

ΣΥΣΤΗΜΑΤΑ ΠΑΡΑΛΛΗΛΗΣ ΕΠΕΞΕΡΓΑΣΙΑΣ

ΑΝΑΦΟΡΑ ΠΡΟΠΑΡΑΣΚΕΥΑΣΤΙΚΗΣ ΑΣΚΗΣΗΣ



Στοιχεία Ομάδας

- Αναγνωριστικό: parlab05
- Μέλος 1^ο: Πέππας Μιχαήλ – Αθανάσιος, Α.Μ: 03121026
- Μέλος 2^ο: Σαουνάτσος Ανδρέας, Α.Μ: 03121197
- Ημερομηνία Παράδοσης Αναφοράς: 21.10.2025

■ Ενότητα 1.4.3 – Το Πρόγραμμα

Διθέντος του αρχικού προγράμματος του Game of Life με τη σειριακή υλοποίηση και έχοντας μελετήσει τα παραδείγματα των εισαγωγικών διαφανειών του εργαστηρίου, συντάξαμε το ακόλουθο παράλληλο πρόγραμμα, με χρήση του OpenMP:

```
a1/life_par.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 #include <stdio.h>
15 #include <stdlib.h>
16 #include <sys/time.h>
17 #include <omp.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 {
31     int N;                      // array dimensions
32     int T;                      // time steps
33     int **current, **previous; // arrays - one for current timestep, one for previous
34     int **swap;                 // array pointer
35     int i, j, t, nbrs;          // helper variables
36
37     double time; // variables for timing
38     struct timeval ts, tf;
39
40     /*Read input arguments*/
41     if (argc != 3)
42     {
43         fprintf(stderr, "Usage: ./exec ArraySize TimeSteps\n");
44         exit(-1);
45     }
46     else
47     {
48         N = atoi(argv[1]);
49         T = atoi(argv[2]);
50     }
51 }
```

```

52  /*Allocate and initialize matrices*/
53  current = allocate_array(N); // allocate array for current time step
54  previous = allocate_array(N); // allocate array for previous time step
55
56  init_random(previous, current, N); // initialize previous array with pattern
57
58 #ifdef OUTPUT
59  print_to_pgm(previous, N, 0);
60#endif
61
62  /*Game of Life*/
63
64  gettimeofday(&ts, NULL);
65
66  gettimeofday(&ts, NULL);
67
68  for (t = 0; t < T; ++t)
69  {
70
71  /* Parallelize rows; implicit barrier at loop end */
72  /* schedule is static
73   (i, j) as loop indices are private
74   (N, previous, current) are shared, as they are enclosed by the loop
75   nbrs must be private
76   by default */
77  /* Kept here for clarity */
78 #pragma omp parallel for schedule(static) private(i, j, nbrs) shared(N, previous, current)
79      for (i = 1; i < N - 1; ++i)
80      {
81          for (j = 1; j < N - 1; ++j)
82          {
83              nbrs =
84                  previous[i + 1][j + 1] + previous[i + 1][j] + previous[i + 1][j - 1] +
85                  previous[i][j - 1] + previous[i][j + 1] +
86                  previous[i - 1][j - 1] + previous[i - 1][j] + previous[i - 1][j + 1];
87
88              current[i][j] = (nbrs == 3 || (previous[i][j] + nbrs == 3)) ? 1 : 0;
89          }
90      } /* implicit barrier here: all threads finished step t */
91
92 #ifdef OUTPUT
93     print_to_pgm(current, N, t + 1); /* single thread here: we're back in serial */
94#endif
95
96     /* Safe to swap: we're outside the parallel region created by 'parallel for' */
97     swap = current;
98     current = previous;
99     previous = swap;
100 }
101
102 gettimeofday(&tf, NULL);
103 time = (tf.tv_sec - ts.tv_sec) + (tf.tv_usec - ts.tv_usec) * 0.000001;
104
105 free_array(current, N);

```

```
106     free_array(previous, N);
107     printf("GameOfLife: Size %d Steps %d Time %lf\n", N, T, time);
108 #ifdef OUTPUT
109     system(FINALIZE);
110 #endif
111 }
112
113 int **allocate_array(int N)
114 {
115     int **array;
116     int i, j;
117     array = malloc(N * sizeof(int *));
118     for (i = 0; i < N; i++)
119         array[i] = malloc(N * sizeof(int));
120     for (i = 0; i < N; i++)
121         for (j = 0; j < N; j++)
122             array[i][j] = 0;
123     return array;
124 }
125
126 void free_array(int **array, int N)
127 {
128     int i;
129     for (i = 0; i < N; i++)
130         free(array[i]);
131     free(array);
132 }
133
134 void init_random(int **array1, int **array2, int N)
135 {
136     int i, pos, x, y;
137
138     for (i = 0; i < (N * N) / 10; i++)
139     {
140         pos = rand() % ((N - 2) * (N - 2));
141         array1[pos % (N - 2) + 1][pos / (N - 2) + 1] = 1;
142         array2[pos % (N - 2) + 1][pos / (N - 2) + 1] = 1;
143     }
144 }
145
146 void print_to_pgm(int **array, int N, int t)
147 {
148     int i, j;
149     char *s = malloc(30 * sizeof(char));
150     sprintf(s, "out%d.pgm", t);
151     FILE *f = fopen(s, "wb");
152     fprintf(f, "P5\n%d %d 1\n", N, N);
153     for (i = 0; i < N; i++)
154         for (j = 0; j < N; j++)
155             if (array[i][j] == 1)
156                 fputc(1, f);
157             else
158                 fputc(0, f);
159     fclose(f);
```

```
160     free(s);  
161 }  
162  
163
```

Όσον αφορά τις λεπτομέρειες της υλοποίησής μας (όσον αφορά το παράλληλο τμήμα του προγράμματος, δηλ. τις γραμμές 69-100), επισημαίνουμε τα εξής:

- Η μεταβλητή *t*, δηλαδή ο αριθμός των γενεών, δεν είναι μια διαδικασία που επιδέχεται παραλληλοποίησης, καθώς αφενός μεν η μία διαδέχεται την άλλη, χωρίς να μπορούμε να πάμε στην επόμενη χωρίς να έχει τελειώσει η προηγούμενη, αφετέρου δε το μεγαλύτερο «κέρδος» προκύπτει από την παραλληλοποίηση των υπολογισμών εντός μιας γενιάς. Αυτό διότι, στο τέλος κάθε γενιάς πρέπει όλα τα threads να «συναντηθούν», ώστε να ενημερώσουν τα κοινά μας δεδομένα (δηλαδή τους πίνακες). Επομένως, στο τέλος κάθε γενιάς, υπάρχει ένα νοητό «φράγμα», στο οποίο συναντιούνται και συγχρονίζονται όλα τα threads, μέχρι όλα να τελειώσουν. Δηλαδή, οι γενιές εκτελούνται σειριακά, ενώ οι υπολογισμοί καθεμίας παράλληλα.
- Η παραλληλοποίηση του προγράμματος γίνεται στην γραμμή:

```
#pragma omp parallel for schedule(static) private(i, j, nbrs) shared(N, previous, current)
```

όπου παραλληλοποιούμε το for loop για το *i*. Σημειώνουμε (υπάρχουν και σχόλια στον κώδικα) ότι το scheduling είναι by default static, οι μεταβλητές του loop {*i*, *j*} είναι by default private και τα {*N*, previous, shared} by default shared, καθώς βρίσκονται εντός του παράλληλου τμήματος. Επομένως, η δήλωση των ανωτέρω δεν χρειάζεται, αλλά γίνεται για λόγους διαφάνειας. Η μεταβλητή *nbrs* πρέπει να δηλωθεί ως private, για να αποφευχθούν race conditions, εφόσον δεν δηλώνεται μέσα στο παράλληλο τμήμα.

- Στο τέλος του nested for loop, τα threads περιμένουν μέχρι να τελειώσουν όλα (βλ. σχόλιο) και να γίνει η ασφαλής-ατομική πρόσβαση στους πίνακες που ενημερώνουμε {previous, current}, καθώς αυτοί βρίσκονται εκτός του παράλληλου τμήματος.

Τέλος, τα αρχεία Makefile, make_on_queue.sh και run_on_queue.sh, έχουν συνταχθεί σε πλήρη αντιστοιχία με τα παραδείγματα των διαφανειών, φέρουν μικρές διαφορές που διευκολύνουν την υλοποίηση και την οργάνωση και παρουσιάζονται, για λόγους πληρότητας (χωρίς να χρειάζεται κάποια επεξήγηση), ακολούθως:

a1/Makefile

```
1 all: life_par
2
3 life_par: life_par.c
4     gcc -O3 -fopenmp -o life_par life_par.c
5
6 clean:
7     rm life_par
8
```

a1/make_on_queue.sh

```
1 #!/bin/bash
2
3 ## Give the Job a descriptive name
4 #PBS -N makejob
5
6 ## Output and error files
7 #PBS -o makejob.out
8 #PBS -e makejob.err
9
10 ## How many machines should we get?
11 #PBS -l nodes=1
12
13 ## Start
14 ## Load appropriate module
15 module load openmp
16
17 ## Run make in the src folder (modify properly)
18 cd /home/parallel/parlab05/a1/
19 make
20
```

a1/run_on_queue.sh

```
1 #!/bin/bash
2
3 ## Give the Job a descriptive name
4 #PBS -N life_par
5
6 ## Output and error files
7 #PBS -o life_par.out
8 #PBS -e life_par.err
9
10 ## How many machines should we get?
11 #PBS -l nodes=1:ppn=8
12
13 ## How long should the job run for?
14 #PBS -l walltime=01:00:00
15
16 ## Module Load
17 module load openmp
18
19 ## Defaults if not passed via -v
20 : "${THREADS:=8}"
21 : "${N:=1024}"
22 : "${STEPS:=1000}"
23
24 ## Start
25 cd /home/parallel/parlab05/a1/ || exit 1
26
27 # --- OpenMP runtime settings ---
28 export OMP_NUM_THREADS="${THREADS}"      # 1,2,4,6,8 per the assignment
29
30 # Run and capture outputs by config
31 RESULT_DIR="benchmarks/N${N}_T${THREADS}"
32 mkdir -p "${RESULT_DIR}"
33
34 ./life_par "${N}" "${STEPS}" \
35   > "${RESULT_DIR}/life_${THREADS}_${N}.out" \
36   2> "${RESULT_DIR}/life_${THREADS}_${N}.err"
37
38
```

■ Ενότητα 1.4.4 – Οι Μετρήσεις

Έχοντας συντάξει το παράλληλο πρόγραμμά μας, το υποβάλλαμε στα μηχανήματα του εργαστηρίου για κάθε συνδυασμό των παραμέτρων {μέγεθος ταμπλό, πυρήνες}. Οι παράμετροι αυτές, πήραν τις ακόλουθες τιμές, όπως ζητούνταν από την εκφώνηση της άσκησης:

- Μέγεθος ταμπλό: {64, 1024, 4096}
- Πυρήνες: {1, 2, 4, 6, 8}

Σύνολο $3 \times 5 = 15$ υποβολές.

Έτσι, ανοίγοντας τα αντίστοιχα αρχεία “filename.out”, λάβαμε τα ακόλουθα αποτελέσματα, τα οποία και παρουσιάζουμε στον κάτωθι πίνακα.

SIZE	THREADS	TIME	SPEEDUP
64	1	0.022613	1.000000
64	2	0.013416	1.685525
64	4	0.009780	2.312168
64	6	0.008942	2.528853
64	8	0.009163	2.467860
1024	1	11.793298	1.000000
1024	2	5.868753	2.009507
1024	4	2.929962	4.025069
1024	6	1.965196	6.001080
1024	8	1.476472	7.987485
4096	1	189.347966	1.000000
4096	2	94.832356	1.996660
4096	4	47.729763	3.967084
4096	6	32.393775	5.845196
4096	8	28.594441	6.621845

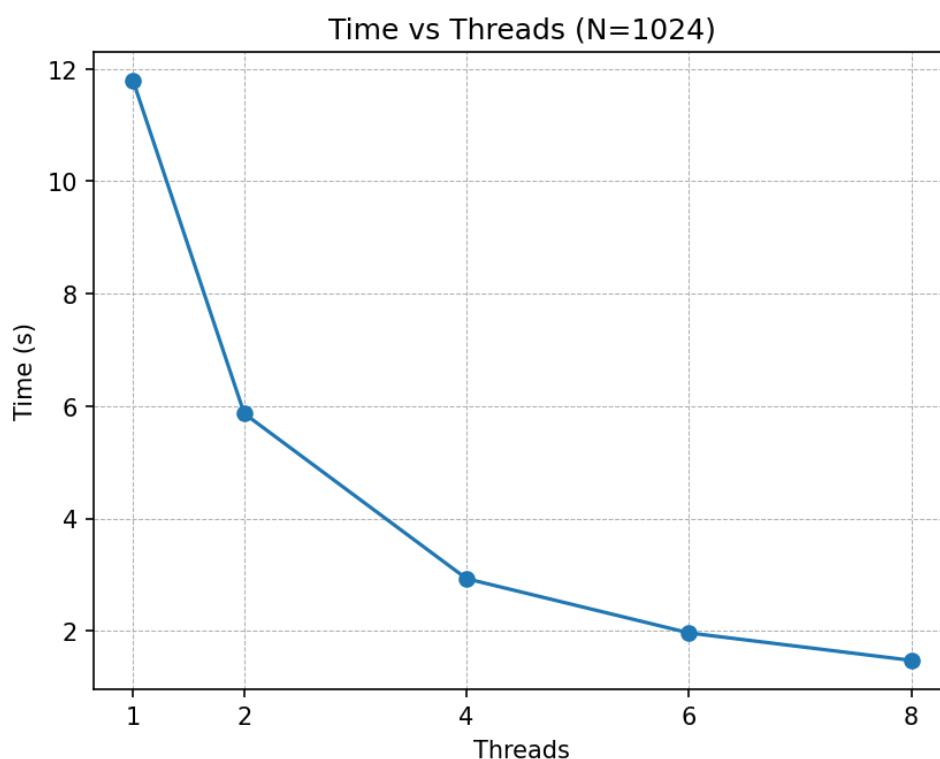
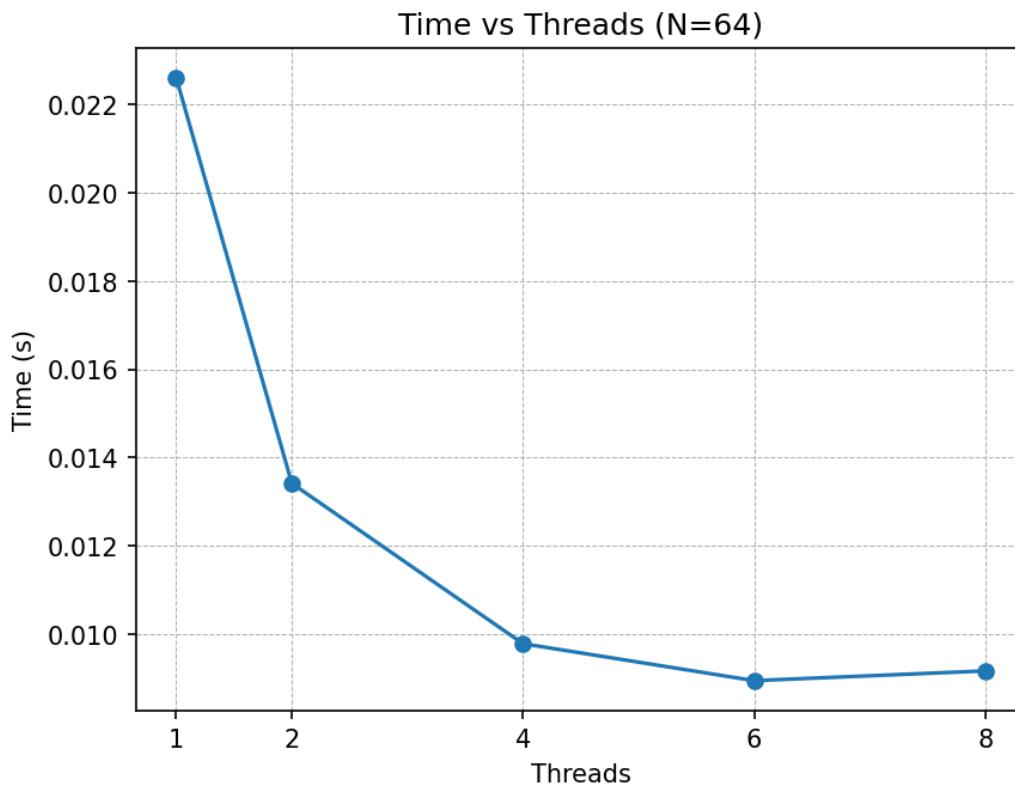
Τα αποτελέσματα αυτά αποτελούν τα δείγματα αναφοράς, με βάση τα οποία θα καταστρώσουμε τα ζητούμενα διαγράμματα (χρόνου και speedup) του επόμενου ερωτήματος. Επισημαίνουμε ότι:

$$Speedup = \frac{Time_of_N_cores}{Time_of_1_core}$$

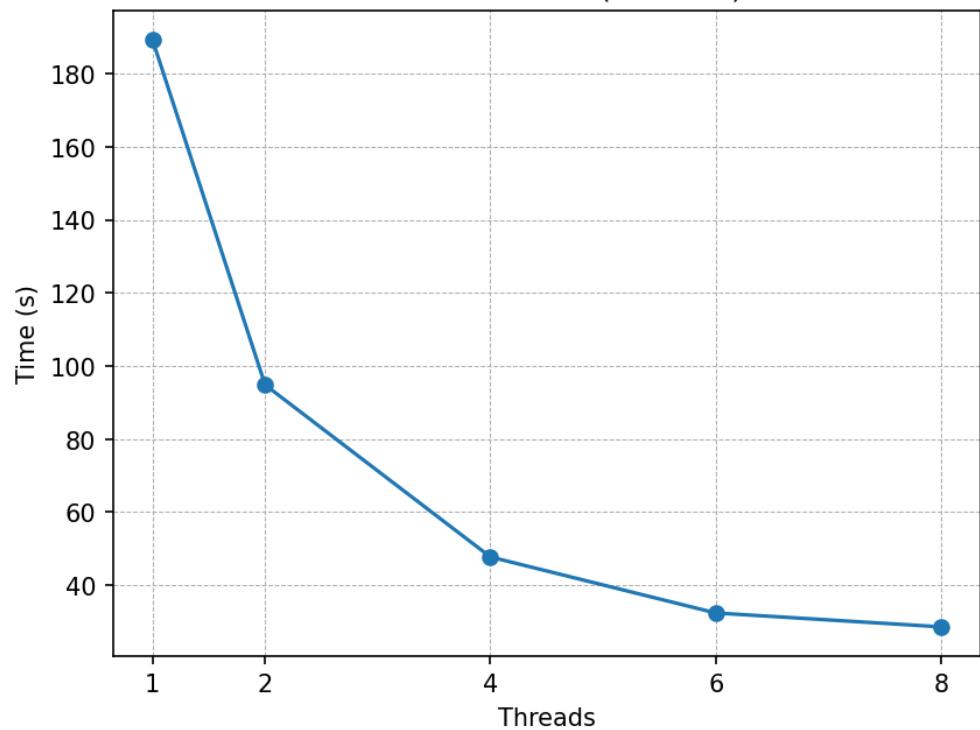
■ Ενότητα 1.4.5 – Τα Διαγράμματα

Σύμφωνα με τις μετρήσεις του παραπάνω πίνακα, καταστρώνουμε τα ακόλουθα διαγράμματα για κάθε μέγεθος ταμπλό (με χρήση ενός προγράμματος Python):

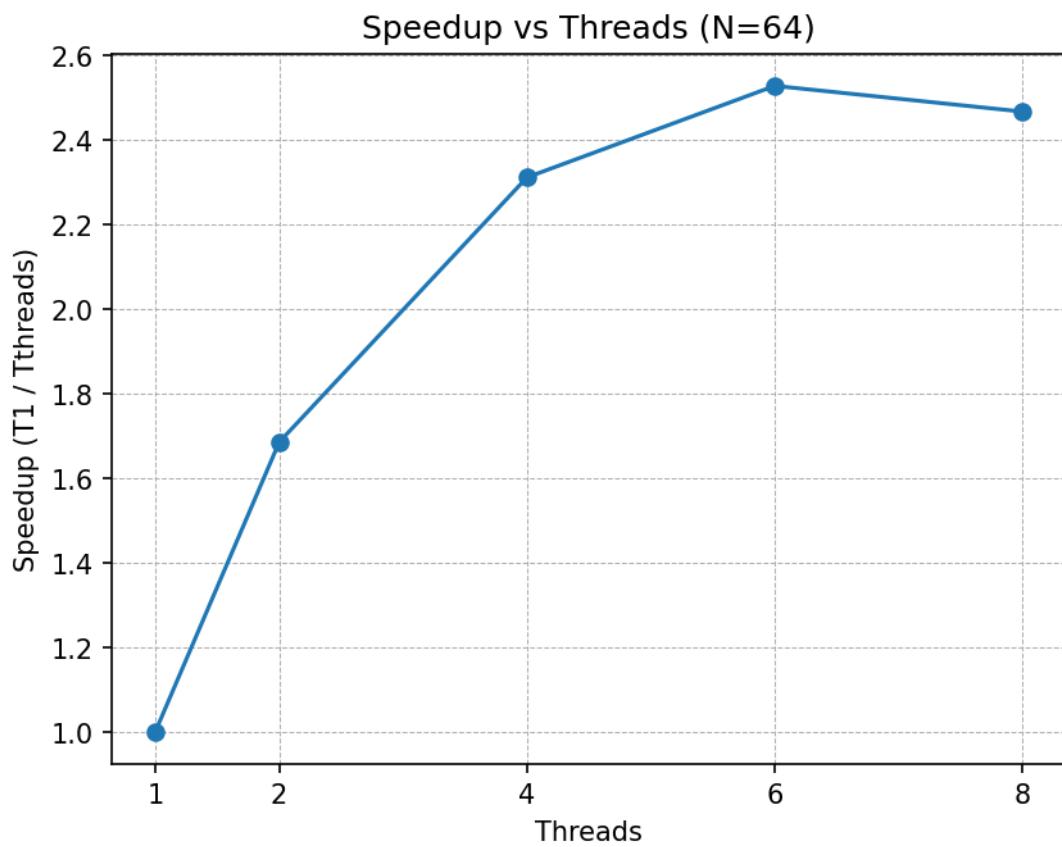
Διαγράμματα Χρόνου



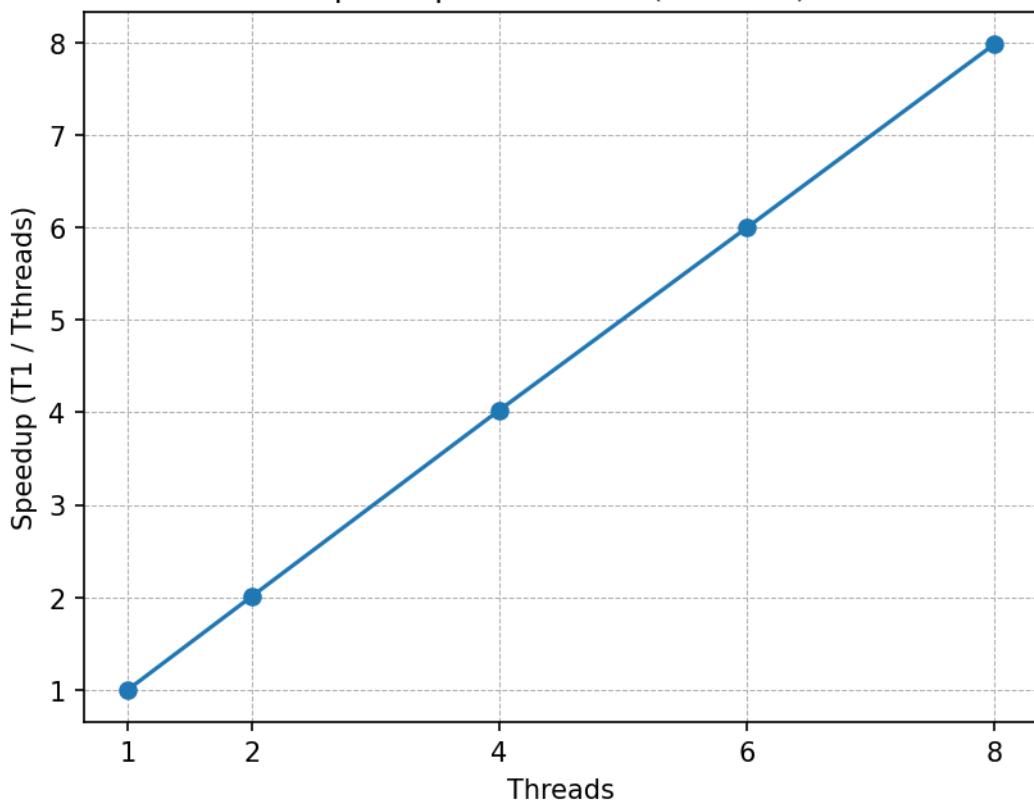
Time vs Threads (N=4096)



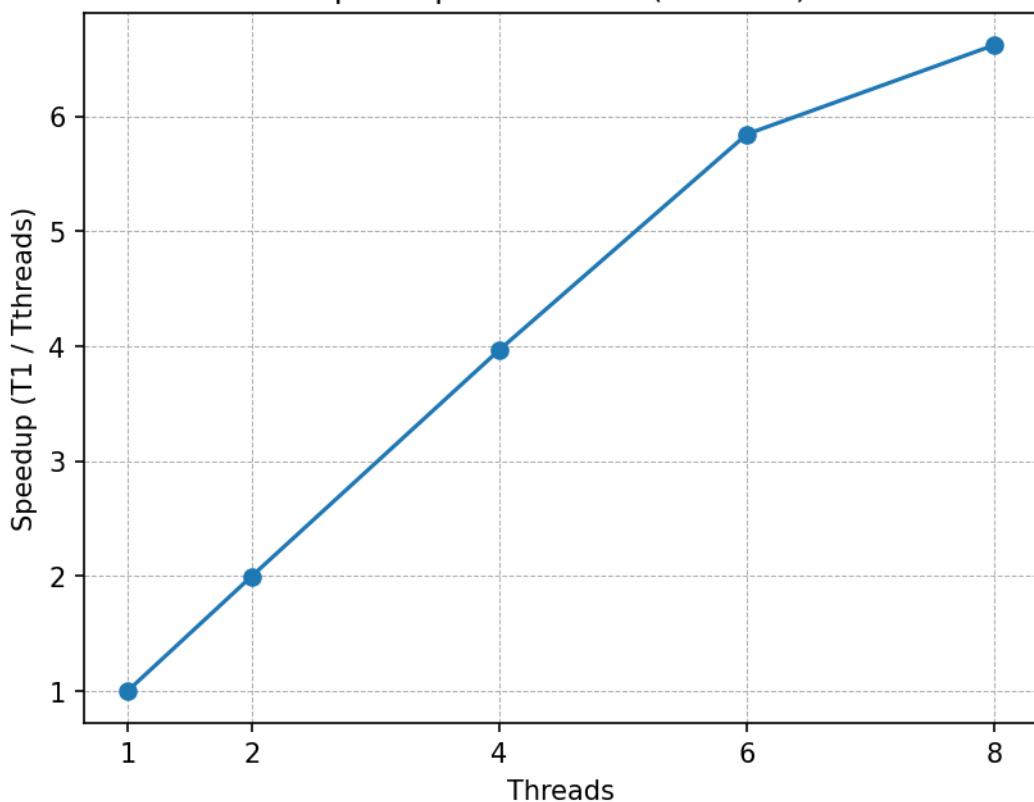
Διαγράμματα Speedup



Speedup vs Threads (N=1024)



Speedup vs Threads (N=4096)



Με βάση τα παραπάνω αποτελέσματα και διαγράμματα παρατηρούμε ότι:

- **Για N=64**, ο χρόνος εκτέλεσης δεν κλιμακώνει πέραν των 2 threads (δηλ. δεν είναι ανάλογος του 1/threads) και από τα 4 threads και έπειτα μειώνεται ελάχιστα με την αύξηση των νημάτων. Το speedup είναι μικρότερο από γραμμικό (κοίλη συνάρτηση) και στα 8 νήματα εμφανίζει μικρή υποχώρηση. Αυτό οφείλεται στο μικρό φορτίο ανά νήμα: το κόστος δημιουργίας και συγχρονισμού των νημάτων αποτελεί σημαντικό ποσοστό του συνολικού χρόνου και άρα δεν έχουμε μεγάλο CPU intensity.
- **Για N=1024**, ο χρόνος μειώνεται σχεδόν γραμμικά με τον αριθμό νημάτων, επιτυγχάνοντας ιδανική κλιμάκωση ($\sim 1/\text{threads}$). Το πρόβλημα είναι αρκετά μεγάλο ώστε να επικρατεί ο υπολογιστικός φόρτος έναντι του overhead δημιουργίας νημάτων, ενώ τα δεδομένα χωρούν πλήρως στην cache, αποφεύγοντας καθυστερήσεις από τη μνήμη. Έτσι, και η επιτάχυνση είναι γραμμική.
- **Για N=4096**, η επιτάχυνση αρχικά είναι σχεδόν γραμμική έως τα 4 νήματα, αλλά στη συνέχεια μειώνεται. Από τα 6 και ειδικά στα 8 νήματα, η βελτίωση της απόδοσης περιορίζεται, καθώς το πρόβλημα γίνεται memory-bound (έχουμε αρχιτεκτονική κοινής μνήμης, αφού τρέχουμε το πρόγραμμα σε 1 node, με πολλά threads): η αυξημένη κίνηση στη μνήμη και το περιορισμένο εύρος ζώνης (bandwidth) επιβραδύνουν την περαιτέρω κλιμάκωση, παρότι ο συνολικός χρόνος συνεχίζει να μειώνεται. Έτσι, η επιτάχυνση αρχικά είναι γραμμική, αλλά στο τέλος γίνεται κούλη.

