 32-39 που πέφτουν μέσα στο cluster 1, μπορούμε να μειώσουμε σημαντικά τον χρόνο για τα 16 νήματα. Από εκεί και πέρα η κλιμάκωση σταματάει. Παραθέτουμε το τελικό script υποβολής ακολούθως:

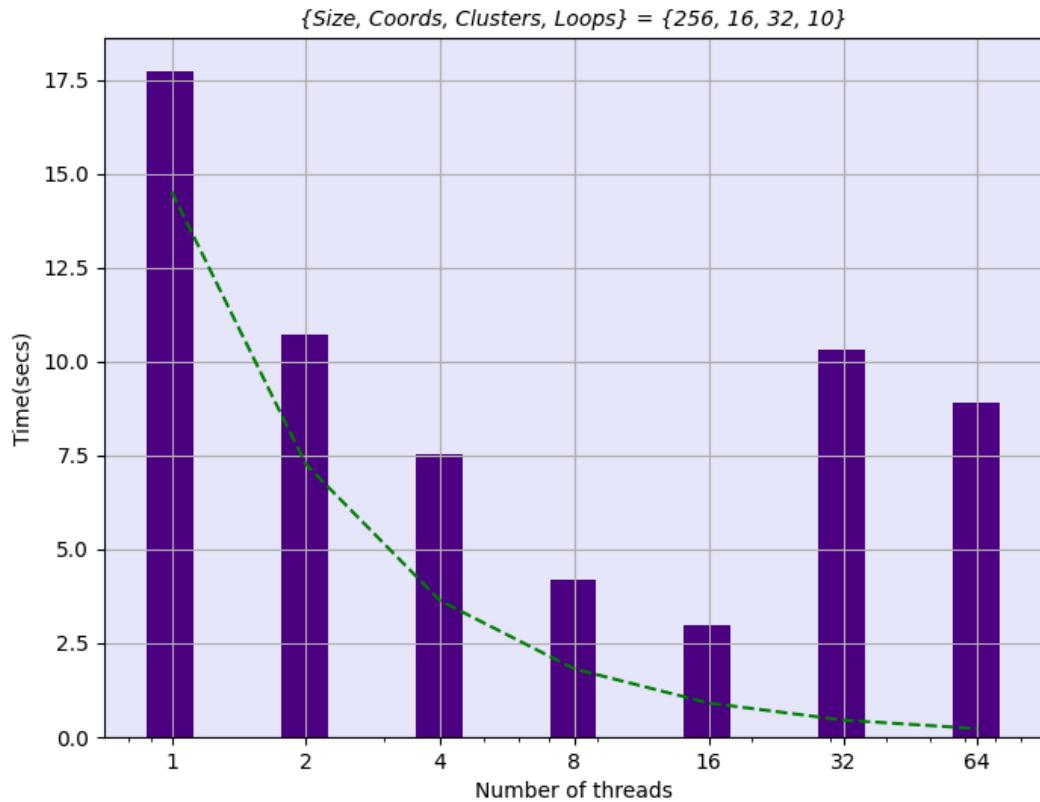
```

#!/bin/bash
## Give the Job a descriptive name
#PBS -N run_kmeans
## Output and error files
#PBS -o gomp_hyper_kmeans.out
#PBS -e gomp_hyper_kmeans.err
## How many machines should we get?
#PBS -l nodes=1:ppn=8
##How long should the job run for?
#PBS -l walltime=00:10:00
## Start
## Run make in the src folder (modify properly)
module load openmp
cd /home/parallel/parlab09/kmeans
Size=256
Coords=16
Clusters=32
Loops=10
for i in 1 2 4 8 16 32 64; do
    export OMP_NUM_THREADS=$i
    if [[ $i -eq 16 ]]; then
        export GOMP_CPU_AFFINITY="$(seq -s, 0 7),$(seq -s, 32 39)"
    else
        export GOMP_CPU_AFFINITY="$(seq -s, 0 $((i - 1)))"
    fi
done
./kmeans_omp_naive -s $Size -n $Coords -c $Clusters -l $Loops
done

```

Αποτελέσματα

Shared Clusters with GOMP_CPU_AFFINITY[0-7][32-40]



2) Copied Clusters & Reduce

Υλοποίηση

Μοιράζουμε σε κάθε νήμα ένα διαφορετικό τμήμα των πινάκων newClusters, newClusterSize, οπότε τα δεδομένα γίνονται private, δεν υπάρχουν race conditions αλλά απαιτείται reduction (με πρόσθεση) στο τέλος για το τελικό αποτέλεσμα (η οποία πραγματοποιείται εδώ από 1 νήμα).

```
omp_reduction_kmeans.c

1 #include <stdio.h>
2 #include <stdlib.h>
3 #include "kmeans.h"
4 /*
5  * TODO: include openmp header file
6  */
7
8 // square of Euclid distance between two multi-dimensional points
9 inline static double euclid_dist_2(int      numdims, /* no. dimensions */
10                                double * coord1,   /* [numdims] */
11                                double * coord2)  /* [numdims] */
12{
13     int i;
14     double ans = 0.0;
15
16     for(i=0; i<numdims; i++)
17         ans += (coord1[i]-coord2[i]) * (coord1[i]-coord2[i]);
18
19     return ans;
20 }
21
22 inline static int find_nearest_cluster(int      numClusters, /* no. clusters */
23                                       int      numCoords,    /* no. coordinates */
24                                       double * object,      /* [numCoords] */
25                                       double * clusters)   /* [numClusters][numCoords] */
26{
27     int index, i;
28     double dist, min_dist;
29
30     // find the cluster id that has min distance to object
31     index = 0;
32     min_dist = euclid_dist_2(numCoords, object, clusters);
33
34     for(i=1; i<numClusters; i++)
35         dist = euclid_dist_2(numCoords, object, &clusters[i*numCoords]);
36     // no need square root
37     if (dist < min_dist) { // find the min and its array index
38         min_dist = dist;
39         index = i;
40     }
41 }
42
43     return index;
44 }
45
46 void kmeans(double * objects,           /* in: [numObjs][numCoords] */
47             int      numCoords,        /* no. coordinates */
48             int      numObjs,          /* no. objects */
49             int      numClusters,       /* no. clusters */
50             double   threshold,        /* minimum fraction of objects that change membership */
51             long     loop_threshold,   /* maximum number of iterations */
52             int      * membership,     /* out: [numObjs] */
53             double   * clusters)       /* out: [numClusters][numCoords] */
54{
55     int i, j, k;
56     int index, loop=0;
57     double timing = 0;
58
59     double delta;           // fraction of objects whose clusters change in each loop
60     int * newClusterSize; // [numClusters]: no. objects assigned in each new cluster
61     double * newClusters; // [numClusters][numCoords]
62     int nthreads;          // no. threads
63
64     nthreads = omp_get_max_threads();
```

```

64  printf("OpenMP Kmeans - Reduction\t(number of threads: %d)\n", nthreads);
65
66 // initialize membership
67 for (i=0; i<numObjs; i++)
68     membership[i] = -1;
69
70 // initialize newClusterSize and newClusters to all 0
71 newClusterSize = (typeof(newClusterSize)) malloc(numClusters, sizeof(*newClusterSize));
72 newClusters = (typeof(newClusters)) malloc(numClusters * numCoords, sizeof(*newClusters));
73
74 // Each thread calculates new centers using a private space. After that, thread 0 does an
75 // array reduction on them.
76 int * local_newClusterSize[nthreads]; // [nthreads][numClusters]
77 double * local_newClusters[nthreads]; // [nthreads][numClusters][numCoords]
78
79 /*
80  * Hint for false-sharing
81  * This is noticed when numCoords is low (and neighboring local_newClusters exist close to
82  * each other).
83  * Allocate local cluster data with a "first-touch" policy.
84  */
85 // Initialize local (per-thread) arrays (and later collect result on global arrays)
86 for (k=0; k<nthreads; k++)
87 {
88     local_newClusterSize[k] = (typeof(*local_newClusterSize)) malloc(numClusters,
89     sizeof(**local_newClusterSize));
90     local_newClusters[k] = (typeof(*local_newClusters)) malloc(numClusters * numCoords,
91     sizeof(**local_newClusters));
92 }
93
94 timing = wtime();
95 do {
96     /* before each loop, set cluster data to 0
97     // #pragma omp parallel for private(i,j)
98     for (i=0; i<numClusters; i++) {
99         for (j=0; j<numCoords; j++)
100             newClusters[i*numCoords + j] = 0.0;
101         newClusterSize[i] = 0;
102     }
103
104     delta = 0.0;
105
106     /* TODO: Initiliaze local cluster data to zero (separate for each thread)
107     */
108
109     #pragma omp parallel for private(k, i, j) shared(local_newClusters, local_newClusterSize)
110     schedule(static)
111     for (k=0; k<nthreads; ++k){
112         for (i=0; i<numClusters; i++) {
113             for (j=0; j<numCoords; j++)
114                 local_newClusters[k][i*numCoords + j] = 0.0;
115             local_newClusterSize[k][i] = 0;
116         }
117     }
118
119     int thread_id;
120
121     #pragma omp parallel for private(i, j, thread_id, index) shared(local_newClusters,
122     local_newClusterSize) reduction(+:delta) schedule(static)
123     for (i=0; i<numObjs; i++)
124     {
125         thread_id = omp_get_thread_num();
126
127         // find the array index of nearest cluster center
128         index = find_nearest_cluster(numClusters, numCoords, &objects[i*numCoords], clusters);
129
130         // if membership changes, increase delta by 1
131         if (membership[i] != index)
132             delta += 1.0;
133
134         // assign the membership to object i
135         membership[i] = index;
136
137     }
138 }
```

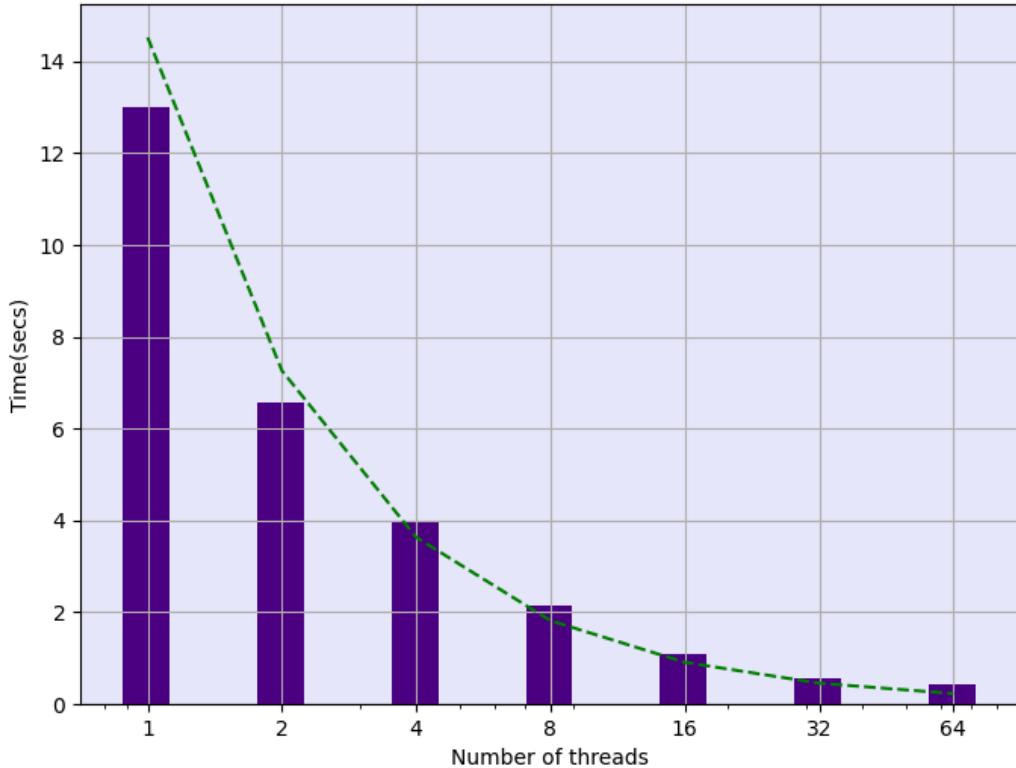
```

131     // update new cluster centers : sum of all objects located within (average will be
132     // performed later)
133     /*
134      * TODO: Collect cluster data in local arrays (local to each thread)
135      * Replace global arrays with local per-thread
136      */
137
138     local_newClusterSize[thread_id][index]++;
139     for (j=0; j<numCoords; j++)
140         local_newClusters[thread_id][index*numCoords + j] += objects[i*numCoords + j];
141 }
142 /*
143  * TODO: Reduction of cluster data from local arrays to shared.
144  * This operation will be performed by one thread
145 */
146
147 for (i=0; i<numClusters; ++i){
148     for (k=0; k<nthreads; ++k){
149         newClusterSize[i] += local_newClusterSize[k][i];
150         for (j=0; j<numCoords; ++j)
151             newClusters[i*numCoords+j] += local_newClusters[k][i*numCoords+j];
152     }
153 }
154
155
156 // average the sum and replace old cluster centers with newClusters
157 // #pragma omp parallel for private(i,j)
158 for (i=0; i<numClusters; i++) {
159     if (newClusterSize[i] > 0) {
160         for (j=0; j<numCoords; j++) {
161             clusters[i*numCoords + j] = newClusters[i*numCoords + j] / newClusterSize[i];
162         }
163     }
164 }
165
166 // Get fraction of objects whose membership changed during this loop. This is used as a
167 // convergence criterion.
168 delta /= numObjs;
169
170 loop++;
171 printf("\r\ntcompleted loop %d", loop);
172 fflush(stdout);
173 } while (delta > threshold && loop < loop_threshold);
174 timing = wtime() - timing;
175 printf("\n          nloops = %3d    (total = %7.4fs)  (per loop = %7.4fs)\n", loop, timing,
176 timing/loop);
177
178 for (k=0; k<nthreads; k++)
179 {
180     free(local_newClusterSize[k]);
181     free(local_newClusters[k]);
182 }
183 free(newClusters);
184 free(newClusterSize);
185 }
```

Αποτελέσματα

Copied Clusters & Reduction

$\{Size, Coords, Clusters, Loops\} = \{256, 16, 32, 10\}$

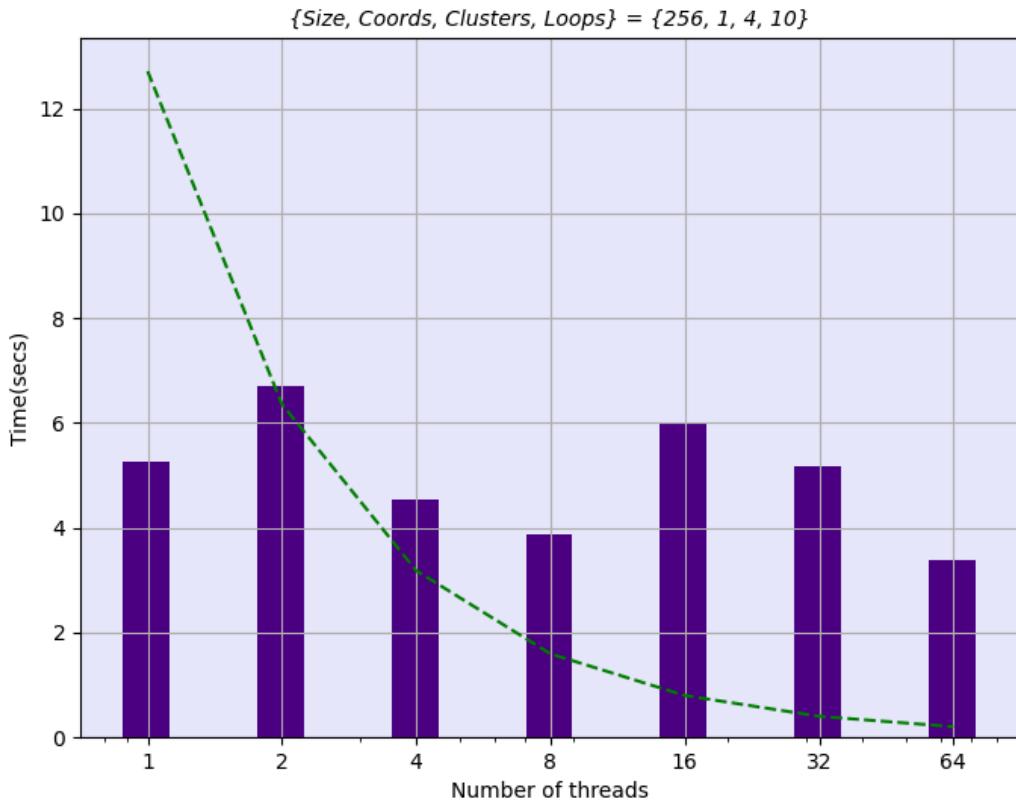


Παρατηρούμε τέλεια κλιμάκωση μέχρι και τα 32 νήματα και αρκετά καλή και στα 64 εφόσον δεν εισάγουμε overheads συγχρονισμού και η σειριακή ενοποίηση (reduction) δεν είναι computational intensive για να καθυστερεί τον αλγόριθμο.

Δοκιμές με μικρότερο dataset

Τα αποτελέσματα δεν είναι ίδια για άλλα μεγέθη πινάκων. Συγκεκριμένα για το επόμενο configuration παρατηρούμε τα εξής:

Copied Clusters & Reduction



Κυρίαρχο ρόλο για αυτήν την συμπεριφορά αποτελεί το φαινόμενο false sharing, που εμφανίζεται σε μικρά datasets (εδώ κάθε object έχει μόνο 1 συντεταγμένη!), όταν σε ένα cache line καταφέρνουν να χωρέσουν παραπάνω από 1 objects και σε κάθε εγγραφή γίνονται πάρα πολλά περιττά invalidations. Μια λύση είναι το padding όμως έχει memory overhead και δεν προτιμάται.

First-touch Policy

Προς αποφυγή των παραπάνω εκμεταλλευόμαστε την πολιτική των linux κατά το mapping των virtual με physical addresses. Η δέσμευση φυσικής μνήμης πραγματοποιείται κατά την 1η εγγραφή του αντικειμένου (η calloc το εξασφαλίζει γράφοντας 0 ενώ η malloc όχι), οπότε εάν το κάθε νήμα γράψει ξεχωριστά στο κομμάτι του πίνακα που του αντιστοιχεί (ουσιαστικά παραλληλοποιώντας την αντιγραφή των shared πινάκων) θα απεικονιστεί στην μνήμη του αυτό και μόνο.

Υλοποίηση

```
omp_reduction_kmeans.c

1 #include <stdio.h>
2 #include <stdlib.h>
3 #include "kmeans.h"
4 /*
5  * TODO: include openmp header file
6  */
7
8 // square of Euclid distance between two multi-dimensional points
9 inline static double euclid_dist_2(int numdims, /* no. dimensions */
10                                 double * coord1, /* [numdims] */
11                                 double * coord2) /* [numdims] */
12 {
13     int i;
14     double ans = 0.0;
15 }
```

```

16     for(i=0; i<numdims; i++)
17         ans += (coord1[i]-coord2[i]) * (coord1[i]-coord2[i]);
18
19     return ans;
20 }
21
22 inline static int find_nearest_cluster(int      numClusters, /* no. clusters */
23                                         int      numCoords,   /* no. coordinates */
24                                         double * object,    /* [numCoords] */
25                                         double * clusters) /* [numClusters][numCoords] */
26 {
27     int index, i;
28     double dist, min_dist;
29
30     // find the cluster id that has min distance to object
31     index = 0;
32     min_dist = euclid_dist_2(numCoords, object, clusters);
33
34     for(i=1; i<numClusters; i++) {
35         dist = euclid_dist_2(numCoords, object, &clusters[i*numCoords]);
36         // no need square root
37         if (dist < min_dist) { // find the min and its array index
38             min_dist = dist;
39             index   = i;
40         }
41     }
42     return index;
43 }
44
45 void kmeans(double * objects,           /* in: [numObjs][numCoords] */
46             int      numCoords,        /* no. coordinates */
47             int      numObjs,          /* no. objects */
48             int      numClusters,       /* no. clusters */
49             double   threshold,        /* minimum fraction of objects that change membership */
50             long    loop_threshold,    /* maximum number of iterations */
51             int    * membership,       /* out: [numObjs] */
52             double  * clusters)        /* out: [numClusters][numCoords] */
53 {
54     int i, j, k;
55     int index, loop=0;
56     double timing = 0;
57
58     double delta;           // fraction of objects whose clusters change in each loop
59     int * newClusterSize; // [numClusters]: no. objects assigned in each new cluster
60     double * newClusters; // [numClusters][numCoords]
61     int nthreads;          // no. threads
62
63     nthreads = omp_get_max_threads();
64     printf("OpenMP Kmeans - Reduction\t(number of threads: %d)\n", nthreads);
65
66     // initialize membership
67     for (i=0; i<numObjs; i++)
68         membership[i] = -1;
69
70     // initialize newClusterSize and newClusters to all 0
71     newClusterSize = (typeof(newClusterSize)) calloc(numClusters, sizeof(*newClusterSize));
72     newClusters = (typeof(newClusters))  calloc(numClusters * numCoords, sizeof(*newClusters));
73
74     // Each thread calculates new centers using a private space. After that, thread 0 does an
75     // array reduction on them.
76     int * local_newClusterSize[nthreads]; // [nthreads][numClusters]
77     double * local_newClusters[nthreads]; // [nthreads][numClusters][numCoords]
78
79     /*
80      * Hint for false-sharing
81      * This is noticed when numCoords is low (and neighboring local_newClusters exist close to
82      * each other).
83      * Allocate local cluster data with a "first-touch" policy.
84      */
85
86     timing = wtime();
87     do {
88         // before each loop, set cluster data to 0
89         for (i=0; i<numClusters; i++) {
90             for (j=0; j<numCoords; j++)

```

```

89         newClusters[i*numCoords + j] = 0.0;
90         newClusterSize[i] = 0;
91     }
92
93     delta = 0.0;
94
95     /*
96      * TODO: Initiliaze local cluster data to zero (separate for each thread)
97      */
98
99     #pragma omp parallel for private(k,i,j) schedule(static)
100    for (k=0; k<nthreads; ++k){
101        local_newClusterSize[k] = (typeof(*local_newClusterSize)) calloc(numClusters,
102        sizeof(**local_newClusterSize));
103        local_newClusters[k] = (typeof(*local_newClusters)) calloc(numClusters * numCoords,
104        sizeof(**local_newClusters));
105
106        for (i=0; i<numClusters; i++) {
107            for (j=0; j<numCoords; j++)
108                local_newClusters[k][i*numCoords + j] = 0.0;
109                local_newClusterSize[k][i] = 0;
110        }
111    int thread_id;
112
113    #pragma omp parallel for private(i, j, thread_id, index) shared(local_newClusters,
114    local_newClusterSize) reduction(+:delta) schedule(static)
115    for (i=0; i<numObjs; i++)
116    {
117        thread_id = omp_get_thread_num();
118
119        // find the array index of nearest cluster center
120        index = find_nearest_cluster(numClusters, numCoords, &objects[i*numCoords], clusters);
121
122        // if membership changes, increase delta by 1
123        if (membership[i] != index)
124            delta += 1.0;
125
126        // assign the membership to object i
127        membership[i] = index;
128
129        // update new cluster centers : sum of all objects located within (average will be
130        // performed later)
131
132        local_newClusterSize[thread_id][index]++;
133        for (j=0; j<numCoords; j++)
134            local_newClusters[thread_id][index*numCoords + j] += objects[i*numCoords + j];
135    }
136
137    for (i=0; i<numClusters; ++i){
138        for (k=0; k<nthreads; ++k){
139            newClusterSize[i] += local_newClusterSize[k][i];
140            for (j=0; j<numCoords; ++j)
141                newClusters[i*numCoords+j] += local_newClusters[k][i*numCoords+j];
142        }
143
144        for (i=0; i<numClusters; i++) {
145            if (newClusterSize[i] > 0) {
146                for (j=0; j<numCoords; j++) {
147                    clusters[i*numCoords + j] = newClusters[i*numCoords + j] / newClusterSize[i];
148                }
149            }
150        }
151        delta /= numObjs;
152
153        loop++;
154        printf("\r\tcompleted loop %d", loop);
155        fflush(stdout);
156    } while (delta > threshold && loop < loop_threshold);
157    timing = wtime() - timing;

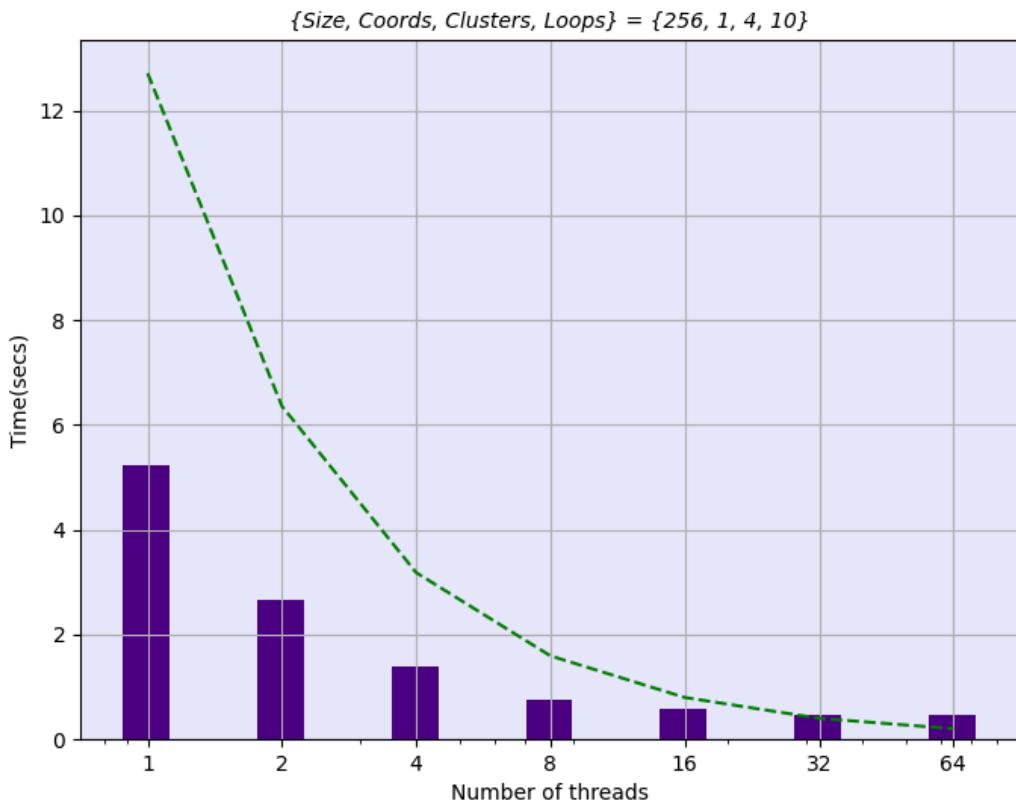
```

```

158     printf("\n          nloops = %3d    (total = %7.4fs)  (per loop = %7.4fs)\n", loop, timing,
159     timing/loop);
160
161     for (k=0; k<nthreads; k++)
162     {
163         free(local_newClusterSize[k]);
164         free(local_newClusters[k]);
165     }
166     free(newClusters);
167     free(newClusterSize);
168 }
```

Αποτελέσματα

Copied Clusters & Reduction with First-Touch Policy



Υπάρχει σαφής βελτίωση και καλή κλιμάκωση μέχρι τα 32 νήματα ακόμα και σε σχέση με την ιδανική εκτέλεση του σειριακού αλγορίθμου. Ο καλύτερος χρόνος σε αυτό το ερώτημα είναι 0.4605s στα 32 νήματα!

Numa-aware initialization

Με βάση όσα αναφέρθηκαν για το pinning σε cores και την πολιτική first-touch, η αρχικοποίηση των shared πινάκων μπορεί να γίνει και αυτή ατομικά από κάθε νήμα σε ένα private τμήμα αυτού. Για την υλοποίηση προσθέτουμε το omp parallel for directive με στατική δρομολόγηση. Αυτή είναι απαραίτητη, ώστε τα νήματα που θα βάλουν τους τυχαίους αριθμούς στα objects, να είναι τα ίδια νήματα με αυτά που θα τα επεξεργαστούν στην main.c με σκοπό να είναι ήδη στις caches και να μην χρειάζεται να τα μεταφέρουν από την κύρια μνήμη ή από άλλα νήματα.

Υλοποίηση

Τροποποιούμε το file_io.c που δίνεται :

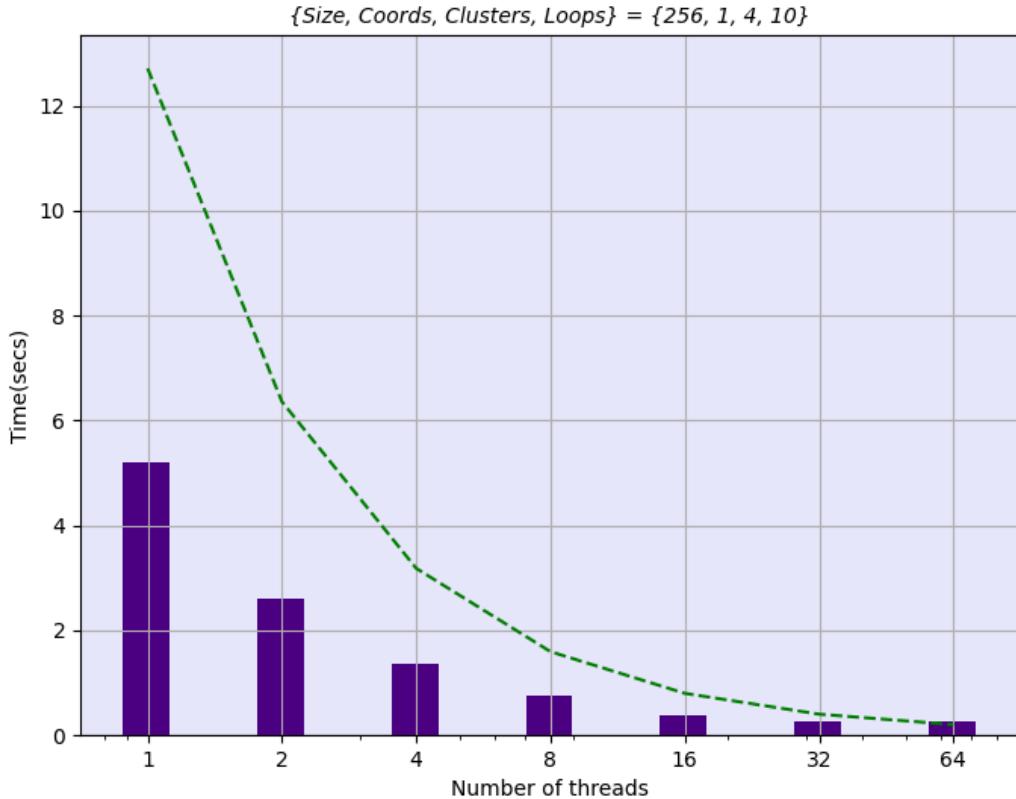
file_io.c

```
1 #include <stdio.h>
2 #include <stdlib.h>
3 #include <string.h>      /* strtok() */
4 #include <sys/types.h>    /* open() */
5 #include <sys/stat.h>
6 #include <fcntl.h>
7 #include <unistd.h>       /* read(), close() */
8 // TODO: remove comment from following line
9 #include <omp.h>
10
11 #include "kmeans.h"
12
13 double * dataset_generation(int numObjs, int numCoords)
14 {
15     double * objects = NULL;
16     long i, j;
17     // Random values that will be generated will be between 0 and 10.
18     double val_range = 10;
19
20     /* allocate space for objects[][] and read all objects */
21     objects = (typeof(objects)) malloc(numObjs * numCoords * sizeof(*objects));
22
23     /*
24      * Hint : Could dataset generation be performed in a more "NUMA-Aware" way?
25      *        Need to place data "close" to the threads that will perform operations on them.
26      *        reminder : First-touch data placement policy
27      */
28     int nthreads = omp_get_max_threads();
29     int chunk = numObjs / nthreads;
30     int thread_id, start_offs, end_offs;
31
32     #pragma omp parallel private(i, j, thread_id, start_offs, end_offs) shared(nthreads, chunk,
33     objects, numObjs, numCoords, val_range)
34     {
35         //set the binding to cores manually
36
37         thread_id = omp_get_thread_num();
38         start_offs = thread_id * chunk;
39         end_offs = (thread_id == nthreads-1) ? numObjs : start_offs + chunk;
40
41         for (i=start_offs; i<end_offs; i++)
42         {
43             unsigned int seed = i;
44             for (j=0; j <numCoords; j++)
45             {
46                 objects[i*numCoords + j] = (rand_r(&seed) / ((double) RAND_MAX)) * val_range;
47                 if (_debug && i == 0)
48                     printf("object[i=%ld][j=%ld]=%f\n", i, j, objects[i*numCoords + j]);
49             }
50         }
51     }
52     return objects;
53 }
```

)

Αποτελέσματα

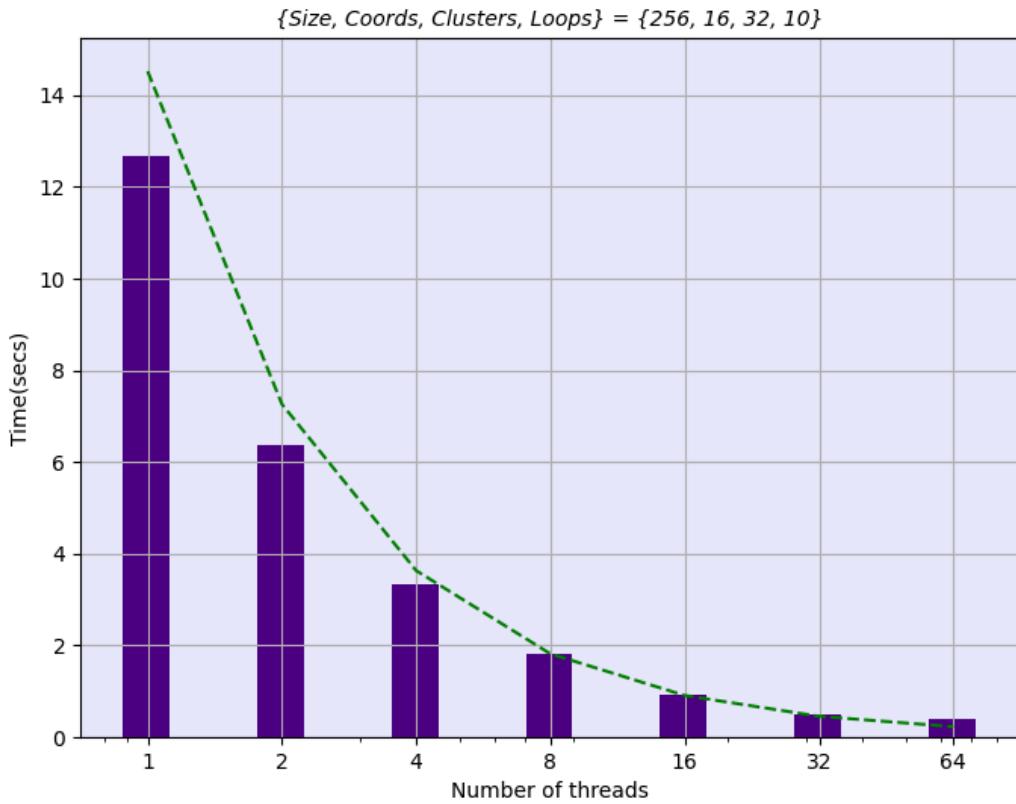
Copied Clusters & Reduction with First-Touch Policy & NUMA-aware initialization



Παρατηρούμε καλύτερη κλιμάκωση μέχρι τα 32 νήματα με χρόνο 0.2667s! Το κυρίαρχο bottleneck σε αυτήν την περίπτωση είναι το overhead της δημιουργίας των νημάτων.

Τέλος με όλες τις προηγούμενες αλλαγές δοκιμάζουμε ξανά το μεγάλο dataset που είχαμε στην αρχή:

Copied Clusters & Reduction with First-Touch Policy & NUMA-aware initialization



Παρατηρούμε πως υπάρχει τέλεια κλιμάκωση του αλγορίθμου. Οπότε bottleneck θα μπορούσε να θεωρηθεί το computive intensity για κάθε object.

FLOYD WARSHALL

1) Recursive

Υλοποίηση

Δημιουργούμε ένα παράλληλο section κατά την πρώτη κλήση, αφού έχουμε ενεργοποιήσει την επιλογή για nested tasks μέσω την `omp_set_nested(1)`. (**Μπορούμε να το θέσουμε και ως environmental variable (OMP_NESTED=TRUE, OMP_MAX_ACTIVE_LEVELS=64)**) Για την διατήρηση των εξαρτήσεων κατά τον υπολογισμό των blocks (A11) -> (A12 A21) -> A22 και αντιστρόφως, τοποθετούμε κατάλληλα barriers έμμεσα με τα taskwait directives.

`fw_sr.c`

```
1  /*
2   * Recursive implementation of the Floyd-Warshall algorithm.
3   * command line arguments: N, B
4   * N = size of graph
5   * B = size of submatrix when recursion stops
6   * works only for N, B = 2^k
7   */
8
9  #include <stdio.h>
10 #include <stdlib.h>
11 #include <sys/time.h>
12 #include "util.h"
13 #include <omp.h>
14
15 inline int min(int a, int b);
16 void FW_SR (int **A, int arow, int acol,
17             int **B, int brow, int bcol,
18             int **C, int crow, int ccol,
19             int myN, int bsize);
20
21 int main(int argc, char **argv)
22 {
23     int **A;
24     int i,j;
25     struct timeval t1, t2;
26     double time;
27     int B=16;
28     int N=1024;
29
30     if (argc !=3){
31         fprintf(stdout, "Usage %s N B \n", argv[0]);
32         exit(0);
33     }
34
35     N=atoi(argv[1]);
36     B=atoi(argv[2]);
37
38     A = (int **) malloc(N*sizeof(int *));
39     for(i=0; i<N; i++) A[i] = (int *) malloc(N*sizeof(int));
40
41     graph_init_random(A,-1,N,128*N);
42     //enable nested task generation
43     omp_set_nested(1);
44     // default is equal to 1
45     omp_set_max_active_levels(64);
46
47     gettimeofday(&t1,0);
48
49     #pragma omp parallel
50     {
51         #pragma omp single
52         {
53             FW_SR(A,0,0, A,0,0,A,0,0,N,B);
54         }
55     }
56     gettimeofday(&t2,0);
```

```

57     time=(double)((t2.tv_sec-t1.tv_sec)*1000000+t2.tv_usec-t1.tv_usec)/1000000;
58     printf("FW_SR,%d,%d,.4f\n", N, B, time);
59
60     /*
61      for(i=0; i<N; i++)
62        for(j=0; j<N; j++) fprintf(stdout,"%d\n", A[i][j]);
63     */
64
65     return 0;
66 }
67
68
69 inline int min(int a, int b)
70 {
71     if(a<=b) return a;
72     else return b;
73 }
74
75 void FW_SR (int **A, int arow, int acol,
76             int **B, int brow, int bcol,
77             int **C, int crow, int ccol,
78             int myN, int bsize)
79 {
80     int k,i,j;
81
82     /*
83      * The base case (when recursion stops) is not allowed to be edited!
84      * What you can do is try different block sizes.
85     */
86
87     if(myN<=bsize)
88         for(k=0; k<myN; k++)
89             for(i=0; i<myN; i++)
90                 for(j=0; j<myN; j++)
91                     A[arow+i][acol+j]=min(A[arow+i][acol+j], B[brow+i][bcol+k]+C[crow+k][ccol+j]);
92     else {
93
94         FW_SR(A,arow, acol,B,brow, bcol,C,crow, ccol, myN/2, bsize); // A00
95         #pragma omp task
96         FW_SR(A,arow, acol+myN/2,B,brow, bcol,C,crow, ccol+myN/2, myN/2, bsize); //A01
97         #pragma omp task
98         FW_SR(A,arow+myN/2, acol,B,brow+myN/2, bcol,C,crow, ccol, myN/2, bsize); //A10
99         #pragma omp taskwait
100        FW_SR(A,arow+myN/2, acol+myN/2,B,brow+myN/2, bcol,C,crow, ccol+myN/2, myN/2, bsize); //
101        A11
102        FW_SR(A,arow+myN/2, acol+myN/2,B,brow+myN/2, bcol+myN/2,C,crow+myN/2, ccol+myN/2, myN/2,
103              bsize); //A11
104        #pragma omp task
105        FW_SR(A,arow+myN/2, acol,B,brow+myN/2, bcol+myN/2,C,crow+myN/2, ccol, myN/2, bsize); //A10
106        #pragma omp task
107        FW_SR(A,arow, acol+myN/2,B,brow, bcol+myN/2,C,crow+myN/2, ccol+myN/2, myN/2, bsize); //A01
108    }
109    // printf("Nested parallelism enabled: %d\n", omp_get_nested());
110 }

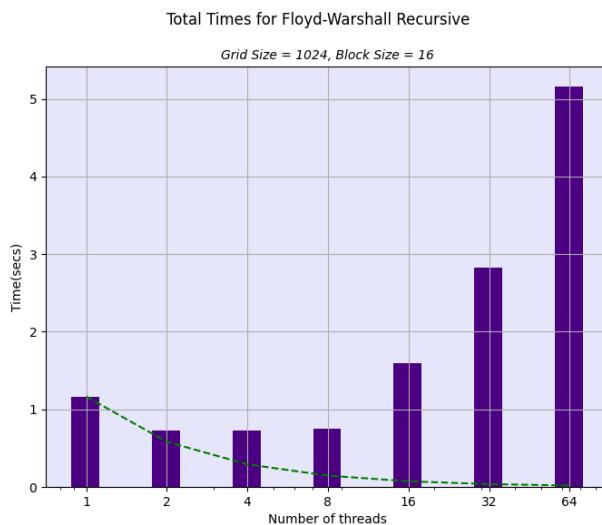
```

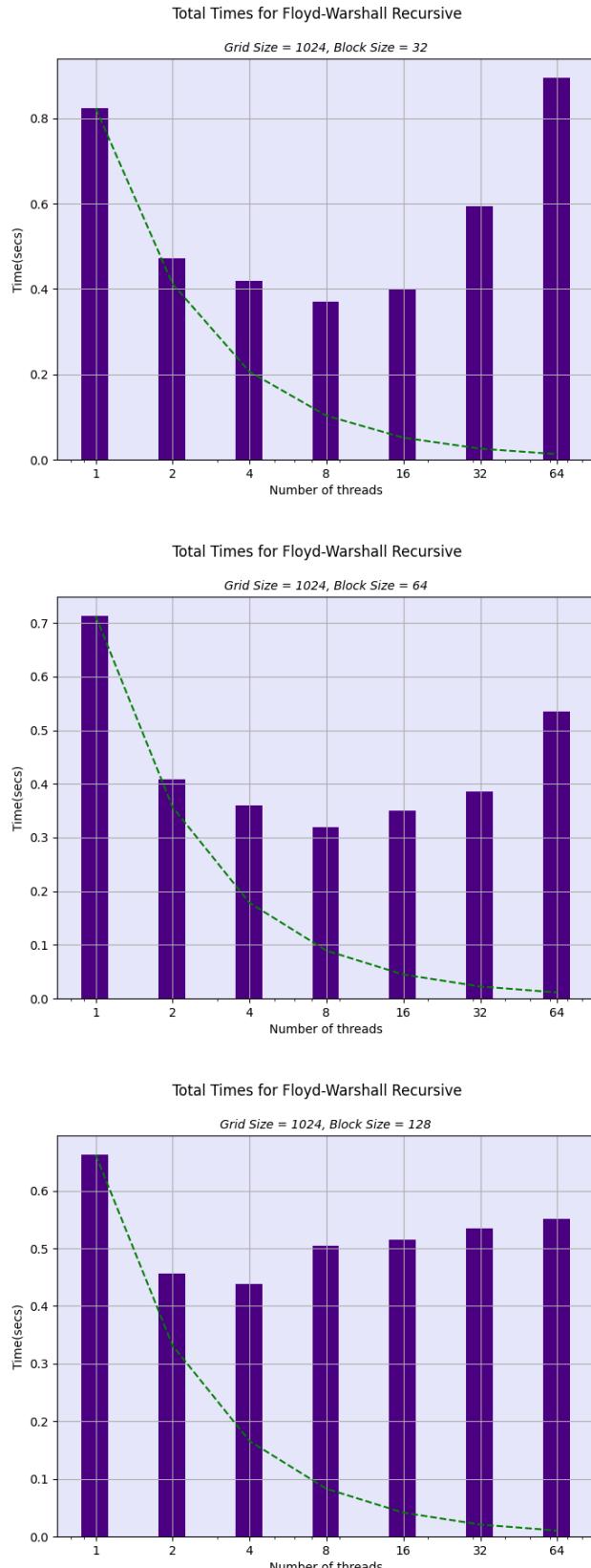
Πειραματιστήκαμε σχετικά με την βέλτιστη τιμή του BSIZE τρέχοντας τις προσομοιώσεις που ακολουθούν. Διαισθητικά η optimal τιμή οφείλει να εκμεταλλεύεται πλήρως το cache size και δεδομένου ότι έχουμε τετράγωνο grid για 1 recursive call που δημιουργεί 4 sub-blocks μεγέθους B θα είναι $B_{opt} = \text{sqrt}(\text{cache size})$. Για τα πειράματα χρησιμοποιήσαμε το ακόλουθο script:

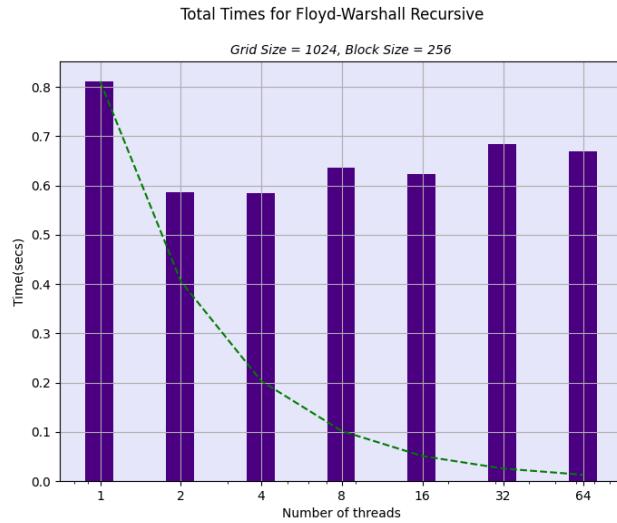
```
#!/bin/bash
## Give the Job a descriptive name #PBS -N run_fw
## Output and error files #PBS -o run_fw_recursive.out #PBS -e run_fw_recursive.err
## How many machines should we get? #PBS -l nodes=1:ppn=8
## How long should the job run for? #PBS -l walltime=00:10:00
## Start
## Run make in the src folder (modify properly)
module load openmp/1.8.3
cd /home/parallel/parlab09/a2/FW
./fw $SIZE
export OMP_NESTED=TRUE
export OMP_MAX_ACTIVE_LEVELS=64
for SIZE in 1024 2048 4096; do
    for BSIZE in 16 32 64 128 256; do
        echo -e "\nBSIZE=${BSIZE}\n"
        for n in 1 2 4 8 16 32 64; do
            export OMP_NUM_THREADS=${n}
            echo -e "\nNumber of threads: ${n}"
            ./fw_sr $SIZE $BSIZE
        done
    done
done
```

Αποτελέσματα

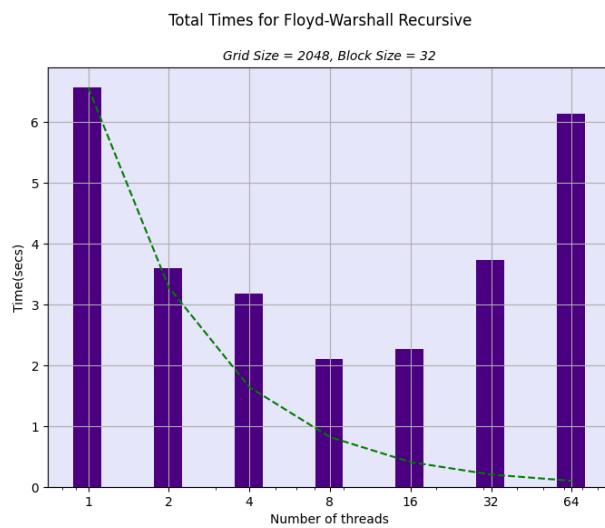
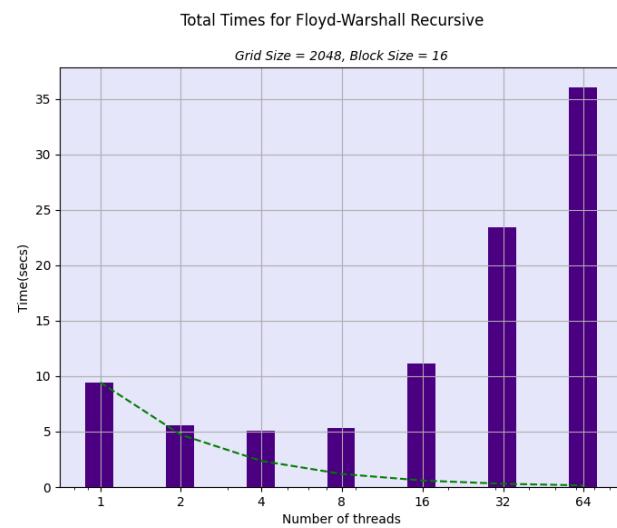
$$\{N = 1024\}$$





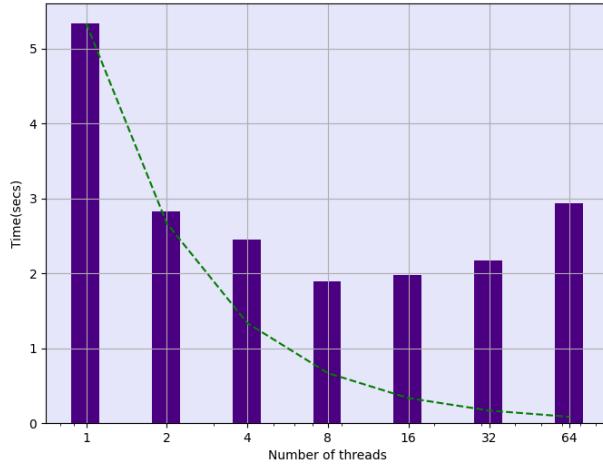


{N = 2048}



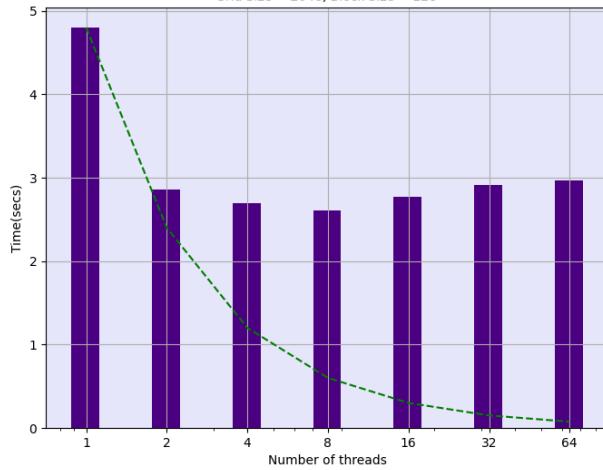
Total Times for Floyd-Warshall Recursive

Grid Size = 2048, Block Size = 64



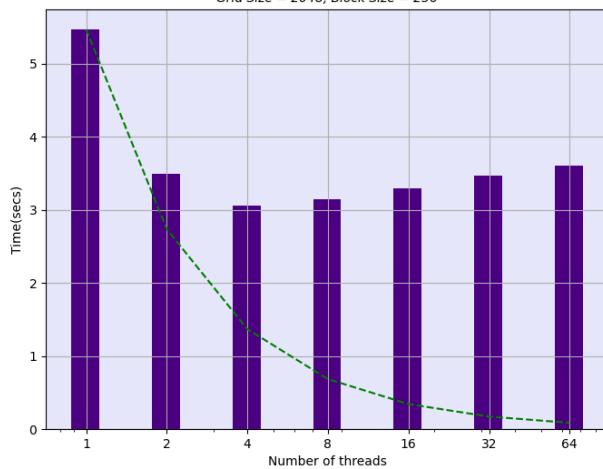
Total Times for Floyd-Warshall Recursive

Grid Size = 2048, Block Size = 128

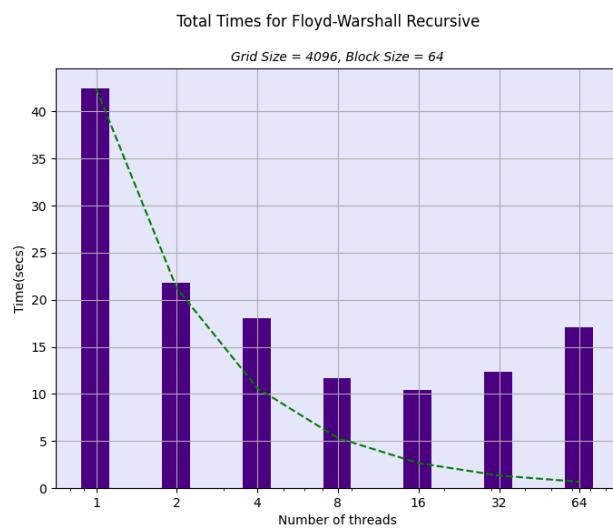
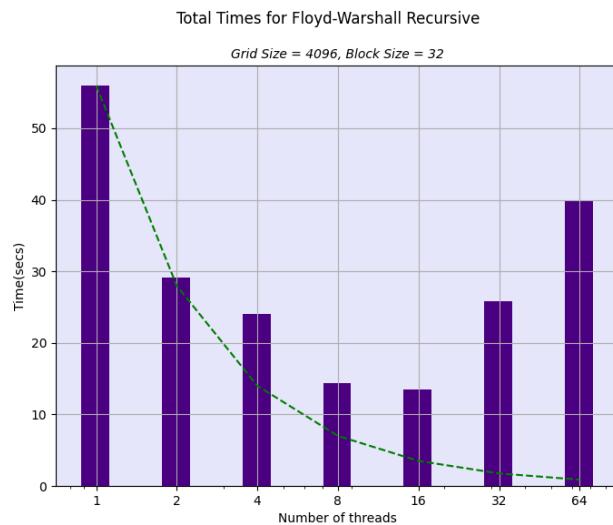
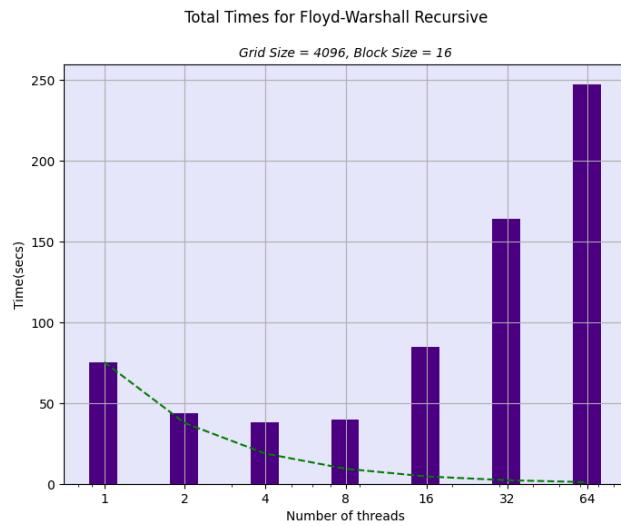


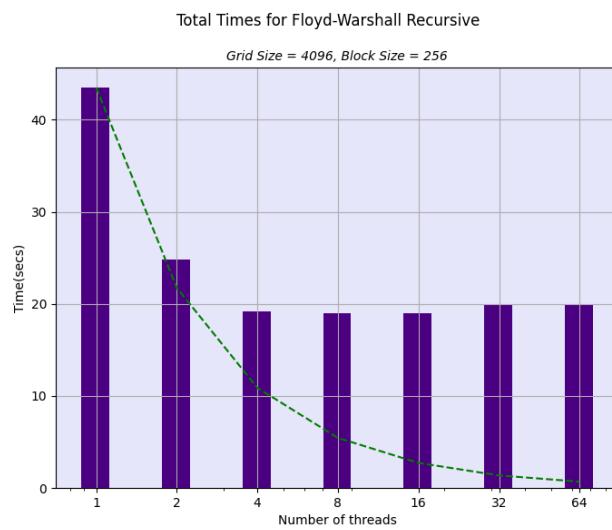
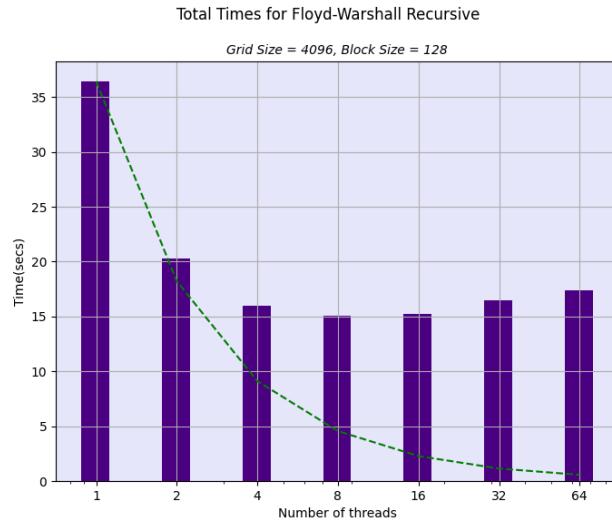
Total Times for Floyd-Warshall Recursive

Grid Size = 2048, Block Size = 256



$$\{ N = 4096 \}$$





Καταλήξαμε πως η ιδανική τιμή είναι $B=64$ και ο καλύτερος χρόνος που πετύχαμε χρησιμοποιώντας αυτήν για 4096 μέγεθος πίνακα ήταν 10.4486 με 16 threads. Από το σημείο αυτό και έπειτα ο αλγόριθμος δεν κλιμακώνει και φανερώνει την αδυναμία του χάρη στην αναδρομή.

2) TILED

Υλοποίηση

Φτιάχνουμε 1 παράλληλο section με κατάλληλα barriers, ώστε να υπολογίζεται πρώτα (single) το κοστό στοιχείο στην διαγώνιο, έπειτα όσα βρίσκονται κατά μήκος του “σταυρού” που σχηματίζεται εκατέρωθεν αυτού, και τέλος τα blocks στοιχείων που απομένουν. Καθένα από τα στάδια 2 και 3 έχει 4 for loops που μπορούν να παραλληλοποιηθούν με parallel for και επειδή είναι ανεξάρτητα μεταξύ τους με παράμετρο nowait. Το collapse(2) πραγματοποιεί flattening για καλύτερη λειτουργία του parallel for για nested loops. Με χρήση μόνο των παραπάνω επιτυγχάνουμε χρόνο εκτέλεσης 2.2 secs.

Για περαιτέρω βελτίωση επιχειρήσαμε να χρησιμοποιήσουμε SIMD εντολές αρχικά μέσω του OpenMP με το αντίστοιχο directive και στην συνέχεια γράφοντας χειροκίνητα τις intrinsics εντολές για AVX μοντέλο που υποστηρίζει 4-size vector operations, καθώς διαπιστώσαμε ότι vector operations μεγαλύτερου μεγέθους (π.χ με 8 στοιχεία AVX2) δεν υποστηρίζεται στο εν λόγω μηχάνημα και λαμβάνουμε σφάλμα Illegal hardware instruction. Στην πρώτη εκδοχή λάβαμε συνολικό χρόνο εκτέλεσης 1.7secs.

Η χρήση των intrinsics απευθείας μας δίνει την δυνατότητα να εκμεταλλευτούμε πλήρως και την αρχιτεκτονική της κρυφής μνήμης μέσω loop unrolling. Συγκεκριμένα, αναγνωρίσαμε ότι το size του cacheline είναι 64bytes, συνεπώς χωράνε 16 integers, ή 4 vectors 4άδων σε όρους AVX. Άρα επιτυγχάνουμε μέγιστο locality exploitation κάνοντας unroll με παράγοντα 4 και αυξάνοντας το j κατά 16 σε κάθε iteration. Ακόμη, παρατηρούμε ότι τα στοιχεία A[i][k] είναι ανεξάρτητα του j και η φόρτωση αυτών των vectors μπορεί να γίνει στο εξωτερικό loop. Ο καλύτερος χρόνος εκτέλεσης που επιτύχαμε αυτήν την εκδοχή είναι **1.39 secs!**

```
fw_smd.c

1 #include <stdio.h>
2 #include <stdlib.h>
3 #include <sys/time.h>
4 #include <immintrin.h> // For SSE2 intrinsics
5 #include <omp.h>
6
7 inline void FW(int **A, int K, int I, int J, int B);
8
9 int main(int argc, char **argv)
10 {
11     int **A;
12     int i, j, k;
13     struct timeval t1, t2;
14     double time;
15     int B = 64;
16     int N = 1024;
17
18     if (argc != 3) {
19         fprintf(stdout, "Usage %s N B\n", argv[0]);
20         exit(0);
21     }
22
23     N = atoi(argv[1]);
24     B = atoi(argv[2]);
25
26     // Allocate memory for A with 32-byte alignment
27     posix_memalign((void**)&A, 32, N * sizeof(int*));
28     for (i = 0; i < N; ++i) {
29         posix_memalign((void**)&A[i], 32, N * sizeof(int));
30     }
31
32     // Initialize the graph with random values
33     graph_init_random(A, -1, N, 128 * N);
34
35     // Start timer
36     gettimeofday(&t1, 0);
37 }
```

```

38 // Main loop of the Floyd-Warshall algorithm with tiling
39 for (k = 0; k < N; k += B) {
40     #pragma omp parallel
41     {
42         #pragma omp single
43         {
44             FW(A, k, k, k, B);
45         }
46         #pragma omp for nowait
47         for (i = 0; i < k; i += B)
48             FW(A, k, i, k, B);
49
50         #pragma omp for nowait
51         for (i = k + B; i < N; i += B)
52             FW(A, k, i, k, B);
53
54         #pragma omp for nowait
55         for (j = 0; j < k; j += B)
56             FW(A, k, k, j, B);
57
58         #pragma omp for nowait
59         for (j = k + B; j < N; j += B)
60             FW(A, k, k, j, B);
61
62         #pragma omp barrier
63
64         #pragma omp for collapse(2) nowait
65         for (i = 0; i < k; i += B)
66             for (j = 0; j < k; j += B)
67                 FW(A, k, i, j, B);
68
69         #pragma omp for collapse(2) nowait
70         for (i = 0; i < k; i += B)
71             for (j = k + B; j < N; j += B)
72                 FW(A, k, i, j, B);
73
74         #pragma omp for collapse(2) nowait
75         for (i = k + B; i < N; i += B)
76             for (j = 0; j < k; j += B)
77                 FW(A, k, i, j, B);
78
79         #pragma omp for collapse(2) nowait
80         for (i = k + B; i < N; i += B)
81             for (j = k + B; j < N; j += B)
82                 FW(A, k, i, j, B);
83
84         #pragma omp barrier
85     }
86 }
87
88 // Stop timer and calculate execution time
89 gettimeofday(&t2, 0);
90 time = (double)((t2.tv_sec - t1.tv_sec) * 1000000 + t2.tv_usec - t1.tv_usec) / 1000000;
91 fprintf(stdout, "FW_TILED,%d,%d,%f\n", N, B, time);
92
93 // Free the memory
94 for (i = 0; i < N; i++) {
95     _mm_free(A[i]); // Free each row
96 }
97 _mm_free(A); // Free the pointer array
98
99 return 0;
100}
101
102 inline void FW(int **A, int K, int I, int J, int B)
103 {
104     int i, j, k;
105
106     // Iterate over a block of tiles (3D loop over the block)
107     for (k = K; k < K + B; k++) {
108         for (i = I; i < I + B; i++) {
109             // _mm_prefetch((const char*)&A[i][j], _MM_HINT_T0);
110             // _mm_prefetch((const char*)&A[k][j], _MM_HINT_T0);
111             // _mm_prefetch((const char*)&A[i][j + 16], _MM_HINT_T0);
112             // _mm_prefetch((const char*)&A[k][j + 16], _MM_HINT_T0);

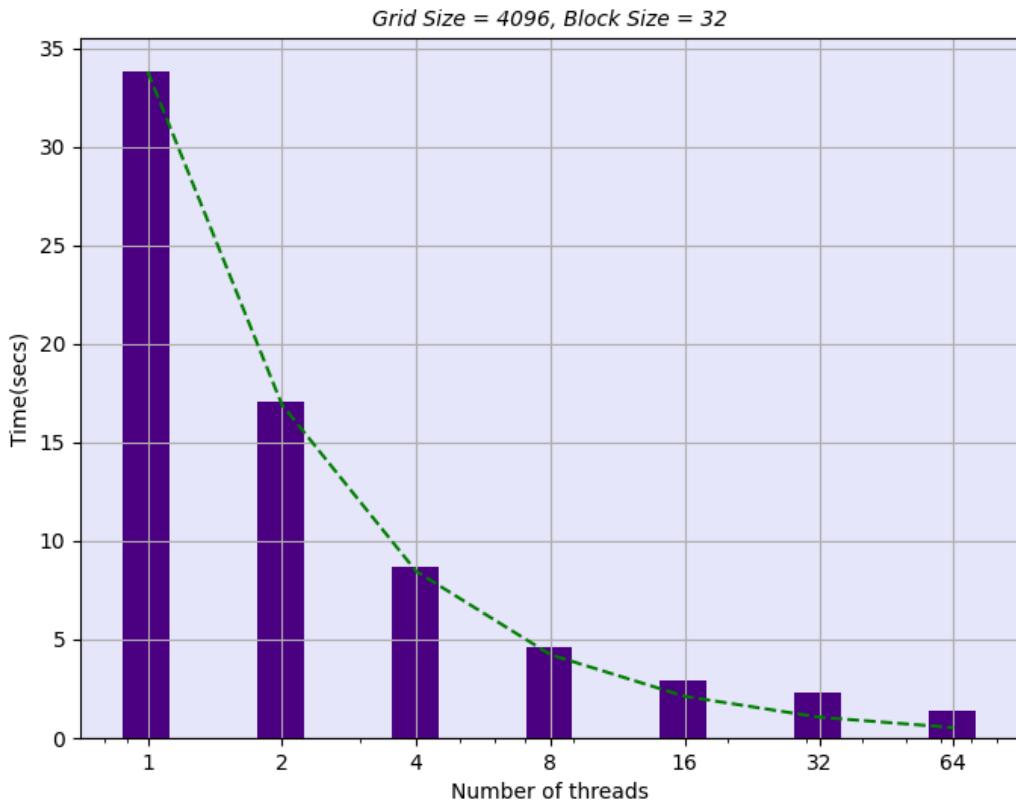
```

```

114     __m128i A_i_k = _mm_load_si128((__m128i*)&A[i][k]);
115
116     for (j = J; j < J + B; j+=16){
117
118         __m128i A_i_j = _mm_load_si128((__m128i*)&A[i][j]);
119         __m128i A_k_j = _mm_load_si128((__m128i*)&A[k][j]);
120
121         __m128i A_plus = _mm_add_epi32(A_i_k, A_k_j);
122         __m128i result = _mm_min_epi32(A_i_j, A_plus);
123
124         _mm_store_si128((__m128i*)&A[i][j], result);
125
126         // next chunk
127         A_i_j = _mm_load_si128((__m128i*)&A[i][j+4]);
128         A_k_j = _mm_load_si128((__m128i*)&A[k][j+4]);
129
130         A_plus = _mm_add_epi32(A_i_k, A_k_j);
131         result = _mm_min_epi32(A_i_j, A_plus);
132
133         _mm_store_si128((__m128i*)&A[i][j+4], result);
134
135         //next chunk
136         A_i_j = _mm_load_si128((__m128i*)&A[i][j+8]);
137         A_k_j = _mm_load_si128((__m128i*)&A[k][j+8]);
138
139         A_plus = _mm_add_epi32(A_i_k, A_k_j);
140         result = _mm_min_epi32(A_i_j, A_plus);
141
142         _mm_store_si128((__m128i*)&A[i][j+8], result);
143
144         //next chunk
145         A_i_j = _mm_load_si128((__m128i*)&A[i][j+12]);
146         A_k_j = _mm_load_si128((__m128i*)&A[k][j+12]);
147
148         A_plus = _mm_add_epi32(A_i_k, A_k_j);
149         result = _mm_min_epi32(A_i_j, A_plus);
150
151         _mm_store_si128((__m128i*)&A[i][j+12], result);
152
153         // if(j == J)
154         //     _mm_prefetch((const char*)&A[i+1][k], _MM_HINT_T0);
155     }
156 }
157 }
158 // if (k + 1 < K + B) {
159 //     _mm_prefetch((const char*)&A[i][k + 1], _MM_HINT_T0);
160 // }
161 }
162 }
```

Αποτελέσματα

Total Times for Floyd-Warshall Tiled



Παραθέτουμε αναλυτικά και τους βέλτιστους χρόνους:

Number of threads: 1

FW_TILED,4096,32,33.8411

Number of threads: 2

FW_TILED,4096,32,17.0405

Number of threads: 4

FW_TILED,4096,32,8.7231

Number of threads: 8

FW_TILED,4096,32,4.5795

Number of threads: 16

FW_TILED,4096,32,2.9022

Number of threads: 32

FW_TILED,4096,32,2.3016

Number of threads: 64

FW_TILED,4096,32,1.3925

Αμοιβαίος Αποκλεισμός-Κλειδώματα

Στο συγκεκριμένο ερώτημα καλούμαστε να αξιολογήσουμε τους διαφορετικούς τρόπους υλοποίησης κλειδωμάτων για αμοιβαίο αποκλεισμό.

Μας δίνονται έτοιμες όλες οι υλοποίησεις των κλειδωμάτων. Για την εκτέλεση του συγκεκριμένου data set (Size = 32, Coords = 16, Clusters = 32, Loops = 10) στον scirouter χρησιμοποιήσαμε το ακόλουθο script :

```
#!/bin/bash
## Give the Job a descriptive name
#PBS -N run_kmeans
## Output and error files
#PBS -o run_kmeans.out
#PBS -e run_kmeans.err
## How many machines should we get?
#PBS -l nodes=1:ppn=8
##How long should the job run for?
#PBS -l walltime=00:10:00
## Start
## Run make in the src folder (modify properly)
module load openmp
cd /home/parallel/parlab09/a2_new/a2/kmeans
export SIZE=32
export COORDS=16
export CLUSTERS=32
export LOOPS=10
for n in 1 2 4 8 16 32 64; do
    export OMP_NUM_THREADS=$n
    echo "Setting OMP_NUM_THREADS=$n" >&2
    export GOMP_CPU_AFFINITY="0-$((($n - 1)))"
    echo "Running ./kmeans_omp_array_lock" >&2
    ./kmeans_omp_array_lock -s $SIZE -n $COORDS -c $CLUSTERS -l $LOOPS
    echo "Running ./kmeans_omp_clh_lock" >&2
    ./kmeans_omp_clh_lock -s $SIZE -n $COORDS -c $CLUSTERS -l $LOOPS
    echo "Running ./kmeans_omp_critical" >&2
    ./kmeans_omp_critical -s $SIZE -n $COORDS -c $CLUSTERS -l $LOOPS
    echo "Running ./kmeans_omp_naive" >&2
    ./kmeans_omp_naive -s $SIZE -n $COORDS -c $CLUSTERS -l $LOOPS
    echo "Running ./kmeans_omp_nosync_lock" >&2
    ./kmeans_omp_nosync_lock -s $SIZE -n $COORDS -c $CLUSTERS -l $LOOPS
    echo "Running ./kmeans_omp_pthread_mutex_lock" >&2
    ./kmeans_omp_pthread_mutex_lock -s $SIZE -n $COORDS -c $CLUSTERS -l $LOOPS
    echo "Running ./kmeans_omp_pthread_spin_lock" >&2
    ./kmeans_omp_pthread_spin_lock -s $SIZE -n $COORDS -c $CLUSTERS -l $LOOPS
    echo "Running ./kmeans_omp_tas_lock" >&2
    ./kmeans_omp_tas_lock -s $SIZE -n $COORDS -c $CLUSTERS -l $LOOPS
    echo "Running ./kmeans_omp_ttas_lock" >&2
    ./kmeans_omp_ttas_lock -s $SIZE -n $COORDS -c $CLUSTERS -l $LOOPS
done
```

Τεχνικές συγχρονισμού

1) pthread_mutex_lock

Σε αυτήν την τεχνική χρησιμοποιείται ένα κλείδωμα αμοιβαίου αποκλεισμού και μεσολαβεί το λειτουργικό σύστημα σε περίπτωση αποτυχίας (context switch). Έτσι, επιτρέπει σε άλλες διεργασίες να τρέχουν μέχρι να ξυπνήσει από κάποιο κατάλληλο signal. Τότε, επιχειρεί εκ νέου να μπει στο κρίσιμο τμήμα.

2) pthread_spin_lock

Σε αυτήν την τεχνική, χρησιμοποείται και πάλι ένα κλείδωμα όμως το κάθε νήμα εκτελεί busy waiting loop για την απόκτηση του. Έτσι, δεν επιτρέπει σε άλλο νήμα να τρέξει στην θέση του (εκτός φυσικά

εάν περάσει χρόνος ίσος με το runtime quantum και το αποσύρει ο scheduler) σπαταλώντας ωφέλιμο CPU time. Αυτό το μοντέλο προσφέρεται για operations που μπλοκάρουν μόνο για λίγους κύκλους, δηλαδή ο χρόνος αναμονής είναι μικρότερος του context switching overhead.

3) tas_lock

Αυτή η τεχνική βασίζεται στην υποστήριξη από το υλικό και χρησιμοποιεί την ατομική εντολή test_and_set που προσφέρει το ISA. Θέτει **ταυτόχρονα** το κλείδωμα (μεταβλητή state) σε 1 και επιστρέφει την προηγούμενη τιμή της (μεταβλητή test). Αν η προηγούμενη τιμή είναι 0 τότε το νήμα δέσμευσε επιτυχώς το κλείδωμα. Κάθε νήμα που προσπαθεί να μπει στο κρίσιμο εκτελεί ένα dummy while loop με την tas. Σε κάθε iteration γράφει στην θέση μνήμης της state, που είναι **μοιραζόμενη**, και στέλνει cache line invalidation στα υπόλοιπα με βάση το πρωτόκολλο MESI για συνάφεια κρυφών μνημών. Συνεπώς, δημιουργείται υπερβολικά μεγάλη και περιττή συμφόρηση στο δίαυλο.

4) ttas_lock

Μοιάζει με την tas, ωστόσο το νήμα δεν γράφει απευθείας την state, αλλά επιχειρεί πρώτα να την διαβάσει (test). Εάν η τιμή της δεν είναι 0, παραμένει μέσα στο busy wit loop και μόλις διαβάσει τιμή 0, προσπαθεί να γράψει σε αυτήν (test_and_set). Τα διαδοχικά reads δεν κοστίζουν σε bandwidth αφού δεν στέλνουν κάποια ενημέρωση μέσω bus. Ο αριθμός των writes μειώνεται σημαντικά όρα και τα συνολικά invalidations. Περαιτέρω βελτίωση γίνεται με εκθετική οπισθοχώρηση (κατά το read και έτσι μειώνονται dummy CPU cycles και αποφεύγονται περισσότερα αποτυχημένα writes), αλλά δεν το εξετάζουμε σε αυτήν την άσκηση.

5) array_lock

Σε αυτήν την τεχνική κάθε νήμα έχει μια δική του μεταβλητή slot, ένα global πίνακα flag και που είναι τώρα το τέλος της ουράς. Κάθε φορά που προσπαθεί ένα νήμα να πάρει το κλείδωμα, ποίρνει το τέλος της ουράς και κάνει ατομική αύξηση κατά 1, θέτει αυτό ως δικό του slot και περιμένει πότε θα γίνει true. Κάθε φορά που ένα νήμα αποδεσμεύει το κλείδωμα, ξαναθέτει το slot του ως false και κάνει το επόμενο true ώστε να πάρει το κλείδωμα αυτός που έχει το επόμενο slot. Αυτή η τεχνική έχει λιγότερη συμφόρηση στο δίαυλο γιατί κάθε νήμα κάνει πάντα 3 αλλαγές για να δεσμεύσει και να αποδεσμεύσει. Επίσης είναι δίκαιη, δηλαδή τα νήματα εκτελούν το κρίσιμο τμήμα με την ίδια σειρά που προσπάθησαν να το δεσμεύσουν. Ωστόσο, ένα νήμα πρέπει να περιμένει στην χειρότερη περίπτωση μια πλήρη περιστροφή του δακτυλίου ώστε να μπει στο κρίσιμο τμήμα, ακόμη και αν είναι το μοναδικό που το επιδιώκει.

6) clh_lock

Σε αυτήν την τεχνική κάθε νήμα έχει ένα κόμβο με ένα κλείδωμα. Κάθε φορά που προσπαθεί να δεσμεύσει το κλείδωμα βάζει το δικό του κλείδωμα να είναι 1, αλλάζει ατομικά τον δείκτη στο κόμβο που αναπαριστά το τέλος της ουράς στον εαυτό του και μετά περιμένει πότε το κλείδωμα του προηγούμενου θα γίνει 0. Αντίστοιχα, όταν αποδεσμεύει το κλείδωμα απλά θέτει το δικό του κλείδωμα σε 0. Το μεγάλο πλεονέκτημα αυτής της τεχνικής είναι πως είναι πολύ κλιμακώσιμη για αρκετά threads, καθώς υπάρχει μόνο 1 κοινή μεταβλητή για τα threads και όχι ολόκληρος πίνακας.

7) pragma omp critical

Θα αξιολογήσουμε και την επίδοση της βιβλιοθήκης του OpenMP για το κρίσιμο τμήμα.

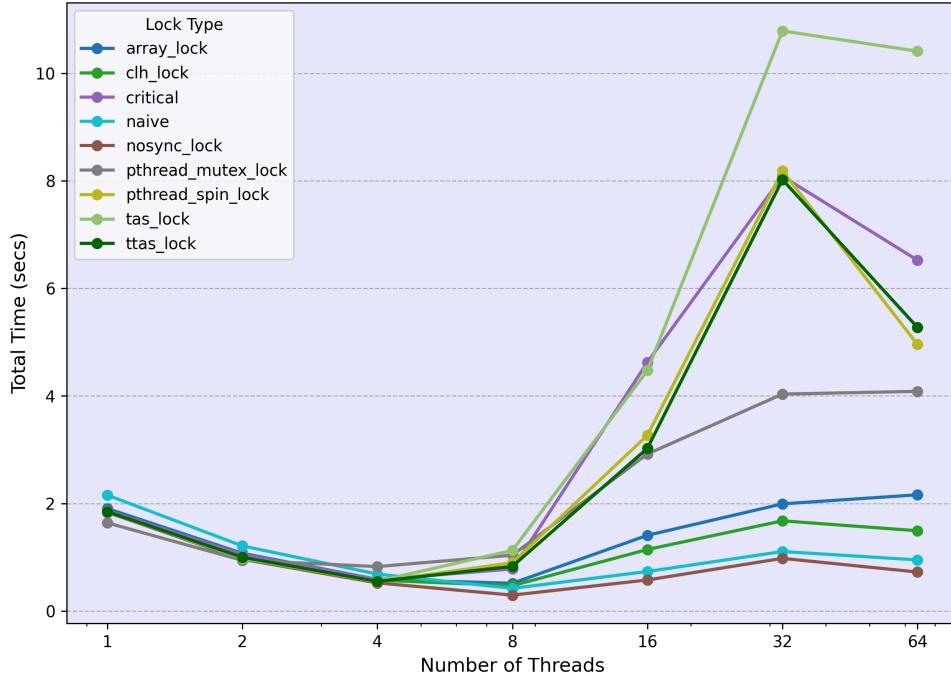
8) pragma omp atomic

Θα αξιολογήσουμε επίσης και την επίδοση μέχρι μόνο 2 ατομικών εντολών και όχι κρίσιμου τμήματος, καθώς μπορεί να μην αλλάζουν τις ίδιες μεταβλητές 2 νήματα.

Αποτελέσματα

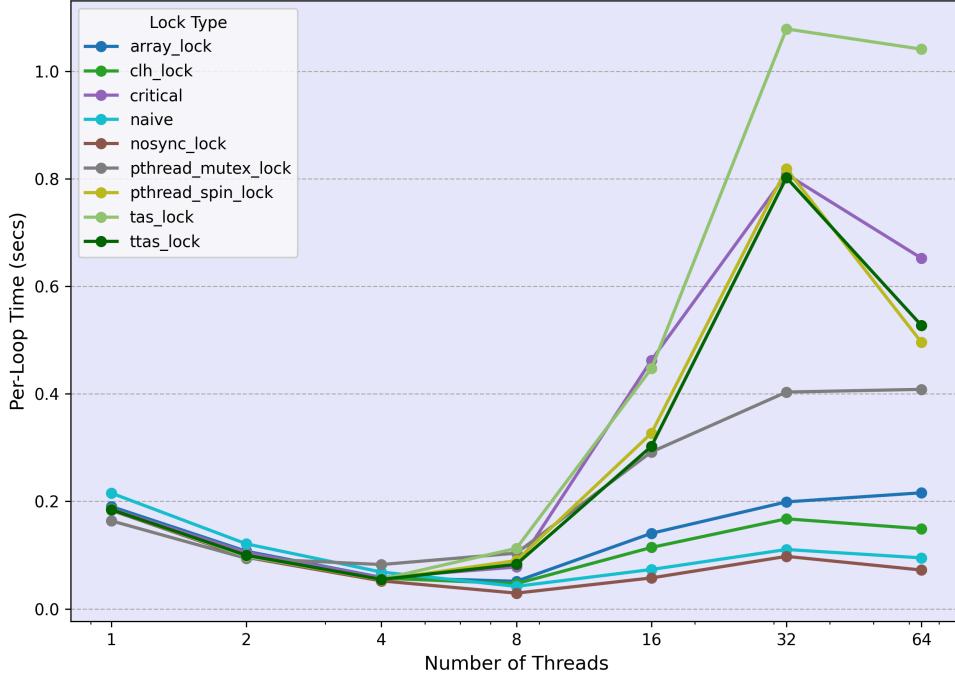
Locking Mechanism vs Time

Configuration = {32, 16, 32, 10}



Locking Mechanism vs Time

Configuration = {32, 16, 32, 10}



Παρατηρήσεις

Γενικά με την αύξηση των threads τα race conditions είναι συχνότερα, οπότε οι συνολικοί χρόνοι εκτέλεσης ανεξαρτήτως μηχανισμού έχουν bottleneck το κόστος συγχρονισμού, όταν η λύση μας πάψει πια να είναι scalable (από τα 8 threads και πάνω όπως φαίνεται στο σχήμα).

naive

Η υλοποίηση naive χρησιμοποιεί atomic add για την εγγραφή στα arrays newClusterSize και newClusters. Συγκεκριμένα, για παράμετρο coords = 16, πραγματοποιεί 16 + 1 (για το newClusterSize) ατομικές εγγραφές. Από 8 threads και πάνω, προσεγγίζει καλύτερα από τις υπόλοιπες την no_sync η οποία δεν χρησιμοποιεί κανένα μηχανισμό συγχρονισμού οπότε παράγει λάθος αποτελέσματα) και χρησιμοποιείται μόνο ως σημείο αναφοράς βέλτιστου χρόνου.

spinlocks

Από 8 threads και πάνω οι μηχανισμοί *tas* και *ttas* δεν κλιμακώνουν εξαιτείας της συμφόρησης που προκαλούν στον δίσκο με τα αλλεπάλληλα cache line invalidations, ειδικότερα όταν τα δεδομένα χρειάζεται να μεταφέρονται πλέον εκτός του NUMA cluster οπότε χρειάζεται να διατηρείται η συνάφεια και μεταξύ των L3 Caches. Ακόμη, όσο περισσότερα γίνονται τα νήματα τόσο αυξάνονται και οι αποτυχημένες προσπάθειες της *test_and_set* και στις δύο περιπτώσεις εξαιτείας της μεγάλης ζήτησης του ίδιου lock. Η *ttas* έχει ωστόσο καλύτερες επιδόσεις, αφού γλιτώνει κάποια περιττά writes, και μάλιστα για λιγότερα από 8 νήματα έχει την καλύτερη επίδοση. Παρόμοια είναι και η συμπεριφορά του *pthread_spinlock* αφού όλα βασίζονται στην λογική των busy-wait loops σπαταλώντας CPU time. Με βάση τη implementation στην glibc, η *pthread_spin_lock* χρησιμοποιεί ένα υβρίδιο των προηγούμενων 2. Στην 1η προσπάθεια, χρησιμοποιεί την ατομική εντολή *atomic_exchange* που θεωρεί ταχύτερη εάν παρέχεται από το υλικό και δεν καλεί την CAS (*compare_and_swap*). Σε περίπτωση αποτυχίας απλά διαβάζει την μεταβλητή όπως η *ttas*. Παραθέτουμε τον αντίστοιχο κώδικα:

`pthread_spin_lock.c`

```
1  /* pthread_spin_lock -- lock a spin lock. Generic version.
2   Copyright (C) 2012-2024 Free Software Foundation, Inc.
3   This file is part of the GNU C Library.
4
5   The GNU C Library is free software; you can redistribute it and/or
6   modify it under the terms of the GNU Lesser General Public
7   License as published by the Free Software Foundation; either
8   version 2.1 of the License, or (at your option) any later version.
9
10  The GNU C Library is distributed in the hope that it will be useful,
11  but WITHOUT ANY WARRANTY; without even the implied warranty of
12  MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU
13  Lesser General Public License for more details.
14
15  You should have received a copy of the GNU Lesser General Public
16  License along with the GNU C Library; if not, see
17  <https://www.gnu.org/licenses/>. */
18
19 #include <atomic.h>
20 #include "pthreadP.h"
21 #include <shlib-compat.h>
22
23 int
24 __pthread_spin_lock (pthread_spinlock_t *lock)
25 {
26     int val = 0;
27
28     /* We assume that the first try mostly will be successful, thus we use
29      atomic_exchange if it is not implemented by a CAS loop (we also assume
30      that atomic_exchange can be faster if it succeeds, see
31      ATOMIC_EXCHANGE_USES_CAS). Otherwise, we use a weak CAS and not an
32      exchange so we bail out after the first failed attempt to change the
33      state. For the subsequent attempts we use atomic_compare_and_exchange
34      after we observe that the lock is not acquired.
35      See also comment in pthread_spin_trylock.
36      We use acquire M0 to synchronize-with the release M0 store in
37      pthread_spin_unlock, and thus ensure that prior critical sections
38      happen-before this critical section. */
39 #if ! ATOMIC_EXCHANGE_USES_CAS
```

```

40  /* Try to acquire the lock with an exchange instruction as this architecture
41  has such an instruction and we assume it is faster than a CAS.
42  The acquisition succeeds if the lock is not in an acquired state. */
43  if (__glibc_likely (atomic_exchange_acquire (lock, 1) == 0))
44  return 0;
45 #else
46  /* Try to acquire the lock with a CAS instruction as this architecture
47  has no exchange instruction. The acquisition succeeds if the lock is not
48  acquired. */
49  if (__glibc_likely (atomic_compare_exchange_weak_acquire (lock, &val, 1)))
50  return 0;
51 #endif
52
53 do
54 {
55  /* The lock is contended and we need to wait. Going straight back
56  to cmpxchg is not a good idea on many targets as that will force
57  expensive memory synchronizations among processors and penalize other
58  running threads.
59  There is no technical reason for throwing in a CAS every now and then,
60  and so far we have no evidence that it can improve performance.
61  If that would be the case, we have to adjust other spin-waiting loops
62  elsewhere, too!
63  Thus we use relaxed MO reads until we observe the lock to not be
64  acquired anymore. */
65  do
66  {
67  /* TODO Back-off. */
68
69  atomic_spin_nop ();
70
71  val = atomic_load_relaxed (lock);
72 }
73 while (val != 0);
74
75 /* We need acquire memory order here for the same reason as mentioned
76 for the first try to lock the spinlock. */
77 }
78 while (!atomic_compare_exchange_weak_acquire (lock, &val, 1));
79
80 return 0;
81 }
82 versioned_symbol (libc, __pthread_spin_lock, pthread_spin_lock, GLIBC_2_34);
83
84 #if OTHER_SHLIB_COMPAT (libpthread, GLIBC_2_2, GLIBC_2_34)
85 compat_symbol (libpthread, __pthread_spin_lock, pthread_spin_lock, GLIBC_2_2);
86#endif