ΣΥΣΤΗΜΑΤΑ ΠΑΡΑΛΛΗΛΗΣ ΕΠΕΞΕΡΓΑΣΙΑΣ

ΑΝΑΦΟΡΑ 1^{ης} ΑΣΚΗΣΗΣ



Στοιχεία Ομάδας

- Αναγνωριστικό: parlab05
- Μέλος 1^ο: Πέππας Μιχαήλ – Αθανάσιος, Α.Μ: 03121026
- Μέλος 2^ο: Σαουνάτσος Ανδρέας, Α.Μ: 03121197
- Ημερομηνία Παράδοσης Αναφοράς: 19.11.2025

▪ **Ενότητα 2.1 – Παραλληλοποίηση και Βελτιστοποίηση του Αλγορίθμου K-means**

Στόχος της άσκησης είναι η ανάπτυξη δύο παράλληλων εκδόσεων του αλγορίθμου K-means στο προγραμματιστικό μοντέλο του κοινού χώρου διευθύνσεων με τη χρήση του προγραμματιστικού εργαλείου OpenMP.

Αρχικά, μελετήσαμε το υλικό και τα αρχεία του εργαστηρίου (στον φάκελο του kmeans που μας δίνεται), όπως και το αντίστοιχο υλικό των διαλέξεων του μαθήματος. Έτσι, τροποποιήσαμε τα περιφερειακά αρχεία (Makefile, make_on_queue.sh, run_on_queue.sh, file_io.sh) που εξυπηρετούν την ορθή μεταγλώττιση και λειτουργία των κυρίων αρχείων με τις παράλληλες εκδόσεις του αλγορίθμου μας (omp_naive_kmeans.c, omp_reduction_kmeans.c), ως εξής:

1. **Makefile:** Μετονομάσαμε τα αρχεία, ώστε να ανταποκρίνονται στο εργαστηριακό υλικό που μας δόθηκε, κάναμε uncomment τα σχόλια που μεταγλωττίζουν τα αρχεία με τον παράλληλο κώδικα και συμπεριλάβαμε το -fopenmp, ώστε να δηλώσουμε ότι τα αρχεία μας χρησιμοποιούν τη βιβλιοθήκη OpenMP.
2. **make_on_queue.sh:** Αλλάξαμε τη διεύθυνση του φακέλου src σε «/home/parallel/parlab05/a2/kmeans», ώστε να εξυπηρετεί τις ανάγκες της άσκησης, όπως ζητήθηκε. Καθώς η αλλαγή αυτή είναι μικρή, το συγκεκριμένο αρχείο δεν θα συμπεριληφθεί στην αναφορά.
3. **run_on_queue.sh:** Το αρχείο προσαρμόστηκε ώστε να επιτρέπει την εκτέλεση των πειραμάτων με διαφορετικές πολιτικές δέσμευσης νημάτων (affinity), μέσω παραμέτρων στην εντολή qsub. Υποστηρίζονται οι δύο κύριες φάσεις που ζητούνται στην άσκηση:
(α) εκτέλεση χωρίς καμία πολιτική δέσμευσης (noaff) και
(β) εκτέλεση με προκαθορισμένη πολιτική affinity (aff), όπου τα N νήματα του OpenMP δένονται ρητά στα N πρώτα λογικά CPU slots του κόμβου (0, 1, ..., N-1).

Η επιλογή αυτή πάρθηκε κατόπιν συζήτησης με τον διδάσκοντα και καθώς συνιστά την standard επιλογή της βιβλιοθήκης (κατόπιν αναζήτησης στο διαδίκτυο). Μια άλλη επιλογή με την οποία πειραματίστηκαμε ήταν ο διαμοιρασμός των threads στα 4 nodes του μηχανήματος ισάριθμα, ώστε να

αξιοποιήσουμε στο μέγιστο το διαθέσιμο memory bandwidth (η εφαρμογή μας είναι memory bound), παρά να είναι τοποθετημένα κοντά, στο ίδιο node και αυξάνοντας τη συμφόρηση στον δίαυλο μνήμης. Ωστόσο, για λόγους απλότητας και έκτασης της άσκησης, αφήσαμε την πολιτική στο default.

Σημειώνουμε ότι η επιλογή της πολιτικής γίνεται αυτόματα κατά την υποβολή του πειράματος, ενώ τα αποτελέσματα αποθηκεύονται σε ξεχωριστούς φακέλους, οργανωμένους ανά εκτελέσιμο και ανά επιλεγμένη πολιτική affinity, ώστε να διευκολύνεται η σύγκριση των μετρήσεων.

4. **file io.c:** Συμπληρώσαμε το header που ζητούνταν, ως: #include <omp.h>. Καθώς η αλλαγή αυτή είναι μικρή, το συγκεκριμένο αρχείο δεν θα συμπεριληφθεί στην αναφορά.

Τα αρχεία αυτά βρίσκονται στον orion και στον scirouter της ομάδας μας και παρουσιάζονται (αυτά που άλλαξαν σημαντικά, για λόγους πληρότητας) ακολούθως:

a2/kmeans/Makefile

```
1 .KEEP_STATE:
2
3 CC = gcc
4
5 CFLAGS = -Wall -Wextra -Wno-unused -O3 -std=gnu11
6 # Compile OpenMP sources with -fopenmp (+CFLAGS)
7 OMPFLAGS = $(CFLAGS) -fopenmp
8
9 # Link step includes -fopenmp so binaries link libgomp if any object used OpenMP
10 LDFLAGS = -fopenmp
11
12 H_FILES = kmeans.h
13 COMM_SRC = file_io.c util.c
14
15 # Build all variants
16 all: seq_kmeans omp_naive_kmeans omp_reduction_kmeans
17 seq_kmeans: main.o file_io.o util.o seq_kmeans.o
18     $(CC) $(CFLAGS) $^ -o $@ $(LDFLAGS)
19
20 omp_naive_kmeans: main.o file_io.o util.o omp_naive_kmeans.o
21     $(CC) $(CFLAGS) $^ -o $@ $(LDFLAGS)
22
23 omp_reduction_kmeans: main.o file_io.o util.o omp_reduction_kmeans.o
24     $(CC) $(CFLAGS) $^ -o $@ $(LDFLAGS)
25
26 main.o: main.c $(H_FILES)
27     $(CC) $(CFLAGS) -c $< -o $@
28
29 seq_kmeans.o: seq_kmeans.c $(COMM_SRC) $(H_FILES)
30     $(CC) $(CFLAGS) -c $< -o $@
31
32 # OpenMP objects use OMPFLAGS so pragmas are honored
33 omp_naive_kmeans.o: omp_naive_kmeans.c $(COMM_SRC) $(H_FILES)
34     $(CC) $(OMPFLAGS) -c $< -o $@
35
36 omp_reduction_kmeans.o: omp_reduction_kmeans.c $(COMM_SRC) $(H_FILES)
37     $(CC) $(OMPFLAGS) -c $< -o $@
38
39 file_io.o: file_io.c
40     $(CC) $(CFLAGS) -c $< -o $@
41
42 util.o: util.c
43     $(CC) $(CFLAGS) -c $< -o $@
44
45 clean:
46     rm -rf *.o seq_kmeans omp_naive_kmeans omp_reduction_kmeans
47
48
```

a2/kmeans/make_on_queue.sh

```
1 #!/bin/bash
2
3 ## How to run (example)
4 ## qsub -q parlab make_on_queue.sh
5
6 ## Give the Job a descriptive name
7 #PBS -N make_kmeans
8
9 ## Output and error files
10 #PBS -o make_kmeans.out
11 #PBS -e make_kmeans.err
12
13 ## How many machines should we get?
14 #PBS -l nodes=1:ppn=1
15
16 ## How long should the job run for?
17 #PBS -l walltime=00:10:00
18
19 ## Start
20 ## Run make in the src folder (modify properly)
21
22 cd /home/parallel/parlab05/a2/kmeans
23 make
24
25
```

```

1 #!/bin/bash
2
3 #PBS -N run_kmeans
4 #PBS -o run_kmeans.out
5 #PBS -e run_kmeans.err
6 #PBS -l nodes=1:ppn=64
7 #PBS -l walltime=01:00:00
8
9 # Submission details
10 # usage--no affinity (default): qsub -q serial -l nodes=sandman:ppn=64 -v
# THREADS=32,BIN=omp_naive_kmeans run_on_queue.sh
11 # with default affinity (bind 0..T-1): qsub -q serial -l nodes=sandman:ppn=64 -v
# THREADS=32,AFFINITY=default,BIN=omp_naive_kmeans run_on_queue.sh
12 # BIN=seq_kmeans|omp_naive_kmeans|omp_reduction_kmeans
13 # optional VARS: SIZE=256,COORDS=16,CLUSTERS=32,LOOPS=10
14
15 set -euo pipefail
16 cd /home/parallel/parlab05/a2/kmeans || exit 1
17
18 : "${BIN:=seq_kmeans}"
19 : "${SIZE:=256}"
20 : "${COORDS:=16}"
21 : "${CLUSTERS:=32}"
22 : "${LOOPS:=10}"
23 : "${THREADS:?Set THREADS via qsub -v THREADS=...}"
24 : "${AFFINITY:=none}"
25
26 export OMP_NUM_THREADS="${THREADS}"
27 AFF_LABEL="noaff"
28 if [[ "${AFFINITY,,}" == "default" ]]; then
29   CPUSERT=$(seq 0 $((THREADS-1)) | paste -sd' ' -)
30   export GOMP_CPU_AFFINITY="${CPUSERT}"
31   AFF_LABEL="aff"
32 else
33   unset GOMP_CPU_AFFINITY || true
34 fi
35
36 BENCH_ROOT="/home/parallel/parlab05/a2/kmeans/benchmarks"
37 case "${BIN}" in
38   *seq*) BENCH_SUBDIR_BASE="serial" ;;
39   *naive*) BENCH_SUBDIR_BASE="naive" ;;
40   *reduction*|*copied*) BENCH_SUBDIR_BASE="reduction" ;;
41   *) BENCH_SUBDIR_BASE="other" ;;
42 esac
43 BENCH_SUBDIR="${BENCH_SUBDIR_BASE}/${AFF_LABEL}"
44
45 RUN_TAG="S${SIZE}_N${COORDS}_C${CLUSTERS}_L${LOOPS}_T${THREADS}"
46 RESULT_DIR="${BENCH_ROOT}/${BENCH_SUBDIR}/${RUN_TAG}"
47 mkdir -p "${RESULT_DIR}"
48 {
49   echo "[run_on_queue] BIN=${BIN}"
50   echo "[run_on_queue] OMP_NUM_THREADS=${OMP_NUM_THREADS}"
51   echo "[run_on_queue] GOMP_CPU_AFFINITY=${GOMP_CPU_AFFINITY:-<unset>}"

```

```
53 echo "[run_on_queue] AFF_LABEL=${AFF_LABEL}"
54 echo "[run_on_queue] Params: -s ${SIZE} -n ${COORDS} -c ${CLUSTERS} -l ${LOOPS}"
55 echo "[run_on_queue] Result dir: ${RESULT_DIR}"
56 } | tee "${RESULT_DIR}/meta.txt"
57
58 "./${BIN}" -s "${SIZE}" -n "${COORDS}" -c "${CLUSTERS}" -l "${LOOPS}" \
59 | tee "${RESULT_DIR}/output.txt"
```

a2/kmeans/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
26      *         them.
27      *         reminder : First-touch data placement policy
28      */
29
30     for (i=0; i<numObjs; i++)
31     {
32         unsigned int seed = i;
33         for (j=0; j<numCoords; j++)
34         {
35             objects[i*numCoords + j] = (rand_r(&seed) / ((double) RAND_MAX)) * val_range;
36             if (_debug && i == 0)
37                 printf("object[i=%ld][j=%ld]=%f\n", i, j, objects[i*numCoords + j]);
38         }
39     }
40
41     return objects;
42 }
```

2.1.1 – Shared Clusters

Στην πρώτη παράλληλη υλοποίηση του αλγορίθμου K-means υιοθετήσαμε το μοντέλο shared clusters, χωρίς καμία βελτιστοποίηση στη συλλογή των μερικών αποτελεσμάτων. Πρόκειται για μια «αφελή» (naive) προσέγγιση, όπου όλα τα νήματα ενημερώνουν απευθείας τους κοινόχρηστους πίνακες newClusters[] και newClusterSize[]. Η πρόσβαση σε αυτά τα κοινόχρηστα δεδομένα απαιτεί συγχρονισμό, προκειμένου να αποφευχθούν πιθανά race conditions, ο οποίος στη συγκεκριμένη εκδοχή υλοποιείται αποκλειστικά με #pragma omp atomic για κάθε ενημέρωση-πρόσβαση. Σημειώνουμε ότι θα μπορούσε να έχει χρησιμοποιηθεί και #pragma omp critical, ωστόσο η εντολή αυτή είναι πιο αργή και υποβέλτιστη σε απλές προσβάσεις-πράξεις, όπως αυτή.

Η προσέγγιση αυτή επιτρέπει την εύκολη και άμεση παραλληλοποίηση του βρόχο, αλλά εισάγει σημαντικό κόστος εξαιτίας των συχνών πράξεων που γίνονται με atomic και της υψηλής πιθανότητας contention, ειδικά για μικρό αριθμό συντεταγμένων (numCoords) ή για μεγάλο αριθμό νημάτων. Δηλαδή, η ευκολία υλοποίησης έρχεται με υψηλό κόστος συγχρονισμού. Έτσι, η 2.1.1 συμβάλλει κυρίως ως σημείο αναφοράς για τη σύγκριση με πιο αποδοτικές τεχνικές συγχώνευσης που αναπτύσσονται στην επόμενη ενότητα (2.1.2 – copied clusters and reduce).

Όσον αφορά τις λεπτομέρειες της υλοποίησής μας, επισημαίνουμε τα εξής:

- Ο παράλληλος βρόχος (#pragma omp parallel for) αναθέτει σε κάθε νήμα ένα υποσύνολο αντικειμένων για επεξεργασία.
- Η αύξηση της μεταβλητής delta προστατεύεται με atomic operation, καθώς εδώ δεν χρησιμοποιείται reduction. Ωστόσο, ακόμα και χωρίς atomic (δεν ζητούταν με σχόλιο) η ορθότητα του προγράμματος δεν θα άλλαζε, αφού ναι μεν η delta θα είχε λάθος τιμή, αλλά σίγουρα $\text{delta} \geq 1 > \text{threshold}$, άρα ο έλεγχος θα ήταν πάντα σωστός και αληθής και θα μπορούσαμε να αποφύγουμε αυτό το atomic (η σχετική συζήτηση έγινε και με τον διδάσκοντα και επιλέξαμε να το βάλουμε).
- Η ενημέρωση των δομών newClusterSize[] και newClusters[] γίνεται επίσης με atomic σε κάθε πρόσβαση, γεγονός που καθιστά την υλοποίηση μεν εύκολη, αλλά δε καθόλου αποδοτική, αφού έχουμε μεγάλο κόστος συγχρονισμού, σε πολλαπλά σημεία.

Ο ολοκληρωμένος κώδικας της υλοποίησης (`omp_naive_kmeans.c`) βρίσκεται στον `scirouter` και στον `orion` της ομάδας μας, αλλά παρατίθεται και στην παρούσα αναφορά για λόγους πληρότητας.

```
a2/kmeans/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 #include <omp.h>
8
9 // square of Euclid distance between two multi-dimensional points
10 inline static double euclid_dist_2(int numdims, /* no. dimensions */
11                                     double *coord1, /* [numdims] */
12                                     double *coord2) /* [numdims] */
13 {
14     int i;
15     double ans = 0.0;
16
17     for (i = 0; i < numdims; i++)
18         ans += (coord1[i] - coord2[i]) * (coord1[i] - coord2[i]);
19
20     return ans;
21 }
22
23 inline static int find_nearest_cluster(int numClusters, /* no. clusters */
24                                         int numCoords, /* no. coordinates */
25                                         double *object, /* [numCoords] */
26                                         double *clusters) /* [numClusters][numCoords] */
27 {
28     int index, i;
29     double dist, min_dist;
30
31     // find the cluster id that has min distance to object
32     index = 0;
33     min_dist = euclid_dist_2(numCoords, object, clusters);
34
35     for (i = 1; i < numClusters; i++)
36     {
37         dist = euclid_dist_2(numCoords, object, &clusters[i * numCoords]);
38         // no need square root
39         if (dist < min_dist)
40             { // find the min and its array index
41                 min_dist = dist;
42                 index = i;
43             }
44     }
45     return index;
46 }
47
48 void kmeans(double *objects, /* in: [numObjs][numCoords] */
49             int numCoords, /* no. coordinates */
50             int numObjs, /* no. objects */
51             int numClusters, /* no. clusters */
```

```

52     double threshold,      /* minimum fraction of objects that change membership */
53     long loop_threshold, /* maximum number of iterations */
54     int *membership,      /* out: [numObjs] */
55     double *clusters)     /* out: [numClusters][numCoords] */
56 {
57     int i, j;
58     int index, loop = 0;
59     double timing = 0;
60
61     double delta;          // fraction of objects whose clusters change in each loop
62     int *newClusterSize; // [numClusters]: no. objects assigned in each new cluster
63     double *newClusters; // [numClusters][numCoords]
64     int nthreads;         // no. threads
65
66     nthreads = omp_get_max_threads();
67     printf("OpenMP Kmeans - Naive\t(number of threads: %d)\n", nthreads);
68
69     // initialize membership
70     for (i = 0; i < numObjs; i++)
71         membership[i] = -1;
72
73     // initialize newClusterSize and newClusters to all 0
74     newClusterSize = (typeof(newClusterSize))calloc(numClusters, sizeof(*newClusterSize));
75     newClusters = (typeof(newClusters))calloc(numClusters * numCoords,
76         sizeof(*newClusters));
76
77     timing = wtime();
78
79     do
80     {
81         // before each loop, set cluster data to 0
82         for (i = 0; i < numClusters; i++)
83         {
84             for (j = 0; j < numCoords; j++)
85                 newClusters[i * numCoords + j] = 0.0;
86             newClusterSize[i] = 0;
87         }
88
89         delta = 0.0;
90
91     /*
92     * TODO: Detect parallelizable region and use appropriate OpenMP pragmas
93     */
94     #pragma omp parallel for private(index, j)
95         for (i = 0; i < numObjs; i++)
96         {
97             // find the array index of nearest cluster center
98             index = find_nearest_cluster(numClusters, numCoords, &objects[i * numCoords],
99             clusters);
100
101             // if membership changes, increase delta by 1
102             if (membership[i] != index)
103             {

```

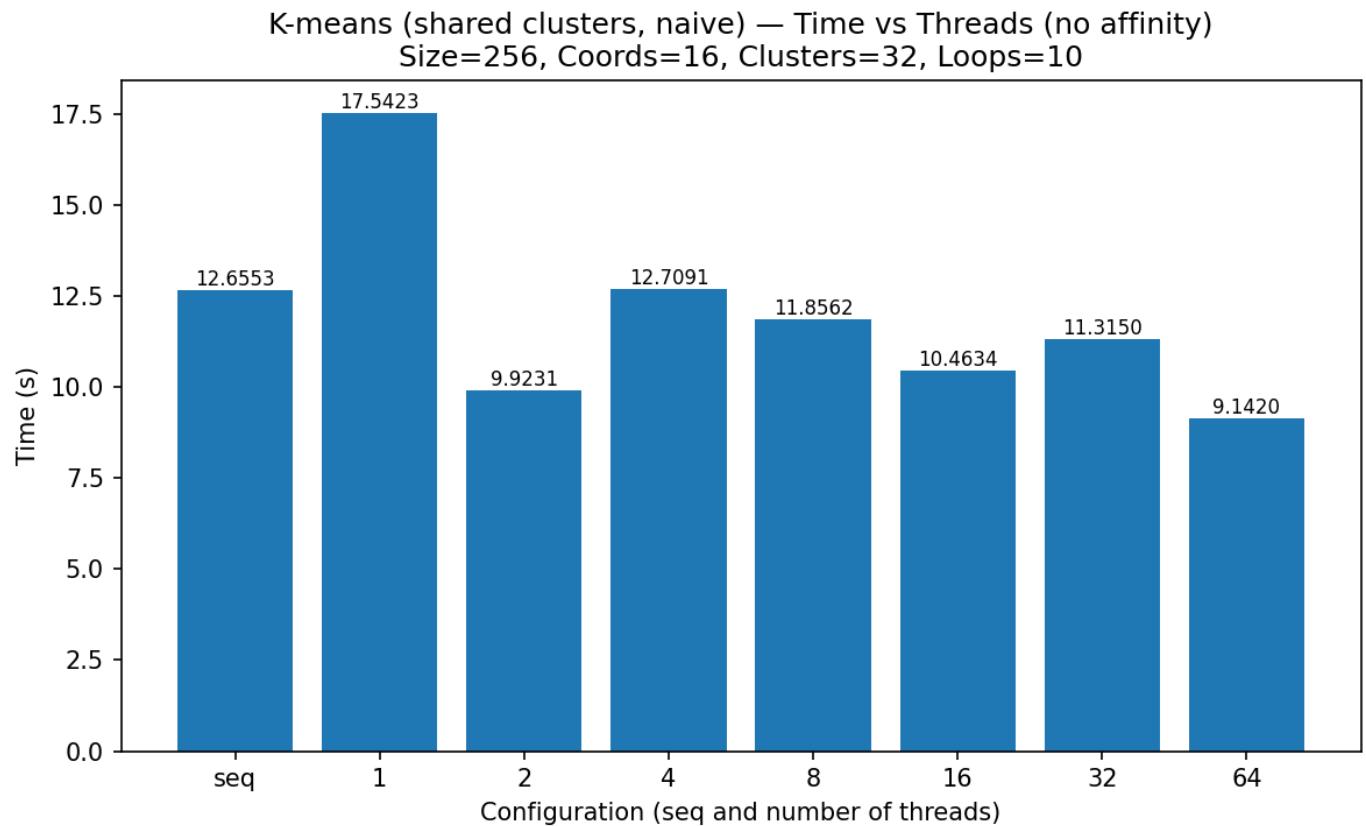
```
104 #pragma omp atomic // protect update on shared "delta" variable
105         delta += 1.0;
106     }
107
108     // assign the membership to object i
109     membership[i] = index;
110
111 // update new cluster centers : sum of objects located within
112 /*
113 * TODO: protect update on shared "newClusterSize" array
114 */
115 #pragma omp atomic
116         newClusterSize[index]++;
117         for (j = 0; j < numCoords; j++)
118     /*
119     * TODO: protect update on shared "newClusters" array
120     */
121 #pragma omp atomic
122         newClusters[index * numCoords + j] += objects[i * numCoords + j];
123     }
124
125     // average the sum and replace old cluster centers with newClusters
126     for (i = 0; i < numClusters; i++)
127     {
128         if (newClusterSize[i] > 0)
129         {
130             for (j = 0; j < numCoords; j++)
131             {
132                 clusters[i * numCoords + j] = newClusters[i * numCoords + j] /
133             newClusterSize[i];
134             }
135         }
136     }
137
138     // Get fraction of objects whose membership changed during this loop. This is used
139     // as a convergence criterion.
140     delta /= numObjs;
141
142     loop++;
143     printf("\r\tcompleted loop %d", loop);
144     fflush(stdout);
145 } while (delta > threshold && loop < loop_threshold);
146     timing = wtime() - timing;
147     printf("\n          nloops = %3d    (total = %7.4fs)  (per loop = %7.4fs)\n", loop,
148 timing, timing / loop);
149
150     free(newClusters);
151     free(newClusterSize);
152 }
```

1. No Affinity

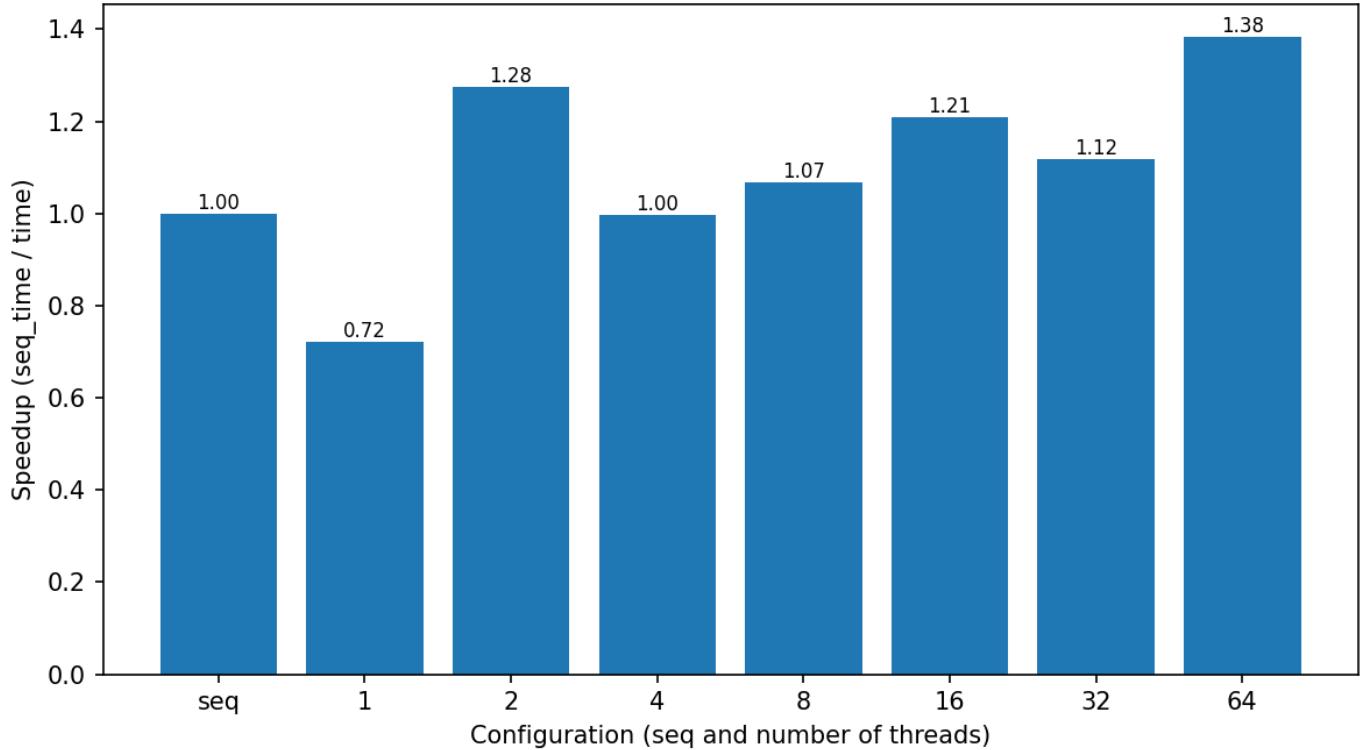
Το παραπάνω παράλληλο πρόγραμμα (omp_naive_kmeans.c) έτρεξε για τις παραμέτρους: {Size, Coords, Clusters, Loops} = {256, 16, 32, 10}, threads = {1, 2, 4, 8, 16, 32, 64} και χωρίς affinity. Τα αποτελέσματα που προέκυψαν παρουσιάζονται στον παρακάτω πίνακα:

THREADS	TIME
seq	12.65
1	17.54
2	9.92
4	12.70
8	11.85
16	10.46
32	11.31
64	9.14

Τα ζητούμενα διαγράμματα (πάντα με βάση τον χρόνο του σειριακού προγράμματος, όπως αναφέρεται) φαίνονται ακολούθως:



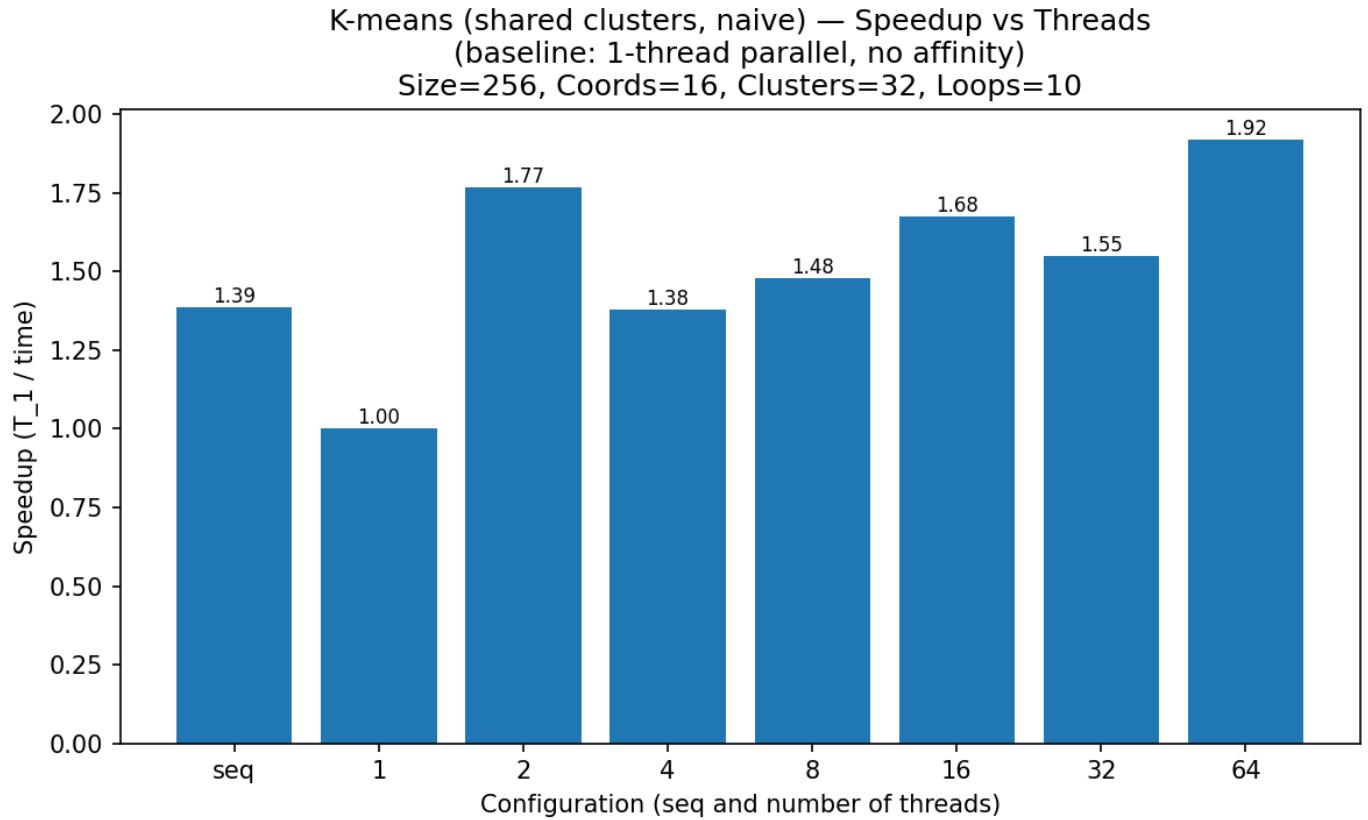
K-means (shared clusters, naive) — Speedup vs Threads (no affinity)
 Size=256, Coords=16, Clusters=32, Loops=10



Σύμφωνα με τα παραπάνω, συμπεραίνουμε ότι υλοποίηση χωρίς affinity δεν κλιμακώνει ικανοποιητικά. Ο σειριακός χρόνος είναι περίπου 12.7s, ενώ η παράλληλη έκδοση με 1 νήμα είναι σαφώς χειρότερη ($\approx 17.5s$, speedup ≈ 0.72), κάπι που αποδίδεται στο κόστος δημιουργίας του παράλληλου προγράμματος-κώδικα (δημιουργία/συγχρονισμός νημάτων, επιπλέον κώδικας OpenMP), όπως αναφέρεται και στις διαφάνειες. Για περισσότερα νήματα, τα διαγράμματα χρόνου και speedup δείχνουν μικρές μόνο βελτιώσεις: γύρω στο 1.28 \times στα 2 νήματα (αν και παρατηρείται κλιμάκωση σχεδόν 2x από το 1 νήμα του παράλληλου προγράμματος), τιμές κοντά στο 1.0–1.2 \times για 4–32 νήματα και μέγιστο περίπου 1.38 \times στα 64 νήματα, πολύ μακριά από την ιδανική γραμμική κλιμάκωση.

Η συμπεριφορά αυτή ταιριάζει ακριβώς με τη θεωρία για synchronization bottlenecks: στη naive shared έκδοση όλες οι ενημερώσεις των πινάκων newClusterSize[] και newClusters[] γίνονται με #pragma omp atomic, άρα πολλές προσπελάσεις σε λίγες κοινόχρηστες μεταβλητές σειριοποιούνται και δημιουργούν έντονο contention, όπως στα παραδείγματα των διαφανειών για fine-grained synchronization. Έτσι, μεγάλο μέρος του χρόνου δαπανάται σε συγχρονισμό αντί για πραγματικό υπολογισμό, ενώ και το σειριακό τμήμα του αλγορίθμου (σύμφωνα με τον νόμο του Amdahl) βάζει χαμηλό άνω φράγμα στο speedup. Τα παραπάνω αποτελέσματα και τα διαγράμματα, λοιπόν, επιβεβαιώνουν ότι η λύση αυτή είναι μεν ορθή και εύκολη στην υλοποίηση, αλλά καθόλου αποδοτική.

Από απλή περιέργεια και χωρίς να ζητείται, καταστρώσαμε και ένα διάγραμμα speedup με βάση όχι τώρα το σειριακό πρόγραμμα, αλλά το παράλληλο με 1 thread. Τα αποτελέσματα φαίνονται ακολούθως:



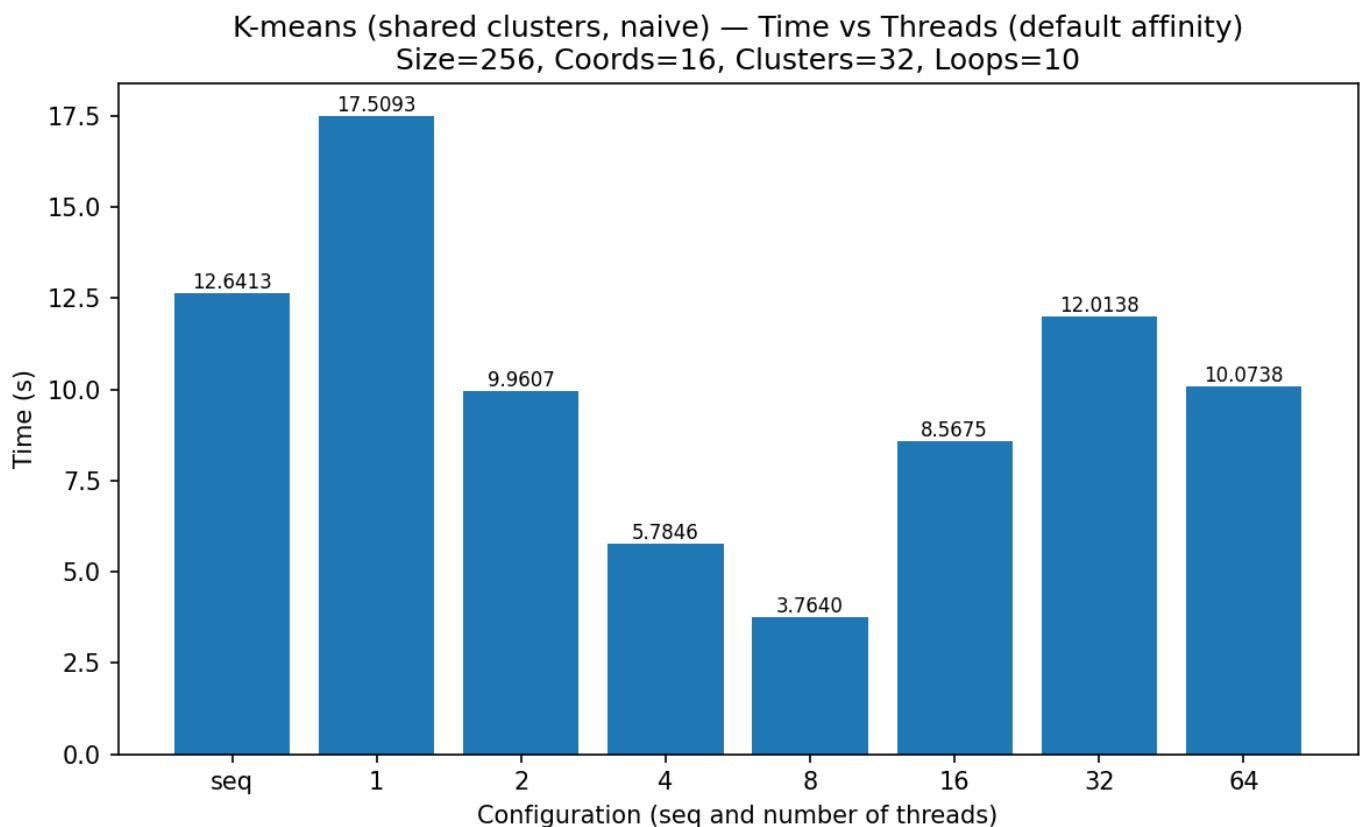
Από το παραπάνω διάγραμμα παρατηρούμε ότι, αν θεωρήσουμε ως βάση το παράλληλο πρόγραμμα με 1 νήμα, όλες οι εκτελέσεις με περισσότερα νήματα εμφανίζουν πλέον speedup μεγαλύτερο της μονάδας. Ενδεικτικά, για 2 νήματα το κέρδος είναι περίπου $1.8\times$ σε σχέση με το 1-thread (σχεδόν γραμμικό και το μόνο ικανοποιητικό), ενώ για 64 νήματα φτάνει σχεδόν το $2\times$ (πολύ κακή κλιμάκωση γενικότερα). Αυτό επιβεβαιώνει ότι ένα σημαντικό τμήμα του κόστους στην περίπτωση του 1 νήματος οφείλεται αποκλειστικά στο parallel overhead του OpenMP (δημιουργία ομάδας νημάτων, συγχρονισμοί κ.λπ.), το οποίο «απλώνεται» σε περισσότερα νήματα και αντισταθμίζεται μερικώς όταν αυξάνουμε τον βαθμό παραλληλίας και ότι η εφαρμογή είναι memory bound. Παρ' όλα αυτά, η κλιμάκωση παραμένει εξαιρετικά κακή και σε απόλυτους όρους ως προς το σειριακό πρόγραμμα και τα κέρδη παραμένουν μικρά, γεγονός που δείχνει ότι το synchronization bottleneck της naive shared υλοποίησης δεν επιτρέπει ουσιαστική κλιμάκωση.

2. Default Affinity

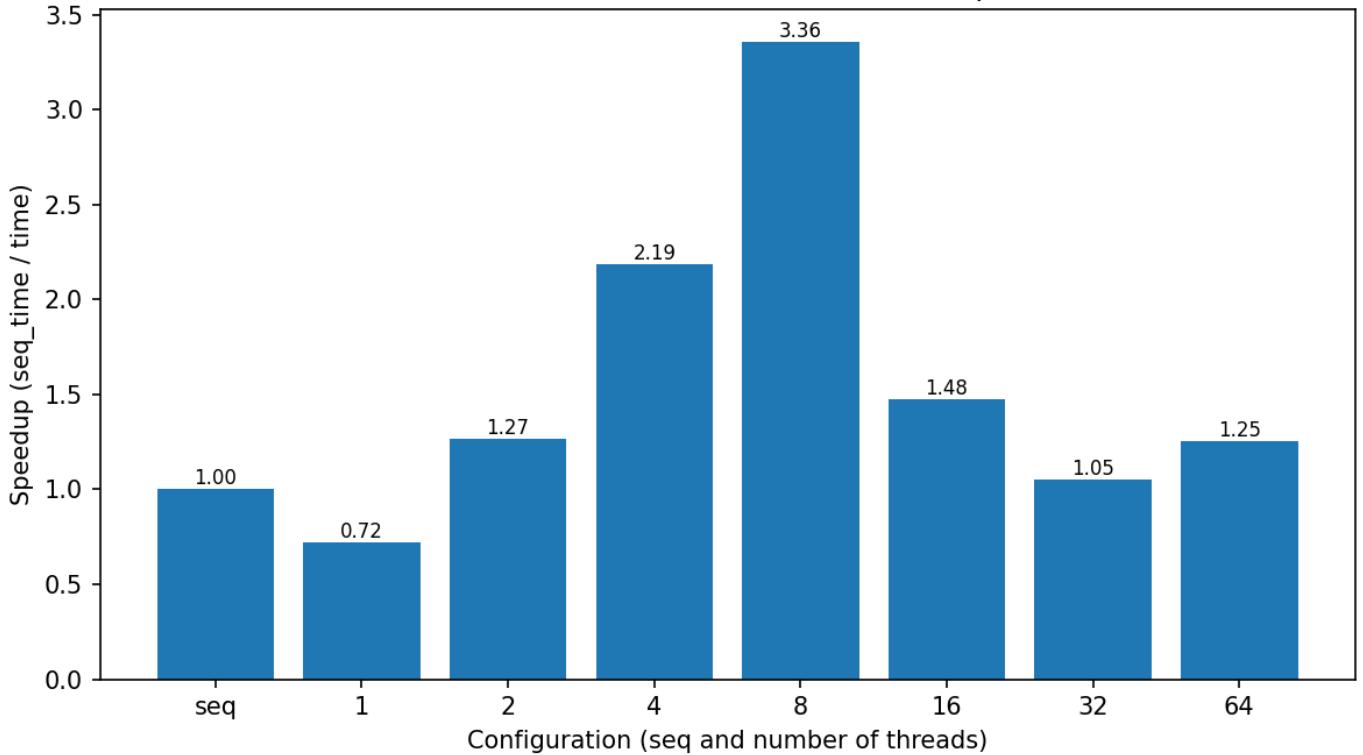
Το παραπάνω παράλληλο πρόγραμμα (omp_naive_kmeans.c) έτρεξε για τις παραμέτρους: {Size, Coords, Clusters, Loops} = {256, 16, 32, 10}, threads = {1, 2, 4, 8, 16, 32, 64, αλλά αυτή τη φορά με affinity. Τα αποτελέσματα που προέκυψαν παρουσιάζονται στον παρακάτω πίνακα:

THREADS	TIME
seq	12.64
1	17.50
2	9.96
4	5.78
8	3.76
16	8.56
32	12.01
64	10.07

Τα ζητούμενα διαγράμματα (πάντα με βάση τον χρόνο του σειριακού προγράμματος, όπως αναφέρεται) φαίνονται ακολούθως:



K-means (shared clusters, naive) — Speedup vs Threads (default affinity)
 Size=256, Coords=16, Clusters=32, Loops=10

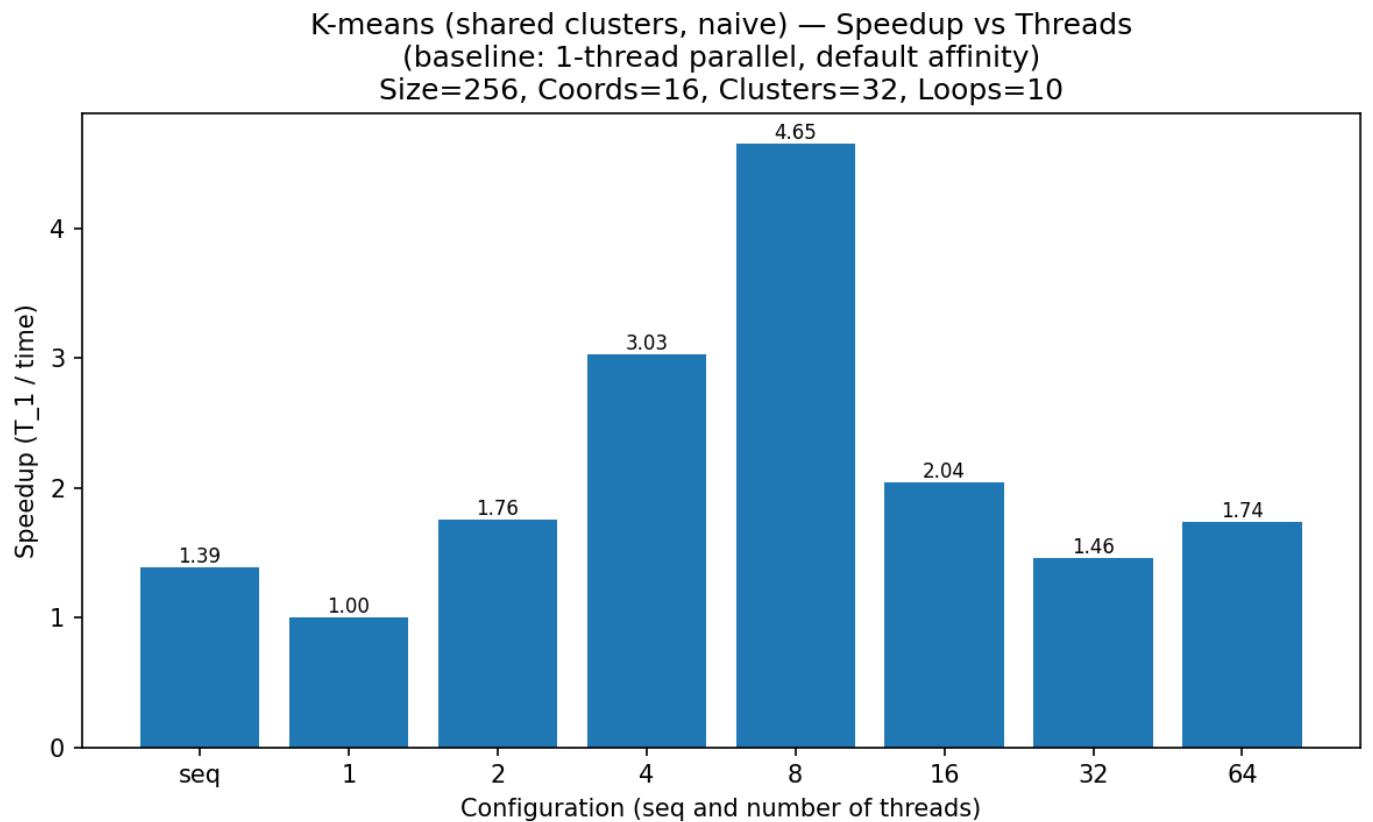


Παρατηρούμε ότι με ενεργοποιημένο το default affinity η συμπεριφορά της naive shared υλοποίησης βελτιώνεται σημαντικά σε σχέση με την περίπτωση χωρίς affinity. Ο χρόνος εκτέλεσης μειώνεται πολύ καλά οριακά γραμμικά ως προς το παράλληλο πρόγραμμα με 1 νήμα) μέχρι τα 8 νήματα (από ~17.5s στο 1 νήμα σε ~3.8s στα 8 νήματα), με αντίστοιχο speedup ~3.4x σε σχέση με το σειριακό πρόγραμμα, γεγονός που δείχνει ότι η δέσμευση των νημάτων σε σταθερούς πυρήνες αξιοποιεί καλύτερα την τοπικότητα cache και μνήμης μέσα στο ίδιο NUMA node, όπως επισημαίνεται και στις διαφάνειες για affinity και locality. Ωστόσο, για 16, 32 και 64 νήματα η επίδοση υποβαθμίζεται (ο χρόνος αυξάνεται ξανά και το speedup πέφτει κοντά στη μονάδα), κάτι που είναι αναμενόμενο για έναν κατά βάση memory-bound αλγόριθμο με έντονο synchronization μέσω atomic πράξεων.

Πιο συγκεκριμένα, όταν ξεπερνάμε τα νήματα που «χωράει άνετα» ένα socket/NUMA node, η εκτέλεση αρχίζει να μοιράζεται σε πολλαπλούς κόμβους μνήμης και αυξάνονται οι απομακρυσμένες προσπελάσεις (remote NUMA accesses) και η συμφόρηση στον δίαυλο μνήμης. Ταυτόχρονα, οι atomic ενημερώσεις στους κοινόχρηστους πίνακες newClusters[] και newClusterSize[] δημιουργούν έντονο contention στις ίδιες cache lines, με αποτέλεσμα η θεωρητική παραλληλία να χάνεται από τον συγχρονισμό, όπως ακριβώς περιγράφεται στις διαφάνειες για synchronization bottlenecks και NUMA αρχιτεκτονικές. Συνολικά, το affinity εκμεταλλεύεται καλά τη δομή του κόμβου μέχρι τα 8 νήματα, αλλά τα αρχιτεκτονικά

και αλγορίθμικά όρια της naive shared λύσης δεν επιτρέπουν ουσιαστική κλιμάκωση πέρα από αυτό το σημείο.

Από απλή περιέργεια και χωρίς να ζητείται, καταστρώσαμε και ένα διάγραμμα speedup με βάση όχι τώρα το σειριακό πρόγραμμα, αλλά το παράλληλο με 1 thread. Τα αποτελέσματα φαίνονται ακολούθως:



Από το τελευταίο διάγραμμα, όπου ως βάση λαμβάνουμε πλέον την παράλληλη εκτέλεση με 1 νήμα, βλέπουμε ότι τα speedups για 2, 4 και 8 νήματα είναι ιδιαίτερα υψηλά (της τάξης του 1.7–1.8x, ~3x και ~4.5x αντίστοιχα). Αυτό δείχνει ότι το σημαντικό parallel overhead της OpenMP (δημιουργία και οργάνωση της ομάδας νημάτων, συγχρονισμοί, barriers) κατανέμεται αποτελεσματικά σε λίγα νήματα όταν αυτά «μένουν» σε κοντινούς πυρήνες του ίδιου NUMA node, με αποτέλεσμα η αύξηση του βαθμού παραλληλίας από 1 σε 2–8 threads να αποδίδει καθαρό κέρδος εντός της ίδιας αρχιτεκτονικής (κάπως κοντά σε γραμμικά, ειδικά αρχικά).

Παρ' όλα αυτά, όταν συνεχίζουμε πέραν από τα 8 νήματα, τα speedups ως προς το 1-thread παράλληλο πρόγραμμα μειώνονται αισθητά, γεγονός που υποδηλώνει ότι έχουμε φτάσει πρακτικά το όριο των πόρων (πυρήνων, cache, memory bandwidth) ενός sockets και αρχίζουμε να «πατάμε» σε δεύτερο NUMA

node ή/και να ενεργοποιούμε hardware multithreading στους ίδιους φυσικούς πυρήνες. Σε έναν αλγόριθμο όπως o K-means, που είναι memory-bound και επιβαρυμένος με atomic operations και συγχρονισμό σε κοινόχρηστες δομές, η περαιτέρω αύξηση των νημάτων δεν μπορεί να εκμεταλλευτεί αποτελεσματικά τις διαθέσιμες memory lanes και οδηγεί σε κορεσμό και σε περισσότερη εμπλοκή μεταξύ των νημάτων. Έτσι, το affinity βελτιώνει σημαντικά την απόδοση μέχρι τα 8 threads, αλλά τα εγγενή NUMA και synchronization bottlenecks της naive shared υλοποίησης εξακολουθούν να περιορίζουν την κλιμάκωση σε μεγαλύτερο αριθμό νημάτων.

2.1.1 – Copied Clusters and Reduce

Στη δεύτερη παράλληλη υλοποίηση του αλγορίθμου K-means υιοθετούμε και πάλι το μοντέλο shared clusters, αλλά αυτή τη φορά με τεχνική copied clusters και reduction για τη συλλογή των μερικών αποτελεσμάτων. Αντί όλα τα νήματα να ενημερώνουν απευθείας τους κοινόχρηστους πίνακες newClusters[] και newClusterSize[], κάθε νήμα διατηρεί δικά του, τοπικά αντίγραφα (local_newClusters[tid], local_newClusterSize[tid]) τα οποία ενημερώνει ελεύθερα, χωρίς atomic operations ή άλλον συγχρονισμό. Στο τέλος του παράλληλου βρόχου, τα τοπικά αυτά αντίγραφα συγχωνεύονται σε έναν κοινό πίνακα μέσω μιας φάσης reduction, η οποία εκτελείται από ένα νήμα (ή σε ένα μικρό, καλά οριοθετημένο σειριακό τμήμα κώδικα).

Η προσέγγιση αυτή αυξάνει λίγο τη χρήση μνήμης και προσθέτει ένα επιπλέον βήμα συγχώνευσης, αλλά μειώνει δραστικά το κόστος συγχρονισμού σε σχέση με τη naïve εκδοχή, καθώς αποφεύγονται οι χιλιάδες ατομικές ενημερώσεις πάνω στις ίδιες cache lines. Έτσι, η 2.1.2 στοχεύει σε πολύ καλύτερη κλιμάκωση με τον αριθμό νημάτων, ειδικά σε NUMA αρχιτεκτονικές, όπου η μείωση του contention στη μνήμη παίζει καθοριστικό ρόλο στην επίδοση.

Όσον αφορά τις λεπτομέρειες της υλοποίησής μας, επισημαίνουμε τα εξής:

- Ο παράλληλος βρόχος (#pragma omp parallel) αναθέτει σε κάθε νήμα ένα υποσύνολο αντικειμένων, όπως και πριν, αλλά οι ενημερώσεις των clusters γίνονται αποκλειστικά στα τοπικά arrays local_newClusterSize[tid] και local_newClusters[tid], χωρίς χρήση atomic.
- Η μεταβλητή delta, που μετράει τις αλλαγές στα memberships, υπολογίζεται πλέον με κατάλληλο reduction μέσα στο parallel, ώστε να αποφεύγονται επιπλέον ατομικές προσπελάσεις και να διατηρείται η ορθότητα του κριτηρίου σύγκλισης.
- Στο τέλος του βρόχου, μια #pragma omp single περιοχή εκτελεί τη φάση reduction, αθροίζοντας τα τοπικά αντίγραφα όλων των νημάτων στους κοινόχρηστους πίνακες newClusterSize[] και newClusters[]. Με αυτόν τον τρόπο συγκεντρώνονται τα μερικά αποτελέσματα με ελάχιστο συγχρονισμό, σε ένα καλά ελεγχόμενο σημείο του προγράμματος.

Ο ολοκληρωμένος κώδικας της υλοποίησης (omp_reduction_kmeans.c) βρίσκεται στον scirouter και στον orion της ομάδας μας, αλλά παρατίθεται και στην παρούσα αναφορά για λόγους πληρότητας.

a2/kmeans/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 #include <omp.h>
8
9 // square of Euclid distance between two multi-dimensional points
10 inline static double euclid_dist_2(int numdims, /* no. dimensions */
11                                 double *coord1, /* [numdims] */
12                                 double *coord2) /* [numdims] */
13 {
14     int i;
15     double ans = 0.0;
16
17     for (i = 0; i < numdims; i++)
18         ans += (coord1[i] - coord2[i]) * (coord1[i] - coord2[i]);
19
20     return ans;
21 }
22
23 inline static int find_nearest_cluster(int numClusters, /* no. clusters */
24                                         int numCoords, /* no. coordinates */
25                                         double *object, /* [numCoords] */
26                                         double *clusters) /* [numClusters][numCoords] */
27 {
28     int index, i;
29     double dist, min_dist;
30
31     // find the cluster id that has min distance to object
32     index = 0;
33     min_dist = euclid_dist_2(numCoords, object, clusters);
34
35     for (i = 1; i < numClusters; i++)
36     {
37         dist = euclid_dist_2(numCoords, object, &clusters[i * numCoords]);
38         // no need square root
39         if (dist < min_dist)
40             { // find the min and its array index
41                 min_dist = dist;
42                 index = i;
43             }
44     }
45     return index;
46 }
47
48 void kmeans(double *objects, /* in: [numObjs][numCoords] */
49             int numCoords, /* no. coordinates */
50             int numObjs, /* no. objects */
51             int numClusters, /* no. clusters */