```

array locks

To array_lock και clh_lock είναι οι πιο scaleable μηχανισμοί, αφού δεν χαρακτηρίζονται από το overhead της atomic σε λιγότερα από 8 νήματα και από το overhead του προωτοκόλλου MESI για περισσότερα από 8 νήματα. Επειδή τα νήματα εισέρχονται στο κρίσιμο τμήμα με ένα καθορισμένο μοτίβο και όλα εν τέλει θα μπουν σε αυτό, κανένα slot στον δακτύλιο δεν μένει αδρανές, ενώ όσο τα νήματα περιμένουν την σειρά τους δεν πραγματοποιούν ανούσιες προβάσεις την μνήμη. Αντιθέτως, εάν το concurrency rate ήταν μικρότερο, ο μηχανισμός θα έπασχε από το overhead διάσχυσης ολόκληρου του δακτυλίου προκειμένου να ικανοποίησε λ.χ. ένα μόνο αίτημα. Αντίστοιχα, για το chl_lock μειώνεται το contention για ένα κοινό global lock και κάθε νήμα εκτελεί spinlock στην μεταβλητή του προηγούμενου node. Συγκρούσεις εξακολουθούμε να έχουμε για την είσοδο στο τέλος της ουράς, παρόλα αυτά η διατήρηση μιας λίστας είναι πιο φθηνή από έναν circular buffer και φαίνεται πως γι' αυτό πετυχαίνει καλύτερους χρόνους.

mutex

Για λιγότερα από 8 νήματα το context switch είναι αρκετά ακριβό (τουλάχιστον για το συκεκριμένο configuration όπου απαιτεί 17 μόνο πράξεις στο κρίσιμο τμήμα που δεν είναι τόσο computational

intense), όμως φαίνεται να υπερτερεί έναντι των spinlocks για 16 και πάνω νήματα όπου το bus traffic γίνεται αφόρητο. Η υλοποίηση του omp critical φαίνεται ότι συνδέεται με την χρήση ενός mutex, ωστόσο έχει χειρότερη επίδοση από το να το καλέσουμε explicitly, που διακιολογούμε εφόσον παρέχει ένα higher level abstraction άρα και επιπλέον overheads. Συγκεκριμένα, διατηρεί μέσω του global context manager ένα mapping για named critical regions το οποίο εισάγει μια πολυπλοκότητα, ενώ τα omp directives απαιτούν κάθε φορά και την κλήση της libomp. Παραθέτουμε τον αντίστοιχο κώδικα:

```
gomp_critical.c

1  /* Copyright (C) 2005-2024 Free Software Foundation, Inc.
2   Contributed by Richard Henderson <rth@redhat.com>.
3
4   This file is part of the GNU Offloading and Multi Processing Library
5   (libgomp).
6
7   Libgomp is free software; you can redistribute it and/or modify it
8   under the terms of the GNU General Public License as published by
9   the Free Software Foundation; either version 3, or (at your option)
10  any later version.
11
12  Libgomp is distributed in the hope that it will be useful, but WITHOUT ANY
13  WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS
14  FOR A PARTICULAR PURPOSE. See the GNU General Public License for
15  more details.
16
17  Under Section 7 of GPL version 3, you are granted additional
18  permissions described in the GCC Runtime Library Exception, version
19  3.1, as published by the Free Software Foundation.
20
21  You should have received a copy of the GNU General Public License and
22  a copy of the GCC Runtime Library Exception along with this program;
23  see the files COPYING3 and COPYING.RUNTIME respectively. If not, see
24  <http://www.gnu.org/licenses/>. */
25
26 /* This file handles the CRITICAL construct. */
27
28 #include "libgomp.h"
29 #include <stdlib.h>
30
31
32 static gomp_mutex_t default_lock;
33
34 void
35 GOMP_critical_start (void)
36 {
37   /* There is an implicit flush on entry to a critical region. */
38   __atomic_thread_fence (MEMMODEL_RELEASE);
39   gomp_mutex_lock (&default_lock);
40 }
41
42 void
43 GOMP_critical_end (void)
44 {
45   gomp_mutex_unlock (&default_lock);
46 }
47
48 #ifndef HAVE_SYNC_BUILTINS
49 static gomp_mutex_t create_lock_lock;
50#endif
51
52 void
53 GOMP_critical_name_start (void **pptr)
54 {
55   gomp_mutex_t *plock;
56
57   /* If a mutex fits within the space for a pointer, and is zero initialized,
58    then use the pointer space directly. */
59   if (GOMP_MUTEX_INIT_0
60       && sizeof (gomp_mutex_t) <= sizeof (void *)
61       && __alignof (gomp_mutex_t) <= sizeof (void *))
62     plock = (gomp_mutex_t *) pptr;
63 }
```

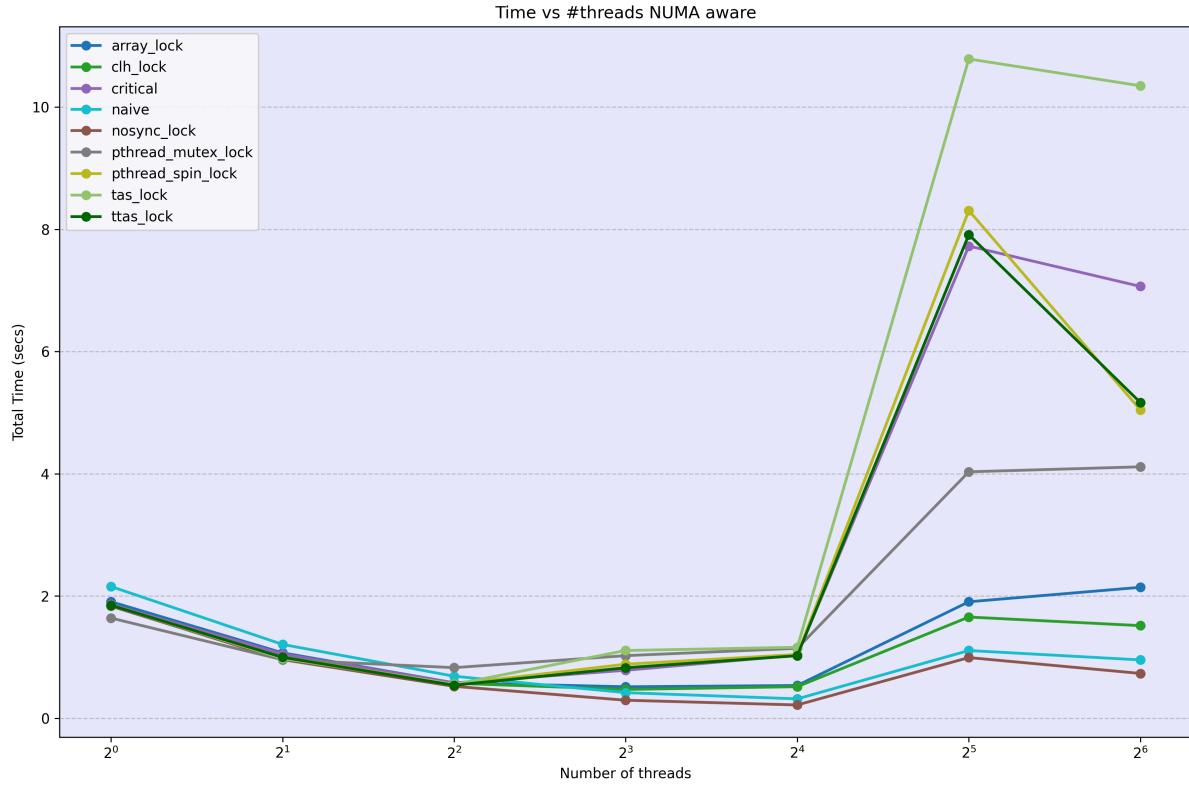
```

64     /* Otherwise we have to be prepared to malloc storage. */
65     else
66     {
67         plock = *pptr;
68
69         if (plock == NULL)
70         {
71 #ifdef HAVE_SYNC_BUILTINS
72             gomp_mutex_t *nlock = gomp_malloc (sizeof (gomp_mutex_t));
73             gomp_mutex_init (nlock);
74
75             plock = __sync_val_compare_and_swap (pptr, NULL, nlock);
76             if (plock != NULL)
77             {
78                 gomp_mutex_destroy (nlock);
79                 free (nlock);
80             }
81         else
82             plock = nlock;
83 #else
84             gomp_mutex_lock (&create_lock_lock);
85             plock = *pptr;
86             if (plock == NULL)
87             {
88                 plock = gomp_malloc (sizeof (gomp_mutex_t));
89                 gomp_mutex_init (plock);
90                 __sync_synchronize ();
91                 *pptr = plock;
92             }
93             gomp_mutex_unlock (&create_lock_lock);
94 #endif
95         }
96     }
97
98     gomp_mutex_lock (plock);
99 }
100
101 void
102 GOMP_critical_name_end (void **pptr)
103 {
104     gomp_mutex_t *plock;
105
106     /* If a mutex fits within the space for a pointer, and is zero initialized,
107      then use the pointer space directly. */
108     if (GOMP_MUTEX_INIT_0
109         && sizeof (gomp_mutex_t) <= sizeof (void *)
110         && __alignof (gomp_mutex_t) <= sizeof (void *))
111         plock = (gomp_mutex_t *) pptr;
112     else
113         plock = *pptr;
114
115     gomp_mutex_unlock (plock);
116 }
117
118 #if !GOMP_MUTEX_INIT_0
119 static void __attribute__((constructor))
120 initialize_critical (void)
121 {
122     gomp_mutex_init (&default_lock);
123 #ifndef HAVE_SYNC_BUILTINS
124     gomp_mutex_init (&create_lock_lock);
125 #endif
126 }
127#endif

```

Τέλος, όταν έχουμε 1 μόνο νήμα, δεν μπλοκάρει σε καμία απόπειρα απόκτησης του lock γι' αυτό όλες οι υλοποιήσεις είναι καλύτερες από την naive, (tas,ttas,array,clh εκτελούν ακριβώς τις ίδεις εντολές - ανάγονται σε 1 ανάγνωση και 1 εγγραφή) και η mutex υπερτερεί γιατί δεν πραγματοποιεί κανένα context switch.

Σημείωση : Αν χρησιμοποιούσαμε hyperthreading για τα 16 νήματα, δηλαδή βάζαμε τα 8 τελευταία νήματα εκτός των 64 λογικών, θα δούμε κλιμάκωση και για 16 νήματα όπως φαίνεται παρακάτω :



Ταυτόχρονες Δομές δεδομένων

Σε αυτό το ερώτημα εξετάζουμε πως κλιμακώνουν διάφορες ταυτόχρονες υλοποιήσεις για μια απλά συνδεδεμένη λίστα.

Οι ταυτόχρονες υλοποιήσεις που θα εξετάσουμε είναι οι εξής :

1) Coarse-grain locking

Σε αυτήν την υλοποίηση υπάρχει ένα γενικό κλείδωμα για όλη την δομή. Για κάθε προσθήκη ή αφαίρεση στοιχείου στη λίστα, το νήμα προσπαθεί να δεσμεύσει το κλείδωμα και να κάνει την κατάλληλη αλλαγή. Είναι πολύ απλό στην υλοποίηση, όμως δεν θα κλιμακώσει καθόλου καθώς όλοι περιμένουν το ίδιο κλείδωμα και δεν εκμεταλλευόμαστε καθόλου παραλληλία σε ανεξάρτητα τμήματα της λίστας.

2) Fine-grain locking

Σε αυτήν την υλοποίηση υπάρχει ένα κλείδωμα για κάθε στοιχείο της λίστας. Ο τρόπος διάσχισης είναι hand-over-hand locking δηλαδή ένα νήμα προσπαθεί να δεσμεύσει τον επόμενο, όταν τα καταφέρει, αφήνει τον προηγούμενο. Αυτό είναι αναγκαστικό προς αποφυγή deadlock, εφόσον για ένα operation απαιτούνται κλειδώματα σε 2 στοιχεία (pred, curr) και θα δημιουργούνταν πρόβλημα αν 2 νήματα επιχειρούσαν να αλλάξουν 2 γειτονικά nodes και προσπαθούσαν να πάρουν τα κλειδώματα με αντίθετη σειρά. Μπορεί να δουλέψει καλύτερα από την coarse grain σε συγκεκριμένες περιπτώσεις, αλλά το σημαντικότερο πρόβλημα είναι πως αν ένα νήμα θέλει να αλλάξει κάτι που βρίσκεται νωρίς στη λίστα, μπλοκάρει όλα τα άλλα νήματα που θέλουν να ψάξουν ή αλλάξουν κάτι που είναι πιο μετά στη λίστα.

3) Optimistic synchronization

Σε αυτήν την υλοποίηση ένα νήμα για κάθε αλλαγή, βρίσκει τον προηγούμενο και τον επόμενο προς αλλαγή, προσπαθεί να τους δεσμεύσει, ελέγχει αν η δομή είναι ακόμη συνεπής (δηλαδή είναι προσβάσιμοι και διαδοχικοί) και κάνει την αλλαγή. Η contains στη συγκεκριμένη υλοποίηση χρησιμοποιεί επίσης κλειδώματα αν και δεν χρειάζεται. Το κύριο πρόβλημα αυτής της υλοποίησης είναι πως η validate διατρέχει όλη την λίστα για να επιβεβαιώσει την συνέπεια και αυτό είναι πάρα πολύ χρονοβόρο.

4) Lazy synchronization

Σε αυτήν την υλοποίηση προσθέτουμε στη δομή μια boolean μεταβλητή που δείχνει αν ο κόμβος βρίσκεται στη λίστα ή έχει διαγραφεί. Η contains διατρέχει τη λίστα χωρίς να κλειδώνει και ελέγχει αυτήν την boolean μεταβλητή, οπότε είναι wait-free. Η validate δεν διατρέχει την λίστα, αλλά κάνει τοπικούς ελέγχους στον προηγούμενο και επόμενο κόμβο, δηλαδή ελέγχει αν ανήκουν στη δομή και οι 2, με την επιπλέον μεταβλητή, και ο next του προηγούμενου είναι ο τωρινός. Η add/remove κάνουν πρώτα λογική και μετά φυσική αλλαγή των κόμβων.

5) Non-blocking

Σε αυτήν την υλοποίηση προσπαθούμε να αφαιρόμε τελείως την ανάγκη για κλειδώματα και να χρησιμοποιήσουμε τις ατομικές εντολές που μας δίνει το instruction set του εκάστοτε επεξεργαστή. Η κεντρική ιδέα είναι να χειριστούμε την boolean μεταβλητή marked και το πεδίο next σαν μία μεταβλητή. Κάνει ατομικό σύνθετο έλεγχο και αλλαγή με 1 εντολή compare and set. Έτσι, η διαγραφή κάνει με 1 εντολή validate και λογική διαγραφή και 1 μόνο προσπάθεια φυσικής διαγραφής. Η find/contains είναι αυτή που εξετάζει αν υπάρχει στοιχείο που έχει διαγραφεί λογικά και όχι φυσικά και το αναλαμβάνει εκείνη. Η προσθήκη αναγκαστικά ξαναπροσπαθεί μέχρι να τα καταφέρει, για να είναι συνεπής η δομή.

Μας δίνονται έτοιμες όλες οι παραπάνω ταυτόχρονες υλοποιήσεις. Για την ζητούμενη εκτέλεση, το σειριακό πρόγραμμα εκτελέστηκε μόνο με 1 thread αλλιώς θα υπάρχει πρόβλημα, για 128 νήματα χρησιμοποιήθηκε oversubscription. Δηλαδή η μεταβλητή MT_CONF τέθηκε σε 0,1,...63,0,1,...63, ώστε να δημιουργηθούν και να γίνουν pinned 128 νήματα σε συγκεκριμένους πυρήνες. Επειδή οι λογικοί πυρήνες του sandman είναι 64, το scheduling των νημάτων πλέον το αναλαμβάνει το λειτουργικό και το software και όχι το ίδιο το υλικό, όπως όταν χρησιμοποιούμε hyperthreading.

```

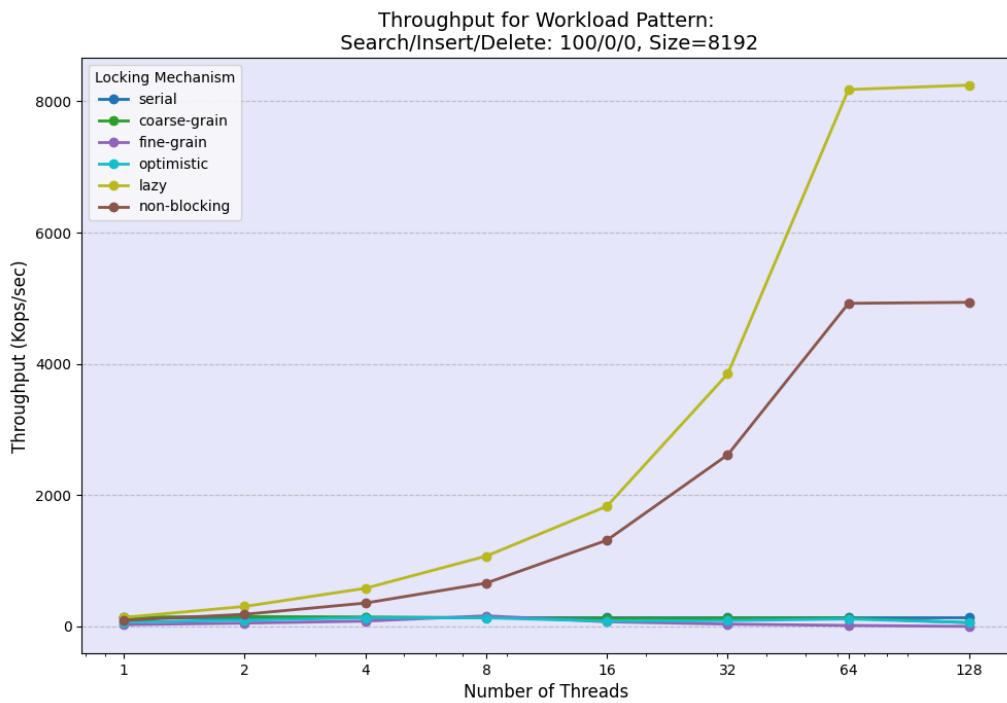
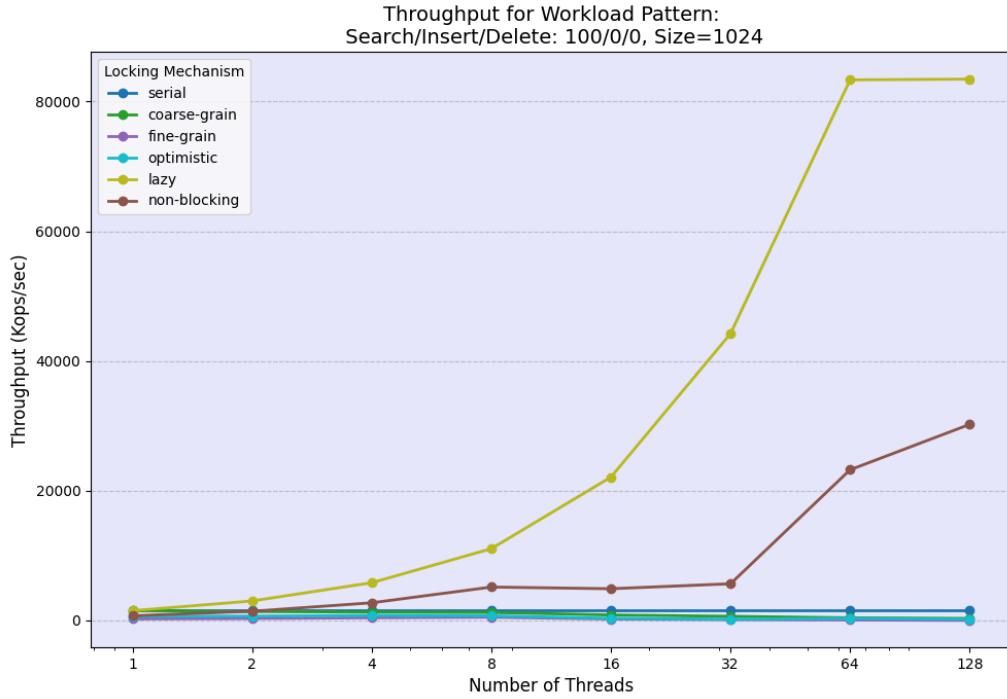
#!/bin/bash
## Give the Job a descriptive name
#PBS -N run_conc_ll
## Output and error files
#PBS -o run_conc_ll.out
#PBS -e run_conc_ll.err
## How many machines should we get?
#PBS -l nodes=1:ppn=8
##How long should the job run for?
#PBS -l walltime=01:00:00
## Start
## Run make in the src folder (modify properly)
module load openmp
cd /home/parallel/parlab09/a2_new/a2/conc_ll
choices=(
    "100 0 0"
    "80 10 10"
    "20 40 40"
    "0 50 50"
)
for lsize in 1024 8192; do
    export LSIZE=$lsize
    echo "LSIZE=$LSIZE"
    for choice in "${choices[@]}"; do
        read -r CONTAINS_PCT ADD_PCT REMOVE_PCT <<< "$choice"
        export MT_CONF="0"
        echo "serial"
        ./x.serial $LSIZE $CONTAINS_PCT $ADD_PCT $REMOVE_PCT
        for n in 1 2 4 8 16 32 64 128; do
            if [ "$n" -eq 128 ]; then
                # Special case for n=128 for over subscription
                export MT_CONF="$((seq -s, 0 63),$(seq -s, 0 63))"
            else
                # Default case for n=1, 2, 4, ..., 64
                export MT_CONF=$((seq -s, 0 $((n-1))) )
            fi
            echo "coarse-grain"
            ./x.cgl $LSIZE $CONTAINS_PCT $ADD_PCT $REMOVE_PCT
            echo "fine-grain"
            ./x.fgl $LSIZE $CONTAINS_PCT $ADD_PCT $REMOVE_PCT
            echo "optimistic"
            ./x.opt $LSIZE $CONTAINS_PCT $ADD_PCT $REMOVE_PCT
            echo "lazy"
            ./x.lazy $LSIZE $CONTAINS_PCT $ADD_PCT $REMOVE_PCT
            echo "non-blocking"
            ./x.nb $LSIZE $CONTAINS_PCT $ADD_PCT $REMOVE_PCT
        done
    done
done

```

Αποτελέσματα

Παρουσιάζονται τα Kops/sec με την μορφή line plots ανά κάθε διαφορετικό workload configuration.

Η serial εκδοχή μας δείχνει την δυναμική του κάθε core, δεν χρησιμοποιεί παραλληλισμό ούτε κλειδώματα και έχει σταθερό throughput ανεξάρτητα από το πλήθος των νημάτων γι'αυτό και την θεωρούμε ως σημείο αναφοράς. Τα queries add / delete είναι υπολογιστικά ισοδύναμα, οπότε μπορούμε να εξάγουμε το συνολικό ποσοστό τους έναντι των search για την ανάλυσή μας.

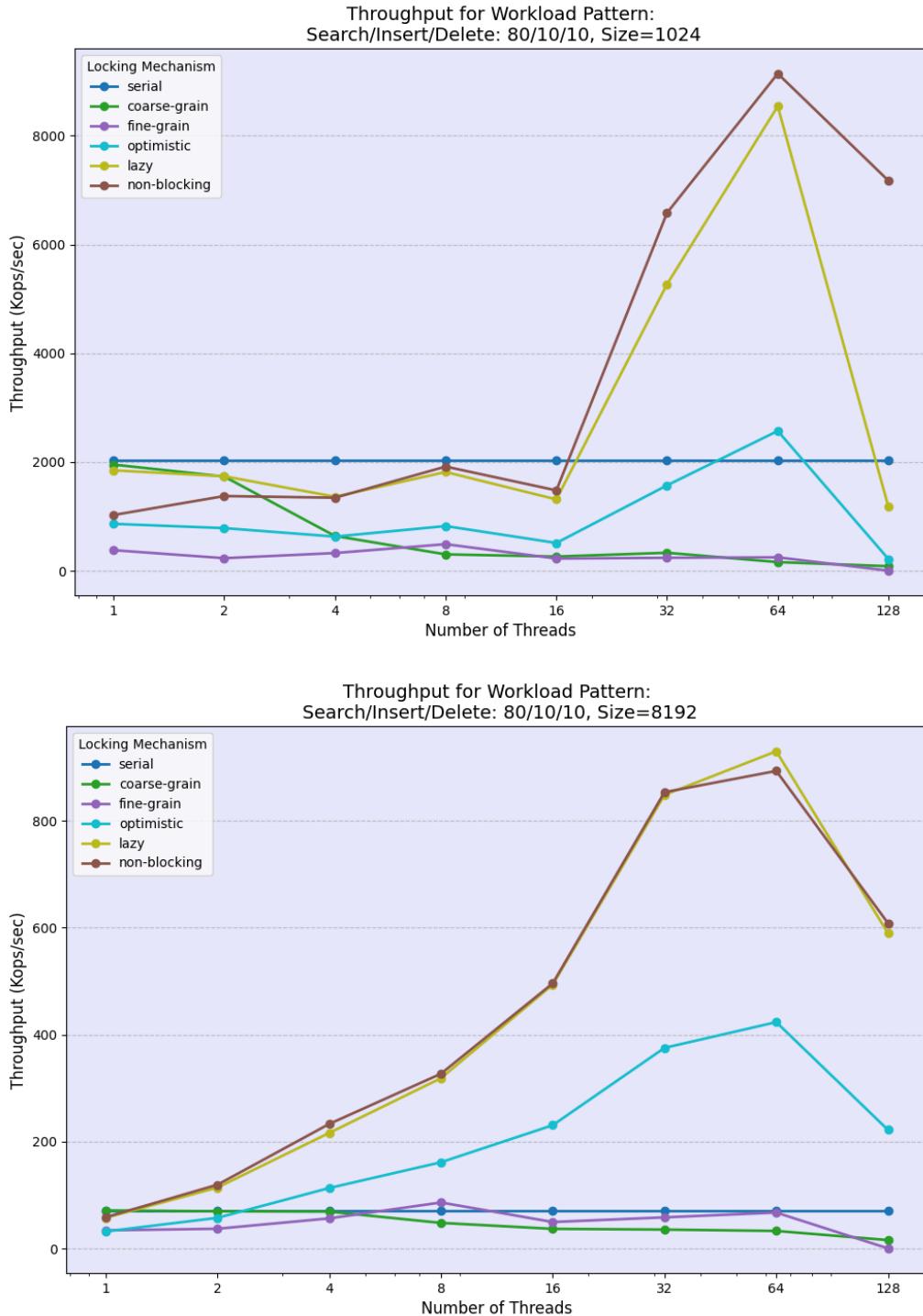


Workload 100/0/0

Στην περίπτωση που το workload αποτελείται εξ'ολοκλήρου από search queries, οι **coarse-grain**, **fine-grain** & και **optimistic** εκδοχές δεν κλιμακώνουν καθόλου για κανένα μέγεθος λίστας αφού μπλοκάρονται όλες σε κάποιο κλείδωμα (είτε είναι global για όλη την δομή στην πρώτη περίπτωση, είτε στην αρχή της λίστας και περνιέται μέσω hand-over-hand για τις άλλες). Ουσιαστικά όλες οι προσβάσεις σειριοποιούνται, οπότε δεν κερδίζουμε τίποτα από τον παραλληλισμό, μονάχα το κόστος δημιουργίας των threads και υλοποίησης των κλειδωμάτων.

Αντίθετα, η **lazy** είναι wait-free στην αναζήτηση επομένως κερδίζουμε throughput χάρη στον παραλληλισμό χωρίς bottlenecks. Στα 64 - 128 νήματα η επίδοση μένει σταθερή γιατί το μηχάνημα

έχει ήδη **100% utilization** (64 logical cores).



Workload 80/10/10

Προσθέτοντας μικρό ποσοστό add / delete βλέπουμε ότι η επιδόσεις όλων είναι καλύτερες, αφού μέρος αυτών των λειτουργιών είναι η συνάρτηση contains που πραγματαποιεί αναζήτηση αλλά με wait free τρόπο. Χειρότερη είναι η **fine-grain** ειδικότερα αν queries στο τέλος της λίστας έπονται από queries στην αρχή της και αναγκάζονται να μπλοκάρουν μέχρι να αποκτήσουν το lock με hand-over-hand τρόπο. Ακόμη, η διαδικασία αυτή εισάγει έξτρα πολυπλοκότητα μέσα από αλλεπάληλες κλήσεις lock / unlock μέχρι να φτάσει στους ζητούμενους κόμβους, οπότε καταλήγει να έχει χειρότερη

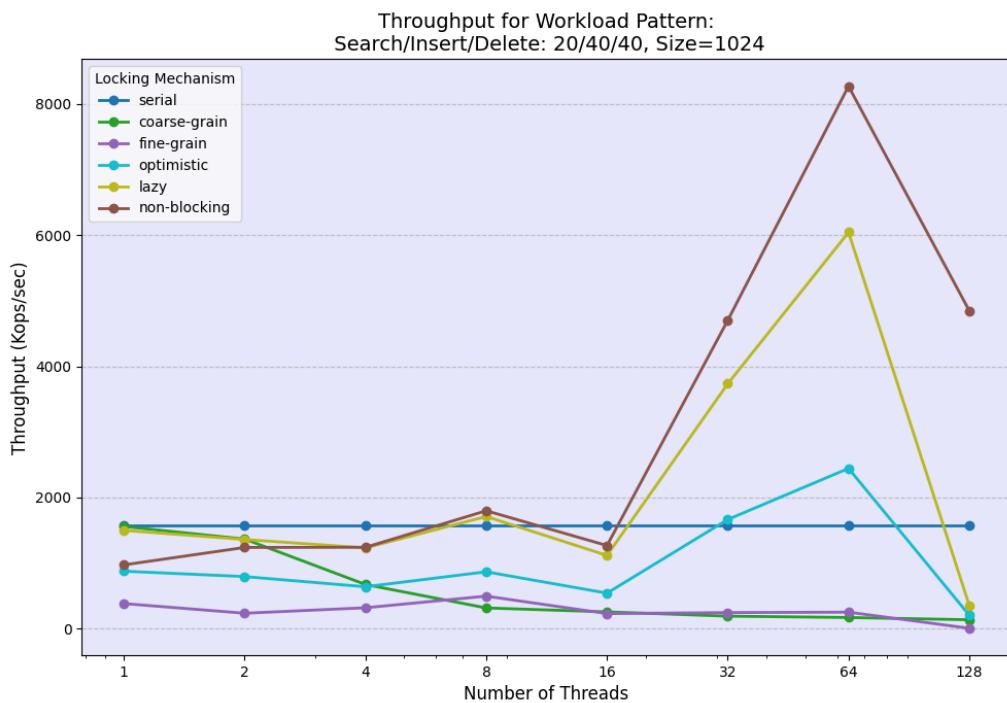
επίδοση από την **coarse-grain**. Η κατάσταση θα μπορούσε να αλλάξει εάν αντιστρέφαμε την σειρά των queries, όμως αυτά παράγονται τυχαία.

Η **optimistic** ακολουθεί την λογική readers-writer lock και διατρέχει την λίστα wait free κάνοντας μετά έναν έλεγχο συνέπειας της δομής (validate). Δεν κλιμακώνει όπως θα περίμεναμε, επειδή η validate έχει γραμμική πολυπλοκότητα και κυριαρχεί το κόστος να διατρέξει την λίστα από την αρχή.

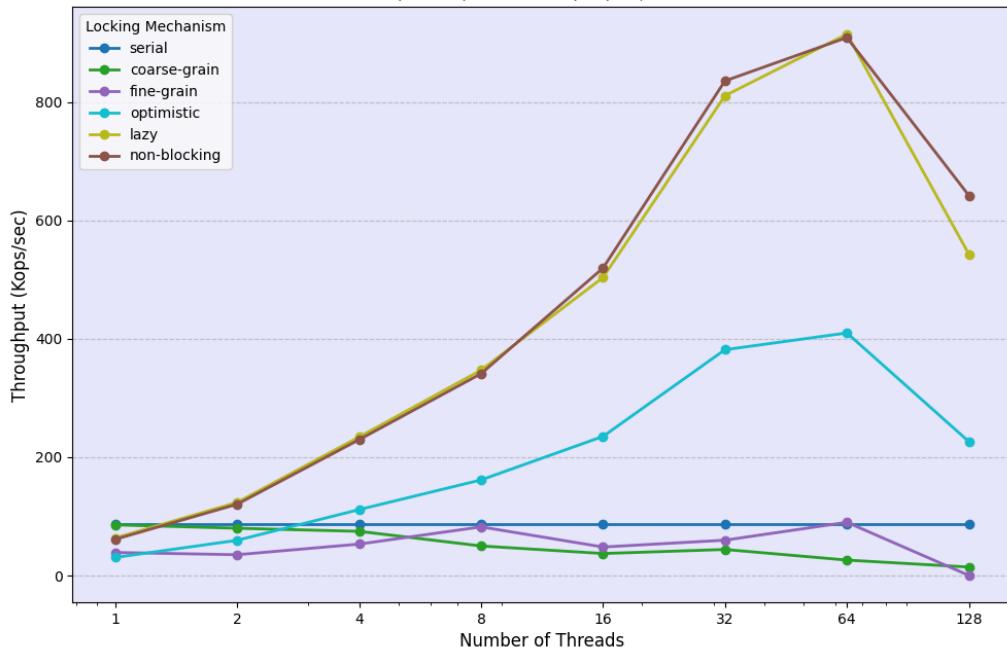
Αντίθετα, η **lazy** υλοποιεί την validate με σταθερό χρόνο, κοιτάζοντας απλά το valid bit για τους κόμβους pred, curr και από 16 threads και πάνω παρουσιάζει τεράστια κλιμάκωση.

Τέλος, η **non-block** είναι εγγενώς wait free και αξιοποιεί packed εντολές που παρέχει το ISA και είναι optimal.

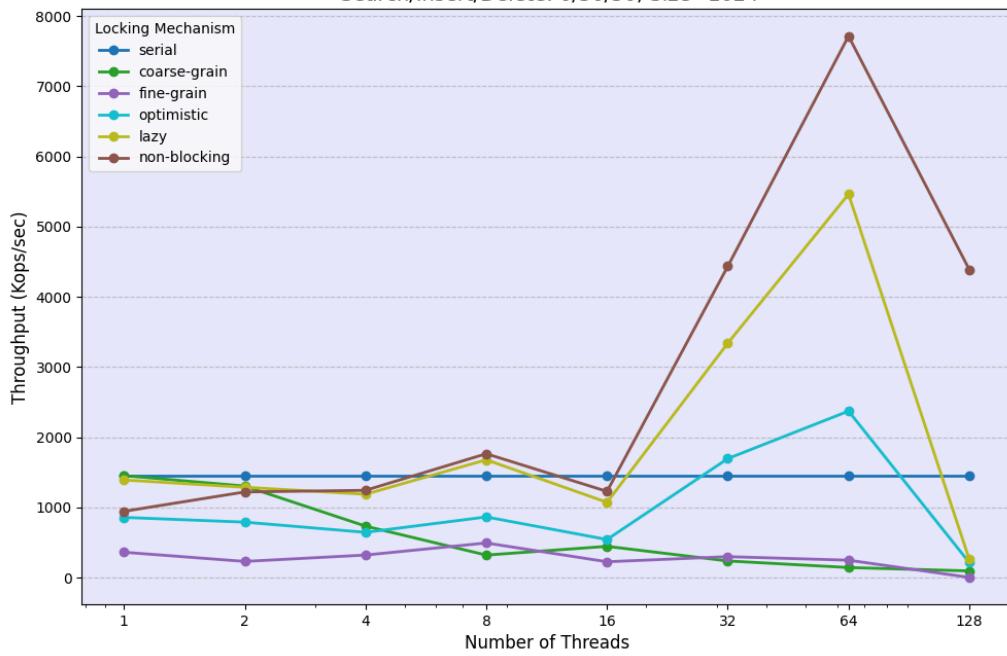
Η μείωση του throughput από τα 8 στα 16 οφείλεται στο ότι τόσο τα locks όσο και τα list data είναι shared στα threads και βγαίνουμε εκτός NUMA cluster. Δεν ισχύει το ίδιο και για μέγεθος 8192 όπου έτσι και αλλιώς η λίστα δεν χωράει ολόκληρη στην cache.