```

```

52     double threshold,      /* minimum fraction of objects that change membership */
53     long loop_threshold, /* maximum number of iterations */
54     int *membership,      /* out: [numObjs] */
55     double *clusters)    /* out: [numClusters][numCoords] */
56 {
57     int i, j, k;
58     int index, loop = 0;
59     double timing = 0;
60
61     double delta;          // fraction of objects whose clusters change in each loop
62     int *newClusterSize; // [numClusters]: no. objects assigned in each new cluster
63     double *newClusters; // [numClusters][numCoords]
64     int nthreads;         // no. threads
65
66     nthreads = omp_get_max_threads();
67     printf("OpenMP Kmeans - Reduction\t(number of threads: %d)\n", nthreads);
68
69     // initialize membership
70     for (i = 0; i < numObjs; i++)
71         membership[i] = -1;
72
73     // initialize newClusterSize and newClusters to all 0
74     newClusterSize = (typeof(newClusterSize))calloc(numClusters, sizeof(*newClusterSize));
75     newClusters = (typeof(newClusters))calloc(numClusters * numCoords,
76         sizeof(*newClusters));
76
77     // Each thread calculates new centers using a private space. After that, thread 0 does
78     // an array reduction on them.
78     int *local_newClusterSize[nthreads]; // [nthreads][numClusters]
79     double *local_newClusters[nthreads]; // [nthreads][numClusters][numCoords]
80
81     /*
82      * Hint for false-sharing
83      * This is noticed when numCoords is low (and neighboring local_newClusters exist
84      * close to each other).
85      * Allocate local cluster data with a "first-touch" policy.
86      */
86     // Initialize local (per-thread) arrays (and later collect result on global arrays)
87     for (k = 0; k < nthreads; k++)
88     {
89         local_newClusterSize[k] = (typeof(*local_newClusterSize))calloc(numClusters,
90             sizeof(**local_newClusterSize));
90         local_newClusters[k] = (typeof(*local_newClusters))calloc(numClusters * numCoords,
91             sizeof(**local_newClusters));
91     }
92
93     timing = wtime();
94     do
95     {
96         // before each loop, set cluster data to 0
97         for (i = 0; i < numClusters; i++)
98         {
99             for (j = 0; j < numCoords; j++)
100                 newClusters[i * numCoords + j] = 0.0;

```

```

101         newClusterSize[i] = 0;
102     }
103
104     // reset delta before each iteration; it will be updated via reduction in the
105     // parallel region
106     delta = 0.0;
107
108     /*
109      * TODO: Initialize local cluster data to zero (separate for each thread)
110      *
111      * We now use an OpenMP parallel region where:
112      * - Each thread zeroes its own local_newClusterSize/local_newClusters.
113      * - The object loop is distributed with 'omp for' and 'reduction(+ : delta)'.
114      * - A single thread reduces the per-thread local arrays into the shared arrays.
115      */
116 #pragma omp parallel private(i, j, k, index)
117 {
118     int tid = omp_get_thread_num();
119     int T   = omp_get_num_threads(); // actual number of threads in this team
120
121     /* per-thread zeroing (first-touch initialization of local cluster data) */
122     for (i = 0; i < numClusters; i++)
123         local_newClusterSize[tid][i] = 0;
124     for (i = 0; i < numClusters * numCoords; i++)
125         local_newClusters[tid][i] = 0.0;
126
127     // Distribute objects across threads and compute per-thread contributions.
128     // delta is accumulated using a reduction to avoid atomics on a shared
129     // variable.
130 #pragma omp for reduction(+ : delta)
131     for (i = 0; i < numObjs; i++)
132     {
133         // find the array index of nearest cluster center
134         index = find_nearest_cluster(numClusters, numCoords,
135                                     &objects[i * numCoords], clusters);
136
137         // if membership changes, increase delta by 1
138         if (membership[i] != index)
139             delta += 1.0;
140
141         // assign the membership to object i
142         membership[i] = index;
143
144         // update new cluster centers : sum of all objects located within (average
145         // will be performed later)
146         /*
147          * TODO: Collect cluster data in local arrays (local to each thread)
148          * Replace global arrays with local per-thread
149          */
150         local_newClusterSize[tid][index]++;
151         for (j = 0; j < numCoords; j++)
152             local_newClusters[tid][index * numCoords + j] += objects[i * numCoords
153 + j];
154     }

```

```

151
152     /*
153      * TODO: Reduction of cluster data from local arrays to shared.
154      * This operation will be performed by one thread
155      *
156      * Here we use 'omp single' so that exactly one thread accumulates
157      * all per-thread local arrays into the shared newClusterSize/newClusters.
158      */
159 #pragma omp single
160 {
161     for (k = 0; k < T; k++) // only sum over the threads actually in this
team
162     {
163         int *srcS = local_newClusterSize[k];
164         double *srcC = local_newClusters[k];
165         if (!srcS || !srcC)
166             continue;
167         for (i = 0; i < numClusters; i++)
168         {
169             newClusterSize[i] += srcS[i];
170             for (j = 0; j < numCoords; j++)
171                 newClusters[i * numCoords + j] += srcC[i * numCoords + j];
172         }
173     }
174     /* implicit barrier after single */
175 } /* end parallel region */

176
177 // average the sum and replace old cluster centers with newClusters
178 for (i = 0; i < numClusters; i++)
179 {
180     if (newClusterSize[i] > 0)
181     {
182         for (j = 0; j < numCoords; j++)
183         {
184             clusters[i * numCoords + j] = newClusters[i * numCoords + j] /
newClusterSize[i];
185         }
186     }
187 }

188
189 // Get fraction of objects whose membership changed during this loop. This is used
as a convergence criterion.
190     delta /= numObjs;

191
192     loop++;
193     printf("\r\tcompleted loop %d", loop);
194     fflush(stdout);
195 } while (delta > threshold && loop < loop_threshold);
196     timing = wtime() - timing;
197     printf("\n nloops = %3d (total = %7.4fs) (per loop = %7.4fs)\n", loop, timing, timing
/ loop);

198
199     for (k = 0; k < nthreads; k++)
200     {

```

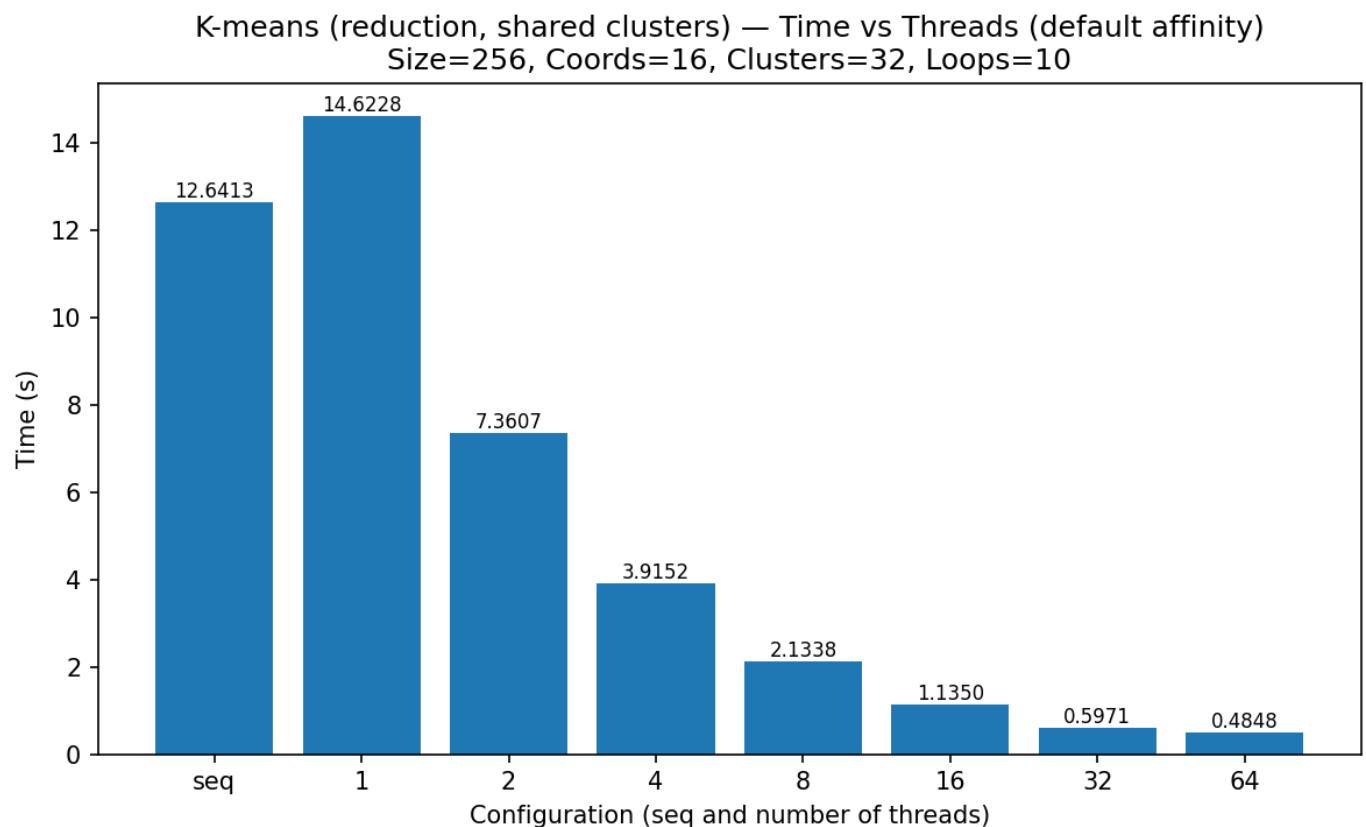
```
201     free(local_newClusterSize[k]);
202     free(local_newClusters[k]);
203 }
204 free(newClusters);
205 free(newClusterSize);
206 }
207
208 }
```

1. Παραλληλοίση για το αρχικό grid size και διαγράμματα

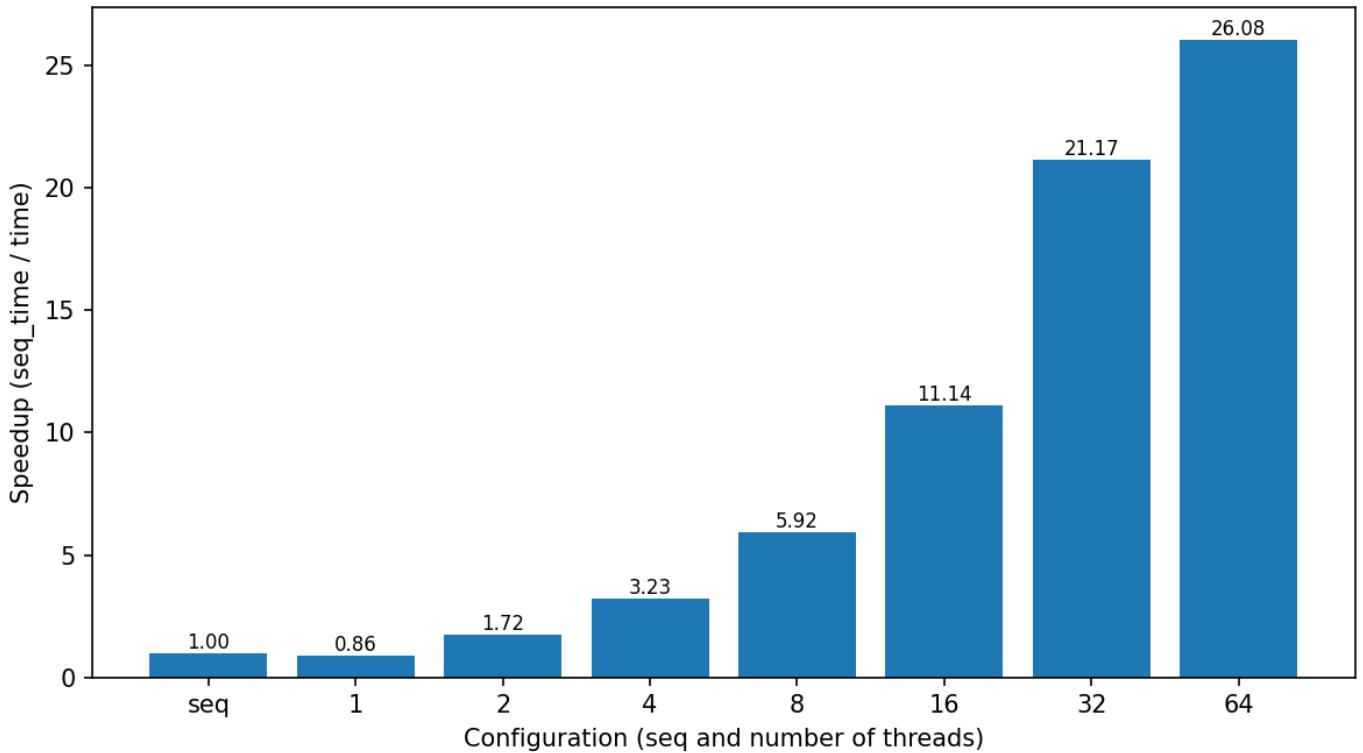
Το παραπάνω παράλληλο πρόγραμμα (omp_reduction_kmeans.c) έτρεξε για τις παραμέτρους: {Size, Coords, Clusters, Loops} = {256, 16, 32, 10}, threads = {1, 2, 4, 8, 16, 32, 64} και με affinity (όπως και πριν). Τα αποτελέσματα που προέκυψαν παρουσιάζονται στον παρακάτω πίνακα:

THREADS	TIME
seq	12.64
1	14.62
2	7.36
4	3.92
8	2.13
16	1.14
32	0.60
64	0.48

Τα ζητούμενα διαγράμματα (πάντα με βάση τον χρόνο του σειριακού προγράμματος, όπως αναφέρεται) φαίνονται ακολούθως:



K-means (reduction, shared clusters) — Speedup vs Threads (default affinity)
Size=256, Coords=16, Clusters=32, Loops=10

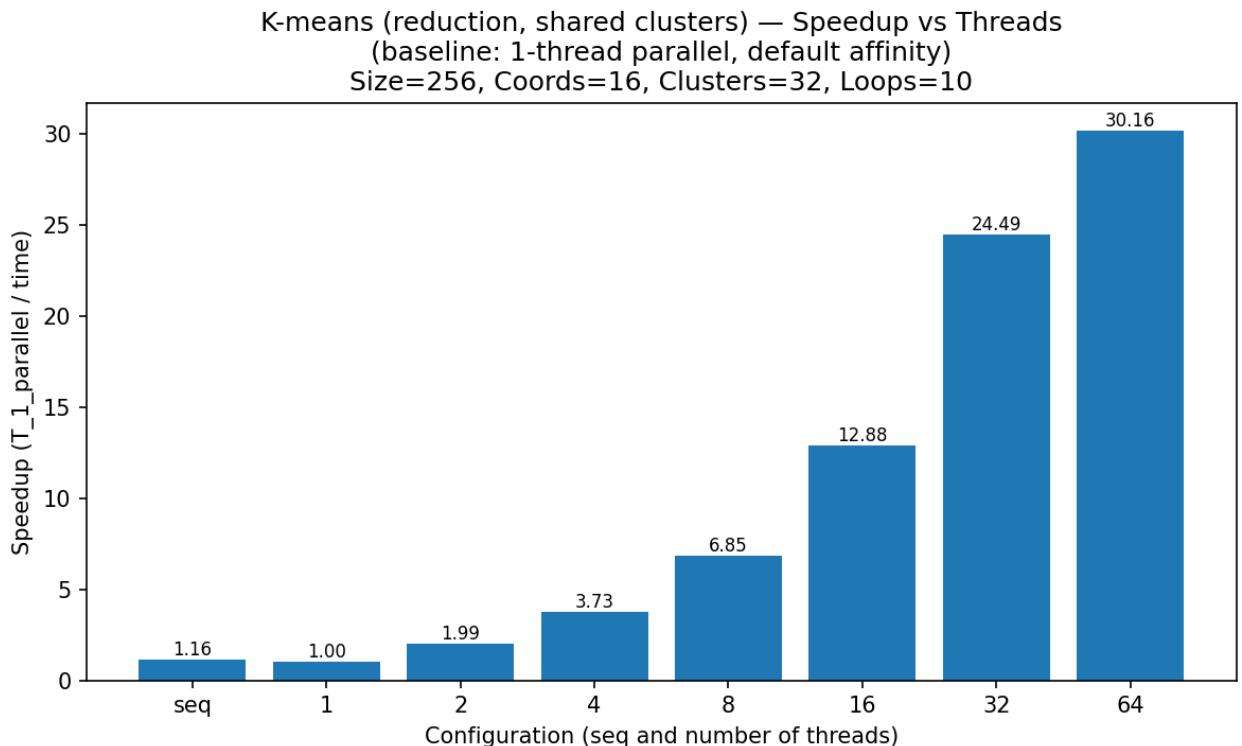


Από τα δύο πρώτα διαγράμματα παρατηρούμε ότι η έκδοση με copied clusters και reduction κλιμακώνει πλέον πολύ ικανοποιητικά σε σχέση με τη naive shared προσέγγιση. Ο σειριακός χρόνος είναι περίπου 12.6s, ενώ η παράλληλη εκτέλεση με 1 νήμα παραμένει λίγο χειρότερη (~14.6s), λόγω του parallel overhead του OpenMP, όπως και πριν. Ωστόσο, από τα 2 νήματα και πάνω ο χρόνος μειώνεται μονοτονικά και σχεδόν γραμμικά: ~7.4s στα 2 threads, ~3.9s στα 4, ~2.1s στα 8, ~1.1s στα 16, ~0.6s στα 32 και ~0.5s στα 64 threads. Αντίστοιχα, το speedup ως προς το σειριακό πρόγραμμα φτάνει περίπου τις 1.7x, 3.2x, 5.9x, 11x, 21x και 26x για 2, 4, 8, 16, 32 και 64 νήματα αντίστοιχα, πολύ κοντά στην ιδανική κλιμάκωση που παρουσιάζεται και στις διαφάνειες.

Η διαφορά σε σχέση με τη naive υλοποίηση εξηγείται από την αρχιτεκτονική της reduction λύσης: κάθε νήμα ενημερώνει αποκλειστικά τα δικά του local αντίγραφα (local_newClusters, local_newClusterSize), αποφεύγοντας atomic ενημερώσεις σε κοινές cache lines και μειώνοντας δραστικά το synchronization bottleneck. Έτσι, το μεγαλύτερο μέρος του χρόνου ξοδεύεται σε πραγματικό υπολογισμό και όχι σε συγχρονισμό, κάτι που επιτρέπει στον αλγόριθμο να κλιμακώνεται πολύ καλύτερα σε ένα NUMA σύστημα όπως ο sandman. Μέχρι τα 32 threads (ένας hardware thread ανά φυσικό πυρήνα) αξιοποιούνται αποτελεσματικά οι πόροι όλων των sockets, ενώ η μικρή “κάμψη” της κλιμάκωσης από τα 32 στα 64 νήματα αποδίδεται κυρίως στο hardware multithreading και στον κορεσμό του

memory bandwidth: δύο λογικά νήματα ανά πυρήνα μοιράζονται την ίδια εκτέλεση και τις ίδιες memory lanes σε έναν ήδη memory-bound αλγόριθμο.

Από απλή περιέργεια και χωρίς να ζητείται, καταστρώσαμε και ένα διάγραμμα speedup με βάση όχι τώρα το σειριακό πρόγραμμα, αλλά το παράλληλο με 1 thread. Τα αποτελέσματα φαίνονται ακολούθως:



Το τελευταίο διάγραμμα, όπου ως βάση λαμβάνουμε την παράλληλη εκτέλεση με 1 νήμα, αναδεικνύει ακόμη πιο καθαρά το όφελος της reduction υλοποίησης. Σε αυτή τη σύγκριση, το σειριακό πρόγραμμα εμφανίζεται ήδη ταχύτερο από το 1-thread parallel (speedup $\approx 1.2 \times$), επιβεβαιώνοντας ότι το κόστος δημιουργίας και οργάνωσης της ομάδας νημάτων είναι σημαντικό όταν χρησιμοποιείται μόνο ένα νήμα. Από τα 2 threads και πάνω, όμως, η κλιμάκωση γίνεται εντυπωσιακή: το speedup φτάνει περίπου τις $2 \times$ στα 2 νήματα, $\sim 3.7 \times$ στα 4, $\sim 6.8 \times$ στα 8, $\sim 12.9 \times$ στα 16, $\sim 24.5 \times$ στα 32 και πάνω από $30 \times$ στα 64 νήματα σε σχέση με το 1-thread parallel.

Τα αποτελέσματα αυτά δείχνουν ότι, μόλις “ξεπληρωθεί” το parallel overhead, η copied clusters and reduction προσέγγιση εκμεταλλεύεται πλήρως την παραλληλία που προσφέρει ο κόμβος: μέχρι τα 32 threads αξιοποιείται ουσιαστικά κάθε φυσικός πυρήνας όλων των NUMA nodes, με πολύ μικρό synchronization cost χάρη στα local arrays, ενώ η περαιτέρω αύξηση στα 64 threads δίνει μικρότερο, αλλά υπαρκτό, πρόσθετο κέρδος λόγω του hardware multithreading. Σε αντίθεση με τη naive shared υλοποίηση, εδώ η κλιμάκωση περιορίζεται κυρίως από το διαθέσιμο memory

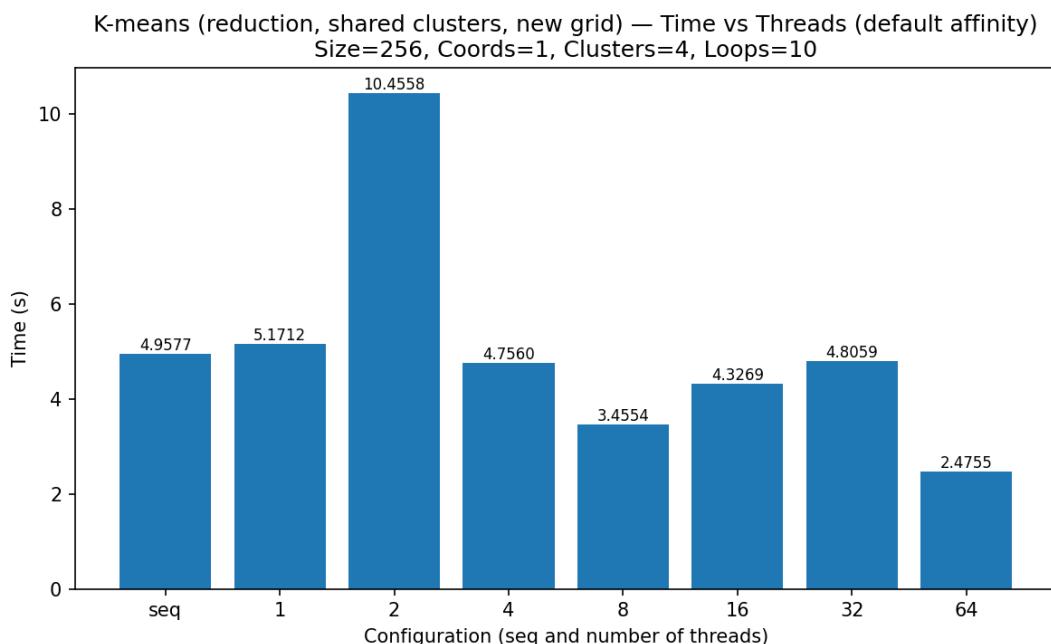
bandwidth και το μικρό σειριακό τμήμα του κώδικα (νόμος του Amdahl), ενώ το synchronization bottleneck έχει ουσιαστικά εξαλειφθεί.

2. Παραλληλοποίηση για το μειωμένο grid size και διαγράμματα

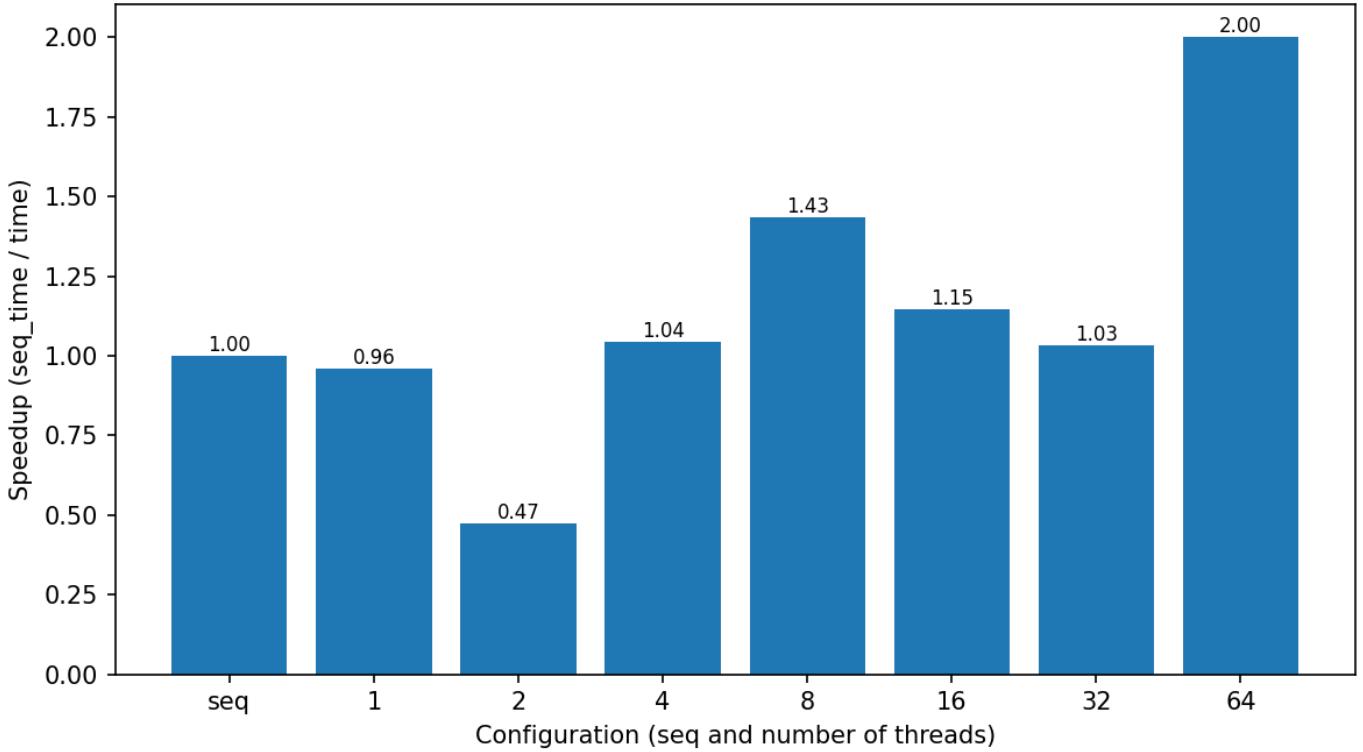
Το παραπάνω παράλληλο πρόγραμμα (omp_reduction_kmeans.c) έτρεξε για τις παραμέτρους: {Size, Coords, Clusters, Loops} = {256, 1, 4, 10}, threads = {1, 2, 4, 8, 16, 32, 64} και με affinity (όπως και πριν). Τα αποτελέσματα που προέκυψαν παρουσιάζονται στον παρακάτω πίνακα:

THREADS	TIME
seq	4.96
1	5.17
2	10.46
4	4.76
8	3.46
16	4.33
32	4.81
64	2.48

Τα ζητούμενα διαγράμματα (πάντα με βάση τον χρόνο του σειριακού προγράμματος, όπως αναφέρεται) φαίνονται ακολούθως:



K-means (reduction, shared clusters, new grid) — Speedup vs Threads (default affinity)
Size=256, Coords=1, Clusters=4, Loops=10

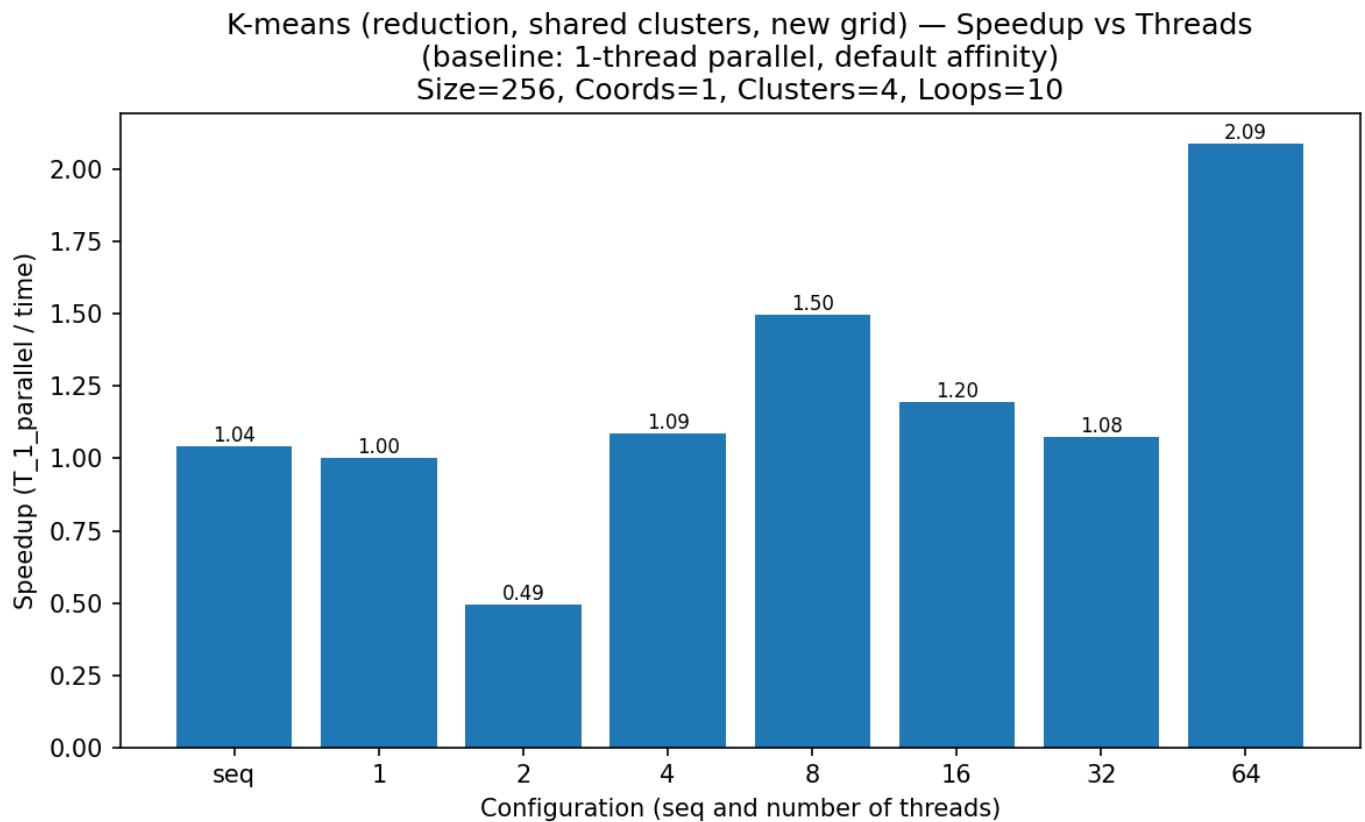


Από τα δύο πρώτα διαγράμματα για το μειωμένο grid παρατηρούμε ότι η συμπεριφορά της έκδοσης με reduction είναι σαφώς χειρότερη από αυτή στο αρχικό, πιο «υπολογιστικά βαρύ» (compute intensive) grid (<{256,16,32,10}). Εδώ ο σειριακός χρόνος είναι περίπου 4.96s και η παράλληλη εκτέλεση με 1 νήμα λίγο χειρότερη (~5.17s), αλλά η κλιμάκωση με περισσότερα νήματα δεν είναι πλέον καλή: στα 2 threads ο χρόνος μάλιστα χειροτερεύει σημαντικά (~10.46s), στα 4 και 8 νήματα έχουμε μια μικρή βελτίωση (4.76s και 3.46s αντίστοιχα), ενώ στα 16 και 32 νήματα ο χρόνος ξανανεβαίνει κοντά στον σειριακό (4.33s και 4.81s) και μόνο στα 64 threads πέφτει στα ~2.48s. Αντίστοιχα, το speedup ως προς το σειριακό πρόγραμμα μόλις που ξεπερνά το 1.4x στα 8 νήματα και φτάνει περίπου το 2x στα 64 νήματα.