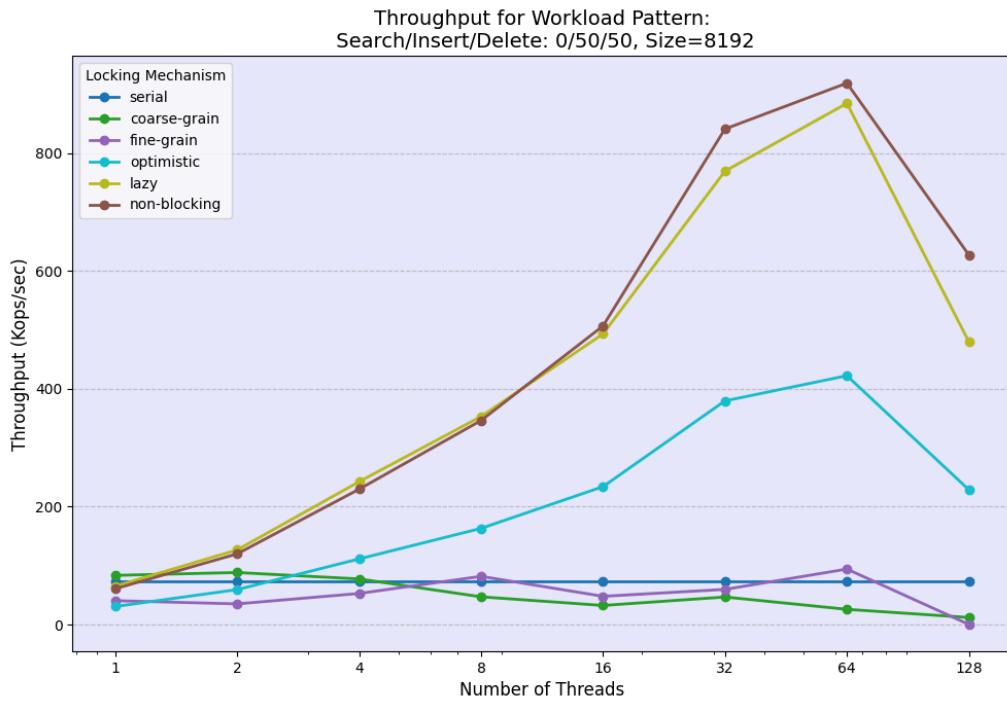


Throughput for Workload Pattern:
Search/Insert/Delete: 20/40/40, Size=8192



Throughput for Workload Pattern:
Search/Insert/Delete: 0/50/50, Size=1024





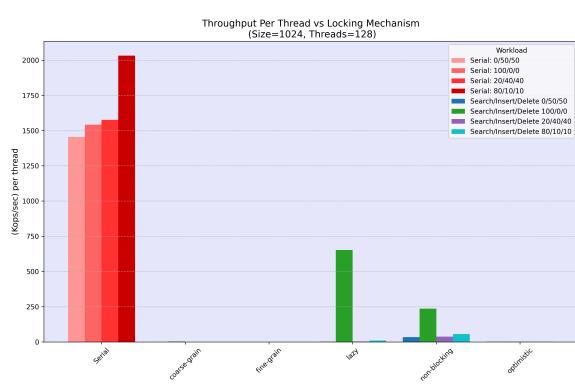
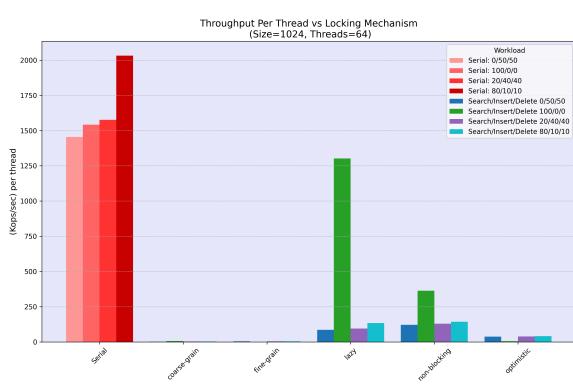
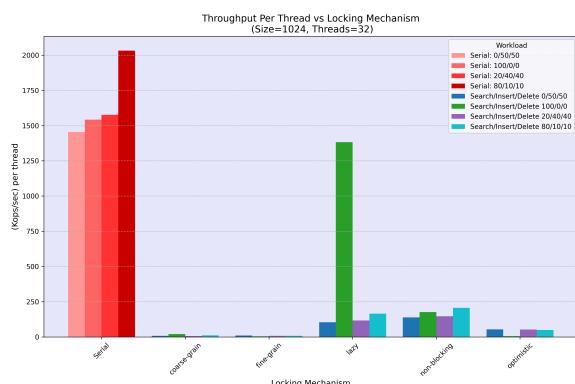
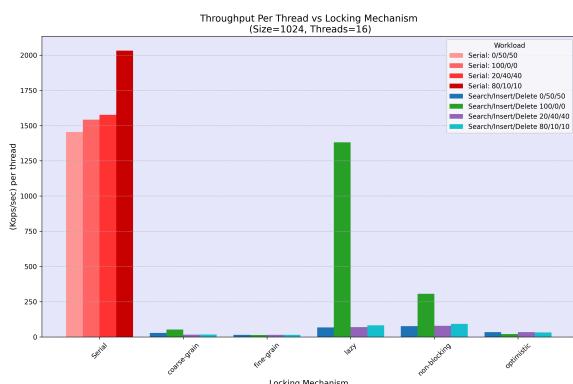
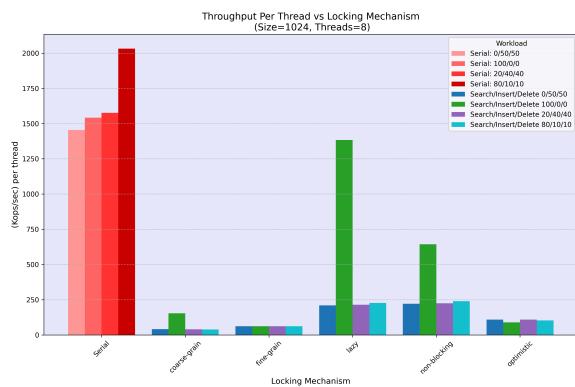
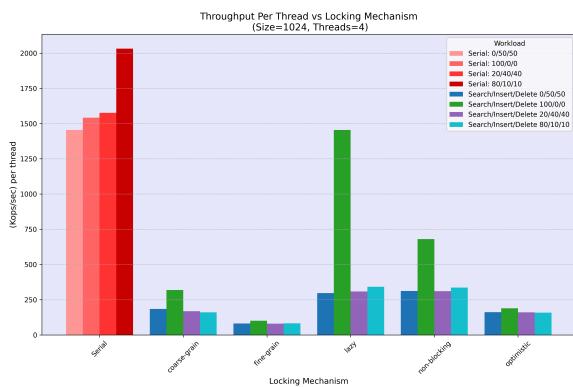
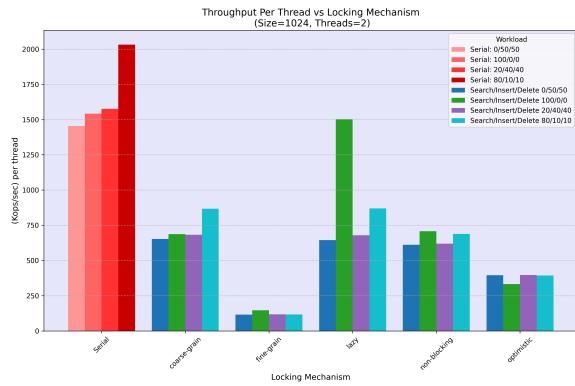
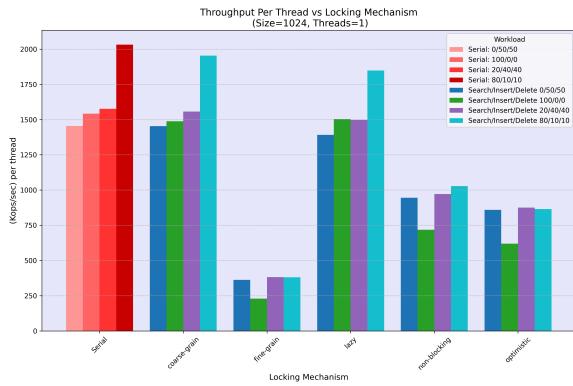
Workload 20/40/40 & Workload 0/50/50

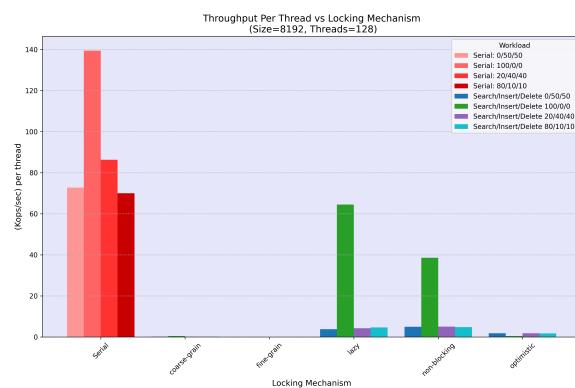
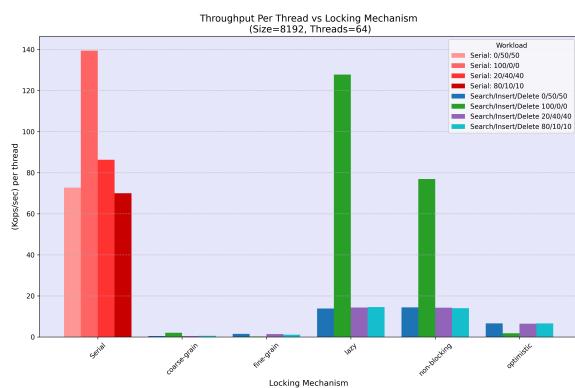
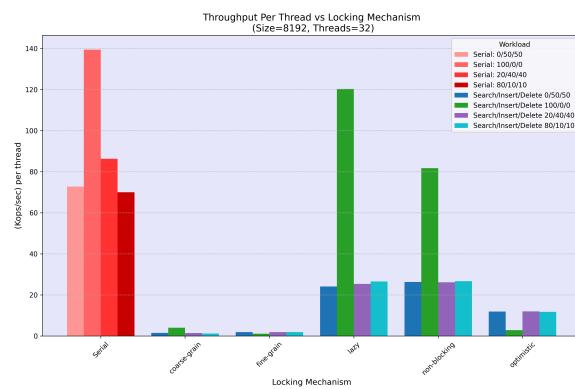
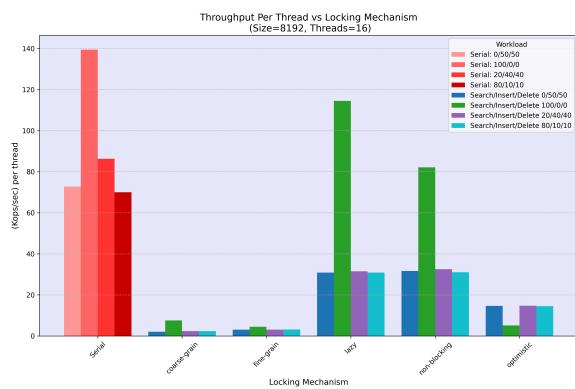
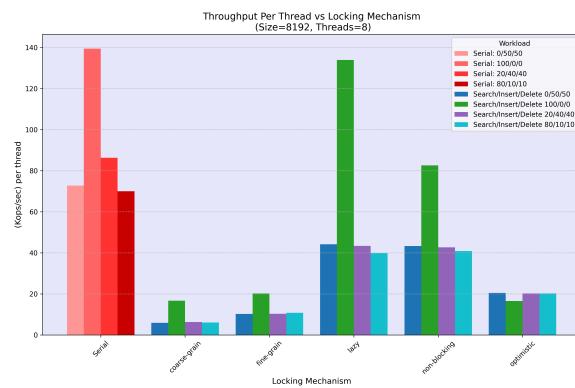
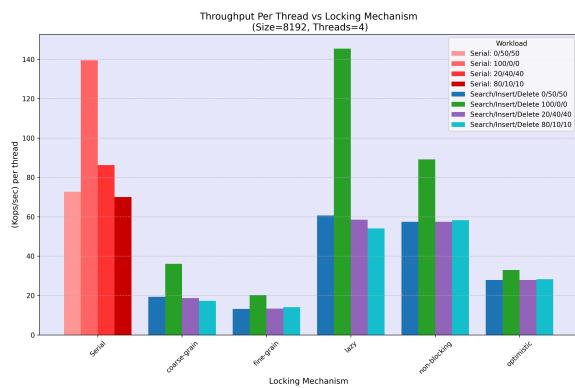
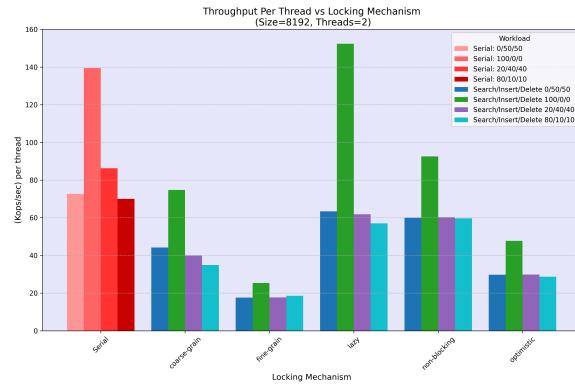
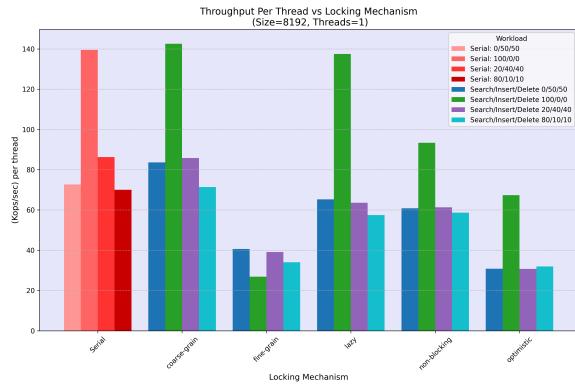
Αυξάνοντας παραπάνω το ποσοστό των add / delete η **lazy** χάνει σε throughput ενώ η **optimistic** κερδίζει στην περίπτωση του μικρού size=1024, και δεν παρουσιάζει αισθητές διαφορές για μεγάλο size=8192. Αυτό συμβαίνει επειδή τα conflicts για μεταβολή κοινών δεδομένων είναι σπανιότερα σε μεγάλο μέγεθος λίστας και εφόσον ο αριθμός των queries μένει σταθερός. Έτσι, το overhead των locking mechanisms μένει σχετικά σταθερό χάρη στο μικρό contention.

Σε όλα τα Workloads,

Το throughput στην πολύ μεγάλη λίστα είναι σημαντικά μικρότερο, αφού όλες οι λειτουργίες είναι γραμμικές ως προς το μέγεθος της λίστας και δεν μπορεί να αξιοποιηθεί πλήρως το data locality. Επιπλέον, η **lazy** τα πάει χειρότερα στο μικρό μήκος από την **non blocking**. Η ειδοποιός διαφορά των δύο αυτών υλοποιήσεων είναι πως η **lazy** στις διαγραφές είναι επίμονη και προσπαθεί να αποκτήσει τα locks των κόμβων που απαιτούνται μέχρι να τα καταφέρει, ενώ η **non blocking** κάνει την λογική διαγραφή και 1 μόνο προσπάθεια φυσικής διαγραφής. Σε workloads με λίγα contains(που κάνει την φυσική αφαίρεση στην **non blocking**) και αρκετά add/remove η **lazy** κάνει συνέχεια και τις φυσικές διαγραφές ενώ η **non blocking** κάνει λίγες και γλυτώνει χρόνο.

Παρακάτω φαίνονται τα κανονικοποιημένα throughputs / thread_count σε σύγκριση με την serial εκδοχή και πως αυτά επηρεάζονται από το workload:





Παραλληλοποίηση και βελτιστοποίηση αλγορίθμων σε επεξεργαστές γραφικών

Σκοπός αυτής της άσκησης είναι η υλοποίηση και η βελτιστοποίηση του αλγορίθμου Kmeans σε μια κάρτα γραφικών της Nvidia με την χρήση της Cuda. Αρχικά, υλοποιούμε μια naïve προσέγγιση και έπειτα αξιολογούμε και συγκρίνουμε άλλες υλοποιήσεις που αφορούν τεχνικές βελτιώσεις σε GPUs.

Naive

Αρχικά υπολογίζουμε το global id στην συνάρτηση `get_id()` ως εξής : $\text{thread_block_size} \times \text{block_id} + \text{local_thread_id}$.

Έπειτα, στη συνάρτηση `euclidean_distance()` υπολογίζεται η ευκλείδια απόσταση μεταξύ δύο σημείων στον n-διάστατο χώρο. Ο πίνακας συντεταγμένων των αντικειμένων καθώς και των clusters είναι μονοδιάστατος. Οπότε, η συντεταγμένη j του cluster i είναι η $\text{cluster}[i \times \text{numCoords} + j]$ καθώς ο δισδιάστατος πίνακας θα είχε διαστάσεις $[\text{numClusters}][\text{numCoords}]$, αντίστοιχα για τα objects. Με ένα for loop υπολογίζεται το άθροισμα όλων των διαφορών στο τετράγωνο. Δεν χρειάζεται να υπολογιστεί η ρίζα, καθώς οι αποστάσεις είναι μη αρνητικές και η ύψωση στο τετράγωνο δεν αλλάζει την μεταξύ τους διάταξη.

Η συνάρτηση `find_nearest_cluster` χρειάζεται να υπολογιστεί μόνο για τα threads που έχουν global id μικρότερο από τον αριθμό των objects, ώστε να μην βρεθεί εκτός ορίων πίνακα και κάθε thread να κάνει έναν υπολογισμό για ένα αντικείμενο. Για να βρεθεί το κοντινότερο cluster, υπολογίζεται αρχικά η απόσταση από το πρώτο cluster και έπειτα υπολογίζεται σειριακά η απόσταση για τα υπόλοιπα. Αν για κάποιο cluster, είναι μικρότερη από την ήδη υπάρχουσα, κρατάμε αυτό στη θέση του προηγούμενου. Ακόμη, γίνεται αύξηση του delta κατά 1, αν διαφέρει από το προηγούμενο κοντινότερο cluster. Χρειάζεται να γίνει με `atomicAdd`, ώστε να έχουμε σωστό συνολικό αποτέλεσμα καθώς είναι κοινό δεδομένο για όλα τα thread blocks και χρειάζεται κατάλληλος συγχρονισμός.

Ο αριθμός των thread blocks υπολογίζεται ως $\frac{\text{numObjects} + \text{blocksize} - 1}{\text{blocksize}}$. Αυτό υπολογίζει την πράξη `ceil` του λόγου του αριθμού των αντικειμένων προς το `blocksize`, ώστε ακόμη και 1 παραπάνω thread να χρειάζεται να δημιουργηθεί καινούριο thread block.

Τέλος αντιγράφουμε τα clusters, memberships, delta χρησιμοποιώντας την εντολή `cudaMemcpy` με κατάλληλο μέγεθος και κατεύθυνση της αντιγραφής.

cuda_kmeans_naive.cu

```
1 #include <stdio.h>
2 #include <stdlib.h>
3
4 #include "kmeans.h"
5 #include "alloc.h"
6 #include "error.h"
7
8 #ifdef __CUDACC__
9 inline void checkCuda(cudaError_t e) {
10     if (e != cudaSuccess) {
11         // cudaGetErrorString() isn't always very helpful. Look up the error
12         // number in the cudaError enum in driver_types.h in the CUDA includes
13         // directory for a better explanation.
14         error("CUDA Error %d: %s\n", e, cudaGetErrorString(e));
15     }
16 }
17
18 inline void checkLastCudaError() {
19     checkCuda(cudaGetLastError());
20 }
21#endif
```

```

22
23 __device__ int get_tid() {
24     return blockDim.x*blockIdx.x + threadIdx.x;
25     //return 0; /* TODO: Calculate 1-Dim global ID of a thread */
26 }
27
28 /* square of Euclid distance between two multi-dimensional points */
29 __host__ __device__ inline static
30 double euclid_dist_2(int numCoords,
31     int numObjs,
32     int numClusters,
33     double *objects,      // [numObjs][numCoords]
34     double *clusters,    // [numClusters][numCoords]
35     int objectId,
36     int clusterId) {
37     int i;
38     double ans = 0.0;
39
40     /* TODO: Calculate the euclid_dist of elem=objectId of objects from elem=clusterId from
41     clusters*/
42     for(i=0; i<numCoords; i++)
43         ans += (objects[objectId*numCoords + i] - clusters[clusterId*numCoords+i]) *
44     (objects[objectId*numCoords + i] - clusters[clusterId*numCoords + i]);
45     return (ans);
46 }
47
48 __global__ static
49 void find_nearest_cluster(int numCoords,
50     int numObjs,
51     int numClusters,
52     double *objects,      // [numObjs][numCoords]
53     double *deviceClusters, // [numClusters][numCoords]
54     int *deviceMembership, // [numObjs]
55     double *devdelta) {
56
57     /* Get the global ID of the thread. */
58     int tid = get_tid();
59
60     /* TODO: Maybe something is missing here... should all threads run this? */
61     if (tid < numObjs) {
62         int index, i;
63         double dist, min_dist;
64
65         /* find the cluster id that has min distance to object */
66         index = 0;
67         /* TODO: call min_dist = euclid_dist_2(...) with correct objectId/clusterId */
68         min_dist = euclid_dist_2(numCoords, numObjs, numClusters, objects, deviceClusters, tid, 0);
69
70         for (i = 1; i < numClusters; i++) {
71             /* TODO: call dist = euclid_dist_2(...) with correct objectId/clusterId */
72             dist = euclid_dist_2(numCoords, numObjs, numClusters, objects, deviceClusters, tid, i);
73             /* no need square root */
74             if (dist < min_dist) { /* find the min and its array index */
75                 min_dist = dist;
76                 index = i;
77             }
78
79             if (deviceMembership[tid] != index) {
80                 /* TODO: Maybe something is missing here... is this write safe? */
81                 atomicAdd(devdelta, 1.0);
82             }
83
84             /* assign the deviceMembership to object objectId */
85             deviceMembership[tid] = index;
86         }
87
88     /**
89     // -----
90     // DATA LAYOUT
91     //
92     // objects      [numObjs][numCoords]
93     // clusters     [numClusters][numCoords]
```

```

94 // newClusters      [numClusters][numCoords]
95 // deviceObjects    [numObjs][numCoords]
96 // deviceClusters   [numClusters][numCoords]
97 // -----
98 //
99 /* return an array of cluster centers of size [numClusters][numCoords] */ 
100 void kmeans_gpu(double *objects,          /* in: [numObjs][numCoords] */
101                 int numCoords,        /* no. features */
102                 int numObjs,         /* no. objects */
103                 int numClusters,     /* no. clusters */
104                 double threshold,   /* % objects change membership */
105                 long loop_threshold,/* maximum number of iterations */
106                 int *membership,    /* out: [numObjs] */
107                 double *clusters,    /* out: [numClusters][numCoords] */
108                 int blockSize) {
109     double timing = wtime(), timing_internal, timer_min = le42, timer_max = 0;
110     double timing_gpu, timing_cpu, timing_transfers, transfers_time = 0.0, cpu_time = 0.0,
111     gpu_time = 0.0;
112     int loop_iterations = 0;
113     int i, j, index, loop = 0;
114     int *newClusterSize; /* [numClusters]: no. objects assigned in each
115                           new cluster */
116     double delta = 0, *dev_delta_ptr;           /* % of objects change their clusters */
117     double **newClusters = (double **) calloc_2d(numClusters, numCoords, sizeof(double));
118
119     double *deviceObjects;
120     double *deviceClusters;
121     int *deviceMembership;
122
123     printf("\n|-----Naive GPU Kmeans-----|\n\n");
124
125     /* initialize membership[] */
126     for (i = 0; i < numObjs; i++) membership[i] = -1;
127
128     /* need to initialize newClusterSize and newClusters[0] to all 0 */
129     newClusterSize = (int *) calloc(numClusters, sizeof(int));
130     assert(newClusterSize != NULL);
131
132     timing = wtime() - timing;
133     printf("t_alloc: %lf ms\n\n", 1000 * timing);
134     timing = wtime();
135
136     int minGridSize, bestblockSize;
137     cudaOccupancyMaxPotentialBlockSize(&minGridSize, &bestblockSize, find_nearest_cluster, 0,
138     0);
139     printf("Naive kmeans, min_grid_size = %d, best_block_size = %d\n\n", minGridSize,
140     bestblockSize);
141
142     //for the first exercise
143     //const unsigned int numThreadsPerClusterBlock = (numObjs > blockSize) ? blockSize : numObjs;
144     //for the upgraded version to find best block size
145     //uncomment properly
146     const unsigned int numThreadsPerClusterBlock = bestblockSize;
147     const unsigned int numClusterBlocks = (numObjs + numThreadsPerClusterBlock - 1) /
148     numThreadsPerClusterBlock; /* TODO: Calculate Grid size, e.g. number of blocks. */
149
150     const unsigned int clusterBlockSharedDataSize = 0;
151
152     checkCuda(cudaMalloc(&deviceObjects, numObjs * numCoords * sizeof(double)));
153     checkCuda(cudaMalloc(&deviceClusters, numClusters * numCoords * sizeof(double)));
154     checkCuda(cudaMalloc(&deviceMembership, numObjs * sizeof(int)));
155     checkCuda(cudaMalloc(&dev_delta_ptr, sizeof(double)));
156
157     timing = wtime() - timing;
158     printf("t_alloc_gpu: %lf ms\n\n", 1000 * timing);
159     timing = wtime();
160
161     checkCuda(cudaMemcpy(deviceObjects, objects,
162                         numObjs * numCoords * sizeof(double), cudaMemcpyHostToDevice));
163     checkCuda(cudaMemcpy(deviceMembership, membership,
164                         numObjs * sizeof(int), cudaMemcpyHostToDevice));
165     timing = wtime() - timing;
166     printf("t_get_gpu: %lf ms\n\n", 1000 * timing);
167     timing = wtime();

```

```

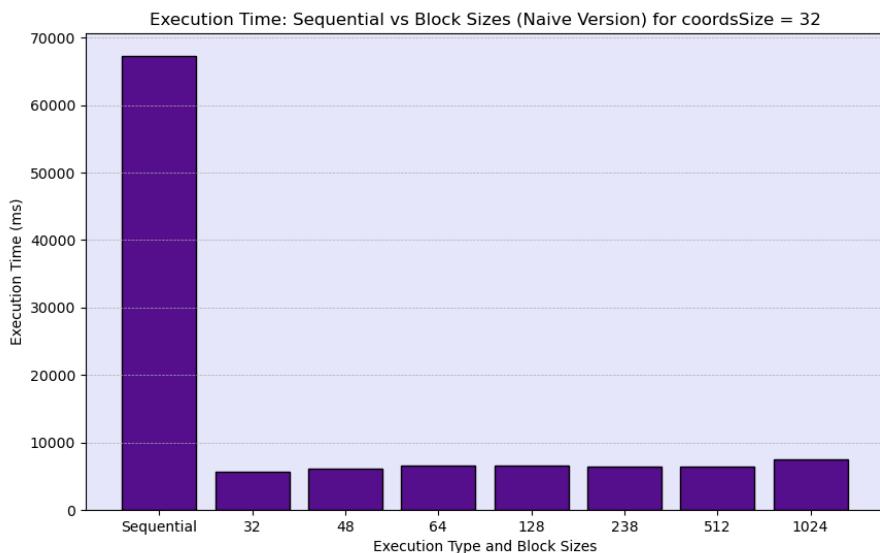
164
165     do {
166         timing_internal = wtime();
167
168         /* GPU part: calculate new memberships */
169
170         timing_transfers = wtime();
171         /* TODO: Copy clusters to deviceClusters*/
172         checkCuda(cudaMemcpy(deviceClusters, clusters, numClusters * numCoords * sizeof(double),
173             cudaMemcpyHostToDevice));
174         transfers_time += wtime() - timing_transfers;
175
176         checkCuda(cudaMemset(dev_delta_ptr, 0, sizeof(double)));
177
178         //printf("Launching find_nearest_cluster Kernel with grid_size = %d,
179         block_size = %d, shared_mem = %d KB\n", numClusterBlocks, numThreadsPerClusterBlock,
180         clusterBlockSharedDataSize/1000);
181         timing_gpu = wtime();
182         find_nearest_cluster
183         <<< numClusterBlocks, numThreadsPerClusterBlock, clusterBlockSharedDataSize >>>
184         (numCoords, numObjs, numClusters,
185          deviceObjects, deviceClusters, deviceMembership, dev_delta_ptr);
186
187         cudaDeviceSynchronize();
188         checkLastCudaError();
189         gpu_time += wtime() - timing_gpu;
190         //printf("Kernels complete for itter %d, updating data in CPU\n", loop);
191
192         timing_transfers = wtime();
193         /* TODO: Copy deviceMembership to membership */
194         checkCuda(cudaMemcpy(membership, deviceMembership, numObjs * sizeof(int),
195             cudaMemcpyDeviceToHost));
196
197         /* TODO: Copy dev_delta_ptr to &delta */
198         checkCuda(cudaMemcpy(&delta, dev_delta_ptr, sizeof(double), cudaMemcpyDeviceToHost));
199         transfers_time += wtime() - timing_transfers;
200
201         /* CPU part: Update cluster centers*/
202         timing_cpu = wtime();
203         for (i = 0; i < numObjs; i++) {
204             /* find the array index of nestest cluster center */
205             index = membership[i];
206
207             /* update new cluster centers : sum of objects located within */
208             newClusterSize[index]++;
209             for (j = 0; j < numCoords; j++) {
210                 newClusters[index][j] += objects[i * numCoords + j];
211             }
212
213             /* average the sum and replace old cluster centers with newClusters */
214             for (i = 0; i < numClusters; i++) {
215                 for (j = 0; j < numCoords; j++) {
216                     if (newClusterSize[i] > 0)
217                         clusters[i * numCoords + j] = newClusters[i][j] / newClusterSize[i];
218                     newClusters[i][j] = 0.0; /* set back to 0 */
219                 }
220                 newClusterSize[i] = 0; /* set back to 0 */
221             }
222
223             delta /= numObjs;
224             //printf("delta is %f - ", delta);
225             loop++;
226             //printf("completed loop %d\n", loop);
227             cpu_time += wtime() - timing_cpu;
228
229             timing_internal = wtime() - timing_internal;
230             if (timing_internal < timer_min) timer_min = timing_internal;
231             if (timing_internal > timer_max) timer_max = timing_internal;
232         } while (delta > threshold && loop < loop_threshold);
233
234         timing = wtime() - timing;
235         printf("nloops = %d : total = %lf ms\n\t-> t_loop_avg = %lf ms\n\t-> t_loop_min = %lf
236         ms\n\t-> t_loop_max = %lf ms\n\t-> t_cpu_avg = %lf ms\n\t-> t_gpu_avg = %lf ms\n\t-> t_transfers_avg = %lf
237         ms\n\n-----|\n",
238

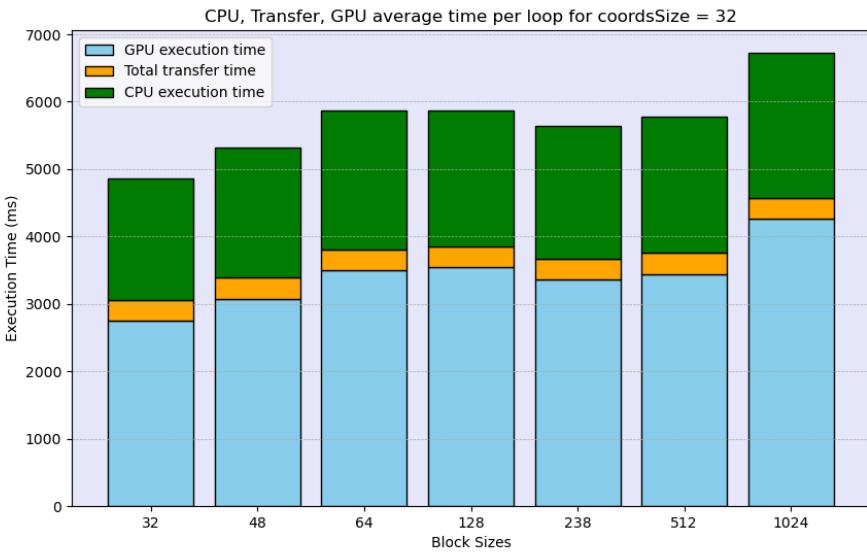
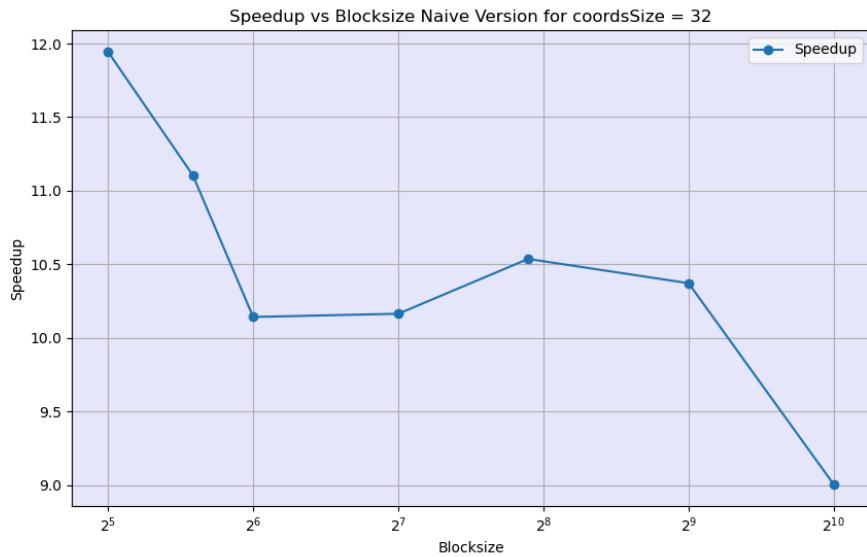
```

```

233     loop, 1000 * timing, 1000 * timing / loop, 1000 * timer_min, 1000 * timer_max,
234     1000 * cpu_time / loop, 1000 * gpu_time / loop, 1000 * transfers_time / loop);
235
236     char outfile_name[1024] = {0};
237     sprintf(outfile_name, "Execution_logs/silver1-V100_Sz-%lu_Coo-%d_Cl-%d.csv",
238             numObjs * numCoords * sizeof(double) / (1024 * 1024), numCoords, numClusters);
239     FILE *fp = fopen(outfile_name, "a+");
240     if (!fp) error("Filename %s did not open successfully, no logging performed\n", outfile_name);
241     fprintf(fp, "%s,%d,%lf,%lf\n", "Naive", blockSize, timing / loop, timer_min, timer_max);
242     fclose(fp);
243     checkCuda(cudaFree(deviceObjects));
244     checkCuda(cudaFree(deviceClusters));
245     checkCuda(cudaFree(deviceMembership));
246
247     free(newClusters[0]);
248     free(newClusters);
249     free(newClusterSize);
250
251     return;
252 }
```

Μετρήθηκαν οι επιδόσεις της σειριακής και της naive έκδοσης για τα διάφορα blocksizes για {Size, Coords, Clusters, Loops} = {1024, 32, 64, 10}. Με βάση αυτές, προκύπτουν τα παρακάτω διαγράμματα χρόνου εκτέλεσης, speedup, χρόνου gpu-cpu-transfer αντίστοιχα.





To speedup του kmeans για την naive έκδοση κυμαίνεται από 9 έως 12 ανάλογα το blocksize. Αυτή η επίδοση είναι αρκετά καλή για την πιο απλή υλοποίηση σε GPU. Ο αλγόριθμος kmeans όμως δεν είναι ιδανικός για GPU, καθώς κάνει αρκετά transfers και allocations. Επίσης το operational intensity δεν είναι το μέγιστο που μπορεί να υποστηρίξει η συγκεκριμένη GPU, καθώς δεν κάνει 6000 πράξεις ανά δεδομένο ώστε να αξιζει πλήρως η μεταφορά του.

Οι χρόνοι μεταφοράς και της CPU είναι προφανώς ίδιοι, οπότε οι αλλαγές του χρόνου της GPU μεταφράζονται άμεσα σε speedup. Το blocksize 32 έχει το καλύτερο, διότι έχει την μεγαλύτερη ευελιξία για το scheduling και κάθε thread block είναι ένα warp. Αφού τα objects είναι τελείως ανεξάρτητα μεταξύ τους για τον υπολογισμό του κοντινότερου cluster, χρησιμοποιούνται στο μέγιστο οι πόροι της GPU. Το blocksize 1024 έχει την μικρότερη ευελιξία, καθώς κάποιοι πόροι δεν αξιοποιούνται στο μέγιστο. Τέλος, το blocksize 238 παρόλο που δεν είναι πολλαπλάσιο του 32 και δεν αξιοποιεί όλα τα warps στο μέγιστο, αφήνονται αναξιοποίητα μόνο 18 threads, οπότε η διαφορά στο σύνολο δεν είναι τόσο μεγάλη και γι' αυτό παρατηρείται μια καλή επίδοση.

Transpose

Οι μονοδιάστατοι πίνακες της GPU είναι πλέον column-based και όχι row-based όπως πριν. Οπότε, ο πίνακας clusters έχει διαστάσεις [numCoords][numClusters]. Η συντεταγμένη j του cluster i είναι η cluster[j × numClusters + i], καθώς πλέον ο πίνακας έχει για κάθε συντεταγμένη μία γραμμή από την τιμή της για κάθε cluster.

transpose_euclid_dist.cu

```
1 __host__ __device__ inline static
2 double euclid_dist_2_transpose(int numCoords,
3                                 int numObjs,
4                                 int numClusters,
5                                 double *objects,           // [numCoords][numObjs]
6                                 double *clusters,          // [numCoords][numClusters]
7                                 int objectId,
8                                 int clusterId) {
9
10    int i;
11    double ans = 0.0, diff;
12
13    /* TODO: Calculate the euclid_dist of elem=objectId of objects from elem=clusterId from
14     clusters, but for column-base format!!! */
15
16    for(i = 0; i < numCoords; i++) {
17        diff = objects[i*numObjs+ objectId] - clusters[i*numClusters + clusterId];
18        ans += diff * diff;
19    }
20
21    return (ans);
22}
```

Η `find_nearest_cluster` μένει απαράλλακτη καθώς αλλάζει μόνο η δομή των δεδομένων. Οι πίνακες `dimObjects`, `dimClusters`, `newClusters` έχουν `numCoords` γραμμές ώστε να είναι column-based και γίνεται κατάλληλο allocation με την `calloc_2d` που μας παρέχεται.

Πριν την εκτέλεση του αλγορίθμου χρειάζεται αντιγραφή των αντικειμένων σε μορφή column-based, οπότε το `dimObjects[j][i] = objects[i][j]`. Ο πίνακας `objects` όμως είναι μονοδιάστατος, ακολουθώντας την ίδια λογική μετατροπής `objects[i][j] = objects[i * numCoords + j]`, καθώς `numCoords` γραμμές θα είχε ο αντίστοιχος δισδιάστατος πίνακας.

transpose allocation.cu

```
1  /* TODO: Transpose dims */
2  double **dimObjects = (double**) calloc_2d(numCoords, numObjs, sizeof(double)); // 
3  calloc_2d(...) -> [numCoords][numObjs]
4  double **dimClusters = (double**) calloc_2d(numCoords, numClusters, sizeof(double)); // 
5  calloc_2d(...) -> [numCoords][numClusters]
6  double **newClusters = (double**) calloc_2d(numCoords, numClusters, sizeof(double)); // 
7  calloc_2d(...) -> [numCoords][numClusters]
8
9  double *deviceObjects;
10 double *deviceClusters;
11 int *deviceMembership;
12
13 printf("\n|-----Transpose GPU Kmeans-----|\n\n");
14
15 // TODO: Copy objects given in [numObjs][numCoords] layout to new
16 // [numCoords][numObjs] layout
17 for (i = 0; i < numObjs; i++) {
18     for (j = 0; j < numCoords; j++) {
19         dimObjects[j][i] = objects[i*numCoords+ j];
20     }
21 }
```

Ο αριθμός των thread blocks δεν αλλάζει, καθώς και οι αντιγραφές πίσω στην CPU μετά τους υπολο-

γισμούς της GPU. Αλλάζει όμως η αντιγραφή των clusters προς την GPU σε dimClusters, καθώς αυτά είναι τα column-based δεδομένα. Ακόμη, αφού ολοκληρωθεί η εκτέλεση του αλγορίθμου χρειάζεται οι συντεταγμένες των τελικών clusters να ξαναγίνουν row-based και κάνουμε την ανάποδη διαδικασία $clusters[i][j] = clusters[i \times numCoords + j] = dimClusters[j][i]$, το οποίο είναι ίδιο με την έκφραση στην αρχή του transpose με κατάλληλη αλλαγή δεικτών. Αναλυτικότερα, για πίνακα m γραμμών και n στηλών το στοιχείο i,j είναι το $i \times n + j$ ή αλλιώς το $j \times m + i$.

transpose_do_while.cu

```

1  do {
2      timing_internal = wtime();
3
4      /* GPU part: calculate new memberships */
5
6      timing_transfers = wtime();
7      /* TODO: Copy clusters to deviceClusters
8      checkCuda(cudaMemcpy(...));
9      checkCuda(cudaMemcpy(deviceClusters, dimClusters[0], numClusters * numCoords *
10         sizeof(double), cudaMemcpyHostToDevice));
11     transfers_time += wtime() - timing_transfers;
12
13     checkCuda(cudaMemset(dev_delta_ptr, 0, sizeof(double)));
14
15     //printf("Launching find_nearest_cluster Kernel with grid_size = %d, block_size
16     = %d, shared_mem = %d KB\n", numClusterBlocks, numThreadsPerClusterBlock,
17     clusterBlockSharedDataSize/1000);
18     timing_gpu = wtime();
19     find_nearest_cluster
20     <<< numClusterBlocks, numThreadsPerClusterBlock, clusterBlockSharedDataSize >>>
21     (numCoords, numObjs, numClusters,
22     deviceObjects, deviceClusters, deviceMembership, dev_delta_ptr);
23
24     cudaDeviceSynchronize();
25     checkLastCudaError();
26     gpu_time += wtime() - timing_gpu;
27     //printf("Kernels complete for itter %d, updating data in CPU\n", loop);
28
29     timing_transfers = wtime();
30     /* TODO: Copy deviceMembership to membership
31     checkCuda(cudaMemcpy(...));
32     checkCuda(cudaMemcpy(membership, deviceMembership, numObjs * sizeof(int),
33     cudaMemcpyDeviceToHost));
34     transfers_time += wtime() - timing_transfers;
35
36     /* CPU part: Update cluster centers*/
37
38     timing_cpu = wtime();
39     for (i = 0; i < numObjs; i++) {
40         /* find the array index of nestest cluster center */
41         index = membership[i];
42
43         /* update new cluster centers : sum of objects located within */
44         newClusterSize[index]++;
45         for (j = 0; j < numCoords; j++)
46             newClusters[j][index] += objects[i * numCoords + j];
47     }
48
49     /* average the sum and replace old cluster centers with newClusters */
50     for (i = 0; i < numClusters; i++) {
51         for (j = 0; j < numCoords; j++) {
52             if (newClusterSize[i] > 0)
53                 dimClusters[j][i] = newClusters[j][i] / newClusterSize[i];
54             newClusters[j][i] = 0.0; /* set back to 0 */
55         }
56         newClusterSize[i] = 0; /* set back to 0 */
57     }
58
59     delta /= numObjs;

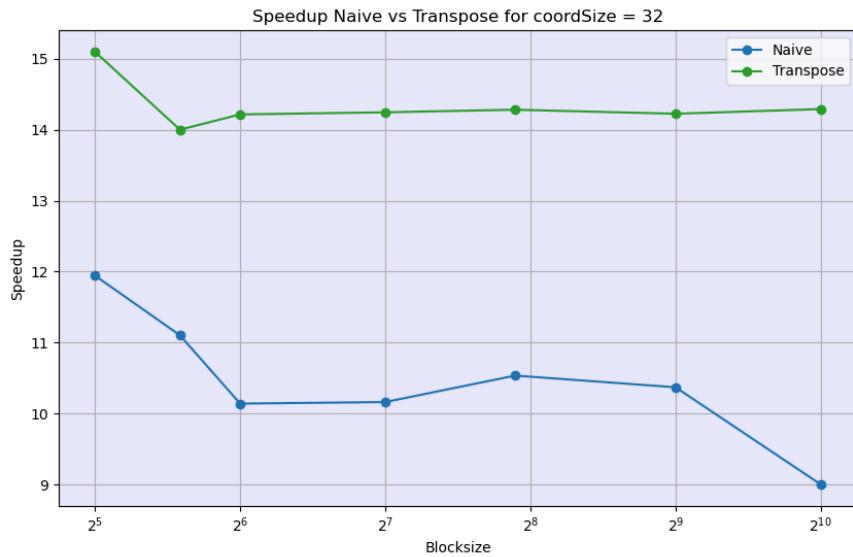
```

```

60     //printf("delta is %f - ", delta);
61     loop++;
62     //printf("completed loop %d\n", loop);
63     cpu_time += wtime() - timing_cpu;
64
65     timing_internal = wtime() - timing_internal;
66     if (timing_internal < timer_min) timer_min = timing_internal;
67     if (timing_internal > timer_max) timer_max = timing_internal;
68 } while (delta > threshold && loop < loop_threshold);
69
70 /*TODO: Update clusters using dimClusters. Be carefull of layout!!! clusters[numClusters]
71 [numCoords] vs dimClusters[numCoords][numClusters] */
72 for (i = 0; i < numClusters; i++) {
73     for (j = 0; j < numCoords; j++) {
74         clusters[i*numCoords + j] = dimClusters[j][i];
75     }

```

Επαναλάβαμε τις μετρήσεις για την transpose εκδοχή και παρακάτω παρουσιάζεται το διάγραμμα speedup σε σύγκριση με την naive εκδοχή.



Παρατηρούμε πως το blocksize παίζει τελείως διαφορετικό ρόλο σε σχέση με την naive περίπτωση. Σε κάθε γραμμή βρίσκεται η τιμή της ίδιας συντεταγμένης για όλα τα objects, clusters. Οπότε, όταν ένα thread κάνει access μια συγκεκριμένη συντεταγμένη θα έρθουν στην cache οι τιμές της αντίστοιχης συντεταγμένης για επόμενα objects, clusters. Καθώς τα warps προχωράνε στις συντεταγμένες στην GPU, θα υπάρχουν στην L1 cache στοιχεία που θα χρησιμοποιήσουν τα επόμενα warps ή επόμενα thread του ίδιου warp. Έτσι, καλύπτεται το global memory latency, αφού περιμένουν πολύ λιγότερα warps στοιχεία για να κάνουν τους υπολογισμούς τους. Όταν έρχεται 1 cache line, θα καλύψει σύγουρα όσα threads υπολογίζουν συντεταγμένες που αντιστοιχούν σε αυτήν την cache line. Αυτό οδηγεί σε πολύ καλύτερη καλύψη των κενών χρόνων μεταξύ των warps και thread blocks, οπότε τα μεγάλα blocksizes έχουν σχεδόν όλα την ίδια επίδοση. Το blocksize 32 συνεχίζει να έχει αισθητά καλύτερα επίδοση λόγω ευελιξίας, όπως και στην naive εκδοχή.

Shared

Χρειάζεται να επεκτείνουμε την `find_nearest_cluster` κατάλληλα. Η κοινή μνήμη έχει εμβέλεια στο thread block και είναι πολύ πιο γρήγορη από την global μνήμη. Γι' αυτό μπορούν να αποθηκευτούν σε αυτήν οι συντεταγμένες των clusters που τις χρειάζεται όλο το thread block, ώστε να έχει γρήγορη πρόσβαση σε αυτές.

Το πρώτο βήμα είναι να αντιγραφούν οι συντεταγμένες των clusters στην shared memory για κάθε thread block. Τα clusters είναι 64 και το ελάχιστο blocksize είναι 32, οπότε σε αυτήν την περίπτωση θα πρέπει κάθε thread μέσα στο thread block να αντιγράψει 2 ολόκληρα clusters. Γενικότερα όμως είναι καλή πρακτική για να τρέχει για οποιοδήποτε μέγεθος blocksize και αριθμό cluster, να θεωρούμε πως κάθε thread έχει υπό την υπόλοιψή του παραπάνω από 1 cluster.

Για να διαχωριστούν σωστά και ισάξια τα clusters κάθε thread ξεκινάει από το local id του και προχωράει με βήμα όσο το thread block, μέχρι να τελειώσει ο αριθμός των clusters. Υπενθυμίζεται πως τα clusters είναι column-based όπως και στην transpose εκδοχή. Οι υπολογισμοί των κοντινότερων clusters πρέπει να ξεκινήσουν αφού αντιγραφούν όλα τα clusters μέσα στο thread block. Οπότε, χρειάζεται ένας συγχρονισμός των νημάτων για να είναι σίγουρο πως θα έχει γίνει αυτό.

Τέλος, για να αξιοποιηθούν σωστά, στην θέση των `deviceClusters` μπαίνει ο πίνακας της διαμοιραζόμενης μνήμης που μόλις γεμίσαμε σαν παράμετρος στην κλήση υπολογισμού της ευκλείδιας απόστασης.

shared_find_nearest_cluster.cu

```
1  __global__ static
2  void find_nearest_cluster(int numCoords,
3                           int numObjs,
4                           int numClusters,
5                           double *objects,
6                           double *deviceClusters,    // [numCoords] [numObjs]
7                           int *deviceMembership,     // [numClusters] [numObjs]
8                           double *devdelta) {
9      extern __shared__ double shmemClusters[];
10
11     /* TODO: Copy deviceClusters to shmemClusters so they can be accessed faster.
12        BEWARE: Make sure operations is complete before any thread continues... */
13     int no_cluster, i;
14
15     //use local_id because shared memory is per thread block
16     for (no_cluster = threadIdx.x; no_cluster < numClusters; no_cluster+=blockDim.x) {
17         for (i = 0; i < numCoords; i++) {
18             shmemClusters[i * numClusters + no_cluster] = deviceClusters[i * numClusters + no_cluster];
19         }
20     }
21     __syncthreads();
22
23     /* Get the global ID of the thread. */
24     int tid = get_tid();
25
26     /* TODO: Maybe something is missing here... should all threads run this? */
27     if (tid < numObjs) {
28         int index;
29         double dist, min_dist;
30
31         /* find the cluster id that has min distance to object */
32         index = 0;
33         /* TODO: call min_dist = euclid_dist_2(...) with correct objectId/clusterId using clusters
34         in shmem*/
35         min_dist = euclid_dist_2_transpose(numCoords, numObjs, numClusters, objects, shmemClusters,
36         tid, 0);
37         for (i = 1; i < numClusters; i++) {
38             /* TODO: call dist = euclid_dist_2(...) with correct objectId/clusterId using clusters
39             in shmem*/
40             dist = euclid_dist_2_transpose(numCoords, numObjs, numClusters, objects, shmemClusters,
41             tid, i);
```

```

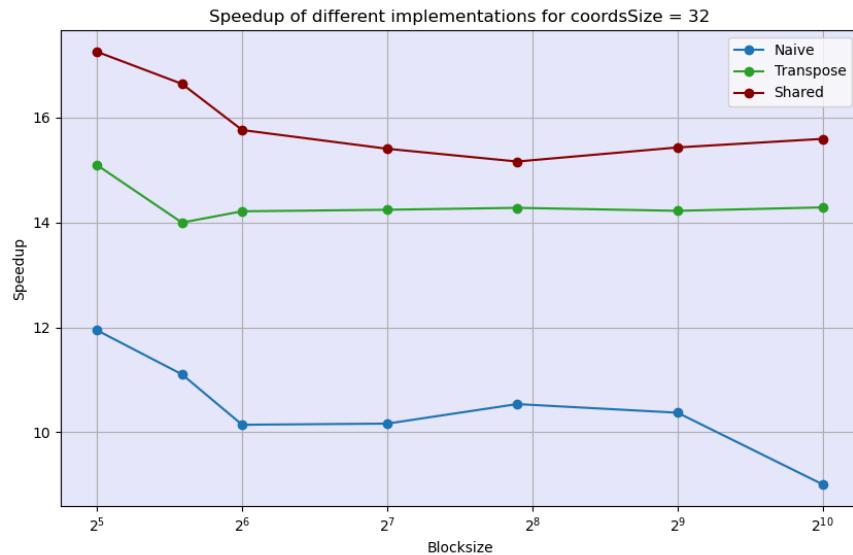
38     /* no need square root */
39     if (dist < min_dist) { /* find the min and its array index */
40         min_dist = dist;
41         index = i;
42     }
43 }
44
45 if (deviceMembership[tid] != index) {
46     /* TODO: Maybe something is missing here... is this write safe? */
47     atomicAdd(devdelta, 1.0);
48 }
49
50 /* assign the deviceMembership to object objectId */
51 deviceMembership[tid] = index;
52 }
53
54 }
55 }
```

Το μέγεθος της διαμοιραζόμενης μνήμης χρειάζεται να δηλωθεί στην κλήση του GPU kernel. Η διαμοιραζόμενη μνήμη θα έχει $\text{numClusters} \times \text{numCoords}$ πραγματικούς αριθμούς. Οπότε:

`const unsigned int clusterBlockSharedDataSize = numClusters * numCoords * sizeof(double);`

Οι υπόλοιπες διαδικασίες αντιγραφής και μετατροπής των clusters παραμένουν ίδιες με την transpose εκδοχή.

Επαναλάβαμε τις μετρήσεις για την shared εκδοχή και παρακάτω παρουσιάζεται το διάγραμμα speedup σε σύγκριση με τις υπόλοιπες εκδοχές.

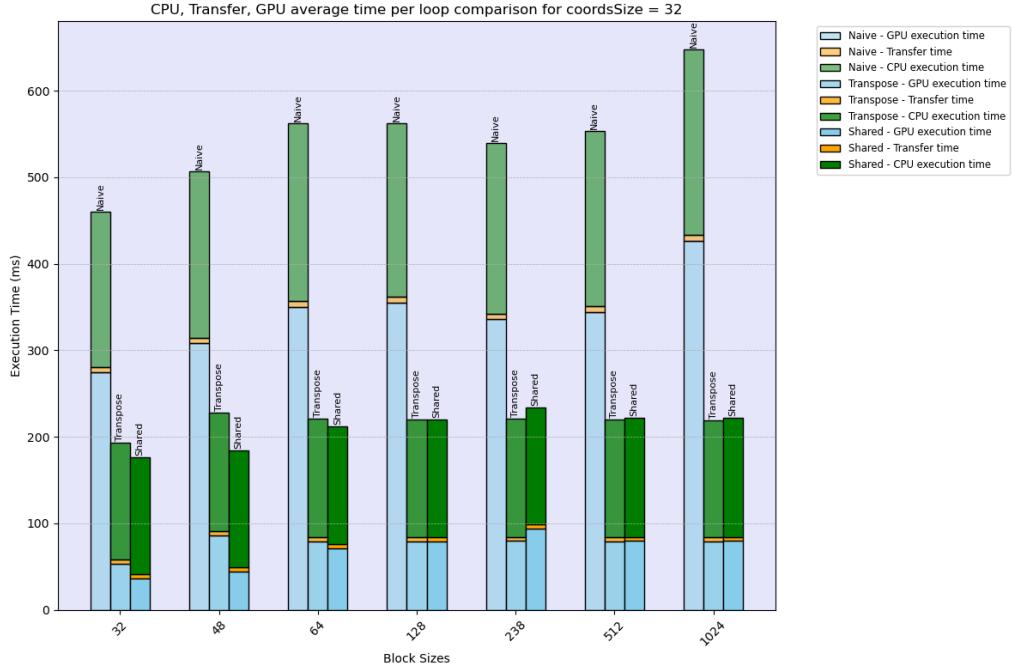


Παρατηρείται μια πτώση επίδοσης για blocksize > 64 και προφανώς, επειδή δεν είναι πολλαπλάσιο του 32 και σπαταλάει πόρους, το blocksize 238 έχει την χειρότερη επίδοση. Υπάρχουν 3 πιθανές εξηγήσεις για αυτό:

- 1)Καθυστέρηση λόγω δυσκολίας συγχρονισμού πολλών threads μέσα στο thread block
- 2)Η διαμοιραζόμενη μνήμη δέχεται υπερβολικά κοινά αιτήματα από τα threads
- 3)Με μεγάλα blocksizes υπάρχουν λίγα thread blocks ανά SM, δεν είναι τόσο ισομερώς κατανημεμένα, οπότε δεν γίνεται η κολύτερη αξιοποίηση των πόρων.

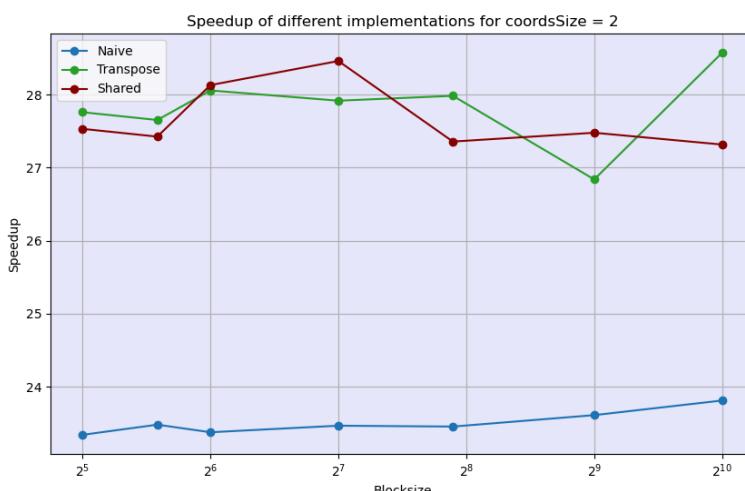
Σύγκριση υλοποιήσεων / bottleneck Analysis

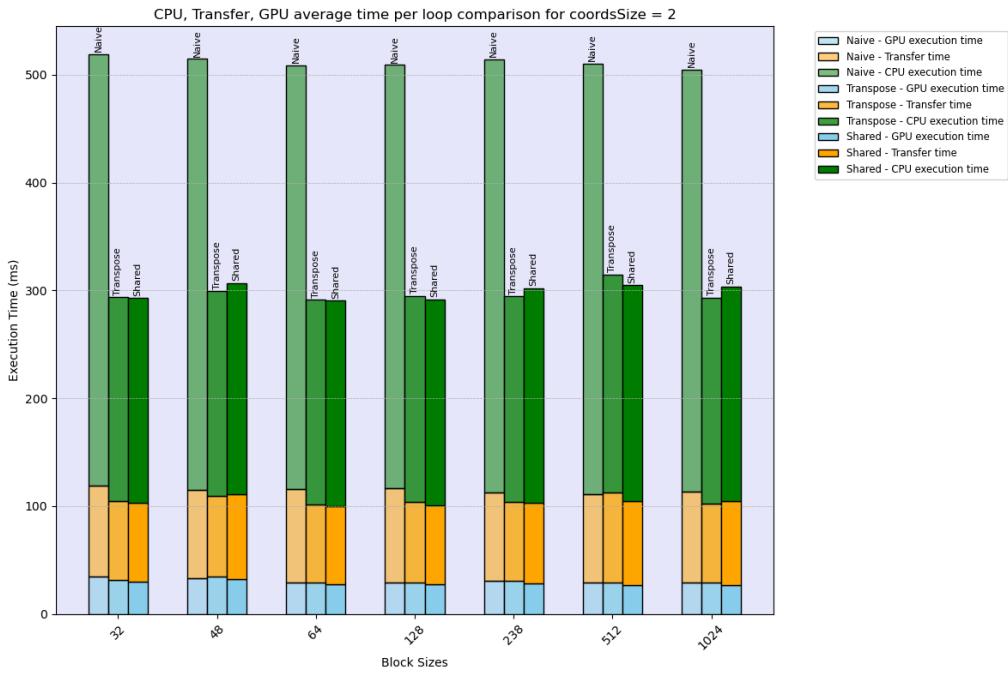
Για καλύτερη ανάλυση των επιμέρους υλοποιήσεων ακολουθεί ένα διάγραμμα ανάλυσης των επιμέρους χρόνων CPU, GPU, transfers ανά επανάληψη.



Όπως είναι λογικό οι χρόνοι μεταφορών και της CPU είναι σχεδόν ίδιοι μεταξύ των υλοποιήσεων. Παρατηρείται πως στα blocksizes με τις καλύτερες επιδόσεις (32, 64), ο χρόνος εκτέλεσης στην GPU είναι σημαντικά μικρότερος από το χρόνο εκτέλεσης στην CPU για τον υπολογισμό των καινούριων clusters. Οπότε, μια ιδέα θα ήταν να μεταφερθεί όλος ο φόρτος των επαναλήψεων στην GPU, ακόμη και να μην είναι ιδανικό για GPU, όπως θα δούμε και παρακάτω.

Επαναλάβαμε τις μετρήσεις για όλες τις εκδοχές για $\{Size, Coords, Clusters, Loops\} = \{1024, 2, 64, 10\}$ για τα πιθανά blocksizes. Το μέγεθος του προβλήματος παραμένει το ίδιο, αλλά μειώθηκε ο αριθμός των συντεταγμένων από 32 σε 2, οπότε αυξήθηκε αντίστοιχα ο αριθμός των αντικειμένων, άρα και των thread blocks. Ακολουθούν τα διαγράμματα για speedup και ανάλυση χρόνου επανάληψης για το καινούριο configuration.





Καθώς οι συντεταγμένες είναι μόνο 2, ένα cache line περιέχει παραπάνω από 1 συντεταγμένη για όλα τα cluster. Σε αυτήν την περίπτωση υπάρχει παρόμοιο locality και χωρίς διαμοιραζόμενη μνήμη, οπότε οι transpose, shared διαφέρουν ελάχιστα μεταξύ τους στις επιδόσεις. Η shared έχει χειρότερη επίδοση σε μεγάλα blocksizes καθώς η ζήτηση για τα cache lines που πλέον είναι λιγότερα αυξάνεται σημαντικά με αποτέλεσμα η μνήμη να μην μπορεί να καλύψει την ζήτηση με την ίδια ταχύτητα. Η παρούσα shared υλοποίηση δεν είναι κατάλληλη για την επίλυση του kmeans για arbitrary configurations. Για να λειτουργήσει σωστά η ιδέα της διαμοιραζόμενης μνήμης δεν πρέπει τα cache lines να περιέχουν υπερβολικά μικρά δεδομένα, όπως είδαμε και στο copied clusters με first-touch policy και την βελτίωση με το numa-aware.

Bonus1: σε όλες τις περιπτώσεις (και στις επόμενες υλοποιήσεις) η cudaOccupancyMaxPotentialBlockSize επιστρέφει 1024 που είναι το μέγιστο blocksize. Για το configuration με τις πολλές συντεταγμένες (32) αυτή είναι ίσως η χειρότερη επιλογή. Ενώ για τις λίγες συντεταγμένες (2) το 1024 είναι το κατάλληλο blocksize μόνο για την transpose εκδοχή και για τις υπόλοιπες ή έχει παρόμοια επίδοση με άλλα blocksizes ή χειρότερη. Επειδή η συνάρτηση δεν λαμβάνει υπόψιν το μέγεθος του προβλήματος, δεν δύναται να δώσει κατάλληλο blocksize για όλες τις λύσεις. Δυστυχώς όμως, δεν δίνει κατάλληλο blocksize για κανένα από τα 2 configurations.

Full-offload (All-GPU)

Υπάρχουν πολλοί τρόποι να γίνει η υλοποίηση της update_centroid. Επιλέξαμε να κάνουμε την πρόσθεση των συντεταγμένων για τα καινούρια clusters στην find_nearest_cluster με atomicAdds, ώστε να αξιοποιηθεί ο ισομερισμός της συνολικής δουλειάς. Επειδή ο αριθμός των clusters \times τον αριθμό των συντεταγμένων είναι επαρκώς μεγάλος για τις περισσότερες περιπτώσεις, τα collisions που θα χρειαστούν όντως συγχρονισμό είναι πολύ λιγότερα απ' όσα φαίνονται αρχικά. Ακόμη, ενημερώνεται με atomicAdd και το μέγεθος του cluster που είναι το πιο κοντινό σε κάθε σημείο.

Στην update centroids κάθε thread αναλαμβάνει μία μόνο συντεταγμένη ενός νέου cluster και την διαιρεί με το μέγεθός του. Έπειτα, μηδενίζει την αντίστοιχη συντεταγμένη των, υπολογισμένων από την find_nearest_cluster, clusters ώστε να μπορεί να ξαναξεκινήσει επανάληψη το do-while από την αρχή σωστά. Χρειάζεται όμως να μηδενιστούν και τα μεγέθη αυτών των clusters. Αυτό μπορεί να γίνει μόνο αφού διαιρεθούν όλες οι συντεταγμένες, οπότε χρειάζεται συγχρονισμός των νημάτων και στο τέλος να μηδενιστούν τα μεγέθη. Σημειώνεται πως επειδή η δομή είναι do-while και όχι απλό while, χρειάζεται με την cudaMemcpy να μηδενιστούν αρχικά τα devicenewClusters, ώστε οι προσθέσεις να είναι valid στην πρώτη κλήση του kernel find_nearest_cluster.

all_gpu_calculations.cu

```
1  __global__ static
2  void find_nearest_cluster(int numCoords,
3                           int numObjs,
4                           int numClusters,
5                           double *deviceobjects,           // [numCoords][numObjs]
6   /*
7      TODO: If you choose to do (some of) the new centroid calculation
8      here, you will need some extra parameters here (from "update_centroids").
9   */
10  int *devicenewClusterSize,           // [numClusters]
11  double *devicenewClusters,          // [numCoords][numClusters]
12  double *deviceClusters,            // [numCoords][numClusters]
13  int *deviceMembership,             // [numObjs]
14  double *devdelta) {
15  extern __shared__ double shmemClusters[];
16  /* TODO: copy me from shared version... */
17  int no_cluster, i;
18
19  //use local_id because shared memory is per thread block
20  for (no_cluster = threadIdx.x; no_cluster < numClusters; no_cluster+=blockDim.x) {
21      for (i = 0; i < numCoords; i++) {
22          shmemClusters[i * numClusters + no_cluster] = deviceClusters[i * numClusters + no_cluster];
23      }
24  }
25  __syncthreads();
26
27  /* Get the global ID of the thread. */
28  int tid = get_tid();
29
30  /* TODO: copy me from shared version... */
31  if (tid < numObjs) {
32      int index;
33      double dist, min_dist;
34
35      /* find the cluster id that has min distance to object */
36      index = 0;
37      /* TODO: call min_dist = euclid_dist_2(...) with correct objectId/clusterId using clusters
38      in shmem*/
39      min_dist = euclid_dist_2_transpose(numCoords, numObjs, numClusters, deviceobjects,
40      shmemClusters, tid, 0);
41      for (i = 1; i < numClusters; i++) {
42          /* TODO: call dist = euclid_dist_2(...) with correct objectId/clusterId using clusters
43          in shmem*/
44          dist = euclid_dist_2_transpose(numCoords, numObjs, numClusters, deviceobjects,
45          shmemClusters, tid, i);
```

```

42     /* no need square root */
43     if (dist < min_dist) { /* find the min and its array index */
44         min_dist = dist;
45         index = i;
46     }
47 }
48
49 if (deviceMembership[tid] != index) {
50     /* TODO: Maybe something is missing here... is this write safe? */
51     atomicAdd(devdelta, 1.0);
52 }
53
54 /* assign the deviceMembership to object objectId */
55 deviceMembership[tid] = index;
56
57 /* TODO: additional steps for calculating new centroids in GPU? */
58 //we chose to update the size and do the add here
59 //the division and the actual new Coords will be in update centroids
60 atomicAdd(&devicenewClusterSize[index], 1);
61 for (i = 0; i < numCoords; i++)
62     atomicAdd(&devicenewClusters[i * numClusters + index], deviceobjects[i * numObjs + tid]);
63
64 }
65 }
66
67 __global__ static
68 void update_centroids(int numCoords,
69                     int numClusters,
70                     int *devicenewClusterSize,           // [numClusters]
71                     double *devicenewClusters,        // [numCoords][numClusters]
72                     double *deviceClusters)          // [numCoords][numClusters])
73 {
74
75     /* TODO: additional steps for calculating new centroids in GPU? */
76     int tid = get_tid();
77
78     if (tid < numCoords * numClusters) {
79         /*run through all the elements, just divide by the size of the clusters
80         indexing of the 1d colummn based devicenewClusters is i*numClusters + j
81         so the index of the current cluster is the j, and i the Coords
82         so the index of the current clusters is (i*numClusters + j) % numClusters
83         here the tid runs all the array increasingly so it is i*numClusters + j
84         */
85         deviceClusters[tid] = devicenewClusters[tid] / devicenewClusterSize[tid % numClusters];
86         //reset devicenewClusters after updating deviceClusters
87         devicenewClusters[tid] = 0.0;
88     }
89     __syncthreads();
90     //reset devicenewClusterSize as well
91     if (tid < numClusters) {
92         devicenewClusterSize[tid] = 0;
93     }
94 }
95 }
```

To kernel `find_nearest_cluster` χρειάζεται να κληθεί με περισσότερες παραμέτρους και ίδιο μέγεθος shared memory, ενώ η κλήση του kernel `update_centroid` δεν χρειάζεται καθόλου shared memory.

all_gpu_do_while.cu

```

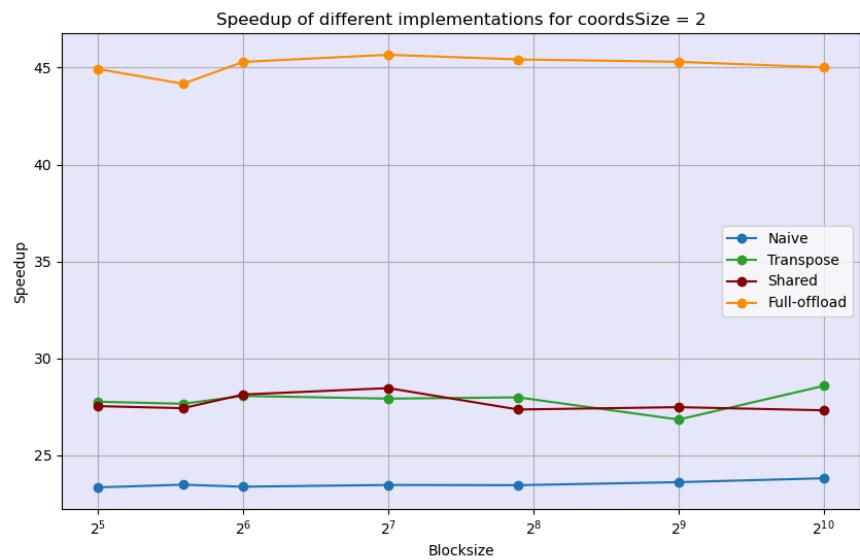
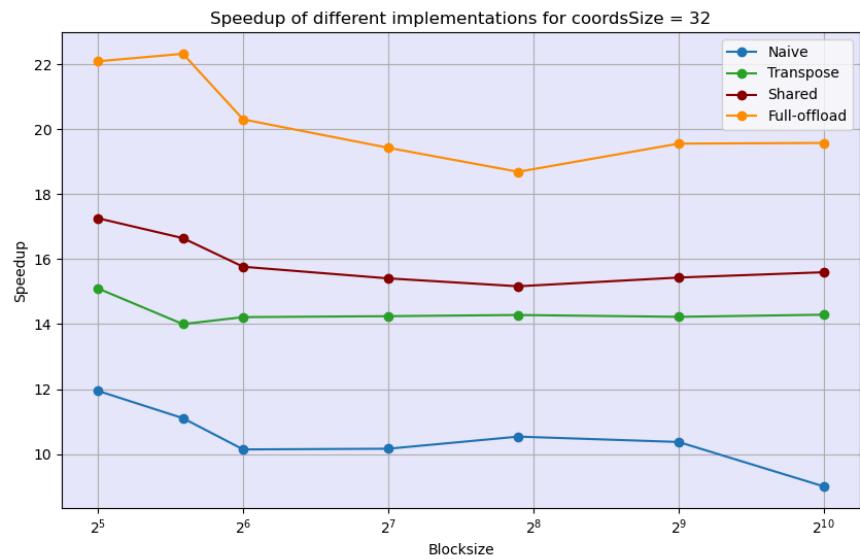
1 do {
2     timing_internal = wtime();
3     checkCuda(cudaMemset(dev_delta_ptr, 0, sizeof(double)));
4     timing_gpu = wtime();
5     //printf("Launching find_nearest_cluster Kernel with grid_size = %d, block_size
6     = %d, shared_mem = %d KB\n", numClusterBlocks, numThreadsPerClusterBlock,
7     clusterBlockSharedDataSize/1000);
8     /* TODO: change invocation if extra parameters needed
9     find_nearest_cluster
10     <<< numClusterBlocks, numThreadsPerClusterBlock, clusterBlockSharedDataSize >>>
11     (numCoords, numObjs, numClusters,
```

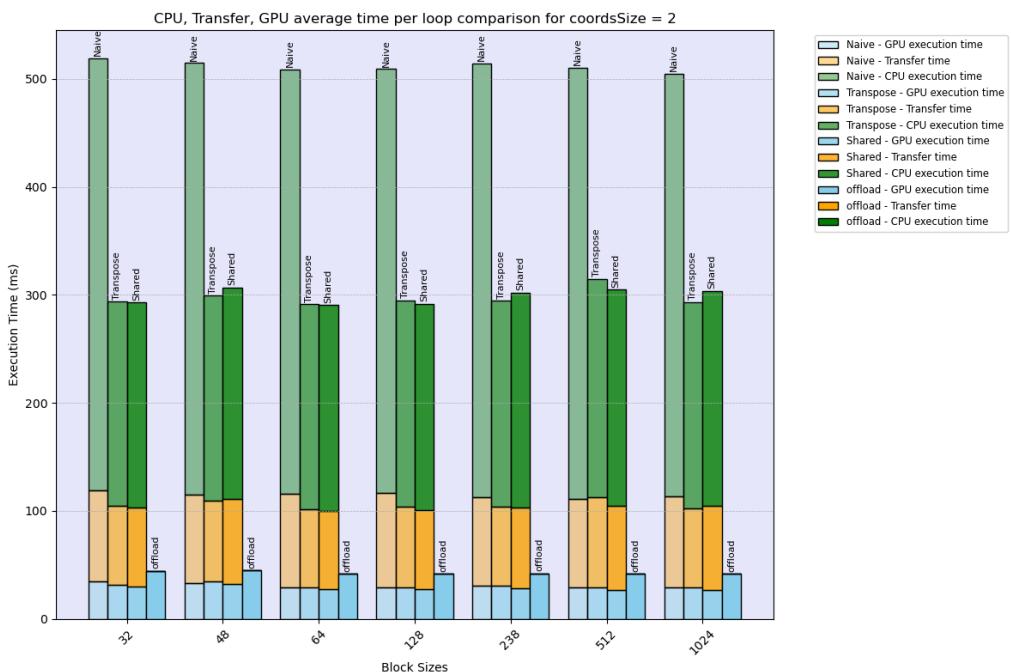
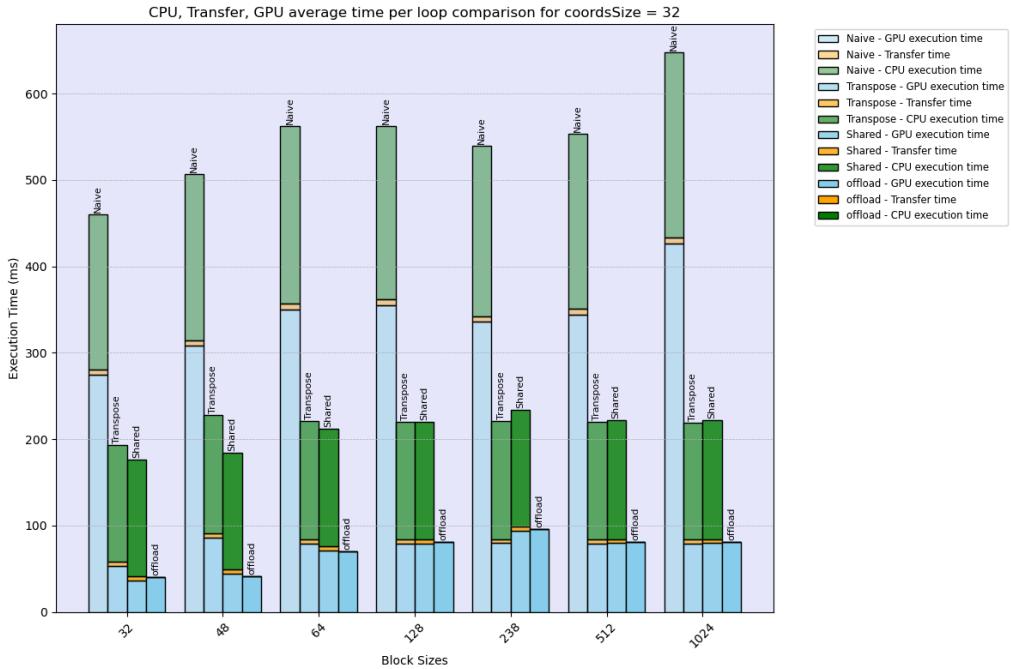
```

10     deviceObjects, devicenewClusterSize, devicenewClusters, deviceClusters,
11     deviceMembership, dev_delta_ptr);
12     */
13     find_nearest_cluster<<< numClusterBlocks, numThreadsPerClusterBlock,
14     clusterBlockSharedDataSize >>>
15     (numCoords, numObjs, numClusters,
16     deviceObjects, devicenewClusterSize, devicenewClusters, deviceClusters,
17     deviceMembership, dev_delta_ptr);
18
19     cudaDeviceSynchronize();
20     checkLastCudaError();
21
22     gpu_time += wtime() - timing_gpu;
23
24     //printf("Kernels complete for itter %d, updating data in CPU\n", loop);
25
26     timing_transfers = wtime();
27     /* TODO: Copy dev_delta_ptr to &delta
28      checkCuda(cudaMemcpy(...));
29      checkCuda(cudaMemcpy(&delta, dev_delta_ptr, sizeof(double), cudaMemcpyDeviceToHost));
30      transfers_time += wtime() - timing_transfers;
31
32      const unsigned int update_centroids_block_sz = (numCoords * numClusters > blockSize) ?
33      blockSize : numCoords *
34
35      * TODO: can use different blocksize here if deemed better */
36      const unsigned int update_centroids_dim_sz = (numCoords * numClusters +
37      update_centroids_block_sz - 1) / update_centroids_block_sz; /* TODO: calculate dim for
38      "update_centroids" */
39      timing_gpu = wtime();
40      /* TODO: use dim for "update_centroids" and fire it
41      update_centroids<<< update_centroids_dim_sz, update_centroids_block_sz, 0 >>>
42      (numCoords, numClusters, devicenewClusterSize, devicenewClusters, deviceClusters); */
43      update_centroids<<< update_centroids_dim_sz, update_centroids_block_sz, 0 >>>
44      (numCoords, numClusters, devicenewClusterSize, devicenewClusters, deviceClusters);
45
46     cudaDeviceSynchronize();
47     checkLastCudaError();
48     gpu_time += wtime() - timing_gpu;
49
50     timing_cpu = wtime();
51     delta /= numObjs;
52     //printf("delta is %f - ", delta);
53     loop++;
54     //printf("completed loop %d\n", loop);
55     cpu_time += wtime() - timing_cpu;
56
57     timing_internal = wtime() - timing_internal;
58     if (timing_internal < timer_min) timer_min = timing_internal;
59     if (timing_internal > timer_max) timer_max = timing_internal;
60   } while (delta > threshold && loop < loop_threshold);

```

Επαναλάβαμε τις μετρήσεις για όλες τις εκδοχές για τα 2 διαφορετικά configurations και παρακάτω παρουσιάζονται τα διαγράμματα για το speedup και την ανάλυση χρόνου εκτέλεσης ανά επανάληψη, ώστε να φαίνονται καλύτερα οι επιδόσεις των διαφόρων εκδοχών.





- 1) Από τα διαγράμματα speedup παρατηρείται πως η Full-offload εκδοχή έχει καλύτερες επιδόσεις στο configuration με τις πολλές συντεταγμένες (32) ενώ έχει πολύ καλύτερες επιδόσεις στο configuration με τις λίγες συντεταγμένες (2).
- 2) To blocksize για τις πολλές συντεταγμένες (32) έχει παρόμοια επιρροή με τις υπόλοιπες εκδοχές και έχει βέλτιστη επίδοση για τα 2 μικρότερα block sizes. Όπως έχει προαναφερθεί, τα μικρά blocksizes έχουν την μέγιστη ευελιξία για το scheduling και είναι λογικό να έχουν καλύτερες επιδόσεις, αφού τα δεδομένα είναι πλήρως ανεξάρτητα. To blocksize για τις λίγες συντεταγμένες (2) δεν επηρεάζει εμφανώς την επίδοση, με εξαίρεση το 48 που λόγω half warps και ότι δεν αποτελεί bottleneck η μνήμη, δεν αξιοποιεί πλήρως τους πόρους της GPU.
- 3) Το κομμάτι update_centroids έχει πολύ χαμηλό computational intensity, οπότε δεν είναι ιδανικό για GPUs. Όμως, είναι πλήρως παραλληλοποιήσιμο, γι' αυτό έχουμε και σαφές speedup σε σχέση

με την χρήση CPU. Μέσω της ανάλυσης χρόνου εκτέλεσης ανά επανάληψη, μπορεί να φανεί πως ο υπολογισμός των καινούριων clusters στην GPU αύξησε ελάχιστα τον χρόνο εκτέλεσης της GPU, ενώ προφανώς μηδενίστηκε ο χρόνος εκτέλεσης της CPU. Σε αυτό οφείλεται η διαφορά επίδοσης, απλά στο speedup συνυπολογίζεται και ο χρόνος allocation και αρχικής μεταφοράς. Γι' αυτό δεν υπάρχει ανάλογο speedup συγκριτικά.

4) Στο configuration με τις λίγες συντεταγμένες (2) ο χρόνος μεταφοράς έχει σημαντικό ποσοστό του συνολικού χρόνου μιας επανάληψης. Καθώς η εκδοχή Full-offload αποφεύγει και τις μεταφορές μέσα στην επανάληψη, το speedup είναι ακόμη μεγαλύτερο.

Bonus 2: Delta Reduction (All-GPU)

Για την υλοποίηση του δενδρικού delta reduction αρχικά χρειάζεται περισσότερη διαμοιραζόμενη μνήμη. Κάθε thread στο thread block πρέπει να έχει τον δικό του delta και μετά να γίνει το reduction. Γι' αυτό χρειάζεται ένας πίνακας από delta και για κάθε thread αντιστοιχεί μια θέση στον πίνακα με όρισμα το local id του.

Κάθε thread αφού βρει το καινούριο κοντινότερο cluster ελέγχει αν είναι το ίδιο με πριν. Αν είναι, τότε βάζει το δικό του delta να είναι 0.0, αλλιώς 1.0. Έπειτα, χρειάζεται συγχρονισμός των threads, πριν εκτελεστεί το reduction ώστε να έχουν υπολογιστεί όλα τα επιμέρους delta.

Το δενδρικό reduction ακολουθεί την λογική ότι σε κάθε επανάληψη οι μισοί προσθέτουν στο δικό τους delta, το delta των υπολοίπων. Οπότε οι πρώτοι μισοί εκτελούν

`delta[local_id] += delta[local_id + size]` και σε κάθε επανάληψη μειώνεται το μέγεθος κατά 2, εξού και δέντρο. Μεταξύ των επαναλήψεων χρειάζεται συγχρονισμός των νημάτων για να έχουν υπολογιστεί τα αποτελέσματα όλων των προσθέσεων. Στο τέλος, το συνολικό αποτέλεσμα θα είναι στο πρώτο thread του thread block, το οποίο χρειάζεται να κάνει μόνο 1 atomicAdd στο global delta.

delta_reduction_find_nearest_cluster.cu

```
1  /*-----< find_nearest_cluster() >-----*/
2  __global__ static
3  void find_nearest_cluster(int numCoords,
4      int numObjs,
5      int numClusters,
6      double *deviceobjects,           // [numCoords][numObjs]
7      int *devicenewClusterSize,      // [numClusters]
8      double *devicenewClusters,      // [numCoords][numClusters]
9      double *deviceClusters,        // [numCoords][numClusters]
10     int *deviceMembership,         // [numObjs]
11     double *devdelta) {
12     extern __shared__ double shmem_total[];
13     double *shmemClusters = shmem_total;
14     double *delta_reduce_buff = shmem_total + numClusters * numCoords;
15     /* TODO: copy me from shared version... */
16     int no_cluster, i;
17
18     //use local_id because shared memory is per thread block
19     for (no_cluster = threadIdx.x; no_cluster < numClusters; no_cluster+=blockDim.x) {
20         for (i = 0; i < numCoords; i++) {
21             shmemClusters[i * numClusters + no_cluster] = deviceClusters[i * numClusters + no_cluster];
22         }
23     }
24     __syncthreads();
25
26     /* Get the global ID of the thread. */
27     int tid = get_tid();
28
29     /* TODO: copy me from shared version... */
30     if (tid < numObjs) {
31
32         /* TODO: copy me from shared version... */
33         int index;
34         double dist, min_dist;
35
36         /* find the cluster id that has min distance to object */
37         index = 0;
38         /* TODO: call min_dist = euclid_dist_2(...) with correct objectId/clusterId using clusters
39         in shmem*/
40         min_dist = euclid_dist_2_transpose(numCoords, numObjs, numClusters, deviceobjects,
41         shmemClusters, tid, 0);
42         for (i = 1; i < numClusters; i++) {
43             /* TODO: call dist = euclid_dist_2(...) with correct objectId/clusterId using clusters
44             in shmem*/
45             dist = euclid_dist_2_transpose(numCoords, numObjs, numClusters, deviceobjects,
46             shmemClusters, tid, i);
47
48             /* no need square root */
```

```

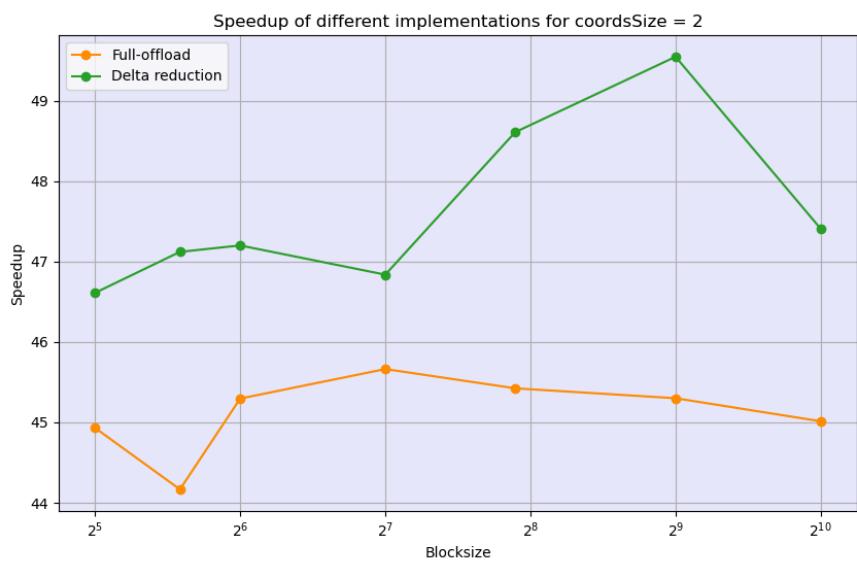
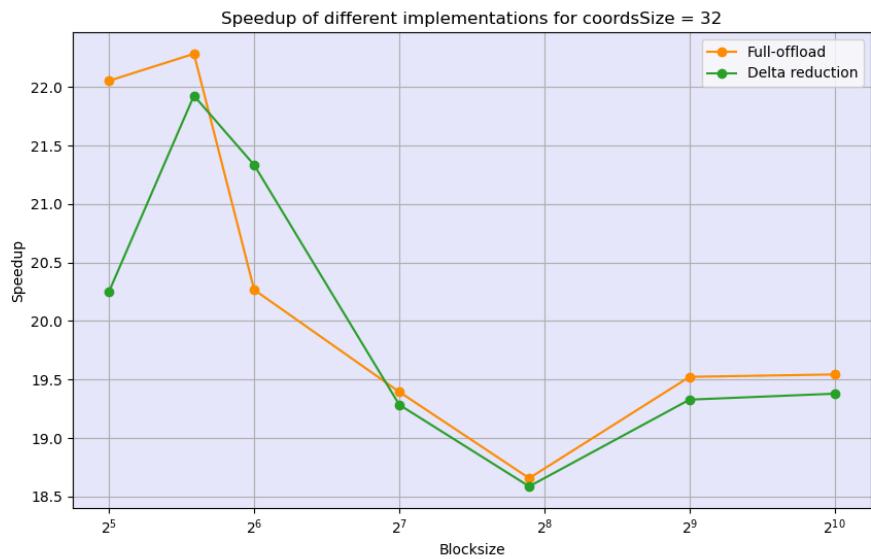
45     if (dist < min_dist) { /* find the min and its array index */
46         min_dist = dist;
47         index = i;
48     }
49 }
50
51 if (deviceMembership[tid] != index) {
52     delta_reduce_buff[threadIdx.x] = 1.0;
53 }
54 else {
55     delta_reduce_buff[threadIdx.x] = 0.0;
56 }
57
58 /* assign the deviceMembership to object objectId */
59 deviceMembership[tid] = index;
60
61 /* TODO: Replacing (*devdelta)+= 1.0; with reduction:
62    - each thread updates the single element of delta_reduce_buff
63    corresponding to its local id (threadIdx.x) -> 1.0 if membership changes, otherwise 0.
64    - Then, ensuring delta_reduce_buff is fully updated, its contents must be summed
65    in delta_reduce_buff[0]
66    either by one thread (lower perf) or with a tree-based reduction (similar to dot reduction
example in slides)
67    - Finally, delta_reduce_buff[0] (local value in block) must be added to devdelta (global
delta value), ensuring write dependencies!
68 */
69
70 /* TODO: additional steps for calculating new centroids in GPU? */
71 atomicAdd(&devicenewClusterSize[index], 1);
72 for (i = 0; i < numCoords; i++) {
73     atomicAdd(&devicenewClusters[i * numClusters + index], deviceobjects[i * numObjs + tid]);
74 }
75
76 __syncthreads();
77 //after everyone in the block is finished do the tree update of delta
78 i = blockDim.x / 2;
79 while (i != 0) {
80     if (threadIdx.x < i) delta_reduce_buff[threadIdx.x] += delta_reduce_buff[threadIdx.x
+ i];
81     __syncthreads();
82     i /= 2;
83 }
84 if (threadIdx.x == 0) atomicAdd(devdelta, delta_reduce_buff[0]);
85 }

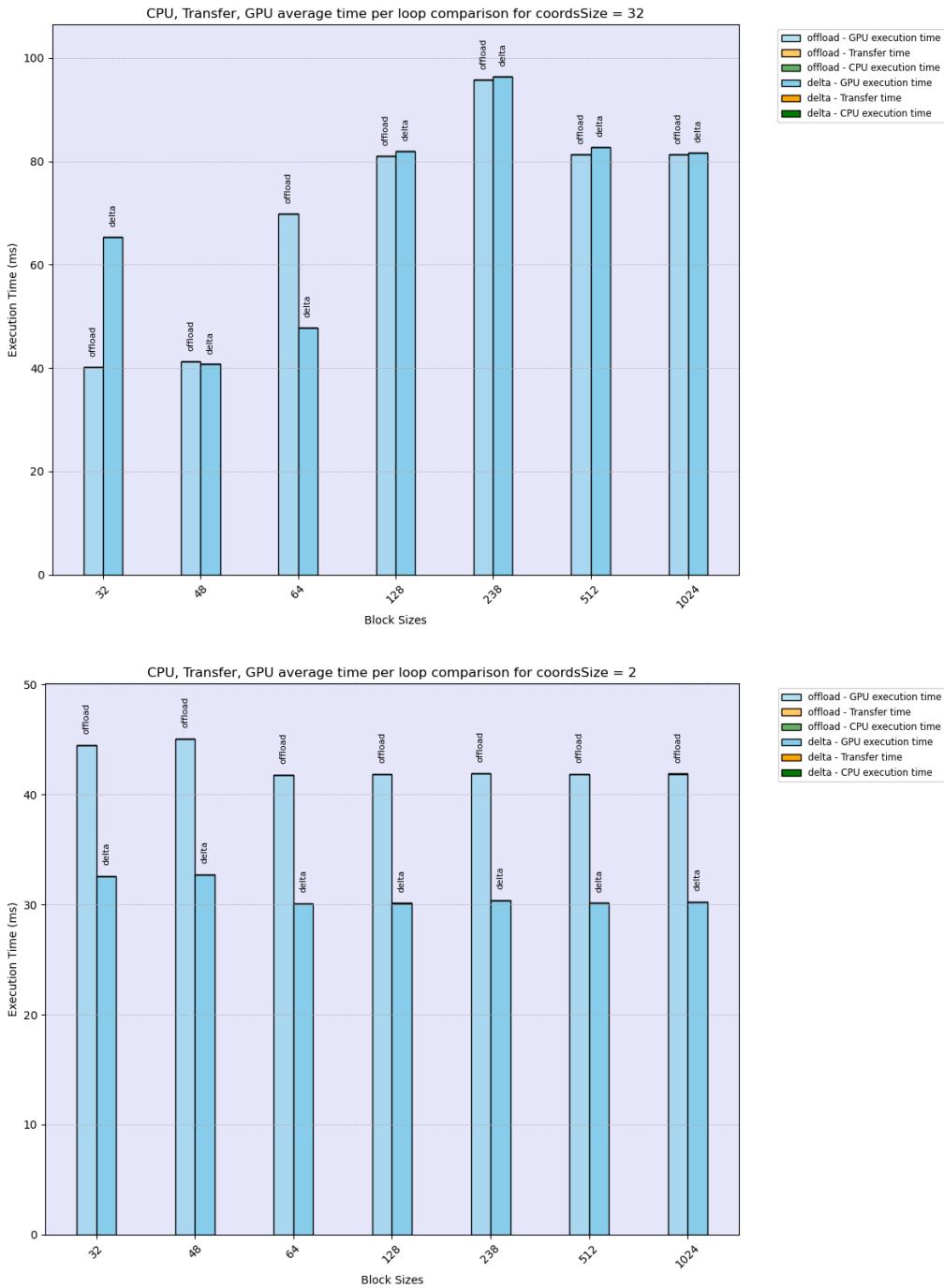
```

Τέλος, η μόνη αλλαγή που χρειάζεται στο υπόλοιπο πρόγραμμα είναι η αύξηση του μεγέθους της διαμοιραζόμενης μνήμης κατά blocksize πραγματικούς. Οπότε:

```
const unsigned int clusterBlockSharedDataSize = numClusters * numCoords * sizeof(double) +
numThreadsPerClusterBlock * sizeof(double);
```

Επαναλάβαμε τις μετρήσεις για την delta reduction εκδοχή και παρακάτω παρουσιάζονται τα διαγράμματα speedup και ανάλυσης χρόνου εκτέλεσης ανά επανάληψη σε σύγκριση με την Full-offload εκδοχή για τα 2 διαφορετικά configurations.