Η βασική διαφορά στα scalability plots, σε σχέση με το προηγούμενο grid, είναι ότι εδώ η κλιμάκωση «επιπεδώνει» πολύ νωρίς και είναι έντονα ακανόνιστη. Στο αρχικό grid με 16 συντεταγμένες και 32 clusters, ο αλγόριθμος ήταν πολύ πιο compute-intensive και η έκδοση με reduction κατάφερνε να φτάσει speedup ~26x στα 64 threads. Αντίθετα, στο μειωμένο grid με μόνο 1 συντεταγμένη και 4 clusters, το workload ανά αντικείμενο είναι πολύ μικρότερο και η εφαρμογή είναι ακόμη πιο έντονα memory-bound: κάθε νήμα κάνει ελάχιστες πράξεις ανά προσπέλαση μνήμης, με αποτέλεσμα το κόστος πρόσβασης σε μνήμη και το overhead του OpenMP να κυριαρχούν. Έτσι, ο διαθέσιμος υπολογισμός δεν αρκεί για να «κρύψει»

τη latency της μνήμης και το scalability περιορίζεται σημαντικά, όπως φαίνεται στα διαγράμματα.

Από απλή περιέργεια και χωρίς να ζητείται, καταστρώσαμε και ένα διάγραμμα speedup με βάση όχι τώρα το σειριακό πρόγραμμα, αλλά το παράλληλο με 1 thread. Τα αποτελέσματα φαίνονται ακολούθως:



Στο τελευταίο διάγραμμα, όπου ως βάση λαμβάνουμε την παράλληλη εκτέλεση με 1 νήμα, γίνεται ακόμη πιο εμφανές ότι στο μειωμένο grid το scaling είναι περιορισμένο και ασταθές. Ο σειριακός κώδικας παραμένει ελαφρώς ταχύτερος από το 1-thread parallel, ενώ τα speedups ως προς το 1-thread πρόγραμμα κινούνται γύρω στη μονάδα για τα περισσότερα T: το 2-thread run είναι σαφώς χειρότερο (speedup < 1), στα 4 και 8 νήματα έχουμε κάποια βελτίωση, ενώ στα 16 και 32 νήματα ουσιαστικά δεν κερδίζουμε τίποτα. Μόνο στα 64 threads παρατηρείται πιο αξιοσημείωτο κέρδος, της τάξης περίπου του 2x, αλλά και πάλι πολύ μακριά από τα speedups που είδαμε στο αρχικό grid και την επιθυμητή κλιμάκωση.

Η διαφορά στα scalability plots σε σχέση με την περίπτωση {256,16,32,10} εξηγείται από το ότι εδώ το πρόβλημα είναι πλέον «πολύ μικρό» υπολογιστικά ανά στοιχείο και κυριαρχείται από τις προσπελάσεις στη μνήμη (memory-bound). Στο αρχικό grid, ο μεγάλος αριθμός συντεταγμένων και clusters παρείχε αρκετό

υπολογισμό ώστε η έκδοση με reduction να εκμεταλλεύεται ουσιαστικά όλους τους πυρήνες μέχρι τα 32 threads και να έχει πολύ καλή κλιμάκωση. Στο νέο grid, το per-object work είναι μικρό και η ταυτόχρονη πρόσβαση πολλών νημάτων στα ίδια δεδομένα καταναλώνει γρήγορα το memory bandwidth, με αποτέλεσμα ο επιπλέον παραλληλισμός να μην μπορεί να μεταφραστεί σε αντίστοιχο speedup. Ουσιαστικά, έχουμε ένα χαρακτηριστικό παράδειγμα από τις διαφάνειες: όσο πιο memory-bound είναι μια εφαρμογή και όσο μικρότερο το διαθέσιμο υπολογιστικό workload, τόσο πιο γρήγορα «σκάει» η κλιμάκωση και τα scalability plots γίνονται ρηχά και ασταθή.

Στο Linux, η πολιτική first-touch σε NUMA συστήματα ορίζει ότι οι φυσικές σελίδες μνήμης δεσμεύονται στο NUMA node του πυρήνα που τις «ακουμπά» πρώτος (πρώτη εγγραφή). Στο πρόγραμμά μας, οι πίνακες local_newClusters και local_newClusterSize δεσμεύονται αρχικά σειριακά (μέσα σε loops στο main thread), οπότε οι αντίστοιχες σελίδες μνήμης τοποθετούνται κατά κανόνα στο NUMA node όπου εκτελείται το main thread. Στη συνέχεια, όταν άλλα νήματα OpenMP που τρέχουν σε διαφορετικούς πυρήνες και nodes προσπελαύνουν αυτές τις δομές, μεγάλο μέρος των προσβάσεων είναι «απομακρυσμένο» (remote NUMA), με αυξημένη latency και μειωμένο effective bandwidth. Αυτό περιορίζει την επίδοση, ειδικά στη reduction υλοποίηση, όπου η πρόσβαση στα τοπικά clusters δεδομένα είναι πολύ συχνή.

Επιπλέον, εμφανίζεται και το φαινόμενο false-sharing: παρόλο που κάθε νήμα ενημερώνει διαφορετικά elements μέσα στο local_newClusters[tid], οι επιμέρους πίνακες για διαφορετικά tid μπορεί να τοποθετούνται σε συνεχόμενες διευθύνσεις και να μοιράζονται τις ίδιες cache lines. Έτσι, όταν δύο threads γράφουν σε διαφορετικά στοιχεία που τυχαίνει να κατοικούν στην ίδια γραμμή cache, οι γραμμές αυτές κάνουν συνεχώς invalidate μεταξύ των πυρήνων, δημιουργώντας σημαντικό overhead, χωρίς να υπάρχει πραγματικό data sharing σε επίπεδο προγράμματος.

Για να αντιμετωπίσουμε προβλήματα NUMA τοποθέτησης (first-touch), μπορούμε να αφήσουμε το κάθε νήμα να κάνει τη δική του δέσμευση μνήμης, π.χ. με malloc μέσα στην παράλληλη περιοχή: κάθε thread εκτελεί το malloc και την αρχικοποίηση των δικών του local_newClusters[tid] και local_newClusterSize[tid], οπότε οι σελίδες του κάθε πίνακα first-touched από το αντίστοιχο νήμα καταλήγουν στον «σωστό» NUMA node. Με αυτόν τον τρόπο, οι επαναλαμβανόμενες προσπελάσεις σε τοπικά δεδομένα γίνονται κυρίως σε τοπική μνήμη, μειώνοντας σημαντικά τα remote NUMA accesses.

Για να περιορίσουμε το false-sharing, κάναμε χρήση padding στα τοπικά arrays: φροντίζουμε κάθε «γραμμή» local_newClusters[tid] να ευθυγραμμίζεται σε μέγεθος cache line (π.χ. 64 bytes) και να μεσολαβεί αρκετό κενό (padding) ανάμεσα στα δεδομένα διαφορετικών νημάτων, ώστε καμία cache line να μην περιέχει ταυτόχρονα δεδομένα από δύο διαφορετικά tid. Με αυτόν τον τρόπο, κάθε γραμμή cache ανήκει ουσιαστικά σε ένα μόνο thread και δεν υπάρχει αλληλοεπικάλυψη που θα προκαλούσε invalidations. Συνδυάζοντας την τεχνική του per-thread malloc (για σωστό first-touch ανά thread και NUMA node) με το κατάλληλο padding, μειώνουμε τόσο τα προβλήματα NUMA τοποθέτησης όσο και τα φαινόμενα false-sharing, όπως προτείνεται και στο hint της εκφώνησης.

Γενικές Παρατηρήσεις

- Η υλοποίηση με reduction αποδείχθηκε σαφώς πιο αποδοτική από αυτή με atomic operations, καθώς μεταφέρει το κόστος συγχρονισμού σε ένα μικρό, καλά οριοθετημένο στάδιο συγχώνευσης (reduction) και αφήνει τον κυρίως βρόχο να εκτελείται χωρίς locks. Αντίθετα, στην παίνε υλοποίηση μεγάλο μέρος του χρόνου χάνεται σε atomic ενημερώσεις πάνω στις ίδιες cache lines. Παρ' όλα αυτά, το reduction δεν είναι πάντα προτιμότερο: σε σενάρια με μικρό πρόβλημα ή λίγες συγκρούσεις στα shared δεδομένα, το επιπλέον κόστος αντιγραφής και συγχώνευσης μπορεί να εξανεμίσει τα κέρδη.
- Πέραν των 32 threads δεν παρατηρείται ποτέ ουσιαστική (σχεδόν γραμμική) βελτίωση, καθώς στο sandman έχουμε 32 φυσικούς πυρήνες και 64 λογικά νήματα μέσω hardware multithreading (hyperthreading). Σε έναν κατά βάση memory-bound αλγόριθμο, όπως ο K-means με shared clusters, δύο λογικά νήματα στον ίδιο πυρήνα μοιράζονται τους ίδιους execution πόρους και τις ίδιες memory lanes, οπότε η περαιτέρω αύξηση των νημάτων δεν μεταφράζεται σε αντίστοιχο speedup και μπορεί να οδηγήσει ακόμη και σε υποβάθμιση της επίδοσης.
- Η επιλογή μιας κατάλληλης πολιτικής affinity βελτιώνει αισθητά την επίδοση. Η δέσμευση των νημάτων σε συγκεκριμένους πυρήνες (ώστε να «μένουν κοντά» σε δεδομένα και cache) μειώνει το κόστος επικοινωνίας με τη μνήμη και εκμεταλλεύεται καλύτερα την τοπικότητα. Παράλληλα, η προσεκτική διασπορά των threads στα NUMA nodes μπορεί να αυξήσει το διαθέσιμο memory bandwidth για memory-bound εφαρμογές. Τα πειράματά μας δείχνουν ξεκάθαρα ότι με ενεργό affinity η επίδοση μέχρι τα 8–16 threads βελτιώνεται σημαντικά σε σχέση με την πλήρως “noaff” περίπτωση.

▪ Ενότητα 2.2 – Παραλληλοποίηση του Αλγορίθμου Floyd-Warshall

- **Υλοποίηση και ανάλυση εξαρτήσεων**

Για την παραλληλοποίηση του recursive Floyd-Warshall άλγορίθμου χρησιμοποιήθηκαν OpenMP tasks. Η βασική πρόκληση προέκυψε από το γεγονός ότι ο αλγόριθμος παίρνει ως ορίσματα τον ίδιο πίνακα (A) τρεις φορές (A,B,C), αν και με διαφορετικά offsets. Αυτό έχει ως αποτέλεσμα, παρόλο που τα A,B,C έχουν διαφορετικό όνομα, να αναφέρονται συχνά στην ίδια θέση μνήμης, δημιουργώντας εξαρτήσεις τύπου RAW (Read After Write) και WAW (Write After Write).

- **Μελέτη των Αναδρομικών Κλήσεων**

Η μελέτη των εξαρτήσεων ανά αναδρομική κλήση βασίστηκε στις ακόλουθες παρατηρήσεις:

1. Κάθε αναδρομική κλήση γίνεται για blocks μισού μεγέθους σε σχέση με αυτά της εισόδου (3 blocks/arrays ίδιων διαστάσεων)
2. Κάθε πίνακας/block χωρίζεται σε 4 ίσα μέρη, επιτρέποντας σε διαφορετικά blocks ίδιων διαστάσεων να μην έχουν μερικές επικαλύψεις.
3. Σε πιο "βαθύ επίπεδο" της αναδρομής, ο αλγόριθμος περιορίζεται σε εγγραφή στον A (και υπό-blocks του) και ανάγνωση από τους B,C (και υπό-blocks τους) στην δεδομένη κλήση. Αυτό εξασφαλίζει ότι σε βαθύτερο επίπεδο δεν υπάρχει επικάλυψη πρόσβασης μνήμης που δεν φαίνεται στο ψηλότερο.

Μελετώντας τις κλήσεις, το πρόβλημα χωρίστηκε σε 4 περιπτώσεις ανάλογα με τις σχέσεις των blocks A,B,C.

- **Σενάρια Παραλληλισμού (Task Chains)**

Θεωρώντας ως call_i την i-οστή σε σειρά αναδρομική κλήση του παρακάτω αλγορίθμου βγάζουμε τα συμπεράσματα του πίνακα:

FWR (A00, B00, C00);

FWR (A01, B00, C01);

FWR (A10, B10, C00);

FWR (A11, B10, C01);

FWR (A11, B10, C01);

FWR (A10, B10, C00);

FWR (A01, B00, C01);

FWR (A00, B00, C00);

Case ID	Σχέση Blocks	Αλυσίδα Εξαρτήσεων	Παραλληλισμός
1	A=B=C (όλα ταυτίζονται)	1 -> {2,3} -> 4 -> 5 -> {6,7}->8	Τα ζευγάρια {2, 3} και {6, 7} εκτελούνται παράλληλα 10.
2/3	A=B ή A=C (αλλά B ≠ C)	Αν A=B: 1->2->7->8. και 3->4->5->6. Αν A=C: 1->3->6->8 και 2->4->5->6->8	Προκύπτουν 2 παράλληλες και εντελώς ανεξάρτητες αλυσίδες εκτέλεσης
4	A≠ B≠ C (όλα διαφορετικά)	1-> 8, 2->7, 3->6, 4->5	Προκύπτουν 4 παράλληλες και εντελώς ανεξάρτητες αλυσίδες

Στην αρχική κλήση ($A, 0, 0, A, 0, 0, A, 0, 0$), εφαρμόζεται η Case 1. Η εξάρτηση RAW των κλήσεων 2, 3 από το 1, και η εξάρτηση του 4 από 2, 3, καθορίζουν τη σειρά εκτέλεσης.

- **Επιλογή block size (B)**

Για την επιλογή του μεγέθους του ελάχιστου block (bsize) όπου γίνεται ο υπολογισμός του min (base case), δοκιμάστηκαν $B=\{16, 32, 64, 128, 256\}$ για $N=1024$:

- ✓ Παρατηρήθηκε ότι η επίδοση βελτιωνόταν σημαντικά με την αύξηση του B.
- ✓ Στο $B=128$ δεν σημειώθηκε βελτίωση, ενώ στο $B=256$ ο χρόνος εκτέλεσης μάλιστα αυξήθηκε.
- ✓ Επιλέχθηκε $B=64$ για την υπόλοιπη άσκηση. Αν και το $B=128$ ήταν αποδοτικό, το $B=64$ προτιμήθηκε για να αξιοποιηθεί περισσότερο η παραλληλία, καθώς μεγαλύτερο B αντιστοιχεί σε λιγότερα layers παράλληλου προγράμματος. Η επιλογή του B επηρεάζει ελάχιστα το speedup, καθώς επηρεάζει με τον ίδιο τρόπο το σειριακό και το παράλληλο πρόγραμμα.

- **Παράμετροι Εκτέλεσης και Περιβάλλοντος**

Η εκτέλεση και οι μετρήσεις του παράλληλου αλγορίθμου Recursive Floyd-Warshall (FW_SR) πραγματοποιήθηκαν μέσω του script run_on_queue.sh στο περιβάλλον PBS (Portable Batch System) του συστήματος sandman. Η εργασία υποβλήθηκε στην ουρά serial ζητώντας 64 πυρήνες (ppn=64) στον κόμβο sandman. Το πρόγραμμα εκτελέστηκε επαναληπτικά για τρία μεγέθη πίνακα (N): 1024, 2048, και 4096, χρησιμοποιώντας την βελτιστοποιημένη τιμή $B=64$ για το μέγεθος του ελάχιστου block. Ο παραλληλισμός OpenMP ελέγχθηκε μέσω της μεταβλητής περιβάλλοντος OMP_NUM_THREADS, η οποία διατρέχθηκε στις τιμές $T=\{1, 2, 4, 8, 16, 32, 64\}$. Κάθε συνδυασμός N και T καταγράφηκε σε ξεχωριστά αρχεία εξόδου (.out και .err), εξασφαλίζοντας ακριβή δεδομένα για την ανάλυση της κλιμάκωσης του αλγορίθμου.

Τα run_on_queue.sh και το κυρίως πρόγραμμα fw_sr_p.c φαίνονται ακολούθως:

a2/FW/run_on_queue.sh

```
1 #!/bin/bash
2
3 #PBS -N run_fw_sr_p
4
5 ## Output error
6 #PBS -o run_fw_sr_p.pbs_out
7 #PBS -e run_fw_sr_p.pbs_err
8
9 ## Sandman, serial queue, 64 threads
10 #PBS -q serial
11 #PBS -l nodes=sandman:ppn=64
12
13 #PBS -l walltime=01:00:00
14
15 cd /home/parallel/parlab05/a2/FW
16
17 module load openmp
18
19 N_VALUES="1024 2048 4096"
20
21 # Block size
22 B=64
23
24 OUTDIR="benchmarks"
25 mkdir -p "$OUTDIR"
26
27 for N in $N_VALUES; do
28     for T in 1 2 4 8 16 32 64; do
29
30         export OMP_NUM_THREADS=$T
31         echo "Running N=$N, B=$B, threads=$T"
32
33         #outputs
34         OUT="${OUTDIR}/fw_sr_p_N${N}_T${T}.out"
35         ERR="${OUTDIR}/fw_sr_p_N${N}_T${T}.err"
36
37         # - stdout → OUT
38         # - stderr → ERR
39         ./fw_sr_p "$N" "$B" >"$OUT" 2>"$ERR"
40     done
41 done
42
43
```

```

a2/FW/fw_sr_p.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 <omp.h>
13 #include "util.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,k;
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     if ((N%B)!=0){
39         fprintf(stdout, "N must be multiple of B\n");
40         exit(0);
41     }
42
43     A = (int **) malloc(N*sizeof(int *));
44     for(i=0; i<N; i++) A[i] = (int *) malloc(N*sizeof(int));
45
46     graph_init_random(A, -1, N, 128*N);
47
48 //-----
49     gettimeofday(&t1,0);
50
51     #pragma omp parallel

```

```

52 #pragma omp single
53 {
54     FW_SR(A,0,0, A,0,0,A,0,0,N,B);
55 }
56
57     gettimeofday(&t2,0);
58
59     time=(double)((t2.tv_sec-t1.tv_sec)*1000000+t2.tv_usec-t1.tv_usec)/1000000;
60     printf("FW_SR,%d,%d,%4f\n", N, B, time);
61
62
63 //    for(i=0; i<N; i++)
64 //        for(j=0; j<N; j++) fprintf(stdout,"%d\n", A[i][j]);
65
66
67     return 0;
68 }
69
70 inline int min(int a, int b)
71 {
72     if(a<=b) return a;
73     else return b;
74 }
75
76 void FW_SR (int **A, int arow, int acol,
77             int **B, int brow, int bcol,
78             int **C, int crow, int ccol,
79             int myN, int bsize)
80 {
81     int k,i,j;
82     /*we use different task paral depending on the blocks A,B,C use.
83      If they use same blocks therre may be future depedencies , else not*/
84
85     //Arrays check
86     if (A!=B || A!=C){
87         printf("Different arrays not supported yet\n");
88         exit (1);
89     }
90     //row check
91     int RAB= arow==brow;
92     int RAC= arow==crow;
93     int RBC= brow==crow;
94     //col check
95     int CAB= acol==bcol;
96     int CAC= acol==ccol;
97     int CBC= bcol==ccol;
98     //case check
99     int case_id;
100    if (RAB&&RAC&&CAB&&CAC) case_id=0; //A,B,C same block
101    else if (RAB&&CAB) case_id=1; //A,B same block
102    else if (RAC&&CAC) case_id=2; //A,C same block
103    else case_id=3; //A separate from B and C
104
105    /*

```

```

106     * The base case (when recursion stops) is not allowed to be edited!
107     * What you can do is try different block sizes.
108     */
109     if(myN<=bsize)
110         for(k=0; k<myN; k++)
111             for(i=0; i<myN; i++)
112                 for(j=0; j<myN; j++)
113                     A[arow+i][acol+j]=min(A[arow+i][acol+j], B[brow+i][bcol+k]+C[crow+k]
114 [ccol+j]);
115     else {
116
117         switch(case_id){
118             case 0: //A,B,C same block
119             {
120                 //call1
121                 FW_SR(A,arow, acol,B,brow, bcol,C,crow, ccol, myN/2, bsize);
122
123                 #pragma omp task firstprivate(arow,acol,brow,bcol,crow,ccol,myN,bsize)
124                 shared(A,B,C)
125                 {
126                     //call2
127                     FW_SR(A,arow, acol+myN/2,B,brow, bcol,C,crow, ccol+myN/2, myN/2,
128 bsize);
129                 }
130                 #pragma omp task firstprivate(arow,acol,brow,bcol,crow,ccol,myN,bsize)
131                 shared(A,B,C)
132                 {
133                     //call3
134                     FW_SR(A,arow+myN/2, acol,B,brow+myN/2, bcol,C,crow, ccol, myN/2,
135 bsize);
136
137                     //call4
138                     FW_SR(A,arow+myN/2, acol+myN/2,B,brow+myN/2, bcol,C,crow, ccol+myN/2,
139 myN/2, bsize);
140
141                     //call5
142                     FW_SR(A,arow+myN/2, acol+myN/2,B,brow+myN/2, bcol+myN/2,C,crow+myN/2,
143 ccol+myN/2, myN/2, bsize);
144
145                     #pragma omp task firstprivate(arow,acol,brow,bcol,crow,ccol,myN,bsize)
146                     shared(A,B,C)
147                     {
148                         //call6
149                         FW_SR(A,arow+myN/2, acol,B,brow+myN/2, bcol+myN/2,C,crow+myN/2, ccol,

```

```

150         #pragma omp taskwait
151
152             //call8
153             FW_SR(A,arow, acol,B,brow, bcol+myN/2,C,crow+myN/2, ccol, myN/2, bsize);
154         }
155         break;
156         case 1: //A,B same block
157         {
158             #pragma omp task firstprivate(arow,acol,brow,bcol,crow,ccol,myN,bsize)
159             shared(A,B,C)
160             {
161                 //call1
162                 FW_SR(A,arow, acol,B,brow, bcol,C,crow, ccol, myN/2, bsize);
163                 //call2
164                 FW_SR(A,arow, acol+myN/2,B,brow, bcol,C,crow, ccol+myN/2, myN/2,
165                 bsize);
166                 //call7
167                 FW_SR(A,arow, acol+myN/2,B,brow, bcol+myN/2,C,crow+myN/2, ccol+myN/2,
168                 myN/2, bsize);
169                 //call8
170                 FW_SR(A,arow, acol,B,brow, bcol+myN/2,C,crow+myN/2, ccol, myN/2,
171                 bsize);
172                 //call9
173                 FW_SR(A,arow+myN/2, acol,B,brow+myN/2, bcol,C,crow, ccol, myN/2,
174                 bsize);
175                 //call14
176                 FW_SR(A,arow+myN/2, acol+myN/2,B,brow+myN/2, bcol,C,crow, ccol+myN/2,
177                 myN/2, bsize);
178                 //call15
179                 FW_SR(A,arow+myN/2, acol+myN/2,B,brow+myN/2, bcol+myN/2,C,crow+myN/2,
180                 ccol+myN/2, myN/2, bsize);
181                 //call16
182                 FW_SR(A,arow+myN/2, acol,B,brow+myN/2, bcol+myN/2,C,crow+myN/2, ccol,
183                 myN/2, bsize);
184             }
185             #pragma omp taskwait
186         }
187         break;
188         case 2: //A,C same block
189         {
190             #pragma omp task firstprivate(arow,acol,brow,bcol,crow,ccol,myN,bsize)
191             shared(A,B,C)
192             {
193                 //call1
194                 FW_SR(A,arow, acol,B,brow, bcol,C,crow, ccol, myN/2, bsize);
195                 //call3
196                 FW_SR(A,arow+myN/2, acol,B,brow+myN/2, bcol,C,crow, ccol, myN/2,
197                 bsize);
198                 //call16

```

```

193         FW_SR(A,arow+myN/2, acol,B,brow+myN/2, bcol+myN/2,C,crow+myN/2, ccol,
194         myN/2, bsize);
195             //call18
196             FW_SR(A,arow, acol,B,brow, bcol+myN/2,C,crow+myN/2, ccol, myN/2,
197             bsize);
198         }
199             //call12
200             FW_SR(A,arow, acol+myN/2,B,brow, bcol,C,crow, ccol+myN/2, myN/2,
201             bsize);
202             //call14
203             FW_SR(A,arow+myN/2, acol+myN/2,B,brow+myN/2, bcol,C,crow, ccol+myN/2,
204             myN/2, bsize);
205             //call15
206             FW_SR(A,arow+myN/2, acol+myN/2,B,brow+myN/2, bcol+myN/2,C,crow+myN/2,
207             ccol+myN/2, myN/2, bsize);
208             //call17
209             FW_SR(A,arow, acol+myN/2,B,brow, bcol+myN/2,C,crow+myN/2, ccol+myN/2,
210             myN/2, bsize);
211         }
212         #pragma omp taskwait
213     }
214     break;
215     case 3: //A separate from B and C
216     #pragma omp task firstprivate(arow,acol,brow,bcol,crow,ccol,myN,bsize)
217     shared(A,B,C)
218     {
219         //call11
220         FW_SR(A,arow, acol,B,brow, bcol,C,crow, ccol, myN/2, bsize);
221         //call18
222         FW_SR(A,arow, acol,B,brow, bcol+myN/2,C,crow+myN/2, ccol, myN/2, bsize);
223     }
224     #pragma omp task firstprivate(arow,acol,brow,bcol,crow,ccol,myN,bsize)
225     shared(A,B,C)
226     {
227         //call12
228         FW_SR(A,arow, acol+myN/2,B,brow, bcol,C,crow, ccol+myN/2, myN/2, bsize);
229         //call17
230         FW_SR(A,arow, acol+myN/2,B,brow, bcol+myN/2,C,crow+myN/2, ccol+myN/2,
231         myN/2, bsize);
232     }
233     #pragma omp task firstprivate(arow,acol,brow,bcol,crow,ccol,myN,bsize)
234     shared(A,B,C)
235     {
236         //call13
237         FW_SR(A,arow+myN/2, acol,B,brow+myN/2, bcol,C,crow, ccol, myN/2, bsize);
238         //call16
239         FW_SR(A,arow+myN/2, acol,B,brow+myN/2, bcol+myN/2,C,crow+myN/2, ccol,
240         myN/2, bsize);
241     }

```

```
235     #pragma omp task firstprivate(arow,acol,brow,bcol,crow,ccol,myN,bsize)
236     shared(A,B,C)
237     {
238         //call4
239         FW_SR(A,arow+myN/2, acol+myN/2,B,brow+myN/2, bcol,C,crow, ccol+myN/2,
240         myN/2, bsize);
241         //call5
242         FW_SR(A,arow+myN/2, acol+myN/2,B,brow+myN/2, bcol+myN/2,C,crow+myN/2,
243         ccol+myN/2, myN/2, bsize);
244         }
245         #pragma omp taskwait
246
247     }
248
249 /*
250 call1
251     FW_SR(A,arow, acol,B,brow, bcol,C,crow, ccol, myN/2, bsize);
252 call2
253     FW_SR(A,arow, acol+myN/2,B,brow, bcol,C,crow, ccol+myN/2, myN/2, bsize);
254 call3
255     FW_SR(A,arow+myN/2, acol,B,brow+myN/2, bcol,C,crow, ccol, myN/2, bsize);
256 call4
257     FW_SR(A,arow+myN/2, acol+myN/2,B,brow+myN/2, bcol,C,crow, ccol+myN/2, myN/2,
258     bsize);
259     FW_SR(A,arow+myN/2, acol+myN/2,B,brow+myN/2, bcol+myN/2,C,crow+myN/2, ccol+myN/2,
260     myN/2, bsize);
261 call6
262     FW_SR(A,arow+myN/2, acol,B,brow+myN/2, bcol+myN/2,C,crow+myN/2, ccol, myN/2,
263     bsize);
264 call7
265     FW_SR(A,arow, acol+myN/2,B,brow, bcol+myN/2,C,crow+myN/2, ccol+myN/2, myN/2,
266     bsize);
267     FW_SR(A,arow, acol,B,brow, bcol+myN/2,C,crow+myN/2, ccol, myN/2, bsize);
268 */
269
```

- **Αποτελέσματα Μετρήσεων Επίδοσης**

Αρχικά δοκιμάσαμε να τρέξουμε τον σειριακό (επαναληπτικό) αλγόριθμο. Ο χρόνος εκτέλεσης για N=1024 ήταν 1.3954.