Η επίδοση της delta reduction εκδοχής είναι χειρότερη από την απλή Full-offload στο configuration με τις πολλές συντεταγμένες (32). Αυτό θα μπορούσε να εξηγηθεί ως προς τον παραπάνω συγχρονισμό που απαιτεί η delta reduction εκδοχή. Επειδή θα γίνουν ούτως ή άλλως 32 atomicAdds για τις συντεταγμένες, δεν θα κάνουν όλα τα threads, που έχουν διαφορετικό καινούριο cluster, ταυτόχρονα atomicAdd στο delta. Δεν θα είναι τόσο ταυτόχρονα στο χρόνο τα atomicAdds, καθώς θα υπάρχουν διαφορετικές μικρές καθυστερήσεις λόγω των 32 προηγούμενων atomicAdds. Έτσι, όχι μόνο δεν υπάρχει βελτίωση, αλλά υπάρχει και μια μικρή επιπρόσθετη καθυστέρηση.

Σε αντίθεση, στο configuration με τις λίγες συντεταγμένες (2) καθώς τα atomicAdds έχουν όντως πολλές κοντινές χρονικά εκτελέσεις και απαιτείται όντως συγχρονισμός, η εκδοχή του delta reduction έχει όντως βελτίωση επίδοσης. Συγκριτικά, θέλει περίπου 25% λιγότερο χρόνο ανά loop για το βέλτιστο blocksize σε σχέση με την απλή Full-offload εκδοχή.

To blocksize έχει διαφορετικό ρόλο μόνο στο configuration με τις λίγες συντεταγμένες (2), καθώς χρειάζεται το blocksize να είναι επαρκώς μεγάλο ώστε η λογαριθμική του πολυπλοκότητα να είναι καλύτερη από την γραμμική που έχει η απλή Full-offload. Σε μικρά blocksizes, η βελτίωση στην επίδοση που υπάρχει είναι μικρότερη καθώς η σταθερά αυτής της πολυπλοκότητας είναι αρκετά μεγάλη.

Παραλληλοποιήση αλγορίθμων με χρήση MPI

Άλλη μια παραλληλοποίηση του K-means

Σκοπός αυτής της άσκησης είναι η υλοποίηση του αλγορίθμου Kmeans πρωτόκολλο ανταλλαγής μηνυμάτων με χρήση του MPI.

Ο σκελετός του αλγορίθμου είναι ίδιος με την υλοποίηση copied clusters, που το κάθε νήμα/διεργασία έχει αντίγραφα των παλιών clusters και έχει δικό του τοπικό πίνακα. Το τελικό βήμα της κάθε επανάληψης είναι το reduction των καινούριων συντεταγμένων των clusters. Σε μοντέλο κοινού χώρου διευθύνσεων, οι τελικές τιμές μετά το reduction γίνονται ορατές, μέσω πρωτοκόλλων συνάφειας, σε όλα τα thread clusters. Σε μοντέλο ανταλλαγής μηνυμάτων, μόνο η διεργασία που πραγματοποιεί το reduction θα αποθηκεύσει στην μνήμη της τις νέες τιμές, καθώς δεν υπάρχει κοινός χώρος διευθύνσεων. Στην περίπτωση του kmeans αυτό δεν αρκεί γιατί πρέπει να τις “δουν” όλες οι διεργασίες, οπότε χρειάζεται να τις μεταφέρει με broadcast η διεργασία που θα αναλάβει το reduction. Αυτό μπορεί να γίνει με την χρήση των εντολών MPI_Reduce στην διεργασία με rank 0 και MPI_Bcast σε όλες τις διεργασίες. Οι δύο αυτές λειτουργίες συνδυάζονται και σε μία συνάρτηση, την MPI_Allreduce.

Η διαδικασία αυτή χρειάζεται να γίνει και για τις συντεταγμένες και για τα μεγέθη των καινούριων clusters, καθώς και για την μεταβλητή delta για κατάλληλο έλεγχο τερματισμού του αλγορίθμου.

kmeans.c

```
1 #include <stdio.h>
2 #include <stdlib.h>
3 #include <mpi.h>
4
5 #include "kmeans.h"
6
7 // square of Euclid distance between two multi-dimensional points
8 inline static double euclid_dist_2(int      numdims, /* no. dimensions */
9                                 double * coord1,   /* [numdims] */
10                                double * coord2)  /* [numdims] */
11 {
12     int i;
13     double ans = 0.0;
14
15     for(i=0; i<numdims; i++)
16         ans += (coord1[i]-coord2[i]) * (coord1[i]-coord2[i]);
17
18     return ans;
19 }
20
21 inline static int find_nearest_cluster(int      numClusters, /* no. clusters */
22                                         int      numCoords,  /* no. coordinates */
23                                         double * object,    /* [numCoords] */
24                                         double * clusters) /* [numClusters][numCoords] */
25 {
26     int index, i;
27     double dist, min_dist;
28
29     // find the cluster id that has min distance to object
30     index = 0;
31     min_dist = euclid_dist_2(numCoords, object, clusters);
32
33     for(i=1; i<numClusters; i++) {
34         dist = euclid_dist_2(numCoords, object, &clusters[i*numCoords]);
35         // no need square root
36         if (dist < min_dist) { // find the min and its array index
37             min_dist = dist;
38             index    = i;
39         }
40     }
41
42     return index;
43 }
```

```

39     }
40 }
41 return index;
42 }
43
44 void kmeans(double * objects,           /* in: [numObjs][numCoords] */
45             int    numCoords,        /* no. coordinates */
46             int    numObjs,         /* no. objects */
47             int    numClusters,      /* no. clusters */
48             double threshold,       /* minimum fraction of objects that change membership */
49             long   loop_threshold,  /* maximum number of iterations */
50             int    * membership,    /* out: [numObjs] */
51             double * clusters)      /* out: [numClusters][numCoords] */
52 {
53     int i, j;
54     int index, loop=0;
55     double timing = 0;
56
57     /* Every variable has its "rank_" version, which is used to store local data,
58      * and its "new" version, which is used to store global data.
59      */
60     double rank_delta, delta = 0;           // fraction of objects whose clusters change
in each loop
61     int * rank_newClusterSize, * newClusterSize; // [numClusters]: no. objects assigned in
each new cluster
62     double * rank_newClusters, *newClusters;    // [numClusters][numCoords]
63
64     // Get rank of this process
65     int rank;
66     MPI_Comm_rank(MPI_COMM_WORLD, &rank);
67
68     // initialize membership
69     for (i=0; i<numObjs; i++)
70         membership[i] = -1;
71
72     // initialize rank_newClusterSize and rank_newClusters to all 0
73     rank_newClusterSize = (typeof(rank_newClusterSize))  calloc(numClusters,
sizeof(*rank_newClusterSize));
74     rank_newClusters    = (typeof(rank_newClusters))   calloc(numClusters * numCoords,
sizeof(*rank_newClusters));
75     newClusterSize       = (typeof(newClusterSize))    calloc(numClusters,
sizeof(*newClusterSize));
76     newClusters          = (typeof(newClusters))     calloc(numClusters * numCoords,
sizeof(*newClusters));
77
78     timing = wtime();
79     do {
80         // before each loop, set cluster data to 0
81         for (i=0; i<numClusters; i++) {
82             for (j=0; j<numCoords; j++)
83                 rank_newClusters[i*numCoords + j] = 0.0;
84             rank_newClusterSize[i] = 0;
85         }
86
87         rank_delta = 0.0;
88
89         for (i=0; i<numObjs; i++) {
90             // find the array index of nearest cluster center
91             index = find_nearest_cluster(numClusters, numCoords, &objects[i*numCoords],
clusters);
92
93             // if membership changes, increase rank_delta by 1
94             if (membership[i] != index)
95                 rank_delta += 1.0;
96
97             // assign the membership to object i
98             membership[i] = index;
99
100            // update new cluster centers : sum of objects located within
101            rank_newClusterSize[index]++;
102            for (j=0; j<numCoords; j++)
103                rank_newClusters[index*numCoords + j] += objects[i*numCoords + j];
104     }

```

```

105
106     /*
107      * TODO: Perform reduction of cluster data (rank_newClusters, rank_newClusterSize)
108      * from local arrays to shared.
109      */
110     MPI_Reduce(rank_newClusterSize, newClusterSize, numClusters, MPI_INT, MPI_SUM,
111                 0, MPI_COMM_WORLD);
112     MPI_Bcast(newClusterSize, numClusters, MPI_INT, 0, MPI_COMM_WORLD);
113
114     MPI_Reduce(rank_newClusters, newClusters, numClusters * numCoords, MPI_DOUBLE, MPI_SUM,
115                 0, MPI_COMM_WORLD);
116     MPI_Bcast(newClusters, numClusters * numCoords, MPI_DOUBLE, 0, MPI_COMM_WORLD);
117
118     // average the sum and replace old cluster centers with newClusters
119     for (i=0; i<numClusters; i++) {
120         if (newClusterSize[i] > 0) {
121             for (j=0; j<numCoords; j++) {
122                 clusters[i*numCoords + j] = newClusters[i*numCoords + j] / newClusterSize[i];
123             }
124         }
125     }
126
127     /*
128      * TODO: Perform reduction from rank_delta variable to delta variable, that will be
129      * used for convergence check.
130      */
131     MPI_Reduce(&rank_delta, &delta, 1, MPI_DOUBLE, MPI_SUM, 0, MPI_COMM_WORLD);
132     MPI_Bcast(&delta, 1, MPI_DOUBLE, 0, MPI_COMM_WORLD);
133
134     // Get fraction of objects whose membership changed during this loop. This is used
135     // as a convergence criterion.
136     delta /= numObjs;
137
138     loop++;
139     //printf("\r\tcompleted loop %d", loop);
140     //fflush(stdout);
141     } while (delta > threshold && loop < loop_threshold);
142
143     timing = wtime() - timing;
144     if (rank == 0) fprintf(stdout, "          nloops = %3d    (total = %7.4fs)  (per loop =
145 %7.4fs)\n", loop, timing, timing/loop);
146
147     free(rank_newClusters);
148     free(rank_newClusterSize);
149     free(newClusters);
150     free(newClusterSize);
151 }

```

Χρειάζεται αρχικά κατάλληλος διαμοιρασμός των objects στις διεργασίες. Η λογική που ακολουθήσαμε ήταν να δώσουμε το πηλίκο, της διαίρεσης πλήθους objects δια size - 1, σε size - 1 διεργασίες και το υπόλοιπο στην τελευταία. Έτσι, καταλήπτουμε την γενική περίπτωση που δεν θα πάρει κάθε διεργασία τον ίδιο αριθμό αντικειμένων. Σημείωση : Δοκιμάσαμε επίσης να μοιράσουμε objects / size διεργασίες σε όλα τα processes και +1 στις πρώτες m διεργασίες όπου m = objects % size, που θεωρητικά εξισορροπεί καλύτερα τον φόρτο εργασίας, όμως δεν διαπιστώσαμε κάποια διαφορά για το συγκεκριμένο dataset.

Το πρώτο rank έχει στην μεταβλητή rank_numObjs την τιμή του πηλίκου της διαίρεσης. Οπότε, για όλες τις διεργασίες εκτός την τελευταίας το send count είναι αυτή η τιμή επί τον αριθμό των συντεταγμένων κάθε αντικειμένου. Αντίστοιχα, για την τελευταία διεργασία είναι το πλήθος αντικειμένων - ó, τι πήραν οι προηγούμενες, ακριβώς δηλαδή ó, τι περισσεύει. Το displs[i] είναι displs[i-1] + τα sendcounts της διεργασίας i.

Μετά τον υπολογισμό των send_counts, displs, γίνονται broadcast ώστε να μπορεί να γίνει κατάλληλο scatter. Με την εντολή MPI_Scatterv μπορεί να γίνει scatter με διαφορετικό αριθμό αντικειμένων σε κάθε διεργασία. Ως παράμετροι χρησιμοποιούνται οι πίνακες send_counts, displs που υπολογίσαμε

παραπάνω.

file_io.c

```
1 #include <stdio.h>
2 #include <stdlib.h>
3 #include <string.h>      /* strtok() */
4 #include <sys/types.h>    /* open() */
5 #include <sys/stat.h>
6 #include <fcntl.h>
7 #include <unistd.h>       /* read(), close() */
8 #include <mpi.h>
9
10 #include "kmeans.h"
11
12 double * dataset_generation(int numObjs, int numCoords, long *rank_numObjs)
13 {
14     double * objects = NULL, * rank_objects = NULL;
15     long i, j, k;
16
17     // Random values that will be generated will be between 0 and 10.
18     double val_range = 10;
19
20     int rank, size;
21     MPI_Comm_rank(MPI_COMM_WORLD, &rank);
22     MPI_Comm_size(MPI_COMM_WORLD, &size);
23
24     /*
25      * TODO: Calculate number of objects that each rank will examine (*rank_numObjs)
26      */
27     if (numObjs % size == 0) *rank_numObjs = numObjs / size; // equally distributed
28     else {
29         if (rank == size-1)
30             *rank_numObjs = numObjs % (size-1);
31         else
32             *rank_numObjs = numObjs / (size-1);
33     }
34
35     /* allocate space for objects[][] and read all objects */
36     int sendcounts[size], displs[size];
37     if (rank == 0) {
38         objects = (typeof(objects)) malloc(numObjs * numCoords * sizeof(*objects));
39         /*
40             * TODO: Calculate sendcounts and displs, which will be used to scatter data to
41             * each rank.
42             * Hint: sendcounts: number of elements sent to each rank
43             *       displs: displacement of each rank's data
44             */
45         int disp_sum = 0;
46         for (i = 0; i < size-1; i++){
47             sendcounts[i] = (*rank_numObjs) * numCoords;
48             displs[i] = disp_sum;
49             disp_sum += sendcounts[i];
50         }
51         sendcounts[size-1] = (numObjs - (size-1)*(*rank_numObjs)) * numCoords;
52         displs[size-1] = disp_sum;
53     }
54
55     /*
56      * TODO: Broadcast the sendcounts and displs arrays to other ranks
57      */
58     MPI_Bcast(sendcounts, size, MPI_INT, 0, MPI_COMM_WORLD);
59     MPI_Bcast(displs, size, MPI_INT, 0, MPI_COMM_WORLD);
60
61
62     /* allocate space for objects[][] (for each rank separately) and read all objects */
63     rank_objects = (typeof(rank_objects)) malloc((*rank_numObjs) * numCoords * sizeof(*rank_objects));
64
65     /* rank 0 will generate data for the objects array. This array will be used later to
66     scatter data to each rank. */
67     if (rank == 0) {
```

```

67     for (i=0; i<numObjs; i++)
68     {
69         unsigned int seed = i;
70         for (j=0; j<numCoords; j++)
71         {
72             objects[i*numCoords + j] = (rand_r(&seed) / ((double) RAND_MAX)) * val_range;
73             if (_debug && i == 0)
74                 printf("object[i=%ld][j=%ld]=%f\n", i, j, objects[i*numCoords + j]);
75         }
76     }
77 }
78 /*
79     * TODO: Scatter objects to every rank. (hint: each rank may receive different number
80     of objects)
81 */
82 MPI_Scatterv(objects, sendcounts, displs, MPI_DOUBLE, rank_objects, sendcounts[rank],
83 MPI_DOUBLE, 0, MPI_COMM_WORLD);
84
85 if (rank == 0)
86     free(objects);
87
88 return rank_objects;
89 }
```

Για την main.c χρειάζεται να κάνουμε ένα αρχικό broadcast θέσεων, καθώς και κατάλληλη συλλογή της τελικής λύσης του αλγορίθμου.

Πιο συγκεκριμένα, αρχικά χρειάζεται να καλέσουμε την συνάρτηση kmeans με rank_numObjs ως παράμετρον, όπου εκεί είναι αποθηκευμένα πόσα αντικείμενα πρέπει να εξετάσει η κάθε διεργασία ξεχωριστά.

Οι πίνακες recv_counts, displs ακολουθούν την ίδια ακριβώς λογική με τους send_counts, displs στο αρχείο file_io.c. Όμως, εδώ χρειάζεται συλλογή μόνο για το membership του κάθε αντικειμένου και όχι για όλες τις συντεταγμένες του. Άρα, για όλες τις διεργασίες είναι rank_numObjs ακέραιοι αριθμοί. Τέλος, γίνεται η συλλογή στην διεργασία με rank 0, με την εντολή MPI_Gatherv με παράμετρους τους πίνακες που υπολογίστηκαν παραπάνω.

main.c

```

1 #include <stdio.h>
2 #include <stdlib.h>
3 #include <string.h>      /* strtok() */
4 #include <sys/types.h>    /* open() */
5 #include <sys/stat.h>
6 #include <fcntl.h>
7 #include <unistd.h>      /* getopt() */
8 #include <mpi.h>
9
10 int _debug;
11 #include "kmeans.h"
12
13 static void usage(char *argv0) {
14     char *help =
15         "Usage: %s [switches]\n"
16         "       -c num_clusters      : number of clusters (must be > 1)\n"
17         "       -s size                : size of examined dataset\n"
18         "       -n num_coords          : number of coordinates\n"
19         "       -t threshold            : threshold value (default : 0.001)\n"
20         "       -l loop_threshold       : iterations threshold (default : 10)\n"
21         "       -d                      : enable debug mode\n"
22         "       -h                      : print this help information";
23     fprintf(stderr, help, argv0);
24     exit(-1);
25 }
```

```

27 int main(int argc, char **argv)
28 {
29     long i, j, opt;
30     extern char* optarg;
31     extern int optind;
32
33     long numClusters=0, numCoords=0, numObjs=0;
34     long rank_numObjs=0;
35     int * membership; // [rank_numObjs] this array will contain membership information
36     for this rank's objects
37         int * tot_membership; // [numObjs] this array will contain membership information
38     for all objects
39         double * objects; // [numObjs * numCoords] data objects
40         double * clusters; // [numClusters * numCoords] cluster center
41         double dataset_size = 0, threshold;
42         long loop_threshold;
43         double io_timing_read;
44
45     /* some default values */
46     _debug = 0;
47     threshold = 0.001;
48     loop_threshold = 10;
49     numClusters = 0;
50
51     while ( (opt = getopt(argc,argv,"n:t:l:c:s:dh")) != EOF) {
52         switch (opt) {
53             case 'c': numClusters = atol(optarg);
54                         break;
55             case 't': threshold=atof(optarg);
56                         break;
57             case 'l': loop_threshold=atol(optarg);
58                         break;
59             case 's': dataset_size=atof(optarg);
60                         break;
61             case 'n': numCoords=atol(optarg);
62                         break;
63             case 'd': _debug = 1;
64                         break;
65             case 'h':
66             default: usage(argv[0]);
67                         break;
68         }
69     }
70     if (numClusters <= 1) {
71         usage(argv[0]);
72     }
73     int rank, size;
74     MPI_Init(&argc,&argv);
75     MPI_Comm_rank(MPI_COMM_WORLD,&rank);
76     MPI_Comm_size(MPI_COMM_WORLD,&size);
77
78     numObjs = (dataset_size*1024*1024) / (numCoords*sizeof(double));
79
80     if (numObjs < numClusters) {
81         if (rank == 0) printf("Error: number of clusters must be larger than the number of
82 data points to be clustered.\n");
83         MPI_Finalize();
84         return 1;
85     }
86     if (rank == 0) printf("dataset_size = %.2f MB      numObjs = %ld      numCoords = %ld
87 numClusters = %ld\n", dataset_size, numObjs, numCoords, numClusters);
88
89     objects = dataset_generation(numObjs, numCoords, &rank_numObjs);
90
91     // Allocate space for clusters (coordinates of cluster centers)
92     clusters = (double*) malloc(numClusters * numCoords * sizeof(double));
93
94     // The first numClusters elements are selected as initial centers. Only rank 0 needs to
95     // calculate this, and later broadcast it to all ranks.
96     if (rank == 0) {
97         for (i=0; i<numClusters; i++)
98             for (j=0; j<numCoords; j++)
99                 clusters[i*numCoords + j] = objects[i*numCoords + j];

```

```

96     // check initial cluster centers for repetition
97     if (check_repeated_clusters(numClusters, numCoords, clusters) == 0) {
98         printf("Error: some initial clusters are repeated. Please select distinct
99         initial centers\n");
100        MPI_Finalize();
101        return 1;
102    }
103    /*
104     printf("Initial cluster centers:\n");
105     for (i=0; i<numClusters; i++) {
106         printf("(0) clusters[%ld] = ",i);
107         for (j=0; j<numCoords; j++)
108             printf(" %6.6f", clusters[i*numCoords + j]);
109         printf("\n");
110     }
111    */
112 }
113 /*
114 * TODO: Broadcast initial cluster positions to all ranks
115 */
116 MPI_Bcast(clusters, numClusters*numCoords, MPI_DOUBLE, 0, MPI_COMM_WORLD);

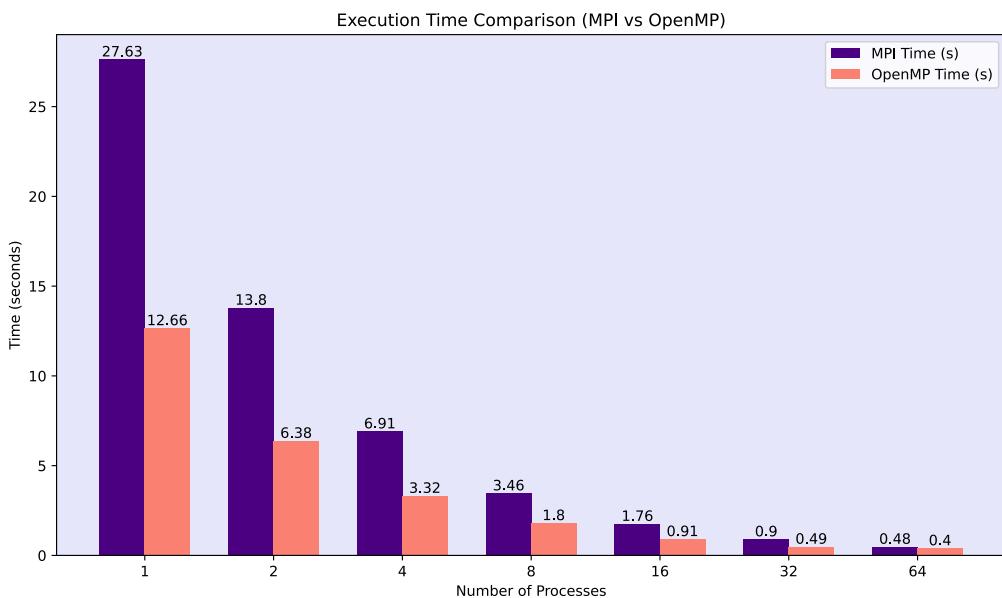
117 // membership: the cluster id for each data object
118 membership = (int*) malloc(rank_numObjs * sizeof(int));
119 tot_membership = (int*) malloc(numObjs * sizeof(int));
120
121 // start the core computation
122 /*
123 * TODO: Fix number of objects that this kmeans function call will process
124 */
125 kmeans(objects, numCoords, rank_numObjs, numClusters, threshold, loop_threshold,
126 membership, clusters);
127
128 /*
129 if (rank == 0) {
130     printf("Final cluster centers:\n");
131     for (i=0; i<numClusters; i++) {
132         printf("clusters[%ld] = ",i);
133         for (j=0; j<numCoords; j++)
134             printf(" %6.6f ", clusters[i*numCoords + j]);
135         printf("\n");
136     }
137 }
138 */
139 /*
140 // Gather membership information from all ranks to tot_membership
141 int recvcounts[size], displs[size];
142 if (rank == 0) {
143     /* TODO: Calculate recvcounts and displs, which will be used to gather data from
144     each rank.
145         * Hint: recvcounts: number of elements received from each rank
146         *       displs: displacement of each rank's data
147     */
148     int disp_sum = 0;
149     for (i = 0; i < size-1; i++){
150         recvcounts[i] = rank_numObjs;
151         displs[i] = disp_sum;
152         disp_sum += recvcounts[i];
153     }
154     recvcounts[size-1] = numObjs - (size-1)*rank_numObjs;
155     displs[size-1] = disp_sum;
156 }
157 */
158 /*
159 * TODO: Broadcast the recvcounts and displs arrays to other ranks.
160 */
161 MPI_Bcast(recvcounts, size, MPI_INT, 0, MPI_COMM_WORLD);
162 MPI_Bcast(displs, size, MPI_INT, 0, MPI_COMM_WORLD);
163
164 /*
165 * TODO: Gather membership information from every rank. (hint: each rank may send different
166 number of objects)

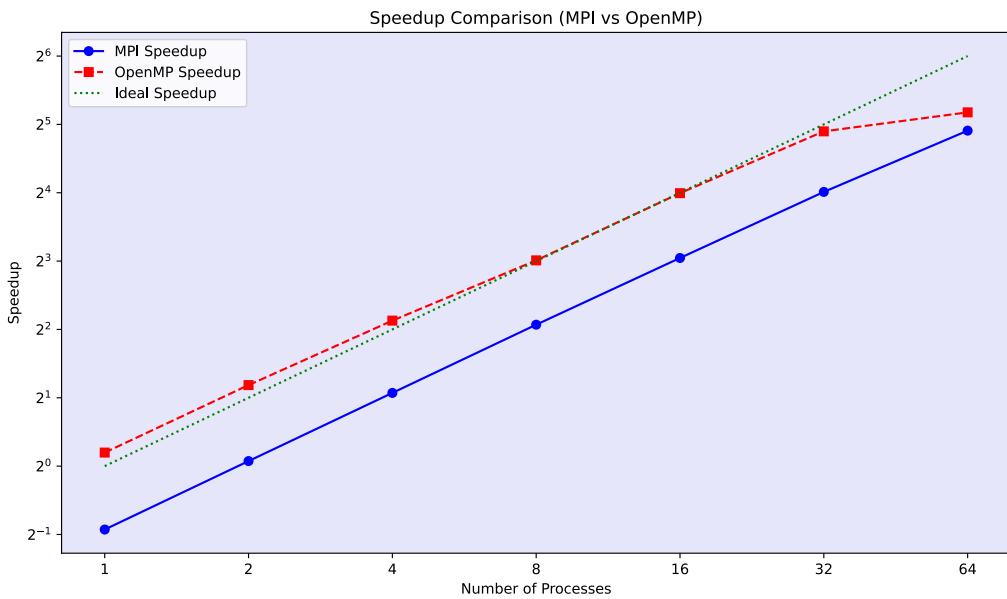
```

```

166     */
167     MPI_Gatherv(membership, rank_numObjs, MPI_INT, tot_membership, recvcounts, displs,
168     MPI_INT, 0, MPI_COMM_WORLD);
168
169     if (_debug && rank == 0)
170         for (i = 0; i < numObjs; ++i)
171             fprintf(stderr, "%d\n", tot_membership[i]);
172
173     free(objects);
174     free(membership);
175     free(tot_membership);
176     free(clusters);
177
178     MPI_Finalize();
179     return 0;
180 }
```

Πραγματοποιήσαμε μετρήσεις για {Size, Coords, Clusters, Loops} = {256, 16, 32, 10} για 1, 2, 4, 8, 16, 32 και 64 MPI διεργασίες. Παρακάτω παρουσιάζονται τα διαγράμματα speedup και ανάλυσης χρόνου εκτέλεσης ως προς τον αριθμό διεργασιών για το mpi, αλλά και συγκριτικά με την υλοποίηση numa-aware με Openmp από την δεύτερη άσκηση. Επιλέξαμε αυτήν την εκδοχή καθώς ήταν η βέλτιστη και επιπλέον έμοιαζε με την λογική του MPI στον καταμερισμό του αρχικού πίνακα, αφού κάθε διεργασία αρχικοποιούσε με first-touch policy στην τοπική της μνήμη το αντίστοιχο εύρος από indices.





Παρατηρείται πως η υλοποίηση στο μοντέλο ανταλλαγής μηνυμάτων έχει τέλεια κλιμάκωση στο πλήθος διεργασιών, όμως είναι αρκετά πιο αργή σε σχέση με την υλοποίηση numa-aware με Openmp. Αυτό συμβαίνει καθώς η numa-aware εκμεταλλεύται πλήρως τον κοινό χώρο διευθύσεων (με χρήση thread pinning και preload αντικειμένων στις caches), ενώ οι διεργασίες στο mpri χρειάζεται να πραγματοποιούν επικοινωνία μεταξύ τους. Επίσης, η υλοποίηση στο μοντέλο ανταλλαγής μηνυμάτων έχει μεγαλύτερο χρόνο εκτέλεσης από την υλοποίηση στο μοντέλο κοινού χώρου διευθύνσεων, καθώς η επικοινωνία μεταξύ των διεργασιών είναι πιο ακριβή.

Heat transfer

Σκοπός της τελευταίας άσκησης είναι η επιτάχυνση των προσεγγιστικών μεθόδων που χρησιμοποιούνται για την επίλυση της μαρικής διαφορικής εξίσωσης της θερμότητας σε 2 διαστάσεις. Οι πιο διαδεδομένοι πυρήνες υπολογισμών είναι οι Jacobi, Gauss-Seidel (SOR) και Red-Black (SOR) τους οποίους καλούμαστε να παραλληλοποιήσουμε.

Jacobi

Για την συγκεκριμένη άσκηση μας δίνεται ήδη ένας σκελετός που καλούμαστε να συμπληρώσουμε. Αρχικά, έχουμε 3 πίνακες διαστάσεων: global, local, global_padded. Ο πρώτος είναι για τις διαστάσεις του προβλήματος που ορίζει ο χρήστης. Ο δεύτερος είναι για τις διαστάσεις του υποπίνακα που έχει κάθε διεργασία, με την παραδοχή πως όλες οι διεργασίες έχουν υποπίνακες ίδιους μεγέθους. Ο τρίτος είναι ένας πίνακας με περισσότερες στήλες και γραμμές από τον global, ώστε να μην χρειάζεται να ελέγχουμε τα όρια του πίνακα στον υπολογισμό των θερμοκρασιών. Επίσης, υπάρχει ένας πίνακας grid που ορίζει τις διαστάσεις της δισδιάστατης απεικόνισης των υπολογιστικών πόρων στον πίνακα.

Για την επικοινωνία μεταξύ των διεργασίων, έχει οριστεί ένας global communicator και ένας καρτεσιανός 2D communicator, για να βρεθεί με τις συναρτήσεις MPI_Cart_create και MPI_Cart_shift η γειτονική διεργασία σε κάθε κατεύθυνση.

Αν η διάσταση του πλέγματος διεργασιών δεν διαιρεί την αντίστοιχη διάσταση του προβλήματος, τότε αυξάνουμε την local διάσταση του πίνακα κατά 1 στην κατεύθυνση που δεν διαιρείται και κάνουμε padding για την κατασκευή του global_padded πίνακα.

Ορίζεται ένας καινούριος τύπος δεδομένων local block ώστε να γίνεται κατάλληλο scatter και gather των δεδομένων στις διεργασίες. Επίσης, στο local block υπάρχουν 2 ghost στήλες και 2 ghost γραμμές για την ανταλλαγή των συνοριακών σημείων με τις γειτονικές διεργασίες.

Κάνουμε scatter των δεδομένων με μοιρασιά ενός local block ανά διεργασία με τα scatteroffsets που υπολογίστηκαν παραπάνω.

Ορίζουμε έναν καινούριο τύπο δεδομένων για την ανταλλαγή στηλών με τις γειτονικές διεργασίες. Χρειάζεται να κρατήσουμε 1 δεδομένο σε κάθε σειρά για κάθε στήλη, οπότε 1 δεδομένο για local[0] + 2 σειρές με stride local[1] + 2.

Με κλήση της συνάρτησης MPI_Cart_shift βρίσκουμε τις γειτονικές διεργασίες προς κάθε κατεύθυνση.

Για να ορίσουμε τα τοπικά όρια υπολογισμών στους υποπίνακες χρειάζεται να λάβουμε υπόψιν τις συνοριακές συνθήκες, που παραμένουν σταθερές καθώς και το padding του global πίνακα. Ακόμη, τα ghost rows και columns περιορίζουν κατά 1 τον υποπίνακα υπολογισμών. Οι διαδικασίες που είναι στην πρώτη σειρά του grid ξεκινάνε από i_min = 2 ενώ οι υπόλοιπες με 1. Οι διαδικασίες που είναι στην πρώτη στήλη του grid ξεκινάνε με j_min = 2 ενώ οι υπόλοιπες με 2. Αντίστοιχα και για τις διαδικασίες που είναι στην τελευταία σειρά ή στήλη. Εκείνες όμως πρέπει να αγνοήσουν και όσες γραμμές έχουμε κάνει padding.

Για την υλοποίηση του for loop, χρησιμοποιήθηκε non blocking επικοινωνία μεταξύ των εργασιών με τις εντολές MPI_Isend, MPI_Irecv και συγχρονισμό με MPI_Waitall. Κάθε διεργασία στέλνει τα στοιχεία της γραμμής 1 στην διεργασία που βρίσκεται βόρεια και λαμβάνει τα στοιχεία της βόρειας διεργασίας και τα αποθηκεύει στην γραμμή 0 για τον επόμενο υπολογισμό. Αντίστοιχα και στις υπόλοιπες κατευθύνσεις. Μετά από την ικανοποίηση των requests με την χρήση της MPI_Waitall, υπολογίζουμε τις καινούριες τιμές για τον υποπίνακα και κάνουμε swap.

Τέλος, αφού τελειώσει το for loop, κάνουμε reduce τους χρόνους με την εντολή MPI_Reduce και gather τις τελικές θερμοκρασίες με MPI_Gatherv.

mpi_skeleton.c

```
1 #include <stdio.h>
2 #include <stdlib.h>
3 #include <math.h>
4 #include <sys/time.h>
5 #include <mpi.h>
6 #include "utils.h"
7
8 int main(int argc, char ** argv) {
9     int rank,size;
10    int global[2],local[2]; //global matrix dimensions and local matrix dimensions (2D-domain,
11 2D-subdomain)
12    int global_padded[2];   //padded global matrix dimensions (if padding is not needed,
13 global_padded=global)
14    int grid[2];           //processor grid dimensions
15    int i,j,t;
16    int global_converged=0,converged=0; //flags for convergence, global and per process
17    MPI_Datatype dummy;      //dummy datatype used to align user-defined datatypes in memory
18    double omega;           //relaxation factor - useless for Jacobi
19
20    struct timeval tts,ttf,tcs,tcf; //Timers: total-> tts,ttf, computation -> tcs,tcf
21    double ttotal=0,tcomp=0,total_time,comp_time;
22
23    double ** U, ** u_current, ** u_previous, ** swap; //Global matrix, local current and
24 previous matrices, pointer to swap between current and previous
25
26
27 MPI_Init(&argc,&argv);
28 MPI_Comm_size(MPI_COMM_WORLD,&size);
29 MPI_Comm_rank(MPI_COMM_WORLD,&rank);
30
31 //----Read 2D-domain dimensions and process grid dimensions from stdin---//
32
33 if (argc!=5) {
34     fprintf(stderr,"Usage: mpirun .... ./exec X Y Px Py");
35     exit(-1);
36 }
37 else {
38     global[0]=atoi(argv[1]);
39     global[1]=atoi(argv[2]);
40     grid[0]=atoi(argv[3]);
41     grid[1]=atoi(argv[4]);
42 }
43
44 //----Create 2D-cartesian communicator---//
45 //----Usage of the cartesian communicator is optional----//
46
47 MPI_Comm CART_COMM;           //CART_COMM: the new 2D-cartesian communicator
48 int periods[2]={0,0};         //periods={0,0}: the 2D-grid is non-periodic
49 int rank_grid[2];             //rank_grid: the position of each process on the new communicator
50
51 MPI_Cart_create(MPI_COMM_WORLD,2,grid,periods,0,&CART_COMM); //communicator creation
52 MPI_Cart_coords(CART_COMM,rank,2,rank_grid);                //rank mapping on the
53 new communicator
54
55 //----Compute local 2D-subdomain dimensions---//
56 //----Test if the 2D-domain can be equally distributed to all processes---//
57 //----If not, pad 2D-domain---//
58
59 for (i=0;i<2;i++) {
60     if (global[i]%grid[i]==0) {
61         local[i]=global[i]/grid[i];
62         global_padded[i]=global[i];
63     }
64 }
65
66 //Initialization of omega
67 omega=2.0/(1+sin(3.14/global[0]));
68
```

```

69     //----Allocate global 2D-domain and initialize boundary values----//
70     //----Rank 0 holds the global 2D-domain----//
71     if (rank==0) {
72         U=allocate2d(global_padded[0],global_padded[1]);
73         init2d(U,global[0],global[1]);
74     }
75
76     //----Allocate local 2D-subdomains u_current, u_previous----//
77     //----Add a row/column on each size for ghost cells----//
78
79     u_previous=allocate2d(local[0]+2,local[1]+2);
80     u_current=allocate2d(local[0]+2,local[1]+2);
81
82     //----Distribute global 2D-domain from rank 0 to all processes----//
83
84     //----Appropriate datatypes are defined here----//
85     //*****The usage of datatypes is optional*****//
86
87     //----Datatype definition for the 2D-subdomain on the global matrix----//
88
89     MPI_Datatype global_block;
90     MPI_Type_vector(local[0],local[1],global_padded[1],MPI_DOUBLE,&dummy);
91     MPI_Type_create_resized(dummy,0,sizeof(double),&global_block);
92     MPI_Type_commit(&global_block);
93
94     //----Datatype definition for the 2D-subdomain on the local matrix----//
95
96     MPI_Datatype local_block;
97     MPI_Type_vector(local[0],local[1],local[1]+2,MPI_DOUBLE,&dummy);
98     MPI_Type_create_resized(dummy,0,sizeof(double),&local_block);
99     MPI_Type_commit(&local_block);
100
101    //----Rank 0 defines positions and counts of local blocks (2D-subdomains) on global
102    //matrix----//
103    int * scatteroffset, * scattercounts;
104    if (rank==0) {
105        scatteroffset=(int*)malloc(size*sizeof(int));
106        scattercounts=(int*)malloc(size*sizeof(int));
107        for (i=0;i<grid[0];i++)
108            for (j=0;j<grid[1];j++) {
109                scattercounts[i*grid[1]+j]=1;
110                scatteroffset[i*grid[1]+j]=(local[0]*local[1]*grid[1]*i+local[1]*j);
111            }
112    }
113
114    //----Rank 0 scatters the global matrix----//
115
116    //*****TODO*****/
117    MPI_Scatterv(U[0],scattercounts,scatteroffset,global_block,&u_current[1]
118 [1],1,local_block,0,MPI_COMM_WORLD);
119
120
121    /*Fill your code here*/
122
123
124
125
126    /*Make sure u_current and u_previous are
127    both initialized*/
128    for(i = 0; i < local[0] + 2; ++i) {
129        for(j = 0; j < local[1] + 2; ++j)
130            u_previous[i][j] = u_current[i][j];
131    }
132
133
134
135
136
137
138
139    //*****//
140
141
142    if (rank==0)
143        free2d(U);

```

```

144
145
146
147 //----Define datatypes or allocate buffers for message passing----//
148
149 //*****TODO*****//
150
151
152
153 /*Fill your code here*/
154 MPI_Datatype column;
155 MPI_Type_vector(local[0] + 2, 1, local[1] + 2, MPI_DOUBLE, &column);
156 MPI_Type_commit(&column);
157
158
159
160
161
162
163
164
165 //*****//
166
167
168 //----Find the 4 neighbors with which a process exchanges messages----//
169
170 //*****TODO*****//
171 int north, south, east, west;
172 MPI_Cart_shift(CART_COMM, 0, 1, &north, &south);
173 MPI_Cart_shift(CART_COMM, 1, 1, &west, &east);
174 //MPI_PROC_NULL if no neighbor exists
175
176
177
178
179 /*Fill your code here*/
180
181
182 /*Make sure you handle non-existing
183 neighbors appropriately*/
184
185
186
187
188
189 //*****//
190
191
192
193 //----Define the iteration ranges per process----//
194 //*****TODO*****//
195
196 int i_min,i_max,j_min,j_max;
197
198 if (rank_grid[0] == 0) i_min = 2;
199 else i_min = 1;
200
201 if (rank_grid[1] == 0) j_min = 2;
202 else j_min = 1;
203
204 // find useless rows and columns due to padding
205 int useless_rows = global_padded[0] - global[0];
206 int useless_columns = global_padded[1] - global[1];
207
208 // the last row and column of processes will have to not take into account the useless
209 // rows and columns
210 if (rank_grid[0] == grid[0] - 1) {
211     i_max = local[0] -useless_rows;
212 } else {
213     i_max = local[0] + 1;
214 }
215
216 if (rank_grid[1] == grid[1] - 1) {
217     j_max = local[1] - useless_columns;
218 } else {
219     j_max = local[1] + 1;
220 }
221
222

```

```

223     /*Fill your code here*/
224
225
226
227
228
229     /*Three types of ranges:
230      -internal processes
231      -boundary processes
232      -boundary processes and padded global array
233 */
234
235
236
237
238
239     //*****
240
241
242
243
244     //----Computational core----/
245     gettimeofday(&tt, NULL);
246     #ifdef TEST_CONV
247         for (t=0;t<T && !global_converged;t++) {
248             #endif
249             #ifndef TEST_CONV
250                 #undef T
251                 #define T 256
252
253                 for (t=0;t<T;t++) {
254                     #endif
255
256                     int req_count = 0;
257                     MPI_Request request[8];
258                     //*****TODO*****
259                     if (north != MPI_PROC_NULL) {
260                         MPI_Isend(&u_previous[1][0], local[1] + 2, MPI_DOUBLE, north, 0, MPI_COMM_WORLD,
261                         &request[req_count++]);
262                     }
263                     if (south != MPI_PROC_NULL) {
264                         MPI_Isend(&u_previous[local[0]][0], local[1] + 2, MPI_DOUBLE, south, 0,
265                         MPI_COMM_WORLD, &request[req_count++]);
266                     }
267                     if (west != MPI_PROC_NULL) {
268                         MPI_Isend(&u_previous[0][1], 1, column, west, 0, MPI_COMM_WORLD, &request[req_count+1]);
269                     }
270                     if (east != MPI_PROC_NULL) {
271                         MPI_Isend(&u_previous[0][local[1]], 1, column, east, 0, MPI_COMM_WORLD,
272                         &request[req_count++]);
273                     }
274                     //receives
275                     if (north != MPI_PROC_NULL) {
276                         MPI_Irecv(&u_previous[0][0], local[1] + 2, MPI_DOUBLE, north, 0, MPI_COMM_WORLD,
277                         &request[req_count++]);
278                     }
279                     if (south != MPI_PROC_NULL) {
280                         MPI_Irecv(&u_previous[local[0]+1][1], local[1] + 2, MPI_DOUBLE, south, 0,
281                         MPI_COMM_WORLD, &request[req_count++]);
282                     }
283                     if (west != MPI_PROC_NULL) {
284                         MPI_Irecv(&u_previous[0][0], 1, column, west, 0, MPI_COMM_WORLD, &request[req_count+1]);
285                     }
286                     if (east != MPI_PROC_NULL) {
287                         MPI_Irecv(&u_previous[0][local[1] + 1], 1, column, east, 0, MPI_COMM_WORLD,
288                         &request[req_count++]);
289                     }
290
291                     MPI_Waitall(req_count, request, MPI_STATUSES_IGNORE);
292
293                     gettimeofday(&tcs,NULL);
294
295                     //impelment the computation

```

```

291         for (i=i_min;i<i_max;i++)
292             for (j=j_min;j<j_max;j++)
293                 u_current[i][j]=(u_previous[i-1][j]+u_previous[i+1][j]+u_previous[i]
294 [j-1]+u_previous[i][j+1])/4.0;
295
296             gettimeofday(&tcf,NULL);
297             tcomp+=(tcf.tv_sec-tcs.tv_sec)+(tcf.tv_usec-tcs.tv_usec)*0.000001;
298             //swap u_previous and u_current
299             swap=u_previous;
300             u_previous=u_current;
301             u_current=swap;
302
303             /*Fill your code here*/
304
305             /*Compute and Communicate*/
306
307             /*Add appropriate timers for computation*/
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328 #ifdef TEST_CONV
329     if (t%C==0) {
330         //*****TODO*****
331         /*Test convergence*/
332         converged = converge(u_previous, u_current, i_min, i_max, j_min, j_max);
333         MPI_Barrier(MPI_COMM_WORLD);
334         MPI_Allreduce(&converged, &global_converged, 1, MPI_INT, MPI_LAND, MPI_COMM_WORLD);
335         if (global_converged) break;
336
337     }
338 #endif
339
340
341
342
343     //*****
344
345
346     }
347     gettimeofday(&ttf,NULL);
348
349     ttotal=(ttf.tv_sec-tts.tv_sec)+(ttf.tv_usec-tts.tv_usec)*0.000001;
350
351     MPI_Reduce(&ttotal,&total_time,1,MPI_DOUBLE,MPI_MAX,0,MPI_COMM_WORLD);
352     MPI_Reduce(&tcomp,&comp_time,1,MPI_DOUBLE,MPI_MAX,0,MPI_COMM_WORLD);
353
354
355
356
357     //----Rank 0 gathers local matrices back to the global matrix---//
358
359     if (rank==0) {
360         U=allocate2d(global_padded[0],global_padded[1]);
361     }
362
363
364     //*****TODO*****
365     MPI_Gatherv(&u_previous[1][1], 1, local_block, U[0], scattercounts, scatteroffset,
366     global_block, 0, MPI_COMM_WORLD);
367
368

```

```

369     /*Fill your code here*/
370
371
372
373
374
375
376
377
378
379
380     //***** ****
381
382
383
384
385
386     //----Printing results----/
387
388     //***** **** TODO: Change "Jacobi" to "GaussSeidelSOR" or "RedBlackSOR" for appropriate
389     printing ****
390     if (rank==0) {
391         printf("Jacobi X %d Y %d Px %d Py %d Iter %d ComputationTime %lf TotalTime %lf
392 midpoint %lf\n",global[0],global[1],grid[0],grid[1],t,comp_time,total_time,U[global[0]/2]
393 [global[1]/2]);
394
395     #ifdef PRINT_RESULTS
396         char * s=malloc(50*sizeof(char));
397         sprintf(s,"resJacobiMPI_%dx%d_%dx%d",global[0],global[1],grid[0],grid[1]);
398         fprintf2d(s,U,global[0],global[1]);
399         free(s);
400     #endif
401
402 }
403 MPI_Finalize();
404 return 0;
405 }
```

Gauss-Seidel SOR

Για την υλοποίηση του Gauss-Seidel αλγορίθμου ορίσαμε 1 παραπάνω τύπο δεδομένων για 1 σειρά του τοπικού υποπίνακα.

Η υπόλοιπη υλοποίηση εκτός του for loop παραμένει ίδια. Για την υλοποίηση της ιδέας του Gauss-Seidel χρησιμοποιούμε πάλι non blocking επικοινωνία. Η λογική που ακολουθήσαμε είναι πως σε κάθε χρονική στιγμή t, μια διεργασία χρειάζεται τα συνοριακά της χρονικής στιγμής t από τα νότια και τα ανατολικά (u_previous), τα συνοριακά της χρονικής στιγμής t + 1 από τα βόρεια και τα δυτικά (u_current). Αντίστοιχα, στέλνει βόρεια και δυτικά τα δικά της συνοριακά κελιά. Αφού ικανοποιηθούν όλα αυτά τα requests με MPI_Waitall, ξεκινάει τους επόμενους υπολογισμούς θερμοκρασίας και έπειτα στέλνει στα νότια και ανατολικά τα καινούρια συνοριακά, για να υπολογίσουν τις καινούριες θερμοκρασίες.

gauss_seidel.c

```

1   MPI_Datatype row_bound;
2   MPI_Type_contiguous(local[1], MPI_DOUBLE, &row_bound);
3   MPI_Type_commit(&row_bound);
4
5   MPI_Datatype col_bound;
6   MPI_Type_vector(local[0], 1, local[1]+2, MPI_DOUBLE, &dummy);
7   MPI_Type_create_resized(dummy, 0, sizeof(double), &col_bound);
8   MPI_Type_commit(&col_bound);
9
10  //----Computational core----/
11  gettimeofday(&tt, NULL);
12  #ifdef TEST_CONV
13      for (t=0;t<T && !global_converged;t++) {
```

```

14     #endif
15     #ifndef TEST_CONV
16     #undef T
17     #define T 256
18     for (t=0;t<T;t++) {
19     #endif
20
21     //*****TODO*****
22
23     /*Add appropriate timers for computation*/
24
25     /*Compute and Communicate*/
26
27     swap=u_previous;
28     u_previous=u_current;
29     u_current=swap;
30
31     MPI_Status prev_status[6], curr_status[2];
32     MPI_Request prev_reqs[6], curr_reqs[2];
33     int prev = 0, curr = 0;
34
35     int err;
36
37     if (north != MPI_PROC_NULL) {
38         err = MPI_Irecv(&(u_current[0][1]), 1, row_bound, north, MPI_ANY_TAG, MPI_COMM_WORLD,
39 &prev_reqs[prev++]);
40         err = MPI_Isend(&(u_previous[1][1]), 1, row_bound, north, 0, MPI_COMM_WORLD,
41 &prev_reqs[prev++]);
42     }
43
44     if (south != MPI_PROC_NULL) {
45         err = MPI_Irecv(&(u_previous[local[0]+1][1]), 1, row_bound, south, MPI_ANY_TAG,
46 MPI_COMM_WORLD, &prev_reqs[prev++]);
47     }
48
49     if (east != MPI_PROC_NULL) {
50         err = MPI_Irecv(&(u_previous[1][local[1]+1]), 1, col_bound, east, MPI_ANY_TAG,
51 MPI_COMM_WORLD, &prev_reqs[prev++]);
52     }
53
54     if (west != MPI_PROC_NULL) {
55         err = MPI_Irecv(&(u_current[1][0]), 1, col_bound, west, MPI_ANY_TAG, MPI_COMM_WORLD,
56 &prev_reqs[prev++]);
57         err = MPI_Isend(&(u_previous[1][1]), 1, col_bound, west, 1, MPI_COMM_WORLD,
58 &prev_reqs[prev++]);
59     }
60
61     //only wait for recvs
62     MPI_Waitall(prev, prev_reqs, prev_status);
63
64     // computation starts here
65     gettimeofday(&tcs, NULL);
66     // Gauss Seidel kernel
67     for (i=i_min;i<i_max;i++)
68     for (j=j_min;j<j_max;j++)
69     u_current[i][j]=u_previous[i][j]+(u_current[i-1][j]+u_previous[i+1][j]+u_current[i]
[j-1]+u_previous[i][j+1]-4*u_previous[i][j])*omega/4.0;
70     // computation ends here
71     gettimeofday(&tcf, NULL);
72     tcomp += (tcf.tv_sec-tcs.tv_sec)+(tcf.tv_usec-tcs.tv_usec)*0.000001;
73
74     // send rest values
75     if (south != MPI_PROC_NULL){
76         err = MPI_Isend(&(u_current[local[0]][1]), 1, row_bound, south, 2, MPI_COMM_WORLD,
77 &curr_reqs[curr++]);
78     }
79
80     if (east != MPI_PROC_NULL) {
81         err = MPI_Isend(&(u_current[1][local[1]]), 1, col_bound, east, 3, MPI_COMM_WORLD,
82 &curr_reqs[curr++]);
83     }
84
85     // now sends must be complete

```

```

79     MPI_Waitall(curr, curr_reqs, curr_status);
80
81 #ifdef TEST_CONV
82     if (t%C==0) {
83         //*****TODO*****
84         /*Test convergence*/
85         gettimeofday(&tcsv, NULL);
86         converged = converge(u_previous, u_current, i_min, i_max, j_min, j_max);
87         // min takes 0 if some proc has not converged locally
88         MPI_Allreduce(&converged, &global_converged, 1, MPI_INT, MPI_LAND, MPI_COMM_WORLD);
89         gettimeofday(&tcvf, NULL);
90         tconv += (tcvf.tv_sec-tcsv.tv_sec)+(tcvf.tv_usec-tcsv.tv_usec)*0.00001;
91     }
92 #endif
93
94     //barrier not needed, waitall is enough
95     //MPI_Barrier(CART_COMM);
96
97     //*****
98 }
99
100
101
102 }
```

Red-Black SOR

Για την παραλληλοποίηση της μεθόδου Red-Black με SOR υλοποιήσαμε 3 διαφορετικές εκδοχές: **Blocking**, **Non-Blocking** και **Overlapping** οι οποίες περιγράφονται παρακάτω:

1) Blocking version (Sendrecv)

Αρχικά, χρησιμοποιήθηκε ο ίδιος σκελετός με την υλοποίηση του Jacobi, με την διαφορά ότι ο υπολογισμός γίνεται σε 2 φάσεις (μία για τα μαύρα και μία για τα κόκκινα σημεία). Παρατηρούμε ότι ο απαιτείται η χρήση των MPI μηνυμάτων για την ανταλλαγή των συνοριακών γραμμών-στηλών μόλις τελειώσει η φάση 1 (μαύρα σημεία), διότι η φάση 2 εκτελεί τον υπολογισμό πάνω στις ανανεωμένες τιμές του grid.

mpi_red_black.c

```

1  //----Computational core----//
2  gettimeofday(&sts, NULL);
3  #ifdef TEST_CONV
4  for (t=0;t<T && !global_converged;t++) {
5  #endif
6  #ifndef TEST_CONV
7  #undef T
8  #define T 256
9  for (t=0;t<T;t++) {
10 #endif
11
12
13     //*****TODO*****
14
15     /*Add appropriate timers for computation*/
16
17     /*Compute and Communicate*/
18
19     swap=u_previous;
20     u_previous=u_current;
21     u_current=swap;
22
23     MPI_Status status;
24     int err;
25     // communication starts here
26     gettimeofday(&tms, NULL);
27 }
```

```

28     if (north != MPI_PROC_NULL) {
29         err = MPI_Sendrecv(&u_previous[1][1], 1, row_bound, north, 0,
30                            &u_previous[0][1], 1, row_bound, north, MPI_ANY_TAG,
31                            MPI_COMM_WORLD, &status);
32         if (err != MPI_SUCCESS) {
33             printf("Process %d failed to communicate with North (rank %d)\n", rank, north);
34             MPI_Abort(MPI_COMM_WORLD, err);
35         }
36     }
37
38     if (south != MPI_PROC_NULL) {
39         err = MPI_Sendrecv(&u_previous[local[0]][1], 1, row_bound, south, 1,
40                            &u_previous[local[0]+1][1], 1, row_bound, south, MPI_ANY_TAG,
41                            MPI_COMM_WORLD, &status);
42         if (err != MPI_SUCCESS) {
43             printf("Process %d failed to communicate with South (rank %d)\n", rank, south);
44             MPI_Abort(MPI_COMM_WORLD, err);
45     }
46 }
47
48     if (east != MPI_PROC_NULL) {
49         err = MPI_Sendrecv(&u_previous[1][local[1]], 1, col_bound, east, 2,
50                            &u_previous[1][local[1]+1], 1, col_bound, east, MPI_ANY_TAG,
51                            MPI_COMM_WORLD, &status);
52         if (err != MPI_SUCCESS) {
53             printf("Process %d failed to communicate with East (rank %d)\n", rank, east);
54             MPI_Abort(MPI_COMM_WORLD, err);
55     }
56 }
57
58     if (west != MPI_PROC_NULL) {
59         err = MPI_Sendrecv(&u_previous[1][1], 1, col_bound, west, 3,
60                            &u_previous[1][0], 1, col_bound, west, MPI_ANY_TAG,
61                            MPI_COMM_WORLD, &status);
62         if (err != MPI_SUCCESS) {
63             printf("Process %d failed to communicate with West (rank %d)\n", rank, west);
64             MPI_Abort(MPI_COMM_WORLD, err);
65     }
66 }
67
68 // communication ends here
69 MPI_Barrier(MPI_COMM_WORLD);
70 gettimeofday(&tmf, NULL);
71
72 // computation starts here
73 gettimeofday(&tcs, NULL);
74 // RED SOR
75 // (i+j) is even
76 for (i=i_min;i<i_max;i++)
77     for (j=j_min;j<j_max;j++)
78         if ((i+j)%2==0)
79             u_current[i][j]=u_previous[i][j]+(omega/4.0)*(u_previous[i-1][j]+u_previous[i+1]
[j]+u_previous[i][j-1]+u_previous[i][j+1]-4*u_previous[i][j]);
80 // computation ends here
81 gettimeofday(&tcf, NULL);
82
83 tcomm += (tmf.tv_sec-tms.tv_sec)+(tmf.tv_usec-tms.tv_usec)*0.000001;
84 tcomp += (tcf.tv_sec-tcs.tv_sec)+(tcf.tv_usec-tcs.tv_usec)*0.000001;
85
86 MPI_Barrier(MPI_COMM_WORLD);
87 gettimeofday(&tms, NULL);
88
89     if (north != MPI_PROC_NULL) {
90         err = MPI_Sendrecv(&u_current[1][1], 1, row_bound, north, 0,
91                            &u_current[0][1], 1, row_bound, north, MPI_ANY_TAG,
92                            MPI_COMM_WORLD, &status);
93         if (err != MPI_SUCCESS) {
94             printf("Process %d failed to communicate with North (rank %d)\n", rank, north);
95             MPI_Abort(MPI_COMM_WORLD, err);
96     }
97 }
98
99     if (south != MPI_PROC_NULL) {
100        err = MPI_Sendrecv(&u_current[local[0]][1], 1, row_bound, south, 1,

```

```

101             &u_current[local[0]+1][1], 1, row_bound, south, MPI_ANY_TAG,
102             MPI_COMM_WORLD, &status);
103     if (err != MPI_SUCCESS) {
104         printf("Process %d failed to communicate with North (rank %d)\n", rank, south);
105         MPI_Abort(MPI_COMM_WORLD, err);
106     }
107 }
108
109 if (east != MPI_PROC_NULL) {
110     err = MPI_Sendrecv(&u_current[1][local[1]], 1, col_bound, east, 2,
111                         &u_current[1][local[1]+1], 1, col_bound, east, MPI_ANY_TAG,
112                         MPI_COMM_WORLD, &status);
113     if (err != MPI_SUCCESS) {
114         printf("Process %d failed to communicate with North (rank %d)\n", rank, east);
115         MPI_Abort(MPI_COMM_WORLD, err);
116     }
117 }
118
119 if (west != MPI_PROC_NULL) {
120     err = MPI_Sendrecv(&u_current[1][1], 1, col_bound, west, 3,
121                         &u_current[1][0], 1, col_bound, west, MPI_ANY_TAG,
122                         MPI_COMM_WORLD, &status);
123     if (err != MPI_SUCCESS) {
124         printf("Process %d failed to communicate with West (rank %d)\n", rank, west);
125         MPI_Abort(MPI_COMM_WORLD, err);
126     }
127 }
128
129 MPI_Barrier(MPI_COMM_WORLD);
130 gettimeofday(&tmf, NULL);
131
132 // computation starts here
133 gettimeofday(&tcs, NULL);
134 // BLACK SOR
135 // (i+j) is odd
136 for (i=i_min;i<i_max;i++)
137     for (j=j_min;j<j_max;j++)
138         if ((i+j) % 2 == 1)
139             u_current[i][j]=u_previous[i][j]+(omega/4.0)*(u_current[i-1][j]+u_current[i+1]
[j]+u_current[i][j-1]+u_current[i][j+1]-4*u_previous[i][j]);
140     // computation ends here
141     gettimeofday(&tcf, NULL);
142
143 tcomm += (tmf.tv_sec-tms.tv_sec)+(tmf.tv_usec-tms.tv_usec)*0.000001;
144 tcomp += (tcf.tv_sec-tcs.tv_sec)+(tcf.tv_usec-tcs.tv_usec)*0.000001;
145
146 #ifdef TEST_CONV
147 if (t%C==0) {
148     //*****TODO*****
149     /*Test convergence*/
150     gettimeofday(&tcv, NULL);
151     converged = converge(u_previous, u_current, i_min, i_max, j_min, j_max);
152     MPI_Allreduce(&converged, &global_converged, 1, MPI_INT, MPI_LAND, MPI_COMM_WORLD);
153     gettimeofday(&tcfv, NULL);
154     tconv += (tcfv.tv_sec-tcv.tv_sec)+(tcfv.tv_usec-tcv.tv_usec)*0.000001;
155 }
156 #endif
157
158
159 //*****
160
161
162
163 }

```

2) Non-Blocking version (Isend/Irecv) και ανταλλαγή N/2 σημείων

Η προηγούμενη εκδοχή κάνει πολλές περιττές ανταλλαγές μηνυμάτων, συγκεκριμένα σε κάθε γύρο μεταφέρονται διπλάσια δεδομένα από όσα απαιτούνται. Γι' αυτό, δημιουργήσαμε ένα νέο datatype με διπλάσιο stride από πριν, ώστε να skip-άρουμε τα κόκκινα σημεία όταν βρισκόμαστε στην φάση 1 (υπολογισμός μαύρων) και αντίστοιχα να skip-άρουμε τα μαύρα όταν βρισκόμαστε στην φάση 2

(υπολογισμός κόκκινων).

Ακόμη, η αποστολή και λήψη των συνοριακών σημείων μπορεί να γίνει ταυτόχρονα προς όλες τις κατευθύνσεις, εφόσον χρησιμοποιεί ξεχωριστά sockets για κάθε γείτονα-process. Οπότε η επιλογή των Isend/Irecv που είναι non-blocking επιταχύνουν την ανταλλαγή. Χρειάζεται προσοχή στην επιλογή των αρχικών διευθύνσεων των send και receive buffers, ώστε τα datatypes που ορίσουμε να τοποθετούν τα δεδομένα σε άρτιες ή περιττές θέσεις αναλόγως την φάση που βρισκόμαστε (κατά σύμβαση even indices -> red point, odd indices -> black points). Πρωτού ξεκινήσουν οι υπολογισμοί, εξασφαλίζουμε με Waitall ότι όλες οι μεταφορές έχουν ολοκληρωθεί. Τέλος, ως επιπλεόν βελτιστοποίηση, τροποποιούμε τον kernel του RedBlack και για τις 2 φάσεις, ώστε να γλιτώσουμε n^2 επαναλήψεις με περιττά branches. Συγκεκριμένα, για κάθε i επιλέγουμε το πρώτο j που δίνει άρτιο (ή περιττό) άθροισμα και έπειτα χρησιμοποιούμε βήμα 2 για το εσωτερικό loop.

mpi_red_black_async.c

```
1   MPI_Datatype RedBlack_row;
2   MPI_Type_vector(local[1]/2, 1, 2, MPI_DOUBLE, &dummy);
3   MPI_Type_create_resized(dummy, 0, sizeof(double), &RedBlack_row);
4   MPI_Type_commit(&RedBlack_row);
5
6   MPI_Datatype RedBlack_col;
7   MPI_Type_vector(local[0]/2, 1, 2*(local[1]+2), MPI_DOUBLE, &dummy);
8   MPI_Type_create_resized(dummy, 0, sizeof(double), &RedBlack_col);
9   MPI_Type_commit(&RedBlack_col);
10
11 //----Computational core----//
12 gettimeofday(&pts, NULL);
13 #ifdef TEST_CONV
14 for (t=0;t<T && !global_converged;t++) {
15 #endif
16 #ifndef TEST_CONV
17 #undef T
18 #define T 256
19 for (t=0;t<T;t++) {
20 #endif
21
22 //*****TODO*****
23 /*Add appropriate timers for computation*/
24
25 /*Compute and Communicate*/
26
27 swap=u_previous;
28 u_previous=u_current;
29 u_current=swap;
30
31 MPI_Status red_status[8], black_status[8];
32 MPI_Request red_reqs[8], black_reqs[8];
33 int red_cnt = 0, black_cnt = 0;
34 int err;
35
36 if (north != MPI_PROC_NULL) {
37     MPI_Irecv(&(u_previous[0][1]), 1, RedBlack_row, north, MPI_ANY_TAG, MPI_COMM_WORLD,
38     &red_reqs[red_cnt++]);
39     MPI_Isend(&(u_previous[1][2]), 1, RedBlack_row, north, 0, MPI_COMM_WORLD,
40     &red_reqs[red_cnt++]);
41 }
42
43 if (south != MPI_PROC_NULL) {
44     MPI_Irecv(&(u_previous[local[0]+1][2]), 1, RedBlack_row, south, MPI_ANY_TAG,
45     MPI_COMM_WORLD, &red_reqs[red_cnt++]);
46     MPI_Isend(&(u_previous[local[0]][1]), 1, RedBlack_row, south, 1, MPI_COMM_WORLD,
47     &red_reqs[red_cnt++]);
48 }
49
50 if (east != MPI_PROC_NULL) {
```

```

49         MPI_Irecv(&(u_previous[2][local[1]+1]), 1, RedBlack_col, east, MPI_ANY_TAG,
50 MPI_COMM_WORLD, &red_reqs[red_cnt++]);
51         MPI_Isend(&(u_previous[1][local[1]]), 1, RedBlack_col, east, 2, MPI_COMM_WORLD,
52 &red_reqs[red_cnt++]);
53     }
54
55     if (west != MPI_PROC_NULL) {
56         MPI_Irecv(&(u_previous[1][0]), 1, RedBlack_col, west, MPI_ANY_TAG, MPI_COMM_WORLD,
57 &red_reqs[red_cnt++]);
58         MPI_Isend(&(u_previous[2][1]), 1, RedBlack_col, west, 3, MPI_COMM_WORLD,
59 &red_reqs[red_cnt++]);
60     }
61
62     MPI_Waitall(red_cnt, red_reqs, red_status);
63
64     // computation starts here
65     gettimeofday(&tcs, NULL);
66     // RED SOR
67     // (i+j) is even
68     for (i=i_min; i<i_max; i++) {
69         if (i & 1) j = (j_min&1) ? j_min : j_min+1;
70         else j = (j_min&1) ? j_min+1 : j_min;
71         for (j; j<j_max; j+=2)
72             u_current[i][j]=u_previous[i][j]+(omega/4.0)*(u_previous[i-1][j]+u_previous[i+1]
73 [j]+u_previous[i][j-1]+u_previous[i][j+1]-4*u_previous[i][j]);
74     }
75     // computation ends here
76     gettimeofday(&tcf, NULL);
77
78     tcomp += (tcf.tv_sec-tcs.tv_sec)+(tcf.tv_usec-tcs.tv_usec)*0.000001;
79
80     //MPI_Barrier(MPI_COMM_WORLD);
81
82     if (north != MPI_PROC_NULL) {
83         MPI_Irecv(&(u_current[0][2]), 1, RedBlack_row, north, MPI_ANY_TAG, MPI_COMM_WORLD,
84 &black_reqs[black_cnt++]);
85         MPI_Isend(&(u_current[1][1]), 1, RedBlack_row, north, 0, MPI_COMM_WORLD,
86 &black_reqs[black_cnt++]);
87     }
88
89     if (south != MPI_PROC_NULL) {
90         MPI_Irecv(&(u_current[local[0]+1][1]), 1, RedBlack_row, south, MPI_ANY_TAG,
91 MPI_COMM_WORLD, &black_reqs[black_cnt++]);
92         MPI_Isend(&(u_current[local[0]][2]), 1, RedBlack_row, south, 1, MPI_COMM_WORLD,
93 &black_reqs[black_cnt++]);
94     }
95
96     if (east != MPI_PROC_NULL) {
97         MPI_Irecv(&(u_current[1][local[1]+1]), 1, RedBlack_col, east, MPI_ANY_TAG,
98 MPI_COMM_WORLD, &black_reqs[black_cnt++]);
99         MPI_Isend(&(u_current[2][local[1]]), 1, RedBlack_col, east, 2, MPI_COMM_WORLD,
100 &black_reqs[black_cnt++]);
101
102     if (west != MPI_PROC_NULL) {
103         MPI_Irecv(&(u_current[2][0]), 1, RedBlack_col, west, MPI_ANY_TAG, MPI_COMM_WORLD,
104 &black_reqs[black_cnt++]);
105         MPI_Isend(&(u_current[1][1]), 1, RedBlack_col, west, 3, MPI_COMM_WORLD,
106 &black_reqs[black_cnt++]);
107
108     MPI_Waitall(black_cnt, black_reqs, black_status);
109
110     // computation starts here
111     gettimeofday(&tcs, NULL);
112     // BLACK SOR
113     // (i+j) is odd
114     for (i=i_min; i<i_max; i++) {
115         if (i & 1) j = (j_min & 1) ? j_min+1 : j_min;
116         else j = (j_min & 1) ? j_min : j_min+1;
117         for (j; j<j_max; j+=2)
118             u_current[i][j]=u_previous[i][j]+(omega/4.0)*(u_current[i-1][j]+u_current[i+1]
119 [j]+u_current[i][j-1]+u_current[i][j+1]-4*u_previous[i][j]);

```

```

108     }
109     // computation ends here
110     gettimeofday(&tcf, NULL);
111
112     tcomp += (tcf.tv_sec-tcs.tv_sec)+(tcf.tv_usec-tcs.tv_usec)*0.000001;
113
114     //MPI_Barrier(MPI_COMM_WORLD);
115
116 #ifdef TEST_CONV
117     if (%==0) {
118         //*****TODO*****
119         /*Test convergence*/
120         gettimeofday(&tcvf, NULL);
121         converged = converge(u_previous, u_current, i_min, i_max, j_min, j_max);
122         MPI_Allreduce(&converged, &global_converged, 1, MPI_INT, MPI_LAND, MPI_COMM_WORLD);
123         gettimeofday(&tcvf, NULL);
124         tconv += (tcvf.tv_sec-tcvf.tv_sec)+(tcvf.tv_usec-tcvf.tv_usec)*0.000001;
125     }
126 #endif
127
128
129 //*****
130
131
132
133 }

```

3) Overlapping version

Η κύρια ιδέα αυτής της εκδοχής είναι η ακόλουθη: Σε κάθε γύρο υπολογίζεται πρώτα η περίμετρος των μαύρων σημείων και στέλνεται αμέσως στους γείτονες με non-blocking τρόπο (Isend/Irecv), ώστε ταυτόχρονα να μπορεί να ξεκινήσει η υπολογισμός των εσωτερικών σημείων. Μόλις οι γείτονες λάβουν τις ανανεωμένες συνοριακές τιμές για τα μαύρα σημεία (Waitall), μπορούν να ξεκινήσουν τον υπολογισμό της περιμέτρου για τα κόκκινα σημεία. Ομοιώς, όταν ολοκληρωθεί στέλνεται με non-blocking τρόπο στους γείτονες και ταυτόχρονα ξεκινάει ο υπολογισμός των εσωτερικών κόκκινων σημείων. Η ορθότητα της μεθόδου βασίζεται στην εξής παρατήρηση: Σε κάθε γύρο (red | black) οι μοναδικές εξαρτήσεις εμφανίζονται στα συνοριακά σημεία μετά το τέλος της κάθε φάσης, αφού στο υπόλοιπο grid οι τιμές έχουν ανανεωθεί από το ίδιο το process και μπορούν να ξαναχρησιμοποιηθούν αμέσως από αυτό στην επόμενη φάση. Επιπλέον, όλα τα σημεία του grid μπορούν να υπολογιστούν με οποιαδήποτε σειρά σε μία φάση, αφού γειτονικά cells δεν υπολογίζονται στον ίδια φάση. Οι πυρήνες υπολογισμών που χρησιμοποιούνται ορίζονται ως inline συναρτήσεις και είναι οι βελτιστοποιημένες εκδοχές που περιγράφηκαν παραπάνω.

mpi_red_black_overlapping.c

```

1  #include <stdio.h>
2  #include <stdlib.h>
3  #include <math.h>
4  #include <sys/time.h>
5  #include "mpi.h"
6  #include "utils.h"
7
8  double omega;
9
10 inline void update_red(double **u_current, double **u_previous, int i, int j){
11     u_current[i][j]=u_previous[i][j]+(omega/4.0)*
12             (u_previous[i-1][j]+u_previous[i+1][j]+u_previous[i][j-1]+u_previous[i]
13 [j+1]-4*u_previous[i][j]);
14 }
15 inline void update_black(double **u_current, double **u_previous, int i, int j){
16     u_current[i][j]=u_previous[i][j]+(omega/4.0)*

```

```

17         (u_current[i-1][j]+u_current[i+1][j]+u_current[i][j-1]+u_current[i]
18 [j+1]-4*u_previous[i][j]);
19
20     int main(int argc, char ** argv) {
21         int rank,size;
22         int global[2],local[2]; //global matrix dimensions and local matrix dimensions (2D-domain,
23 2D-subdomain)
23         int global_padded[2]; //padded global matrix dimensions (if padding is not needed,
24 global_padded=global)
24         int grid[2]; //processor grid dimensions
25         int i,j,t;
26         int global_converged=0,converged=0; //flags for convergence, global and per process
27         MPI_Datatype dummy; //dummy datatype used to align user-defined datatypes in memory
28         //double omega; //relaxation factor - useless for Jacobi
29         // make this global
30
31         struct timeval tts,ttf,tcs,tcf,tcvs,tcvf; //Timers: total-> tts,ttf, computation -
31 > tcs,tcf
32         // convergence-> tcvs,tcvf
33         double ttotal=0,tcomp=0,tconv=0,total_time,comp_time,conv_time;
34
35         double ** U, ** u_current, ** u_previous, ** swap; //Global matrix, local current and
35 previous matrices, pointer to swap between current and previous
36
37
38         MPI_Init(&argc,&argv);
39         MPI_Comm_size(MPI_COMM_WORLD,&size);
40         MPI_Comm_rank(MPI_COMM_WORLD,&rank);
41
42         //----Read 2D-domain dimensions and process grid dimensions from stdin----//
43
44         if (argc!=5) {
45             fprintf(stderr,"Usage: mpirun .... ./exec X Y Px Py");
46             exit(-1);
47         }
48         else {
49             global[0]=atoi(argv[1]);
50             global[1]=atoi(argv[2]);
51             grid[0]=atoi(argv[3]);
52             grid[1]=atoi(argv[4]);
53         }
54
55         //----Create 2D-cartesian communicator----//
56         //----Usage of the cartesian communicator is optional----//
57
58         MPI_Comm CART_COMM; //CART_COMM: the new 2D-cartesian communicator
59         int periods[2]={0,0}; //periods={0,0}: the 2D-grid is non-periodic
60         int rank_grid[2]; //rank_grid: the position of each process on the new communicator
61
62         MPI_Cart_create(MPI_COMM_WORLD,2,grid,periods,0,&CART_COMM); //communicator creation
62         MPI_Cart_coords(CART_COMM,rank,2,rank_grid); //rank mapping on the
63 new communicator
64
65         //----Compute local 2D-subdomain dimensions----//
66         //----Test if the 2D-domain can be equally distributed to all processes----//
67         //----If not, pad 2D-domain----//
68
69         for (i=0;i<2;i++) {
70             if (global[i]%grid[i]==0) {
71                 local[i]=global[i]/grid[i];
72                 global_padded[i]=global[i];
73             }
74             else {
75                 local[i]=(global[i]/grid[i])+1;
76                 global_padded[i]=local[i]*grid[i];
77             }
78         }
79
80         //Initialization of omega
81         omega=2.0/(1+sin(3.14/global[0]));
82
83         //----Allocate global 2D-domain and initialize boundary values----//
84         //----Rank 0 holds the global 2D-domain----//

```

```

85     if (rank==0) {
86         U=allocate2d(global_padded[0],global_padded[1]);
87         init2d(U,global[0],global[1]);
88     }
89
90     //----Allocate local 2D-subdomains u_current, u_previous----//
91     //----Add a row/column on each size for ghost cells----//
92
93     u_previous=allocate2d(local[0]+2,local[1]+2);
94     u_current=allocate2d(local[0]+2,local[1]+2);
95
96     //----Distribute global 2D-domain from rank 0 to all processes----//
97
98     //----Appropriate datatypes are defined here----//
99     //*****The usage of datatypes is optional*****/
100
101    //----Datatype definition for the 2D-subdomain on the global matrix----//
102
103    MPI_Datatype global_block;
104    MPI_Type_vector(local[0],local[1],global_padded[1],MPI_DOUBLE,&dummy);
105    MPI_Type_create_resized(dummy,0,sizeof(double),&global_block);
106    MPI_Type_commit(&global_block);
107
108    //----Datatype definition for the 2D-subdomain on the local matrix----//
109
110    MPI_Datatype local_block;
111    MPI_Type_vector(local[0],local[1],local[1]+2,MPI_DOUBLE,&dummy);
112    MPI_Type_create_resized(dummy,0,sizeof(double),&local_block);
113    MPI_Type_commit(&local_block);
114
115    //----Rank 0 defines positions and counts of local blocks (2D-subdomains) on global
116    //matrix----//
117    int * scatteroffset, * scattercounts;
118    double *Uaddr;
119    if (rank==0) {
120        Uaddr = &(U[0][0]);
121        scatteroffset=(int*)malloc(size*sizeof(int));
122        scattercounts=(int*)malloc(size*sizeof(int));
123        for (i=0;i<grid[0];i++)
124            for (j=0;j<grid[1];j++) {
125                scattercounts[i*grid[1]+j]=1;
126                scatteroffset[i*grid[1]+j]=(local[0]*local[1]*grid[1]*i+local[1]*j);
127            }
128    }
129
130    //----Rank 0 scatters the global matrix----//
131
132    //*****TODO*****//
133
134    //excluded boundaries as they are used for communication
135    MPI_Scatterv(Uaddr, scattercounts, scatteroffset, global_block,
136                  &(u_current[1][1]), 1, local_block, 0, MPI_COMM_WORLD);
137    MPI_Scatterv(Uaddr, scattercounts, scatteroffset, global_block,
138                  &(u_previous[1][1]), 1, local_block, 0, MPI_COMM_WORLD);
139
140    /*Make sure u_current and u_previous are
141     both initialized*/
142
143
144    if (rank==0)
145        free2d(U);
146
147
148
149    //----Define datatypes or allocate buffers for message passing----//
150
151    //*****TODO*****//
152
153
154    MPI_Datatype RedBlack_row;
155    MPI_Type_vector(local[1]/2, 1, 2, MPI_DOUBLE, &dummy);
156    MPI_Type_create_resized(dummy, 0, sizeof(double), &RedBlack_row);
157    MPI_Type_commit(&RedBlack_row);
158
159    MPI_Datatype RedBlack_col;

```

```

160 MPI_Type_vector(local[0]/2, 1, 2*(local[1]+2), MPI_DOUBLE, &dummy);
161 MPI_Type_create_resized(dummy, 0, sizeof(double), &RedBlack_col);
162 MPI_Type_commit(&RedBlack_col);
163
164 //*****Find the 4 neighbors with which a process exchanges messages----/
165
166
167 //----Find the 4 neighbors with which a process exchanges messages----/
168 //*****TODO*****//
169 int north, south, east, west;
170
171 /*Make sure you handle non-existing
172 neighbors appropriately*/
173
174 // MPI_PROC_NULL will be returned in such a case
175 MPI_Cart_shift(CART_COMM, 0, 1, &north, &south);
176 MPI_Cart_shift(CART_COMM, 1, 1, &west, &east);
177
178 //*****Define the iteration ranges per process----/
179 //*****TODO*****/
180
181
182
183 //----Define the iteration ranges per process----/
184 //*****TODO*****/
185
186 int i_min,i_max,j_min,j_max;
187
188 /*Three types of ranges:
189 -internal processes
190 -boundary processes
191 -boundary processes and padded global array
192 */
193
194 if (rank_grid[0] == 0) { // boundary proc (first row)
195     i_min = 2;
196     i_max = local[0]+1;
197 }
198 else if (rank_grid[0] == grid[0]-1) { //boundary proc (last row)
199     i_max = local[0] - (global_padded[0] - global[0]); //check for row padding
200     i_min = 1;
201 }
202 else {
203     i_min = 1; // 0 is for messages
204     i_max = local[0]+1;
205 }
206
207 if (rank_grid[1] == 0) { //boundary proc (first col)
208     j_min = 2;
209     j_max = local[1]+1;
210 }
211 else if (rank_grid[1] == grid[1]-1) { //boundary proc (last col)
212     j_max = local[1] - (global_padded[1] - global[1]); //check for col padding
213     j_min = 1;
214 }
215 else {
216     j_min = 1; // 0 is used for messages
217     j_max = local[1]+1;
218 }
219
220 //*****Define the iteration ranges per process----/
221
222 MPI_Status red_status[8], black_status[8];
223 MPI_Request red_reqs[8], black_reqs[8];
224 int red_cnt = 0, black_cnt = 0;
225
226 MPI_Barrier(MPI_COMM_WORLD);
227
228 //----Computational core----/
229 gettimeofday(&tts, NULL);
230
231 // exchange black borders
232 if (north != MPI_PROC_NULL) {
233     MPI_Irecv(&(u_previous[0][1]), 1, RedBlack_row, north, MPI_ANY_TAG, MPI_COMM_WORLD,
234     &black_reqs[black_cnt++]);

```

```

234         MPI_Isend(&(u_previous[1][2]), 1, RedBlack_row, north, 0, MPI_COMM_WORLD,
235         &black_reqs[black_cnt++]);
236     }
237
238     if (south != MPI_PROC_NULL) {
239         MPI_Irecv(&(u_previous[local[0]+1][2]), 1, RedBlack_row, south, MPI_ANY_TAG,
240         MPI_COMM_WORLD, &black_reqs[black_cnt++]);
241         MPI_Isend(&(u_previous[local[0]][1]), 1, RedBlack_row, south, 1, MPI_COMM_WORLD,
242         &black_reqs[black_cnt++]);
243     }
244
245     if (east != MPI_PROC_NULL) {
246         MPI_Irecv(&(u_previous[2][local[1]+1]), 1, RedBlack_col, east, MPI_ANY_TAG,
247         MPI_COMM_WORLD, &black_reqs[black_cnt++]);
248         MPI_Isend(&(u_previous[1][local[1]]), 1, RedBlack_col, east, 2, MPI_COMM_WORLD,
249         &black_reqs[black_cnt++]);
250     }
251
252     if (west != MPI_PROC_NULL) {
253         MPI_Irecv(&(u_previous[1][0]), 1, RedBlack_col, west, MPI_ANY_TAG, MPI_COMM_WORLD,
254         &black_reqs[black_cnt++]);
255         MPI_Isend(&(u_previous[2][1]), 1, RedBlack_col, west, 3, MPI_COMM_WORLD,
256         &black_reqs[black_cnt++]);
257     }
258
259 #ifdef TEST_CONV
260     for (t=0;t<T && !global_converged;t++) {
261 #endif
262 #ifndef TEST_CONV
263 #undef T
264 #define T 256
265     for (t=0;t<T;t++) {
266 #endif
267
268 //*****TODO*****
269
270         MPI_Waitall(black_cnt, black_reqs, black_status);
271
272         swap=u_previous;
273         u_previous=u_current;
274         u_current=swap;
275
276         black_cnt = 0; red_cnt = 0;
277
278         gettimeofday(&tcs, NULL);
279         //compute red borders
280
281         if (north != MPI_PROC_NULL) {
282             for (j=1; j<=local[1]; j+=2)
283                 //first row
284                 update_red(u_current, u_previous, 1, j);
285         }
286
287         if (south != MPI_PROC_NULL) {
288             for (j=1; j<=local[1]; j+=2)
289                 //last row
290                 update_red(u_current, u_previous, local[0], j+1);
291         }
292
293         if (west != MPI_PROC_NULL) {
294             for (i=2; i<local[0]; i+=2)
295                 // first column
296                 update_red(u_current, u_previous, i+1, 1);
297         }
298
299         if (east != MPI_PROC_NULL) {
300             for (i=2; i<local[0]; i+=2)
301                 //last column
302                 update_red(u_current, u_previous, i, local[1]);
303         }
304
305         gettimeofday(&tcf, NULL);
306         tcomp += (tcf.tv_sec-tcs.tv_sec)+(tcf.tv_usec-tcs.tv_usec)*0.000001;