Στη συνέχεια, τρέχοντας το πρόγραμμα με τις παραμέτρους της προηγούμενης παραγράφου στο sandman πήραμε τα εξής αποτελέσματα:

Χρόνοι εκτέλεσης

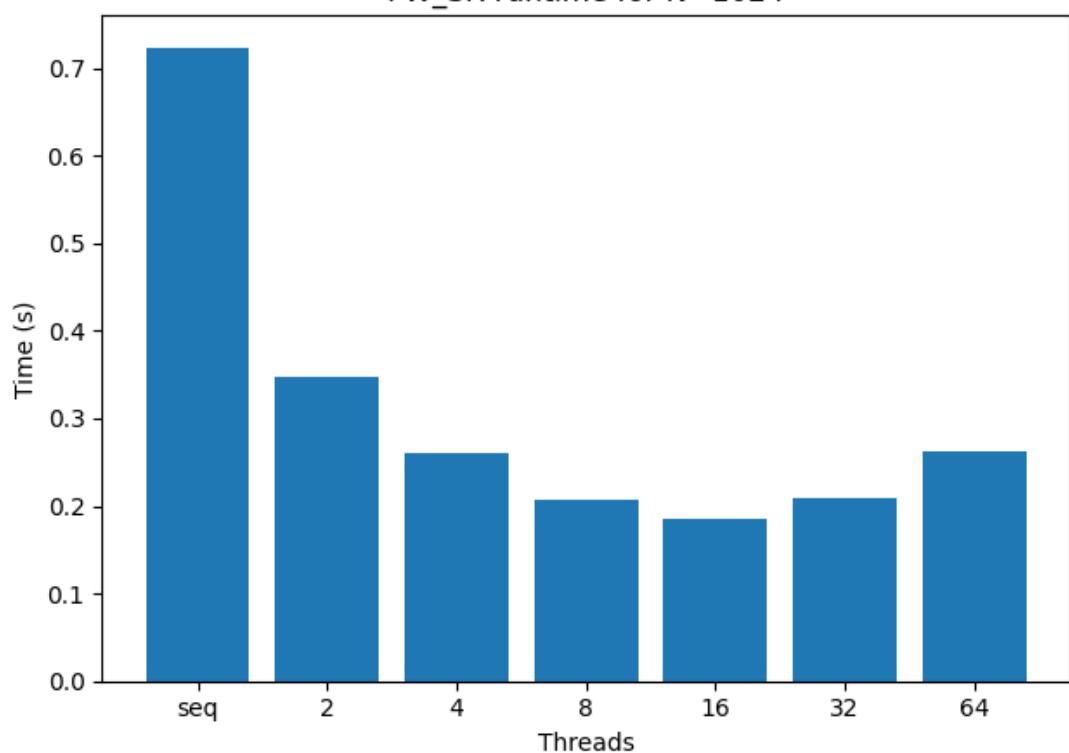
N	Tseq (T=1)	T=2	T=4	T=8	T=16	T=32	T=64
1024	0.7237	0.3467	0.2604	0.2078	0.1850	0.2097	0.2627
2048	5.1359	2.6523	1.9513	1.1305	0.7720	0.7496	0.9577
4096	43.4211	21.7275	14.3879	7.9806	4.4654	3.0423	3.1596

Συντελεστής επιτάχυνσης (speedup)

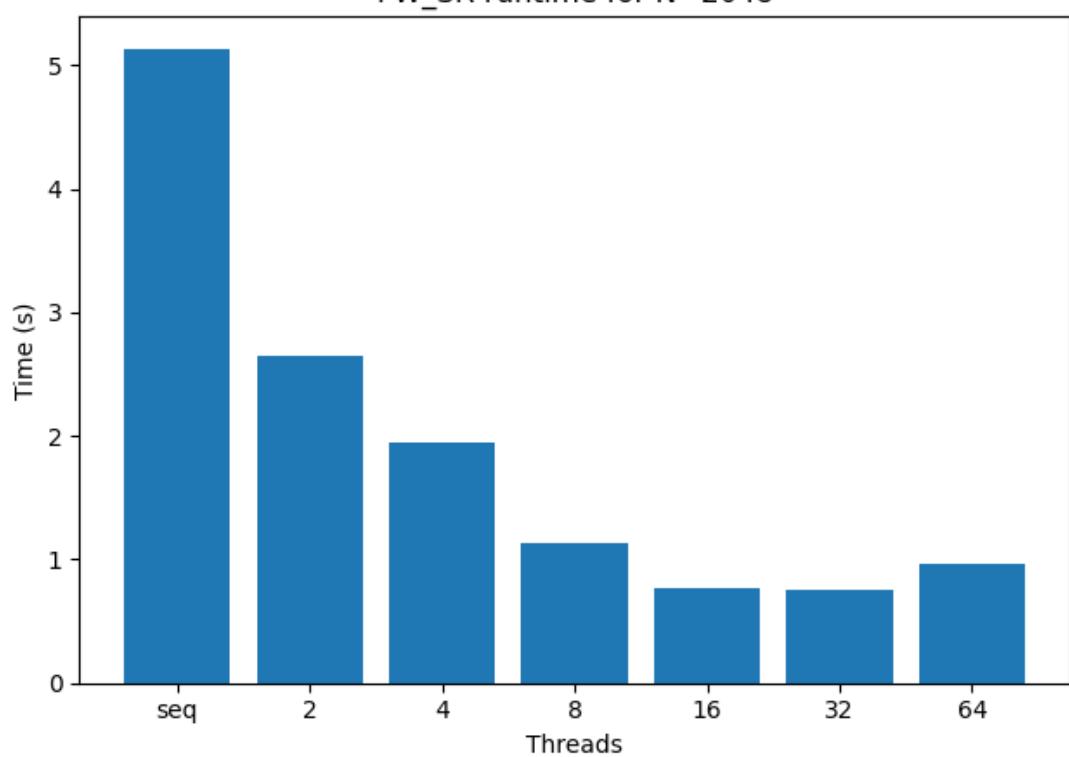
N	T=2	T=4	T=8	T=16	T=32	T=64	Max S
1024	2.09	2.78	3.48	3.91	3.45	2.75	3.91
2048	1.94	2.63	4.54	6.65	6.85	5.36	6.85
4096	2.00	3.02	5.44	9.72	14.27	13.74	14.27

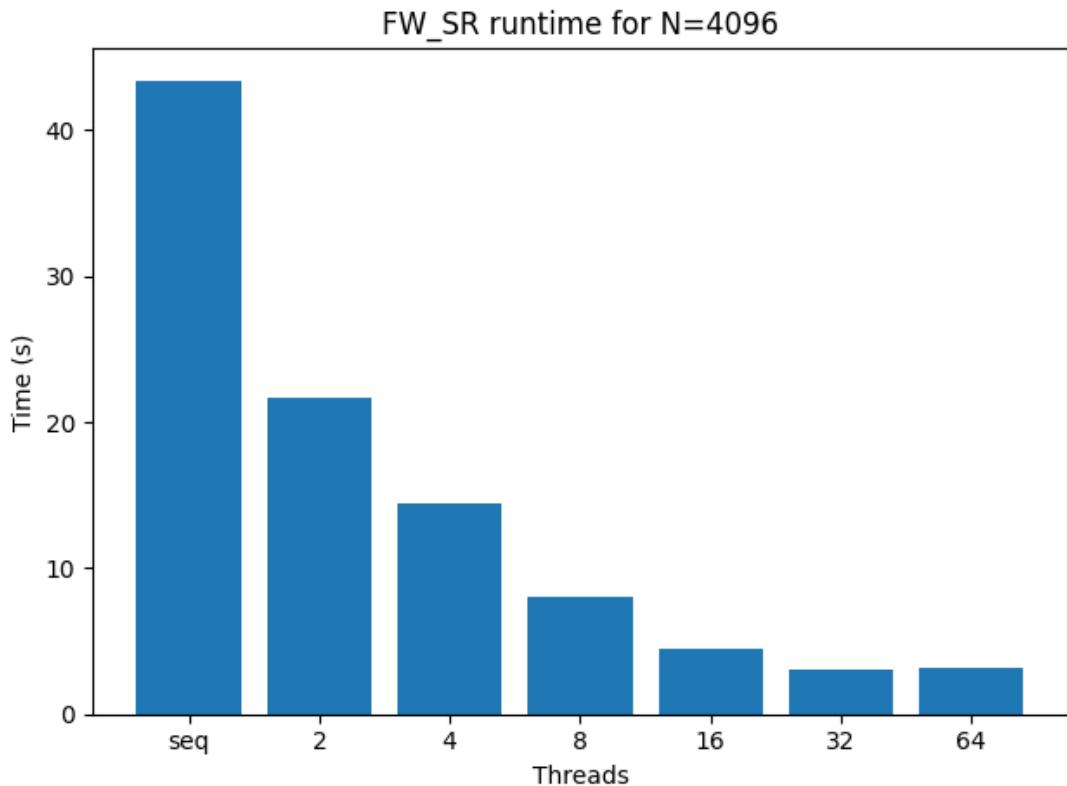
Ακολουθούν και τα barplots για τα runs των διαφορετικών N:

FW_SR runtime for N=1024



FW_SR runtime for N=2048





- **Συμπεράσματα και παρατηρήσεις**

- ✓ Αρχικά παρατηρούμε ότι η χρήση του σειριακού recursive αλγορίθμου έφερε από μόνη της σημαντική βελτίωση (από 1.3954 σε 0.7237). Αυτή οφείλεται πιθανώς στην βελτίωση της cache locality που επιτυγχάνεται με τη recursive δομή του αλγορίθμου. Αυτή η αναδρομική διάσπαση του πίνακα σε μικρότερα blocks επιτρέπει την πιο αποδοτική χρήση της ιεραρχίας της κρυφής μνήμης (cache hierarchy), μειώνοντας δραστικά τα misses.
- ✓ Η κλιμάκωση (speedup) είναι καλύτερη για το μεγαλύτερο dataset ($N=4096$), φτάνοντας το $14.27\times$ στους 32 threads. Αυτό συμβαίνει επειδή, για μεγάλο N , ο τεράστιος όγκος των υπολογισμών ($O(N^3)$) κυριαρχεί έναντι του overhead του OpenMP tasking και των καθυστερήσεων της μνήμης, επιτρέποντας την καλύτερη αξιοποίηση της παραλληλίας.

- ✓ Το ταβάνι στο scalability (Max Speedup) οφείλεται στους thread overheads και στη δομή του taskgraph, η οποία εισάγει περιορισμένη παραλληλία λόγω των εξαρτήσεων ιδιαίτερα στο case 1, όπου οι σειριακές εκτελέσεις μειώνονται απλά σε 6 από τις 8 αρχικές.
- ✓ Η αύξηση των threads στους 64 οδηγεί σε αύξηση του χρόνου εκτέλεσης σε όλες τις περιπτώσεις, υποδηλώνοντας ότι το overhead του tasking υπερβαίνει το όφελος της παραλληλοποίησης.

Σ.Η.Μ.Μ.Υ. Ε.Μ.Π.
Νοέμβριος 2025

ΣΥΣΤΗΜΑΤΑ ΠΑΡΑΛΛΗΛΗΣ ΕΠΕΞΕΡΓΑΣΙΑΣ

ΑΝΑΦΟΡΑ 2^{ης} ΑΣΚΗΣΗΣ



Στοιχεία Ομάδας

- Αναγνωριστικό: parlab05
- Μέλος 1^ο: Πέππας Μιχαήλ – Αθανάσιος, Α.Μ: 03121026
- Μέλος 2^ο: Σαουνάτσος Ανδρέας, Α.Μ: 03121197
- Ημερομηνία Παράδοσης Αναφοράς: 25.11.2025

▪ Αμοιβαίος Αποκλεισμός – Κλειδώματα

1. Εισαγωγή και Αρχεία

Στην παρούσα άσκηση μελετάμε διάφορους μηχανισμούς locks σε πολυνηματικές εφαρμογές, μέσω του αλγορίθμου ταξινόμησης K-means (στη shared έκδοσή του). Στόχος μας, είναι να κατανοήσουμε πώς διαφορετικές στρατηγικές συγχρονισμού (απουσία κλειδώματος, κλασικά mutex και spinlocks, locks τύπου TAS/TTAS, ιεραρχικά κλειδώματα τύπου array και CLH, καθώς και η εντολή critical του OpenMP) επηρεάζουν τον χρόνο εκτέλεσης και τη δυνατότητα κλιμάκωσης της εφαρμογής, όταν αυτή εκτελείται σε ένα πολυπύρηνο σύστημα με μέχρι και 64 λογικά νήματα (το sandman έχει $4 \times 8 = 32$ φυσικούς πυρήνες και $32 \times 2 = 64$ λογικούς). Το κυρίως σώμα του προγράμματος παραμένει το ίδιο και μεταβάλλεται μόνο ο μηχανισμός του lock, ώστε να μπορούμε να απομονώσουμε και να συγκρίνουμε το καθαρό κόστος συγχρονισμού κάθε κλειδώματος.

Ο δοσμένος κώδικας βασίζεται στον K-means από την προηγούμενη άσκηση και αποτελείται από τα κοινά αρχεία file_io.c και util.c, το header file kmeans.h και το πρόγραμμα main.c, το οποίο αναλαμβάνει να διαβάσει τις παραμέτρους και να καλέσει την υλοποίηση του αλγορίθμου. Το αρχείο seq_kmeans.c περιέχει τη σειριακή έκδοση του K-means, ενώ το omp_naive_kmeans.c παρέχει μια απλή παράλληλη προσέγγιση με OpenMP. Για τη μελέτη των κλειδωμάτων χρησιμοποιούνται δύο επιπλέον αρχεία: το omp_critical_kmeans.c, όπου ο συγχρονισμός υλοποιείται με #pragma omp critical, και το omp_lock_kmeans.c, το οποίο ορίζει την παράλληλη εκδοχή που βασίζεται σε μια διεπαφή με κλειδώματα, ως παράμετρο. Η υλοποίηση των επιμέρους κλειδωμάτων γίνεται στον φάκελο locks/, μέσω των αρχείων nosync_lock.c, pthread_mutex_lock.c, pthread_spin_lock.c, tas_lock.c, ttas_lock.c, array_lock.c και clh_lock.c, τα οποία μοιράζονται το κοινό header lock.h (ορισμός του τύπου lock_t και των συναρτήσεων lock_init, lock_acquire, lock_release, lock_free) και το αρχείο alloc.h για τις βοηθητικές δεσμεύσεις μνήμης. Έτσι, με κατάλληλο linking στο Makefile προκύπτουν τα διαφορετικά εκτελέσιμα (kmeans_omp_nosync_lock, kmeans_omp(pthread_mutex_lock, kmeans_omp(pthread_spin_lock, kmeans_omp(tas_lock, kmeans_omp(ttas_lock, kmeans_omp(array_lock, kmeans_omp(clh_lock), τα οποία μοιράζονται τον ίδιο κώδικα του K-means αλλά χρησιμοποιούν διαφορετικό μηχανισμό lock.

Το Makefile του φακέλου έχει προσαρμοστεί ώστε να συνθέτει όλα τα παραπάνω εκτελέσιμα, αξιοποιώντας τις κοινές πηγές file_io.c και util.c και προσθέτοντας κάθε φορά την αντίστοιχη υλοποίηση κλειδώματος από τον υποφάκελο locks/. Για τα κλειδώματα που βασίζονται σε POSIX mutex ή spinlocks ενεργοποιείται επιπλέον η παράμετρος -pthread, ενώ όλα τα παράλληλα εκτελέσιμα μεταγλωττίζονται με τις σημαίες του OpenMP (-fopenmp). Παράλληλα, το πρόγραμμα make_on_queue.sh έχει τροποποιηθεί, ώστε να μετακινείται στον σωστό φάκελο της άσκησης (cd /home/parallel/parlab05/a3) και να εκτελεί εκεί την εντολή make. Με αυτόν τον τρόπο εξασφαλίζουμε ότι όλες οι εκδόσεις του K-means (σειριακή, naive, critical και με κλειδώματα) μεταγλωττίζονται απευθείας στο περιβάλλον του sandman, πριν από τη λήψη των μετρήσεων.

Τέλος, το πρόγραμμα run_on_queue.sh είναι υπεύθυνο για την αυτοματοποιημένη εκτέλεση των πειραμάτων στο μηχάνημα του sandman. Στην τελική του μορφή το script τρέχει όλα τα εκτελέσιμα κλειδωμάτων και για όλους τους αριθμούς νημάτων που ζητούνται (1, 2, 4, 8, 16, 32, 64). Για κάθε συνδυασμό κλειδώματος και αριθμό νημάτων, θέτει κατάλληλα τη μεταβλητή OMP_NUM_THREADS και ορίζει την πολιτική δέσμευσης πυρήνων μέσω της GOMP_CPU_AFFINITY, ώστε τα νήματα να κατανέμονται στους πρώτους φυσικούς πυρήνες του συστήματος (εκτέλεση πάντα «με affinity», όπως απαιτείται). Τα αποτελέσματα κάθε εκτέλεσης αποθηκεύονται σε ξεχωριστό κατάλογο της μορφής benchmarks/<lock_name>/S32_N16_C32_L10_T<threads>/, όπου δημιουργούνται αρχεία meta.txt με τα μεταδεδομένα της εκτέλεσης (εκτελέσιμο, τύπος κλειδώματος, παράμετροι K-means, αριθμός νημάτων, τιμές affinity) και output.txt με την πλήρη έξοδο του προγράμματος, συμπεριλαμβανομένης της γραμμής με τον χρόνο εκτέλεσης και τον αριθμό επαναλήψεων. Έτσι, οι επόμενες ενότητες μπορούν να επεξεργαστούν τα δεδομένα των benchmarks με τρόπο αντίστοιχο της προηγούμενης άσκησης και να παράγουν συγκρίσιμα διαγράμματα για όλες τις υλοποίησεις κλειδωμάτων.

Για λόγους πληρότητας, το τροποποιημένο run_on_queue.sh (μόνο αυτό διαφοροποιήθηκε σημαντικά) παρουσιάζεται ακολούθως:

a3/run_on_queue.sh

```

1  #!/bin/bash
2
3  #PBS -N run_kmeans_locks
4  #PBS -o run_kmeans_locks.out
5  #PBS -e run_kmeans_locks.err
6  #PBS -l nodes=1:ppn=64
7  #PBS -l walltime=01:00:00
8
9  ## How to submit (runs all locks x all thread configs on sandman):
10 ##   qsub -q serial -l nodes=sandman:ppn=64 run_on_queue.sh
11 ##
12 ## Defaults (can be overridden via -v):
13 ##   SIZE=32
14 ##   COORDS=16
15 ##   CLUSTERS=32
16 ##   LOOPS=10
17
18 set -euo pipefail
19
20 # Work in the directory where qsub was executed (your a3 folder)
21 cd "${PBS_O_WORKDIR:-.}" || exit 1
22
23 # Fixed configuration required by the exercise (override with env if needed)
24 SIZE="${SIZE:-32}"
25 COORDS="${COORDS:-16}"
26 CLUSTERS="${CLUSTERS:-32}"
27 LOOPS="${LOOPS:-10}"
28
29 # Thread configurations to test
30 THREADS_LIST=(1 2 4 8 16 32 64)
31
32 # Lock variants (names as they appear in the binary targets)
33 LOCKS=(
34   "nosync_lock"
35   "pthread_mutex_lock"
36   "pthread_spin_lock"
37   "tas_lock"
38   "ttas_lock"
39   "array_lock"
40   "clh_lock"
41 )
42
43 run_one() {
44   local lock_name="$1"
45   local threads="$2"
46   local bin=""
47
48   if [[ "$lock_name" == "critical" ]]; then
49     # OpenMP critical version
50     bin="kmeans_omp_critical"
51   else

```

```

52     # Lock-based versions (built from omp_lock_kmeans.c + one lock object)
53     bin="kmeans_omp_${lock_name}"
54   fi
55
56   if [[ ! -x "./${bin}" ]]; then
57     echo "[WARN] Skipping lock='${lock_name}', threads=${threads}: binary '${bin}' not
58   found"
59   return
60   fi
61
62   # OpenMP settings
63   export OMP_NUM_THREADS="${threads}"
64
65   # Always use thread binding (affinity) as required
66   local affinity=""
67   for ((i=0; i<threads; i++)); do
68     affinity+="${i} "
69   done
70   affinity="${affinity%% }"
71   export GOMP_CPU_AFFINITY="${affinity}"
72
73   # Result directory:
74   # benchmarks/<lock_name>/S32_N16_C32_L10_T8/
75   local
76   result_dir="benchmarks/${lock_name}/S${SIZE}_N${COORDS}_C${CLUSTERS}_L${LOOPS}_T${threads}"
77   mkdir -p "${result_dir}"
78
79   {
80     echo "[run_on_queue] BIN=${bin}"
81     echo "[run_on_queue] LOCK=${lock_name}"
82     echo "[run_on_queue] OMP_NUM_THREADS=${OMP_NUM_THREADS}"
83     echo "[run_on_queue] GOMP_CPU_AFFINITY=${GOMP_CPU_AFFINITY}"
84     echo "[run_on_queue] Params: -s ${SIZE} -n ${COORDS} -c ${CLUSTERS} -l ${LOOPS}"
85     echo "[run_on_queue] Result dir: ${result_dir}"
86   } > "${result_dir}/meta.txt"
87
88   echo "[INFO] Running lock='${lock_name}', threads=${threads}, bin='${bin}'"
89   ./"${bin}" -s "${SIZE}" -n "${COORDS}" -c "${CLUSTERS}" -l "${LOOPS}" \
90   | tee "${result_dir}/output.txt"
91
92   # 1) Run all lock implementations (omp_lock_kmeans.c + locks/)
93   for lock in "${LOCKS[@]}"; do
94     for t in "${THREADS_LIST[@]}"; do
95       run_one "${lock}" "${t}"
96     done
97   done
98
99   # 2) Run the critical version (omp_critical_kmeans.c → kmeans_omp_critical)
100  for t in "${THREADS_LIST[@]}"; do
101    run_one "critical" "${t}"
102  done
103

```

2. Λήψη Μετρήσεων και Διαγράμματα

Στη συνέχεια παρουσιάζουμε τους πίνακες με τα αποτελέσματα όλων των μετρήσεων που προέκυψαν από την εκτέλεση του K-means για τις παραμέτρους εισόδου SIZE=32, COORDS=16, CLUSTERS=32 και LOOPS=10 και για αριθμό νημάτων $T \in \{1, 2, 4, 8, 16, 32, 64\}$. Παραθέτουμε έναν πίνακα για κάθε μηχανισμό (συμπεριλαμβανομένου του critical). Με αυτόν τον τρόπο μπορούμε να συγκρίνουμε άμεσα το κόστος συγχρονισμού κάθε μηχανισμού και πώς αυτό κλιμακώνεται καθώς αυξάνεται ο βαθμός παραλληλισμού:

nosync_lock:

Threads	Total Time
1	1.3238
2	0.7005
4	0.4084
8	0.2984
16	0.7346
32	1.3143
64	0.9590

pthread_mutex_lock:

Threads	Total Time
1	1.3241
2	0.8372
4	0.7302
8	0.8456
16	2.6544
32	3.7170
64	3.8797

pthread_spin_lock:

Threads	Total Time
1	1.3174
2	0.7447
4	0.5589
8	0.8860
16	3.3755
32	7.7281
64	3.7769

tas_lock:

Threads	Total Time
1	1.3291
2	0.7396
4	0.5305
8	1.1465
16	4.5320
32	10.6944
64	9.3531

ttas_lock:

Threads	Total Time
1	1.3151
2	0.7513
4	0.5621
8	0.8613
16	3.2426
32	7.3978
64	4.3841

array_lock:

Threads	Total Time
1	1.3585
2	0.8212
4	0.5121
8	0.4942
16	1.5424
32	2.2050
64	2.1843

clh_lock:

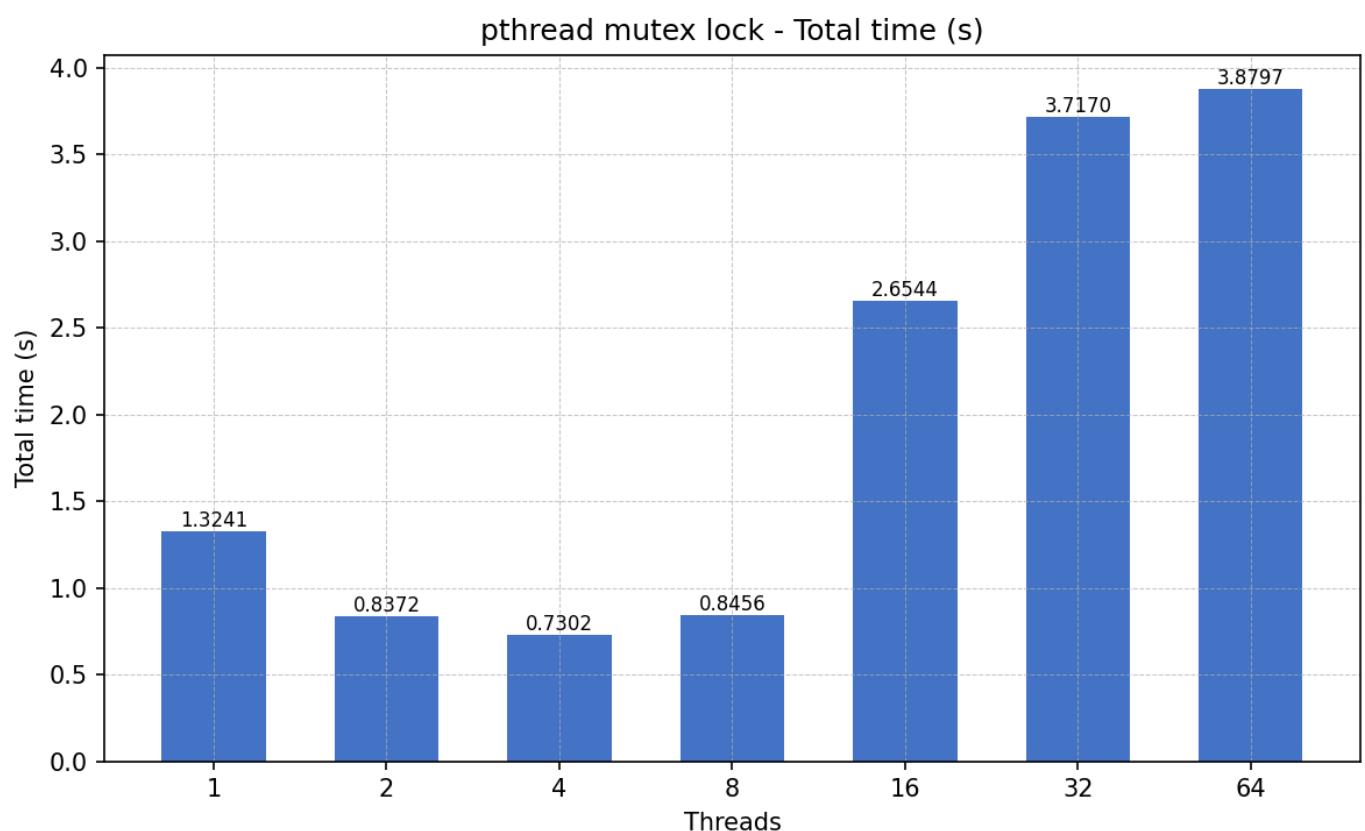
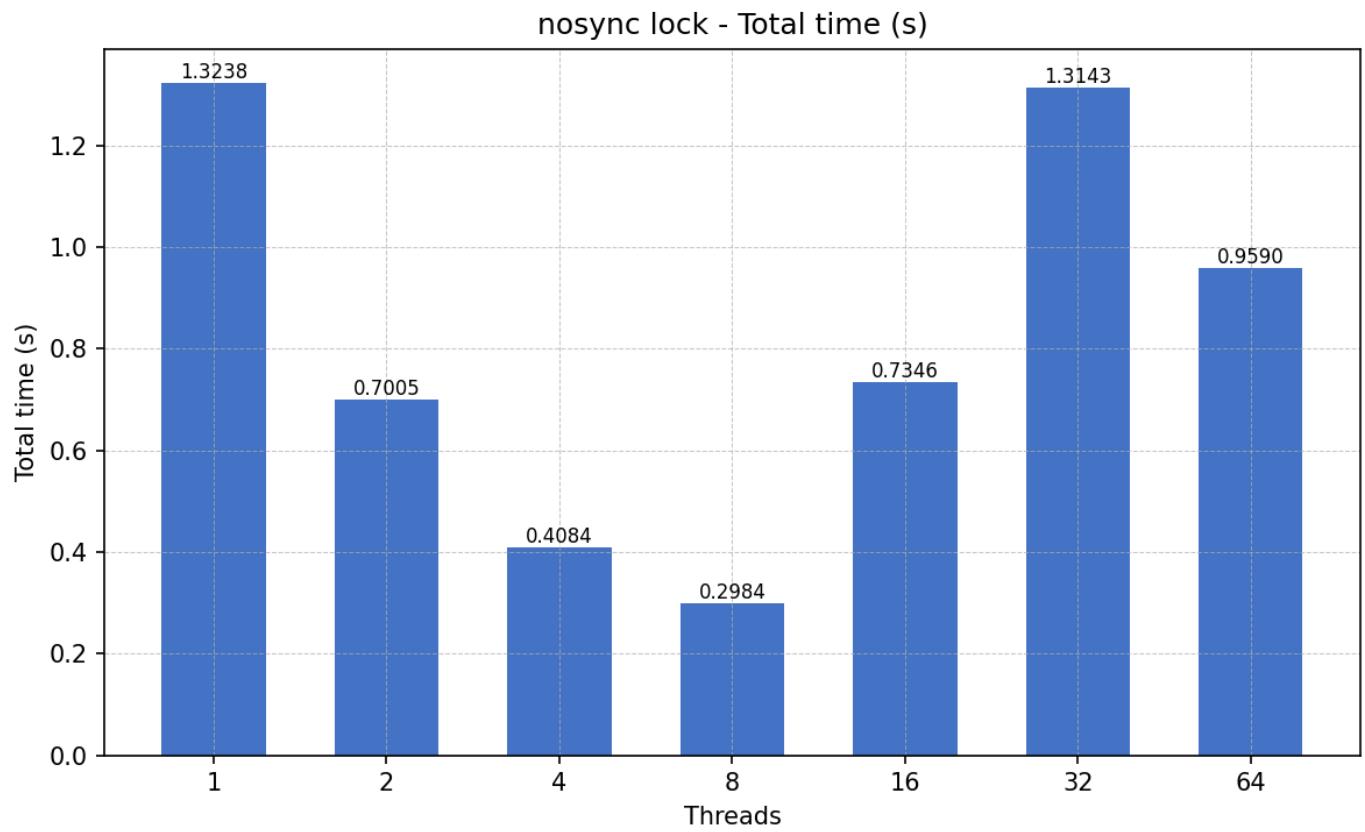
Threads	Total Time
1	1.3182
2	0.7995
4	0.4884
8	0.4319
16	1.2522
32	1.6519
64	1.4726

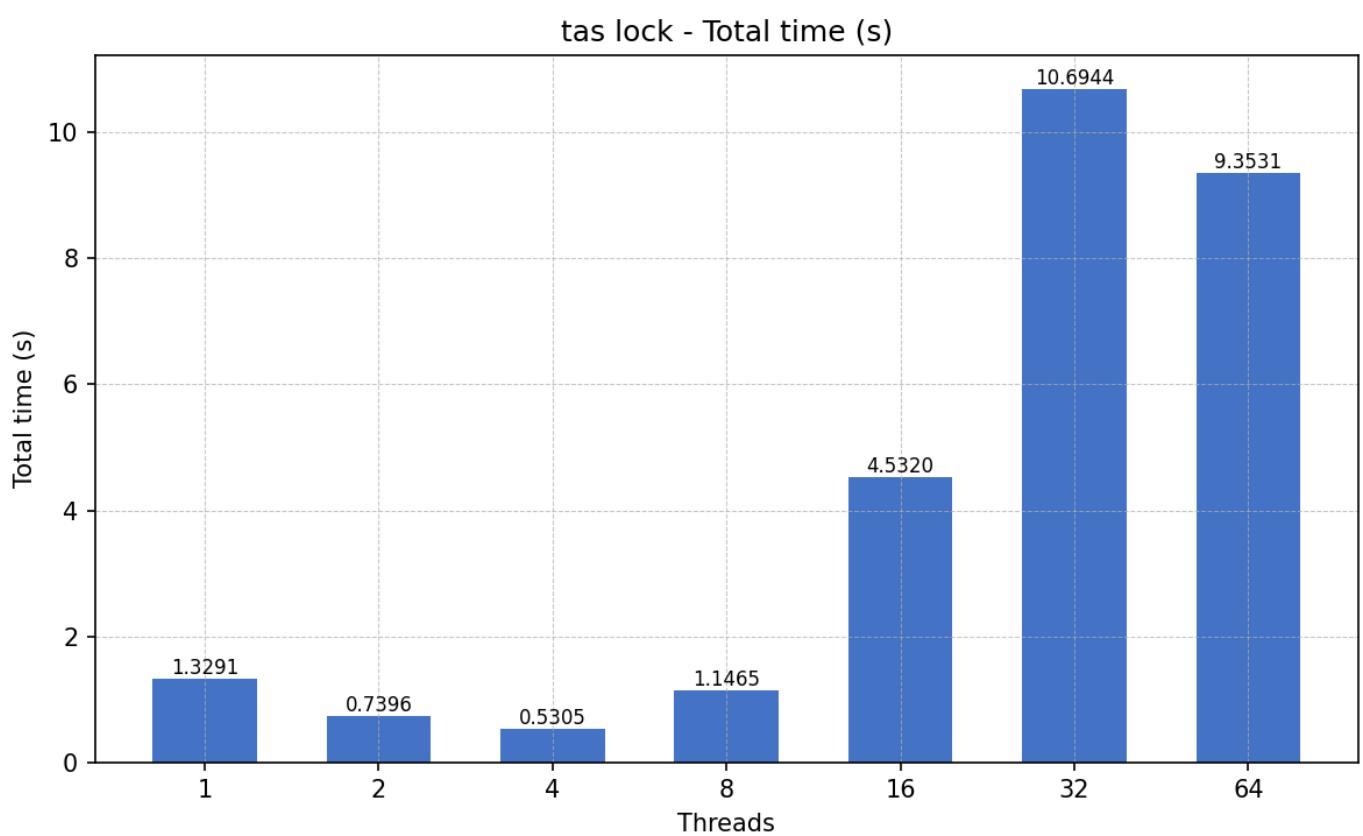
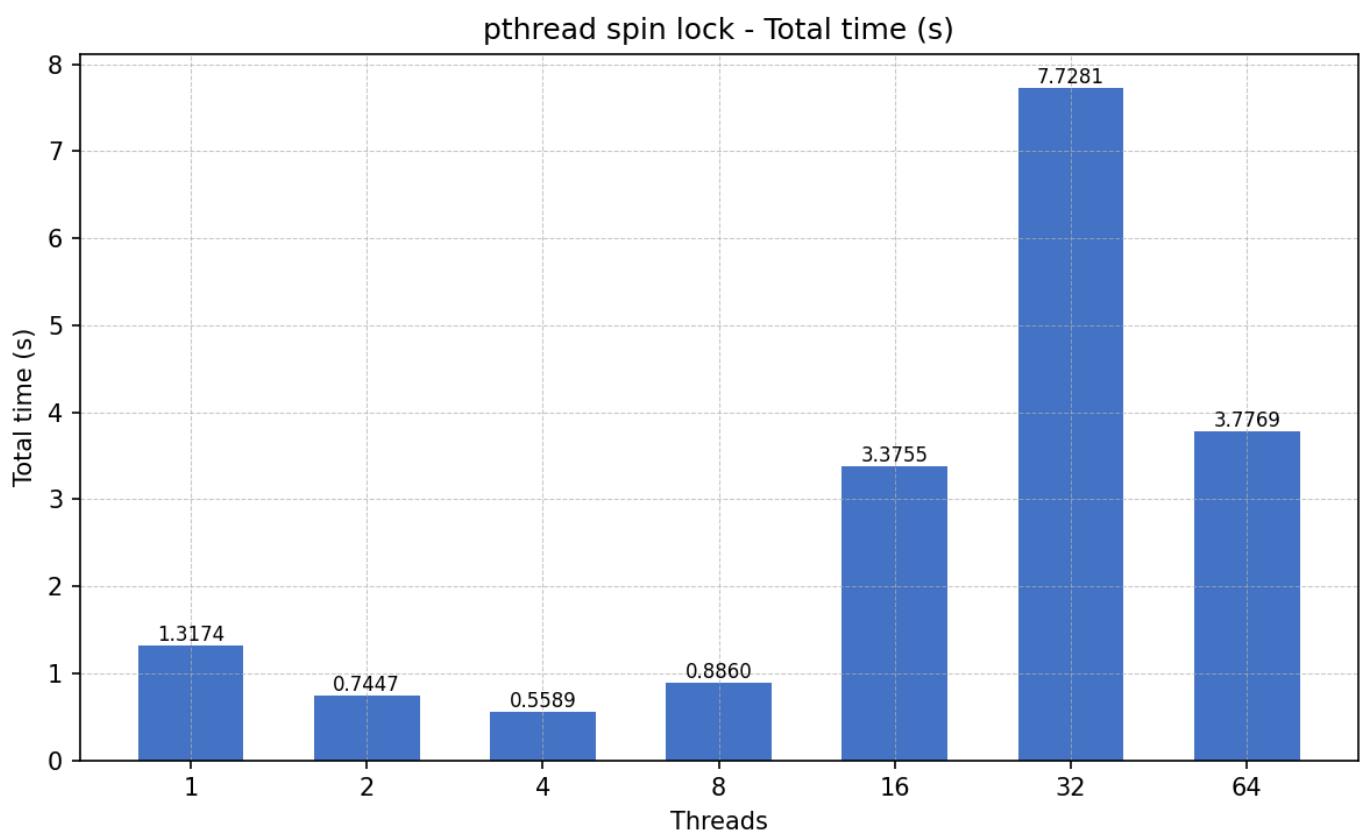
critical:

Threads	Total Time
1	1.3187
2	0.7827
4	0.4916
8	0.7655
16	3.4095
32	7.9911
64	6.1087

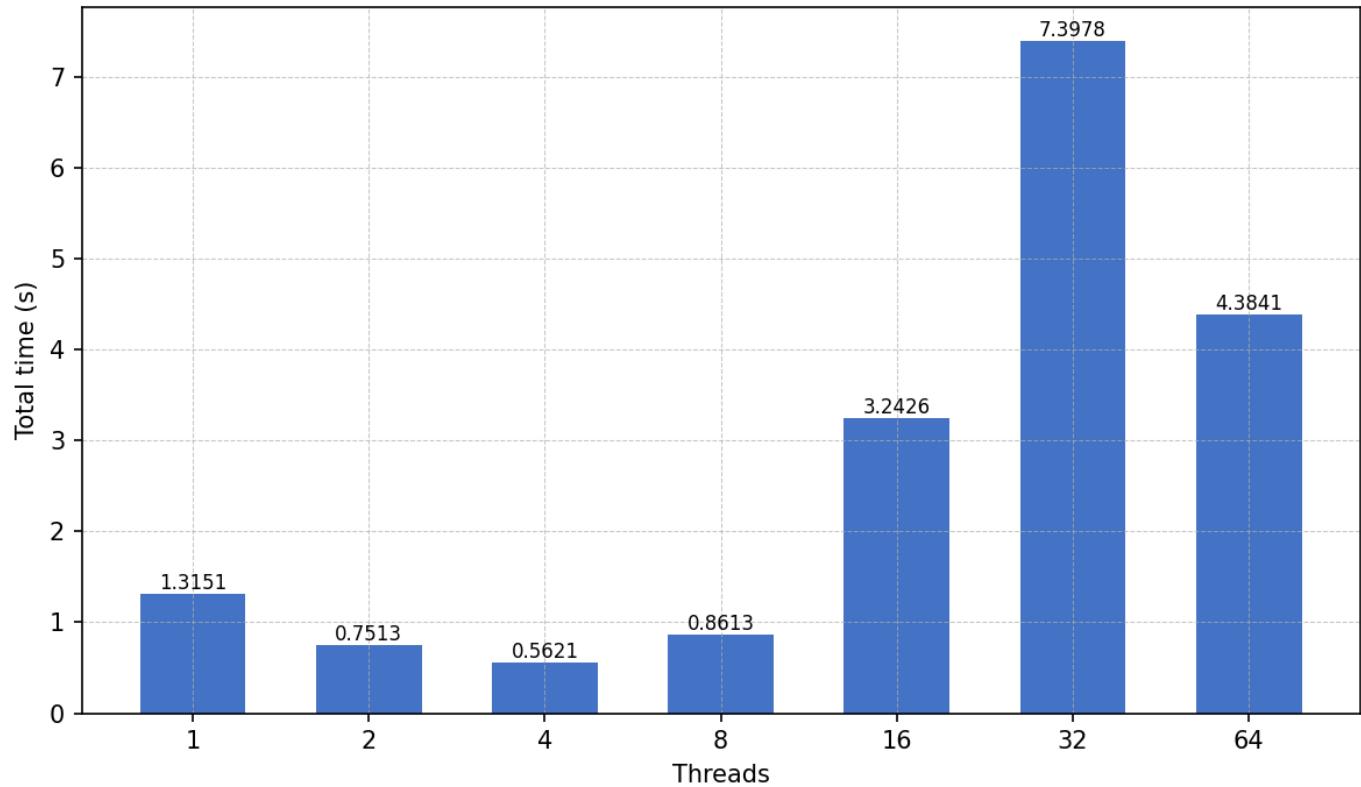
Για να είναι πιο ευδιάκριτες οι διαφορές μεταξύ των locks, παρουσιάζουμε τα αποτελέσματα και σε διαγράμματα. Στα διαγράμματα που ακολουθούν απεικονίζουμε τον χρόνο εκτέλεσης σε συνάρτηση με τον αριθμό νημάτων, για κάθε υλοποίηση κλειδώματος, καθώς και ένα συγκεντρωτικό διάγραμμα για όλους τους τύπους μαζί. Τα διαγράμματα αυτά μας επιτρέπουν να αξιολογήσουμε ποια

κλειδώματα παραμένουν αποδοτικά υπό αυξημένη συμφόρηση, ποια εμφανίζουν κορεσμό λόγω κόστους συγχρονισμού, καθώς και πώς συγκρίνεται η χρήση critical sections του OpenMP με τις υπόλοιπες υλοποιήσεις κλειδώματος:

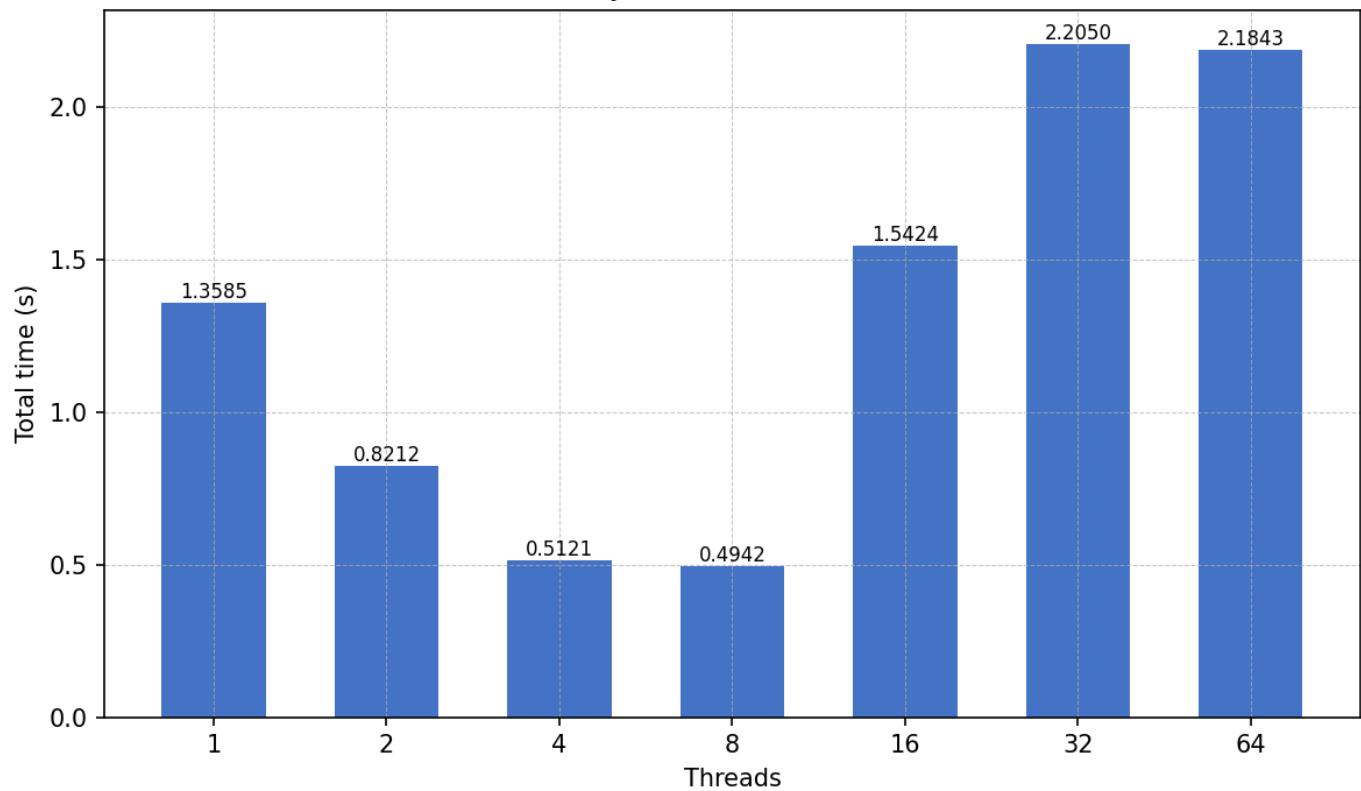


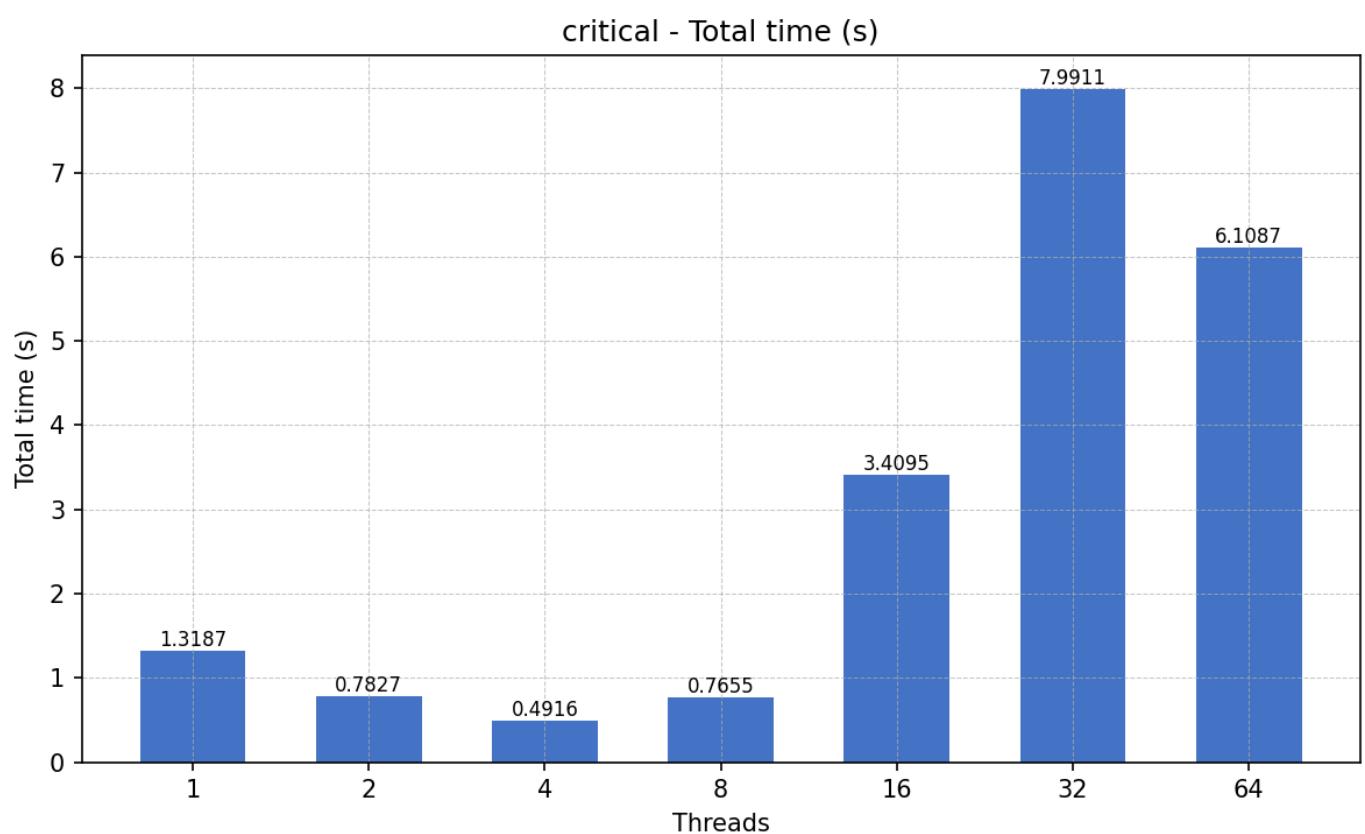
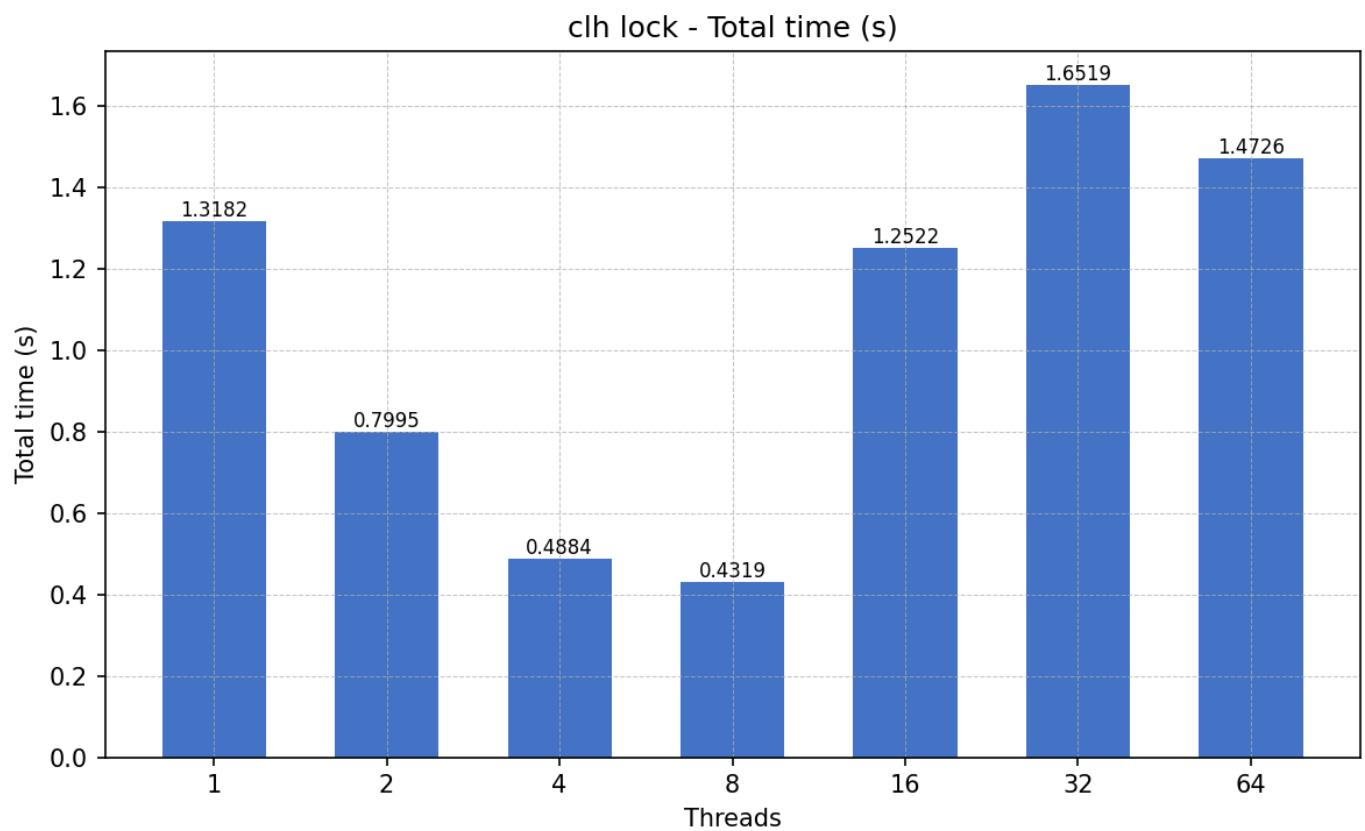


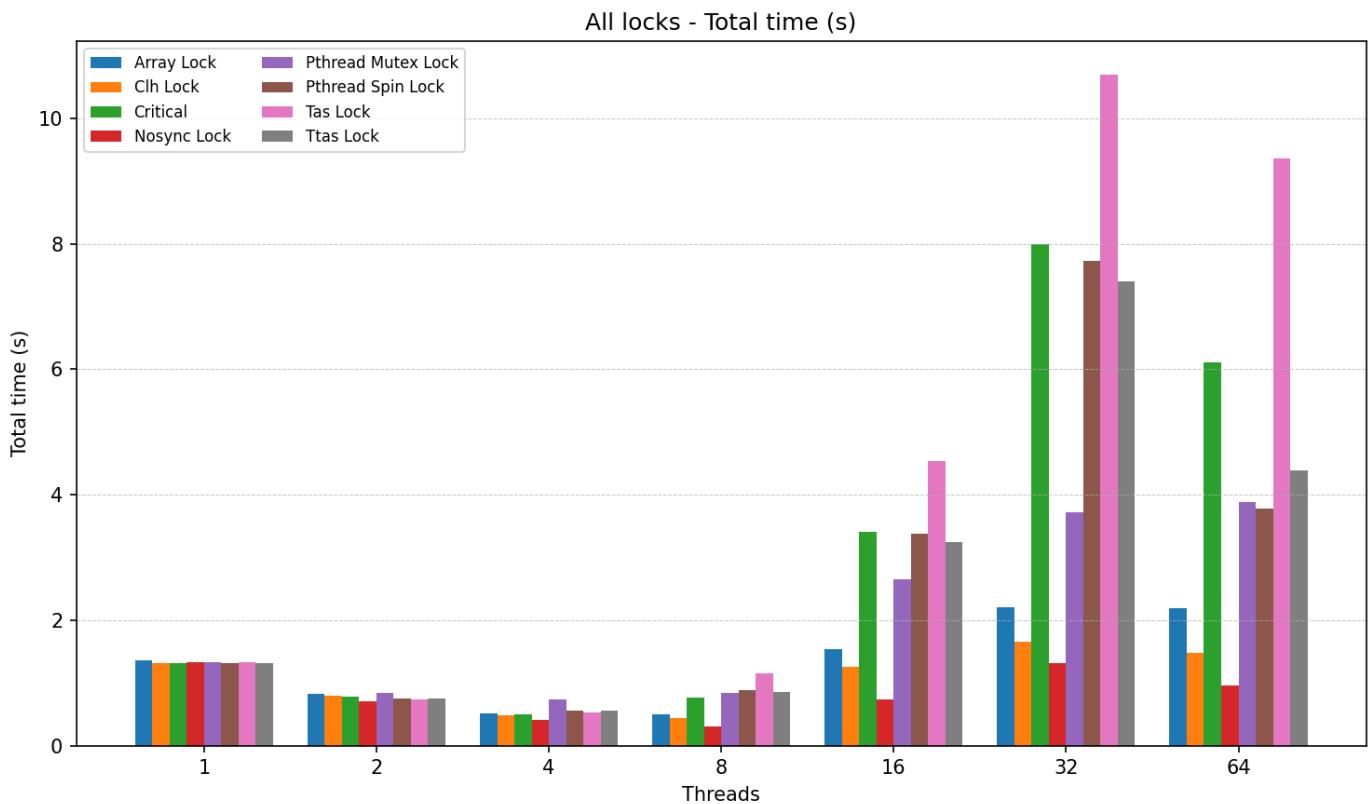
ttas lock - Total time (s)



array lock - Total time (s)







3. Σύγκριση Αποτελεσμάτων

Ξεκινώντας από τη σύγκριση μεταξύ των κλειδωμάτων, για $T=1$ όλοι οι μηχανισμοί (nosync, mutex, spin, TAS/TTAS, array, CLH, omp_critical) έχουν πολύ κοντινούς χρόνους, αφού δεν υπάρχει ουσιαστικό contention και το κόστος του lock είναι αμελητέο. Καθώς αυξάνουμε τα νήματα, ο «ιδανικός» δείκτης nosync_lock και τα queue-based locks (array_lock, clh_lock) βελτιώνουν τον χρόνο μέχρι περίπου τα 8 νήματα και από εκεί και πέρα χειροτερεύουν ήπια. Αντίθετα, τα pthread_mutex_lock, pthread_spin_lock, TAS/TTAS και το omp_critical_kmeans αρχίζουν να πέφτουν γρήγορα: για 8–16 νήματα ο χρόνος σταθεροποιείται ή αυξάνεται και για περισσότερα αυξάνεται ραγδαία, καθώς το κρίσιμο τμήμα σειριοποιείται και το lock γίνεται το βασικό bottleneck.