```

```

802
803     //exchange red borders
804     if (north != MPI_PROC_NULL) {
805         MPI_Irecv(&(u_current[0][2]), 1, RedBlack_row, north, MPI_ANY_TAG, MPI_COMM_WORLD,
806         &red_reqs[red_cnt++]);
807             MPI_Isend(&(u_current[1][1]), 1, RedBlack_row, north, 0, MPI_COMM_WORLD,
808         &red_reqs[red_cnt++]);
809     }
810
811     if (south != MPI_PROC_NULL) {
812         MPI_Irecv(&(u_current[local[0]+1][1]), 1, RedBlack_row, south, MPI_ANY_TAG,
813         MPI_COMM_WORLD, &red_reqs[red_cnt++]);
814         MPI_Isend(&(u_current[local[0]][2]), 1, RedBlack_row, south, 1, MPI_COMM_WORLD,
815         &red_reqs[red_cnt++]);
816     }
817
818     if (east != MPI_PROC_NULL) {
819         MPI_Irecv(&(u_current[1][local[1]+1]), 1, RedBlack_col, east, MPI_ANY_TAG,
820         MPI_COMM_WORLD, &red_reqs[red_cnt++]);
821         MPI_Isend(&(u_current[2][local[1]]), 1, RedBlack_col, east, 2, MPI_COMM_WORLD,
822         &red_reqs[red_cnt++]);
823     }
824
825     if (west != MPI_PROC_NULL) {
826         MPI_Irecv(&(u_current[2][0]), 1, RedBlack_col, west, MPI_ANY_TAG, MPI_COMM_WORLD,
827         &red_reqs[red_cnt++]);
828         MPI_Isend(&(u_current[1][1]), 1, RedBlack_col, west, 3, MPI_COMM_WORLD,
829         &red_reqs[red_cnt++]);
830     }
831
832     gettimeofday(&tcs, NULL);
833     // computer red interior
834     // RED SOR
835     // (i+j) is even
836     for (i=2; i<local[0]; i++) {
837         for (j=(i&1)?3:2; j<local[1]; j+=2)
838             update_red(u_current, u_previous, i, j);
839     }
840     gettimeofday(&tcf, NULL);
841     tcomp += (tcf.tv_sec-tcs.tv_sec)+(tcf.tv_usec-tcs.tv_usec)*0.000001;
842
843     // wait for red borders
844     MPI_Waitall(red_cnt, red_reqs, red_status);
845
846     gettimeofday(&tcs, NULL);
847     // compute black borders
848     if (north != MPI_PROC_NULL){
849         for (j=1; j<=local[1]; j+=2)
850             // first row
851             update_black(u_current, u_previous, 1, j+1);
852     }
853     if (south != MPI_PROC_NULL) {
854         for (j=1; j<=local[1]; j+=2)
855             //last row
856             update_black(u_current, u_previous, local[0], j);
857     }
858
859     if (west != MPI_PROC_NULL) {
860         for (i=2; i<local[0]; i+=2)
861             //first column
862             update_black(u_current, u_previous, i, 1);
863     }
864
865     if (east != MPI_PROC_NULL) {
866         for (i=2; i<local[0]; i+=2)
867             //last column
868             update_black(u_current, u_previous, i+1, local[1]);
869     }
870
871     gettimeofday(&tcf, NULL);
872     tcomp += (tcf.tv_sec-tcs.tv_sec)+(tcf.tv_usec-tcs.tv_usec)*0.000001;
873
874     //exchange black borders

```

```

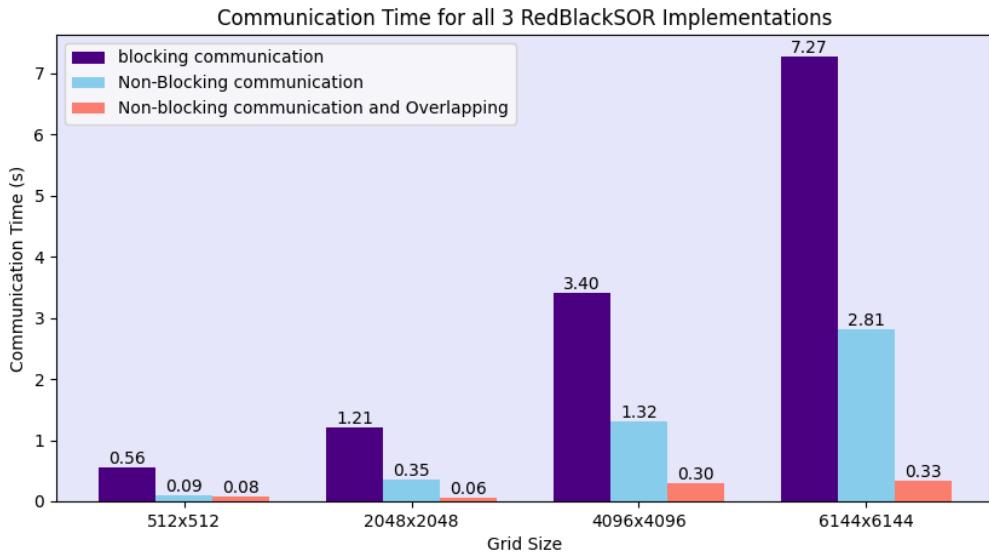
368         if (north != MPI_PROC_NULL) {
369             MPI_Irecv(&(u_current[0][1]), 1, RedBlack_row, north, MPI_ANY_TAG, MPI_COMM_WORLD,
370             &black_reqs[black_cnt++]);
371             MPI_Isend(&(u_current[1][2]), 1, RedBlack_row, north, 0, MPI_COMM_WORLD,
372             &black_reqs[black_cnt++]);
373         }
374
375         if (south != MPI_PROC_NULL) {
376             MPI_Irecv(&(u_current[local[0]+1][2]), 1, RedBlack_row, south, MPI_ANY_TAG,
377             MPI_COMM_WORLD, &black_reqs[black_cnt++]);
378             MPI_Isend(&(u_current[local[0]][1]), 1, RedBlack_row, south, 1, MPI_COMM_WORLD,
379             &black_reqs[black_cnt++]);
380         }
381
382         if (east != MPI_PROC_NULL) {
383             MPI_Irecv(&(u_current[2][local[1]+1]), 1, RedBlack_col, east, MPI_ANY_TAG,
384             MPI_COMM_WORLD, &black_reqs[black_cnt++]);
385             MPI_Isend(&(u_current[1][local[1]]), 1, RedBlack_col, east, 2, MPI_COMM_WORLD,
386             &black_reqs[black_cnt++]);
387         }
388
389         if (west != MPI_PROC_NULL) {
390             MPI_Irecv(&(u_current[1][0]), 1, RedBlack_col, west, MPI_ANY_TAG, MPI_COMM_WORLD,
391             &black_reqs[black_cnt++]);
392             MPI_Isend(&(u_current[2][1]), 1, RedBlack_col, west, 3, MPI_COMM_WORLD,
393             &black_reqs[black_cnt++]);
394         }
395
396         gettimeofday(&tcs, NULL);
397         // compute black interior
398         // BLACK SOR
399         // (i+j) is odd
400         for (i=2; i<local[0]; i++) {
401             for (j=(i&1)?2:3; j<local[1]; j+=2)
402                 update_black(u_current, u_previous, i, j);
403         }
404         gettimeofday(&tcf, NULL);
405         tcomp += (tcf.tv_sec-tcs.tv_sec)+(tcf.tv_usec-tcs.tv_usec)*0.000001;
406
407 #ifdef TEST_CONV
408     if (t%C==0) {
409         //*****TODO*****
410         /*Test convergence*/
411         gettimeofday(&tcvf, NULL);
412         converged = converge(u_previous, u_current, i_min, i_max, j_min, j_max);
413         MPI_Allreduce(&converged, &global_converged, 1, MPI_INT, MPI_LAND, MPI_COMM_WORLD);
414         gettimeofday(&tcfv, NULL);
415         tconv += (tcfv.tv_sec-tcvf.tv_sec)+(tcfv.tv_usec-tcvf.tv_usec)*0.000001;
416     }
417 #endif
418
419 //*****
420 }
421
422 }
423
424
425
426 //----Rank 0 gathers local matrices back to the global matrix----//
427
428 if (rank==0) {
429     U=allocate2d(global_padded[0],global_padded[1]);
430     Uaddr = &(U[0][0]);
431 }
432
433

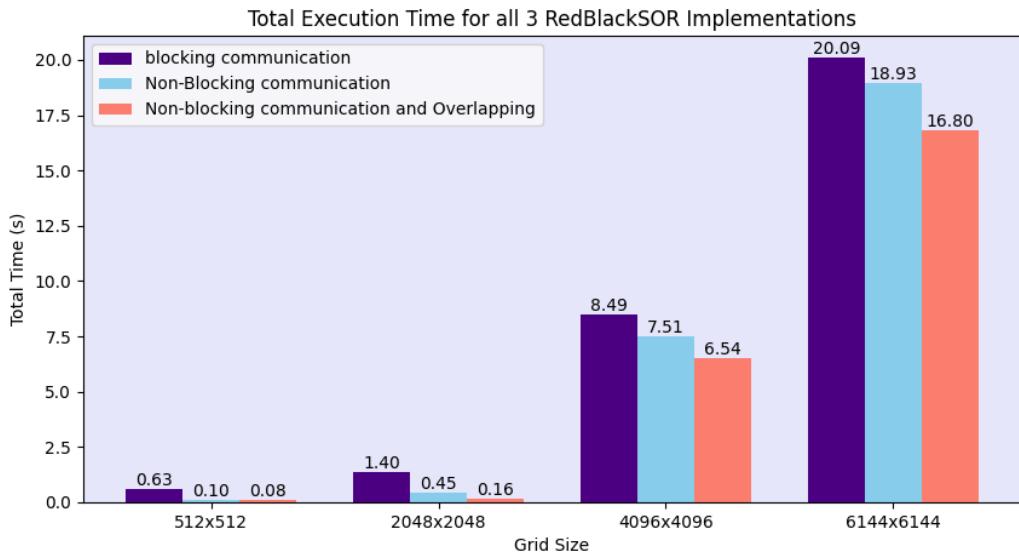
```

```

434 //*****TODO*****
435
436     MPI_Gatherv(&u_current[1][1], 1, local_block, Uaddr, scattercounts, scatteroffset,
437 global_block, 0, MPI_COMM_WORLD);
438 //*****
439
440     //----Printing results----/
441
442     //*****TODO: Change "Jacobi" to "GaussSeidelSOR" or "RedBlackSOR" for appropriate
443     printing*****
444     if (rank==0) {
445         printf("RedBlackSOR X %d Y %d Px %d Py %d Iter %d TotalTime: %lf midpoint
446 %lf\n",global[0],global[1],grid[0],grid[1],t,total_time,U[global[0]/2][global[1]/2]);
447         printf("ComputationTime: %lf CommunicationTime: %lf ", comp_time, (total_time-
448 comp_time));
449         #ifdef TEST_CONV
450             printf("Convergence Time: %lf ", conv_time);
451         #endif
452         #ifdef PRINT_RESULTS
453             char * s=malloc(50*sizeof(char));
454             sprintf(s,"resRedBlackSORMPI_%dx%d_%dx%d",global[0],global[1],grid[0],grid[1]);
455             fprintf2d(s,U,global[0],global[1]);
456             free(s);
457         #endif
458     }
459     MPI_Finalize();
460     return 0;
461 }
```

Οι συγκρίσεις των χρόνων επικοινωνίας και συνολικών χρόνων εκτέλεσης μέχρι το σημείο της σύγκλισης για σταθερό grid μεγέθους 512x512 φαίνονται παρακάτω:

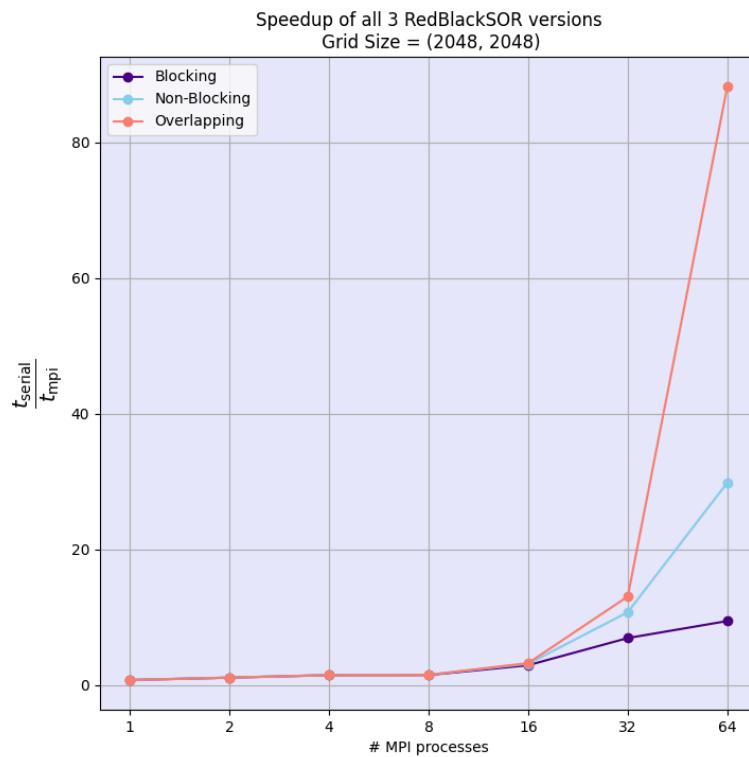


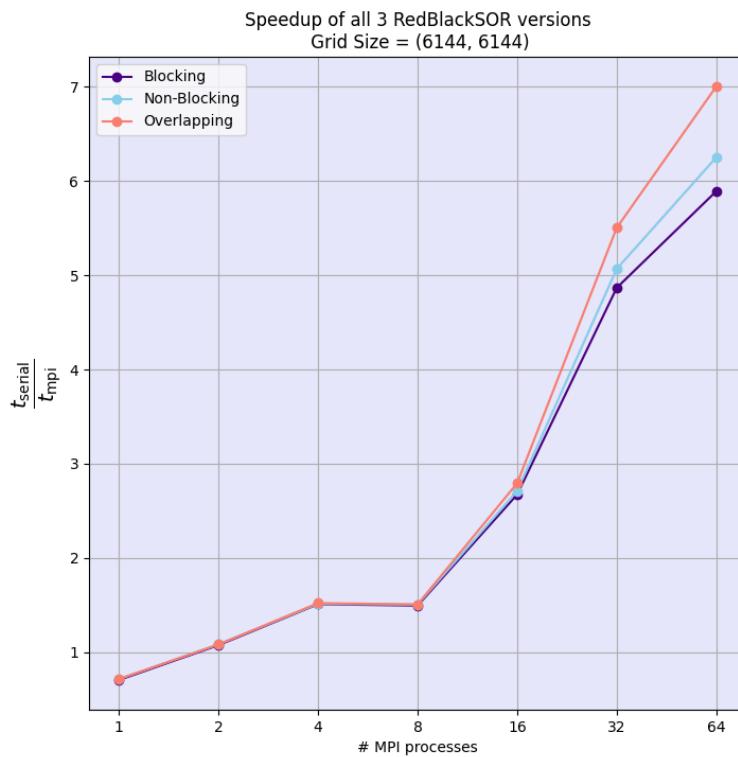
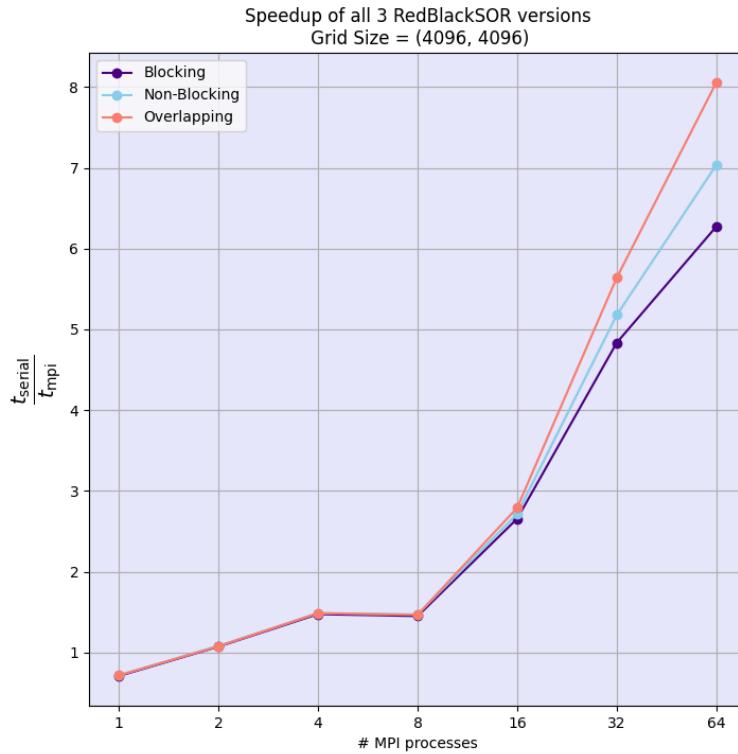


Η overlapping εκδοχή έχει αμελητέο κόστος επικοινωνίας, αφού ταυτόχρονα εκτελεί υπολογισμούς σε κάθε στάδιο του αλγορίθμου. Από το σημείο αυτό και έπειτα η εφαρμογή γίνεται εξ'ολοκληρού compute-bound και δεν βλέπουμε δυνατότητα περαιτέρω βελτιστοποίησης.

Σημείωση: Παρά τα optimisations στο computational part, δεν βλέπουμε μεγάλη μείωση στον χρόνο υπολογισμού (επειδή μεταγλωττίστηκαν όλες οι εκδόσεις με -O3 πιθανά ο compiler κάνει μόνος του κάποια από τα tricks που γράψαμε explicitly στις εκδόσεις 2 κ' 3).

Τα αντίστοιχα speedup plots κρατώντας σταθερό των αριθμό επαναλήψεων T=256 για τα 3 ζητούμενα μεγέθη grid παρατίθενται ακολούθως για λόγους σύγρκισης των μεθόδων:



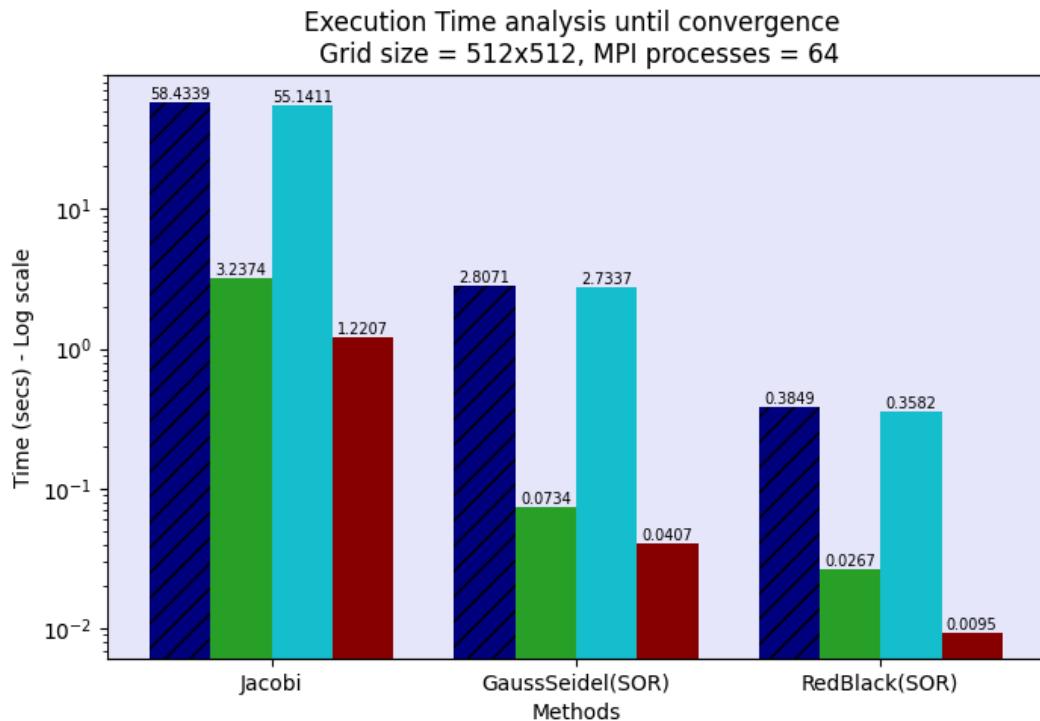


Συνολική Μελέτη επιδόσεων όλων των μεθόδων

Στην συνέχεια μελετάμε μόνο την καλύτερη από τις 3 υλοποιήσεις, δηλ. την Red-Black(SOR)-Overlapping.

Σενάριο Σύγκλισης

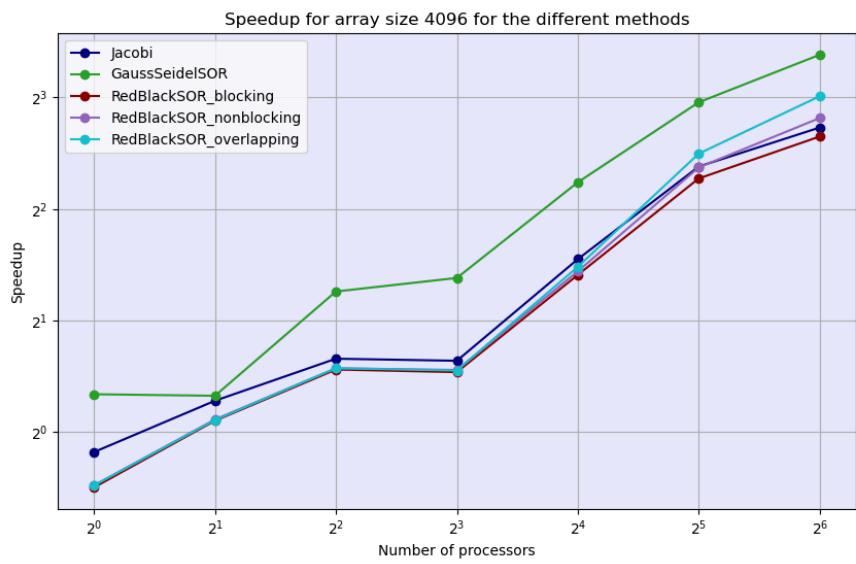
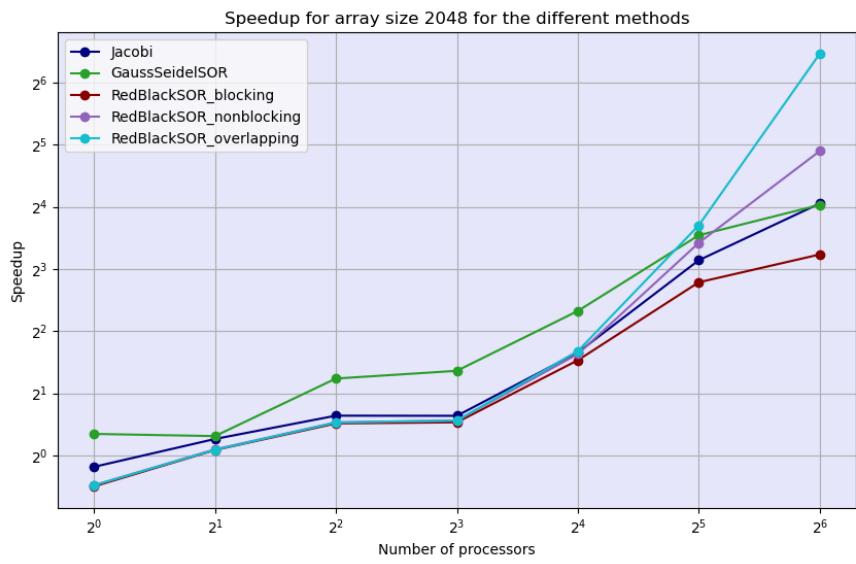
Για το σενάριο σύγκλισης δημιουργούμε τα 3 εκτελέσιμα περνώντας δυναμικά το DCONV Flag κατά την μεταγλώττιση και εξετάζουμε σταθερό μέγεθος πίνακα 512x512 με 64 MPI διεργασίες. Προκύπτει το παρακάτω διάγραμμα με 4 μπάρες ανά αλγόριθμο, οι οποίες με την σειρά αναπαριστούν τον συνολικό χρόνο εκτέλεσης, το χρόνο υπολογισμών, το χρόνο επικουνωνίας και τον χρόνο σύγκλισης αντίστοιχα.

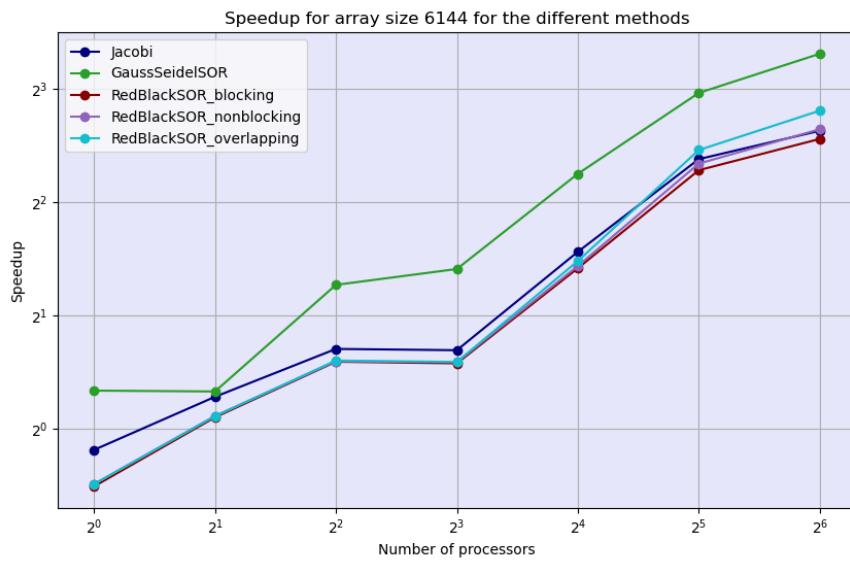


Παρατηρούμε πως ο jacobi έχει την χειρότερη επίδοση με διαφορά με τεράστιο χρόνο να σπαταλάται σε επικοινωνία. Επίσης, χρειάζεται πολύ μεγαλύτερο αριθμό επαναλήψεων για να συγκλίνει σε αντίθεση με τους άλλους αλγορίθμους. Η καλύτερη υλοποίηση για τον red-black sor είναι ελάχιστα καλύτερη σε επίδοση από την gauss-seidel. Η red-black sor έχει λιγότερο computational χρόνο οπότε για μεγαλύτερους πίνακες μάλλον θα είναι προτιμότερη. Εξαρτάται όμως και από τις συνοριακές συνθήκες, καθώς η gauss-seidel έχει γρηγορότερο ρυθμό σύγκλισης.

Σενάριο σταθερού αριθμού επαναλήψεων $T = 256$

Για το σενάριο σταθερού αριθμού επαναλήψεων, δημιουργούμε εκ νέου τα 3 εκτελέσιμα χωρίς το DCONC Flag και εξετάζουμε τις επιδόσεις τους για μεταβήτο αριθμό MPI διεργασιών (1, 2, 4, 8, 16, 32, 64) και μεταβλητά μεγέθη πίνακα (2048x2048, 4096x4096, 6144x6144). Τα διαγράμματα επιτάχυνσης για κάθε μέγεθος πίνακα φαίνονται παρακάτω. Επισημαίνεται πως οι σειριακές μέθοδοι δεν είχαν τον ίδιο χρόνο εκτέλεσης για τους 3 αλγορίθμους.



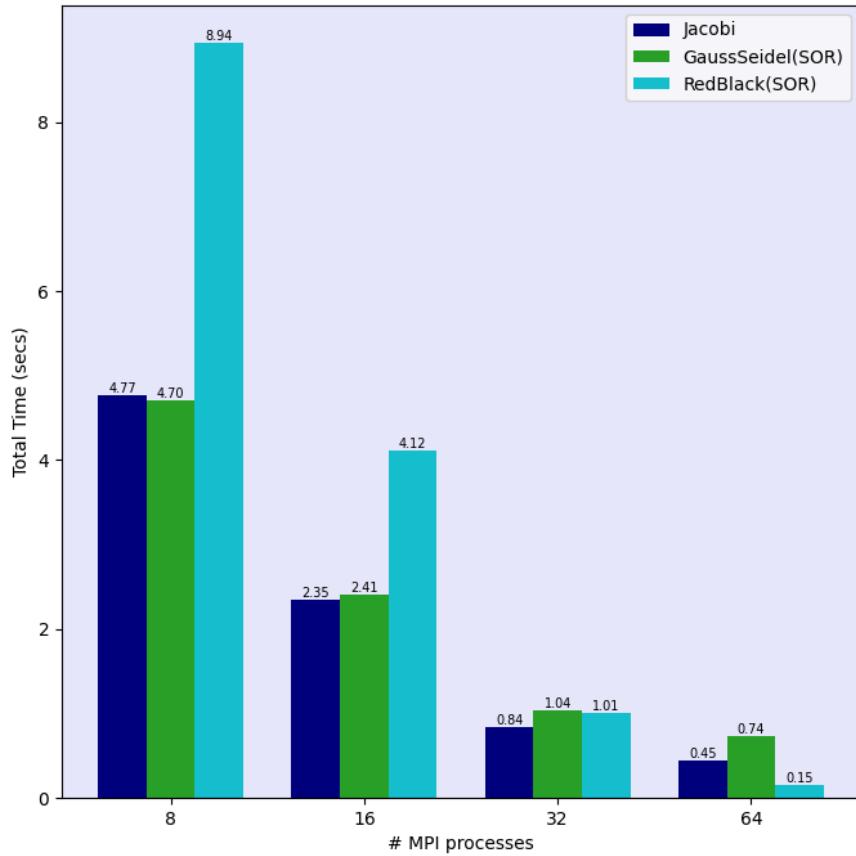


Παρατηρούμε πως για το μικρό μέγεθος 2048, οι υλοποιήσεις red-black sor έχουν την μεγαλύτερη επιτάχυνση, ενώ για τα 2 μεγαλύτερη μεγέθη πινάκων η gauss-seidel έχει την μεγαλύτερη επιτάχυνση, ενώ οι όλες υπόλοιπες έχουν σχεδόν ίδια.

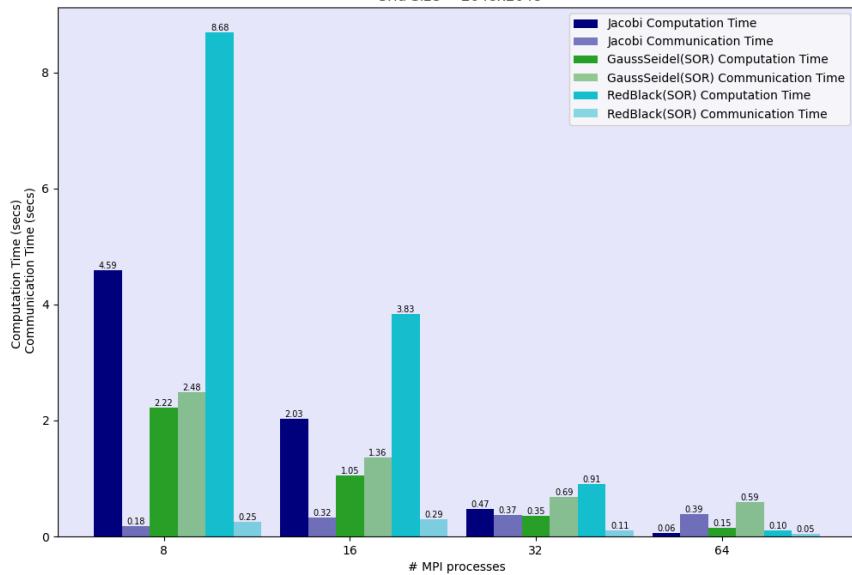
Συνολικά, η red-black sor είναι η καλύτερη υλοποίηση για μικρό μέγεθος πίνακα, ενώ η gauss-seidel είναι καλύτερη για μεγαλύτερους πίνακες. Η jacobi είναι η χειρότερη υλοποίηση σε όλα τα σενάρια.

Παρακάτω παρουσιάζονται τα διάγραμμα χρόνου υπολογισμού/εκτέλεσης για 8, 16, 32, 64 διεργασίες για κάθε μέγεθος πίνακα ξεχωριστά καθώς και ένα συνολικό διάγραμμα για καλύτερη εποπτεία των συνολικών χρόνων.

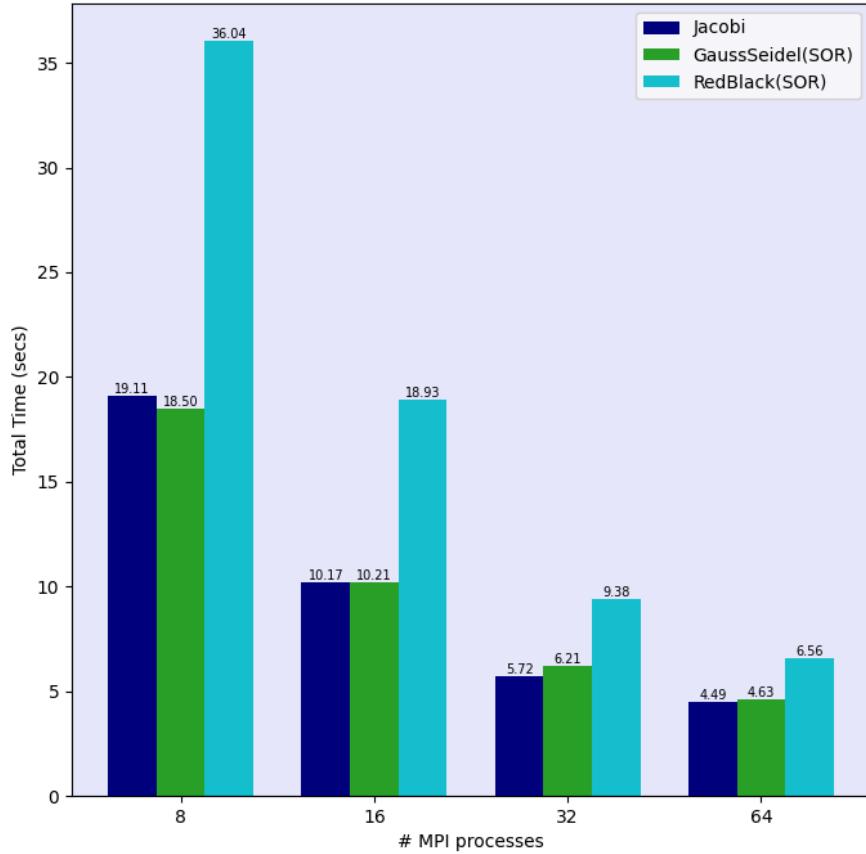
Execution time breakdown for fixed Iteration count T=256
Grid Size = 2048x2048



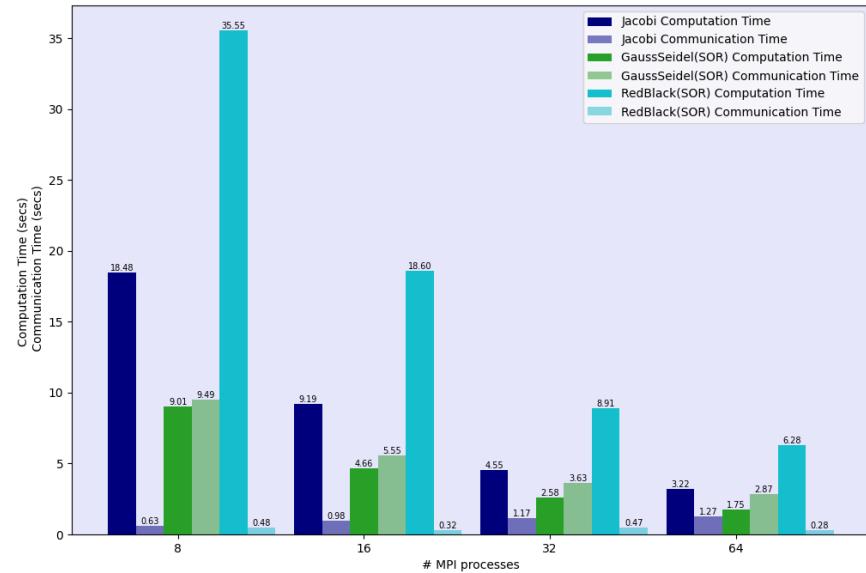
Execution time analysis for fixed Iteration count T=256
Grid Size = 2048x2048



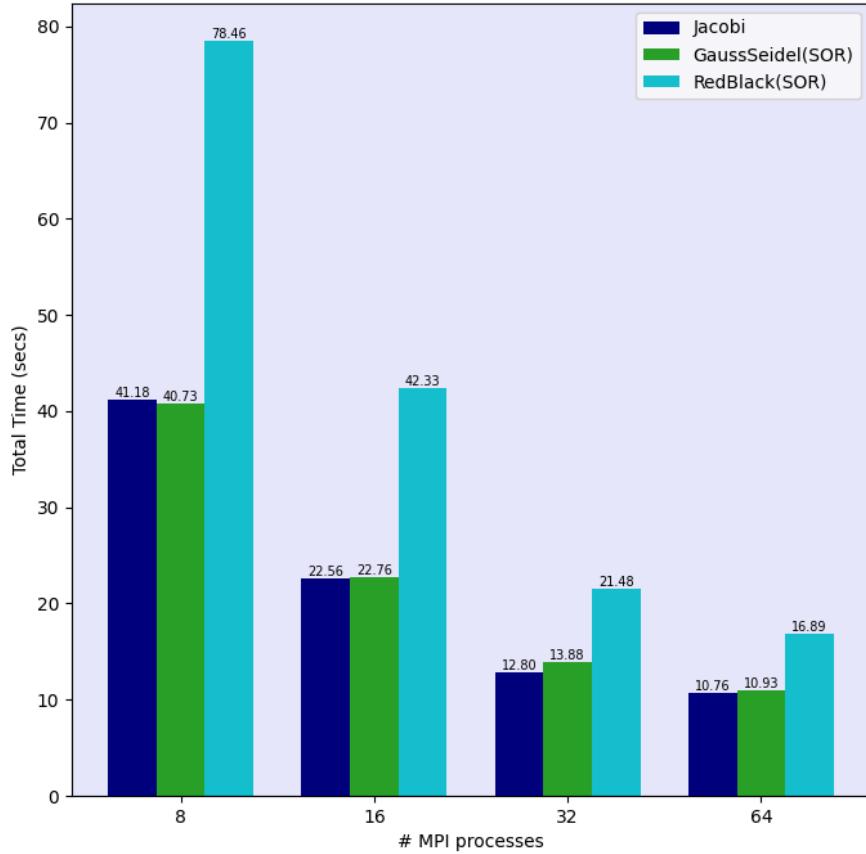
Execution time breakdown for fixed Iteration count T=256
Grid Size = 4096x4096



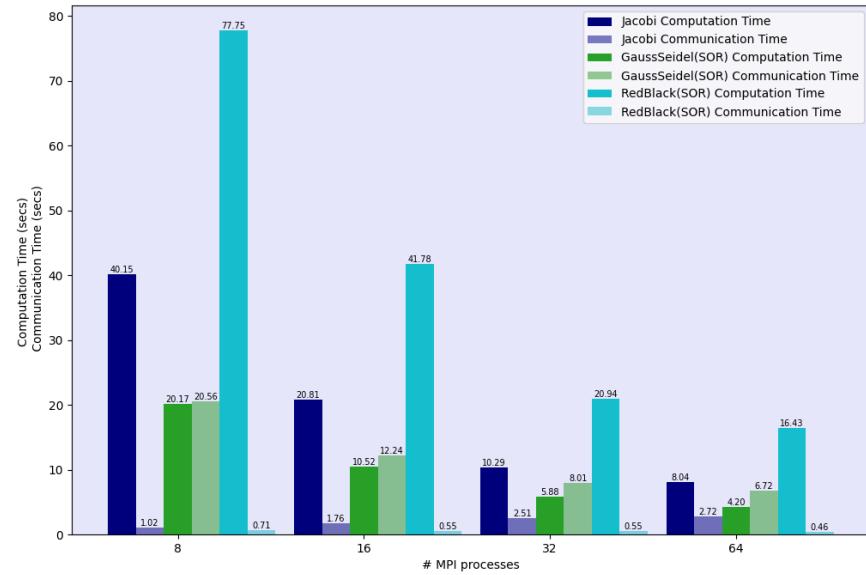
Execution time analysis for fixed Iteration count T=256
Grid Size = 4096x4096

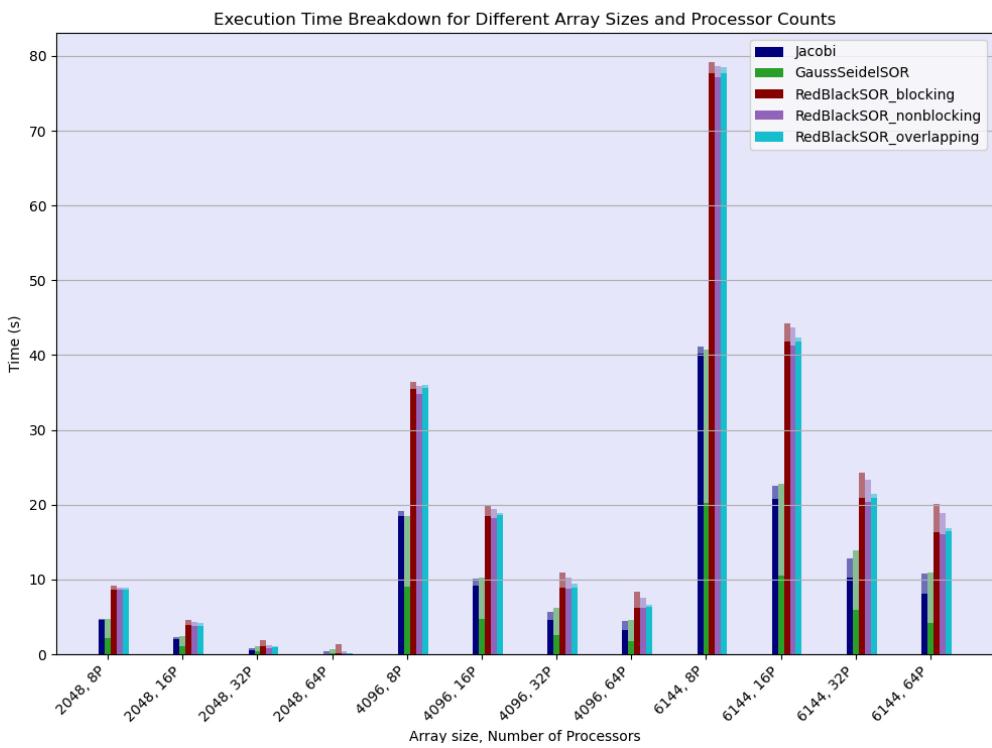


Execution time breakdown for fixed Iteration count T=256
Grid Size = 6144x6144



Execution time analysis for fixed Iteration count T=256
Grid Size = 6144x6144





Παρατηρούμε πως η gauss-seidel έχει περίπου ίδιο χρόνο εκτέλεσης με την jacobi για μεγάλους πίνακες, ενώ συγκλίνει εξαιρετικά πιο γρήγορα. Οπότε, είναι σίγουρα η προτιμητέα μέθοδος για μεγάλους πίνακες. Για μικρά μεγέθη πινάκων, η overlapping red black sor υλοποίηση είναι σίγουρα η καλύτερη καθώς έχει τους μικρότερους χρόνους επικοινωνίας. Σε μικρά μεγέθη πινάκων, οι χρόνοι υπολογισμών είναι μικροί οπότε έχει μεγάλη βαρύτητα ο χρόνος επικοινωνίας.