Η συμπεριφορά αυτή συνδέεται και με τη φύση του ίδιου του προβλήματος: στο configuration {32,16,32,10} ο K-means είναι έντονα memory-bound, με περιορισμένο υπολογιστικό φορτίο ανά στοιχείο (compute intensity). Όταν προσθέτουμε ένα «βαρύ» lock πάνω σε ένα μικρό, κυρίως memory-bound workload, το κόστος συγχρονισμού κυριαρχεί, ειδικά όταν πολλά νήματα προσπαθούν να μπουν στο ίδιο critical section. Τα queue locks περιορίζουν καλύτερα το contention

(λιγότερη τυχαία κίνηση σε κοινές cache lines) και γι' αυτό αντέχουν περισσότερο σε υψηλό παραλληλισμό από ό,τι τα απλά spinlocks ή το critical section.

Σε σχέση με την προηγούμενη άσκηση, τα ευρήματα είναι απολύτως συνεπή. Η shared-clusters υλοποίηση με `#pragma omp atomic` είχε λίγο κέρδος μέχρι περίπου τα 8 νήματα και μετά υπόκειται σε κορεσμό, ακριβώς επειδή πολλές `atomic` ενημερώσεις σε λίγες κοινόχρηστες δομές σειριαποιούσαν μεγάλο μέρος της δουλειάς. Στην τωρινή άσκηση, το `omp_critical_kmeans` και τα απλά spin/TAS locks εμφανίζουν την ίδια συμπεριφορά: λειτουργικά σωστά, αλλά με περιορισμένη κλιμάκωση λόγω έντονου synchronization και memory contention.

Αντίθετα, η λύση της προηγούμενης άσκησης με copied clusters και explicit reduction πέτυχε πολύ καλύτερη κλιμάκωση γιατί μείωσε δραστικά το fine-grained synchronization: κάθε νήμα δούλευε σε ιδιωτικές δομές και συγχρονιζόταν μόνο στη φάση του reduction. Τα queue locks της τωρινής άσκησης κινούνται προς αυτή την κατεύθυνση (περιορίζουν το contention και βελτιώνουν την shared υλοποίηση), αλλά δεν μπορούν να φτάσουν το επίπεδο της αλγορίθμικής βελτίωσης του copied clusters + reduction, ιδίως σε ένα μικρό και τόσο memory-bound πρόβλημα.

'Όπως τονίζεται και στις διαφάνειες του μαθήματος, η επίδοση κάθε μηχανισμού κλειδώματος εξαρτάται κρίσιμα από τον αριθμό νημάτων, το μέγεθος του κρίσιμου/μη κρίσιμου τμήματος και το επίπεδο συμφόρησης. Δεν υπάρχει ένα «καθολικά καλύτερο» κλείδωμα, αλλά διαφορετικοί μηχανισμοί που αποδίδουν καλύτερα σε διαφορετικά σενάρια.

4. Τα Κλειδώματα

Κάθε κλείδωμα υλοποιεί τον αμοιβαίο αποκλεισμό με διαφορετικό τρόπο και αυτό έχει άμεσο αντίκτυπο στην επίδοση, ειδικά σε έντονα memory-bound πρόβλημα, όπως το K-means της άσκησης.

Το `nosync_lock` δεν είναι πραγματικό κλείδωμα. Ουσιαστικά δεν κάνει καμία προστασία του κρίσιμου τμήματος και χρησιμοποιείται μόνο ως θεωρητικό baseline. Γι' αυτό εμφανίζει τους καλύτερους χρόνους: δεν υπάρχει overhead συγχρονισμού, άρα όλα τα νήματα γράφουν χωρίς συντονισμό πάνω στις κοινόχρηστες δομές. Στην πράξη όμως η υλοποίηση αυτή δεν είναι ορθά συγχρονισμένη και μπορεί να οδηγήσει σε λανθασμένα αποτελέσματα. Γι' αυτό χρησιμοποιείται μόνο για να απομονώσουμε το «καθαρό» κόστος του locking.

To `pthread_mutex_lock` υλοποιεί ένα blocking lock σε επίπεδο POSIX. Όταν υπάρχει μικρό contention, η απόδοσή του είναι σχετικά καλή, αλλά μόλις πολλαπλά νήματα αρχίσουν να μπλοκάρουν πάνω στο ίδιο mutex, ο μηχανισμός αναγκάζεται να κάνει system calls και πιθανές μεταβάσεις σε kernel space, με context switches κ.λπ. Σε ένα σενάριο με πολλές, μικρής διάρκειας κρίσιμες περιοχές (όπως οι ενημερώσεις στους μετρητές του K-means) αυτό το overhead γίνεται δυσανάλογα μεγάλο και εξηγεί γιατί το mutex δεν κλιμακώνεται καλά για $T \geq 8$. Το `pthread_spin_lock`, από την άλλη, είναι spin lock: τα νήματα δεν αποκοιμούνται αλλά κάνουν busy-wait σε μια μεταβλητή. Αυτό αποφεύγει τις ακριβές μεταβάσεις στο kernel όταν το lock κρατιέται για πολύ λίγο χρόνο, αλλά υπό έντονο contention το busy-wait καταναλώνει CPU cycles και δημιουργεί μεγάλη πίεση στο memory system, αφού όλα τα νήματα διαβάζουν/γράφουν την ίδια cache line – γι' αυτό βλέπουμε χειροτέρευση χρόνου για πολλά νήματα.

Τα `tas_lock` και `ttas_lock` είναι απλές υλοποιήσεις spin lock σε επίπεδο user-space. To TAS (test-and-set) κάνει συνεχές atomic γράψιμο σε μια κοινή μεταβλητή μέχρι να πάρει το lock, με αποτέλεσμα να προκαλεί μεγάλο traffic στο bus και να «πετάει» την cache line από πυρήνα σε πυρήνα. To TTAS (test-and-test-and-set) βελτιώνει την κατάσταση: τα νήματα πρώτα κάνουν απλό read (χωρίς write) και μόνο όταν φαίνεται ότι το lock είναι ελεύθερο επιχειρούν atomic test-and-set. Αυτό μειώνει κάπως τον αριθμό των writes, αλλά όλα τα νήματα εξακολουθούν να γυρίζουν γύρω από την ίδια cache line. Συνεπώς, σε χαμηλό contention τα TAS/TTAS είναι σχετικά ελαφριά, ενώ σε υψηλό contention έχουν πολύ κακή κλιμάκωση λόγω τρικυμίας στο πρωτόκολλο συνοχής της μνήμης.

Τα `array_lock` και `clh_lock` είναι queue-based locks και εδώ φαίνεται η διαφορά τους στην επίδοση. Στο array lock κάθε νήμα παίρνει μια θέση σε έναν «κυκλικό» πίνακα και περιμένει σε δικό του κελί, έτσι ώστε μόνο ένας μικρός αριθμός cache lines να αλλάζει χέρια κάθε φορά. Στο CLH lock κάθε νήμα δημιουργεί έναν κόμβο σε μια λογική ουρά και σπινάρει πάνω σε μια μεταβλητή που ανήκει στον «προκάτοχό» του. Και στις δύο περιπτώσεις, το spinning γίνεται πάνω σε τοπικά ή σχεδόν τοπικά δεδομένα, με λιγότερη ανταλλαγή cache lines ανά απόκτηση lock και με εγγενή δικαιοσύνη (FIFO σειρά). Αυτό εξηγεί γιατί τα queue locks είχαν πολύ καλύτερη συμπεριφορά για $T=8-32$: μειώνουν δραστικά το lock contention που βλέπαμε στα TAS/TTAS και spin locks, και δεν «πνίγουν» τον αλγόριθμο με άσκοπο coherence traffic.

Τέλος, το `omp_critical_kmeans` χρησιμοποιεί την εντολή `#pragma omp critical`, η οποία αντιστοιχεί σε ένα implicit global lock γύρω από το κρίσιμο τμήμα. Η

λειτουργία του είναι εννοιολογικά παρόμοια με έναν mutex: μόνο ένα νήμα κάθε φορά περνάει μέσα στο critical section, όλα τα υπόλοιπα περιμένουν. Ο απλός αυτός μηχανισμός είναι εύχρηστος προγραμματιστικά, αλλά, όπως και το pthread_mutex_lock, γίνεται γρήγορα bottleneck όταν πολλά νήματα προσπαθούν να ενημερώσουν την ίδια κοινόχρηστη δομή. Έτσι, οι χρόνοι του omp_critical_kmeans αντανακλούν την ίδια εικόνα με το naive shared-clusters της προηγούμενης άσκησης: σωστή αλλά βαριά μορφή συγχρονισμού, που περιορίζει την παραλληλία και δεν κλιμακώνεται καλά σε υψηλό contention, ειδικά σε ένα μικρό και έντονα memory-bound πρόβλημα.

Τα κλειδώματα τύπου TAS/TTAS και spin lock ανήκουν στην κατηγορία των απλών spinlocks, ενώ τα array και CLH συγκαταλέγονται στα scalable queue locks, τα οποία – όπως φαίνεται και από τις μετρήσεις μας – μειώνουν το coherence traffic και κλιμακώνονται καλύτερα υπό έντονο contention.

**Σ.Η.Μ.Μ.Υ. Ε.Μ.Π.
Νοέμβριος 2025**

ΣΥΣΤΗΜΑΤΑ ΠΑΡΑΛΛΗΛΗΣ ΕΠΕΞΕΡΓΑΣΙΑΣ

ΑΝΑΦΟΡΑ 3^{ης} ΑΣΚΗΣΗΣ



Στοιχεία Ομάδας

- Αναγνωριστικό: parlab05
- Μέλος 1^ο: Πέππας Μιχαήλ – Αθανάσιος, Α.Μ: 03121026
- Μέλος 2^ο: Σαουνάτσος Ανδρέας, Α.Μ: 03121197
- Ημερομηνία Παράδοσης Αναφοράς: 20.10.2025

▪ Ταυτόχρονες Δομές Δεδομένων

1. Υλοποιήσεις

Στην άσκηση μελετάμε ταυτόχρονες υλοποιήσεις μίας ταξινομημένης απλά συνδεδεμένης λίστας με βασικές λειτουργίες `contains()`, `add()` και `remove()`. Οι υλοποιήσεις διαφέρουν μόνο στον μηχανισμό συγχρονισμού:

Coarse-grain locking (cgl)

Χρησιμοποιείται ένα κοινό lock (`mutex`) για ολόκληρη τη λίστα. Κάθε λειτουργία κλειδώνει τη δομή στην αρχή και την ξεκλειδώνει στο τέλος, άρα σε κάθε χρονική στιγμή μόνο ένα νήμα μπορεί να εκτελεί οποιαδήποτε λειτουργία στη λίστα.

Fine-grain locking (fgl)

Υπάρχει lock ανά κόμβο και η διάσχιση γίνεται με hand-over-hand (lock coupling): διατηρούνται κλειδωμένοι διαδοχικά ο πρόγονος (`pred`) και ο τρέχων κόμβος (`curr`), και καθώς προχωράμε ξεκλειδώνεται ο προηγούμενος κόμβος και κλειδώνεται ο επόμενος. Αυτό επιτρέπει παραλληλία σε διαφορετικά τμήματα της λίστας, αλλά αυξάνει το overhead από πολλά lock/unlock.

Optimistic synchronization (opt)

Οι λειτουργίες πρώτα διατρέχουν τη λίστα χωρίς locks για να εντοπίσουν το σημείο ενημέρωσης. Στη συνέχεια κλειδώνουν τοπικά τους κόμβους `pred` και `curr` και εκτελούν validation (επαλήθευση) ότι η δομή δεν άλλαξε ενδιάμεσα. Το validation περιλαμβάνει επαναδιάσχιση από την αρχή για να επιβεβαιωθεί ότι το ζεύγος (`pred,curr`) παραμένει έγκυρο, αλλιώς η λειτουργία επαναλαμβάνεται (`retry`).

Lazy synchronization (lazy)

Η `contains()` εκτελείται χωρίς locks και αγνοεί κόμβους που έχουν μαρκαριστεί ως διαγραμμένοι. Η `remove()` υλοποιείται σε δύο φάσεις: πρώτα λογική διαγραφή (θέτοντας `marked=true`) και έπειτα φυσική αφαίρεση (ενημέρωση `pred.next=curr.next`) με τοπικό locking και validation. Έτσι μειώνεται ο χρόνος που κρατιούνται locks και αποφεύγονται συγκρούσεις με αναζητήσεις.

Non-blocking (lock-free) synchronization (nb)

Δεν χρησιμοποιούνται locks. Οι ενημερώσεις γίνονται με ατομικές εντολές týπου CAS πάνω σε δείκτες (συχνά «πακετάροντας» και το mark μαζί με το next). Οι λειτουργίες επαναπροσπαθούν (retry) όταν μια CAS αποτύχει λόγω ταυτόχρονης ενημέρωσης από άλλο νήμα. Η `contains()` είναι wait-free (δεν μπλοκάρει), ενώ `add/remove` είναι lock-free (πάντα κάποιο νήμα προοδεύει).

2. Περιβάλλον Εκτέλεσης και Ρυθμίσεις Παραμέτρων

Τα πειράματα εκτελέστηκαν στο μηχάνημα sandman (ουρά serial) με χρήση `qsub`, σύμφωνα με τις οδηγίες της áskησης. Για κάθε εκτελέσιμο χρησιμοποιήθηκε η ίδια `main.c` διεπαφή και μετρήθηκε το throughput σε Kops/sec (χιλιάδες λειτουργίες ανά δευτερόλεπτο) για σταθερό χρόνο εκτέλεσης.

Ο αριθμός νημάτων και η αντιστοίχισή τους σε λογικούς πυρήνες καθορίστηκε αποκλειστικά από τη μεταβλητή περιβάλλοντος `MT_CONF`, ώστε να επιτυγχάνεται pinning και αναπαραγωγιμότητα. Σε εκτελέσεις έως 64 νήματα τα threads «δέθηκαν» σε διαδοχικούς πυρήνες (π.χ. `MT_CONF=0,1,2,3` για 4 νήματα). Για 64 και 128 νήματα αξιοποιήθηκε hyperthreading και (στην περίπτωση των 128) oversubscription, σύμφωνα με τις οδηγίες της εκφώνησης.

Οι παράμετροι που εξετάστηκαν ήταν:

- Threads: 1, 2, 4, 8, 16, 32, 64, 128
- Μέγεθος λίστας: 1024, 8192
- Workloads: 100-0-0, 80-10-10, 20-40-40, 0-50-50 (contains-add-remove)

To `run_on_queue.sh` φαίνεται, για λόγους πληρότητας, ακολούθως:

a4/conc_ll/run_on_queue.sh

```

1 #!/bin/bash
2
3 ## Job Name
4 #PBS -N run_conc_ll
5
6 ## Output and error of PBS (not the runs)
7 #PBS -o run_conc_ll.pbs_out
8 #PBS -e run_conc_ll.pbs_err
9
10 ## Sandman, serial queue, 64 threads available
11 #PBS -q serial
12 #PBS -l nodes=sandman:ppn=64
13
14 ## Maximum walltime (adjust if necessary)
15 #PBS -l walltime=01:00:00
16
17 ## Go to the directory where qsub was executed
18 # CHANGE THIS TO YOUR ACTUAL DIRECTORY
19 cd $HOME/a2/conc_ll
20
21 # --- Define Core Parameters ---
22 IMPLEMENTATIONS="serial cgl fgl opt lazy nb"
23 NTHREADS="1 2 4 8 16 32 64 128"
24 LIST_SIZES="1024 8192"
25
26 # Workloads: (Contains, Add, Remove)
27 # Format: "C_A_R"
28 WORKLOADS="100_0_0 80_10_10 20_40_40 0_50_50"
29
30 # Directory for results
31 OUTDIR="results_conc_ll"
32 mkdir -p "$OUTDIR"
33
34 # --- Helper Function to generate MT_CONF for thread pinning ---
35 # Generates a comma-separated list of logical core IDs.
36 # Assumes sandman has 64 logical cores (0-63).
37 # For N > 64, it cycles through the 64 available logical cores (oversubscription).
38 get_mt_conf() {
39     local N=$1
40     local CONFIG=""
41     local MAX_LOGICAL_CORES=64
42
43     for i in $(seq 0 $((N - 1))); do
44         # Core ID cycles through 0, 1, ..., 63, 0, 1, ...
45         local CORE_ID=$((i % MAX_LOGICAL_CORES))
46
47         CONFIG="${CONFIG}${CORE_ID}"
48         if [ $i -lt $((N - 1)) ]; then
49             CONFIG="${CONFIG},"
50         fi
51     done

```

```

52     echo "$CONFIG"
53 }
54
55 # --- Main Execution Loop ---
56 for IMPL in $IMPLEMENTATIONS; do
57     EXECUTABLE="./$x.$IMPL"
58
59     for S in $LIST_SIZES; do
60
61         for T in $NTHREADS; do
62
63             # For the serial implementation, only run T=1 (to establish baseline)
64             if [ "$IMPL" == "serial" ] && [ $T -gt 1 ]; then
65                 continue
66             fi
67
68             # --- MT_CONF Setting for Thread Pinning (pthreads) ---
69             if [ $T -gt 1 ]; then
70                 MT_CONF=$(get_mt_conf $T)
71                 export MT_CONF
72             else
73                 # Unset MT_CONF for single-threaded execution (T=1)
74                 unset MT_CONF
75             fi
76
77             echo "Running $IMPL: ListSize=$S, Nthreads=$T, MT_CONF=$MT_CONF"
78
79             for W in $WORKLOADS; do
80                 # Split the workload string (e.g., 100_0_0) into C, A, R variables
81                 IFS='_' read -r C A R <<< "$W"
82
83                 # Input arguments for the executable: <list_size> <contains_pct> <add_pct>
<remove_pct>
84                 ARGS="$S $C $A $R"
85
86                 # Output files for this run
87                 OUT="${OUTDIR}/conc_ll_${IMPL}_${S}${S}_T${T}_W${W}.out"
88                 ERR="${OUTDIR}/conc_ll_${IMPL}_${S}${S}_T${T}_W${W}.err"
89
90                 # Run the program:
91                 # - stdout → OUT
92                 # - stderr → ERR
93                 $EXECUTABLE $ARGS >"$OUT" 2>"$ERR"
94             done
95         done
96     done
97 done
98
99 echo "Execution finished. Results are in the $OUTDIR directory."
100

```

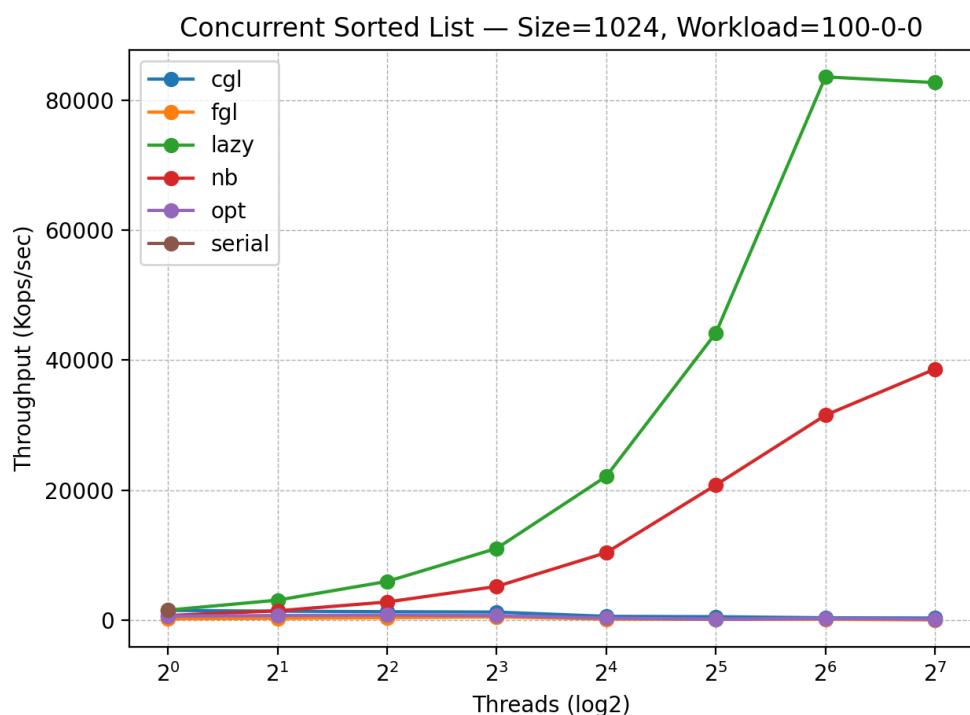
3. Αποτελέσματα και Σχολιασμός

Στα διαγράμματα που ακολουθούν παρουσιάζεται το throughput (Kops/sec), αλλά και το speedup ως συνάρτηση του αριθμού νημάτων, για διαφορετικά μεγέθη λίστας και διαφορετικά workloads. Κάθε γράφημα αντιστοιχεί σε σταθερό συνδυασμό μεγέθους λίστας και workload, ενώ οι καμπύλες αναπαριστούν τις διαφορετικές υλοποιήσεις συγχρονισμού. Για κάθε περίπτωση workload έχουμε :

A. Workload 100-0-0 (μόνο αναζητήσεις)

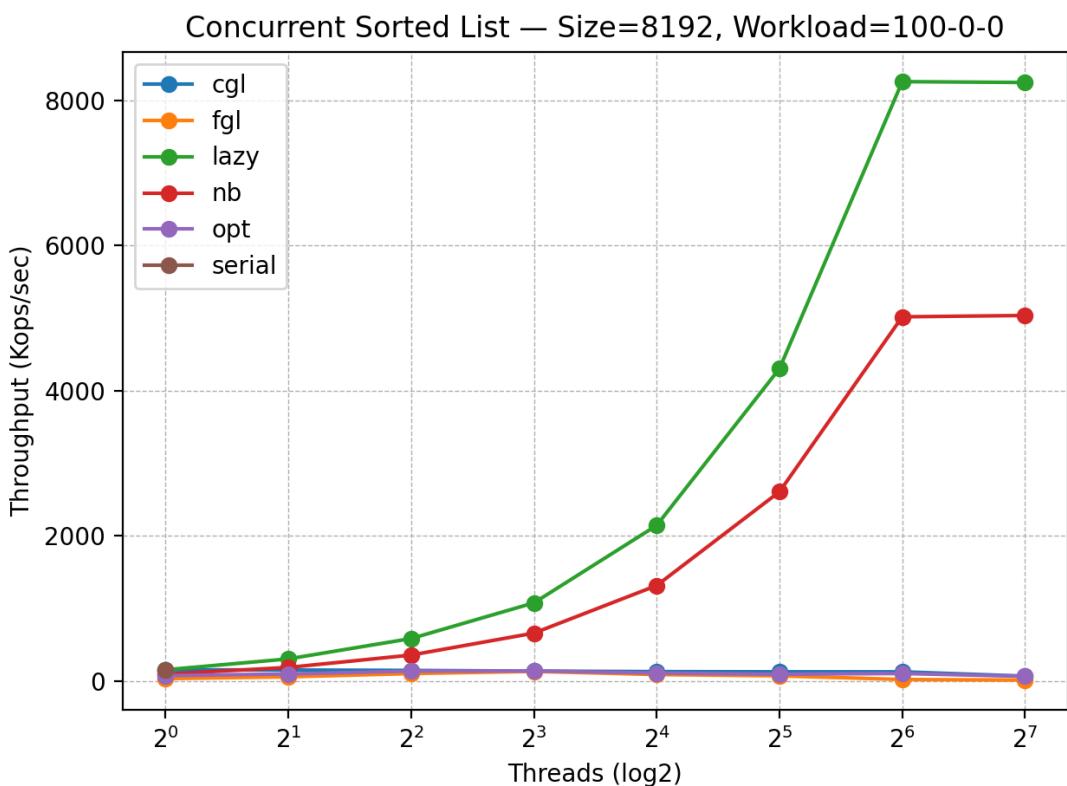
Στο συγκεκριμένο workload όλες οι λειτουργίες είναι contains(), χωρίς καμία εισαγωγή ή διαγραφή. Πρόκειται επομένως για ένα αμιγώς read-only σενάριο, στο οποίο δεν εμφανίζονται συγκρούσεις εγγραφών και η απόδοση εξαρτάται κυρίως από το αν και πόσο blocking απαιτεί η υλοποίηση για τις αναζητήσεις και το κόστος διάσχισης της λίστας (εξαρτώμενο από το μέγεθός της).

Throughput ως προς τον αριθμό νημάτων



Για μέγεθος λίστας $S = 1024$. Οι υλοποιήσεις lazy synchronization και non-blocking (lock-free) παρουσιάζουν τη σαφώς καλύτερη απόδοση και κλιμακώνονται έντονα με την αύξηση του αριθμού νημάτων. Το lazy synchronization επιτυγχάνει το υψηλότερο throughput, με σχεδόν γραμμική αύξηση έως περίπου 64 νήματα και ελαφρά σταθεροποίηση στη συνέχεια. Η non-blocking υλοποίηση ακολουθεί παρόμοια τάση, αν και με χαμηλότερες απόλυτες τιμές throughput.

Αντίθετα, η coarse-grain locking υλοποίηση εμφανίζει σχεδόν σταθερό και χαμηλό throughput ανεξαρτήτως αριθμού νημάτων, καθώς όλες οι αναζητήσεις σειριοποιούνται μέσω ενός καθολικού lock. Η fine-grain locking παρουσιάζει μικρή βελτίωση σε χαμηλό αριθμό νημάτων, ωστόσο η απόδοσή της υποβαθμίζεται καθώς αυξάνεται το concurrency, λόγω του αυξημένου κόστους από τα πολλαπλά lock/unlock σε κάθε βήμα της διάσχισης. Η optimistic synchronization παραμένει σε ενδιάμεσες τιμές, χωρίς να μπορεί να ανταγωνιστεί τις lazy και non-blocking υλοποιήσεις στο συγκεκριμένο σενάριο.



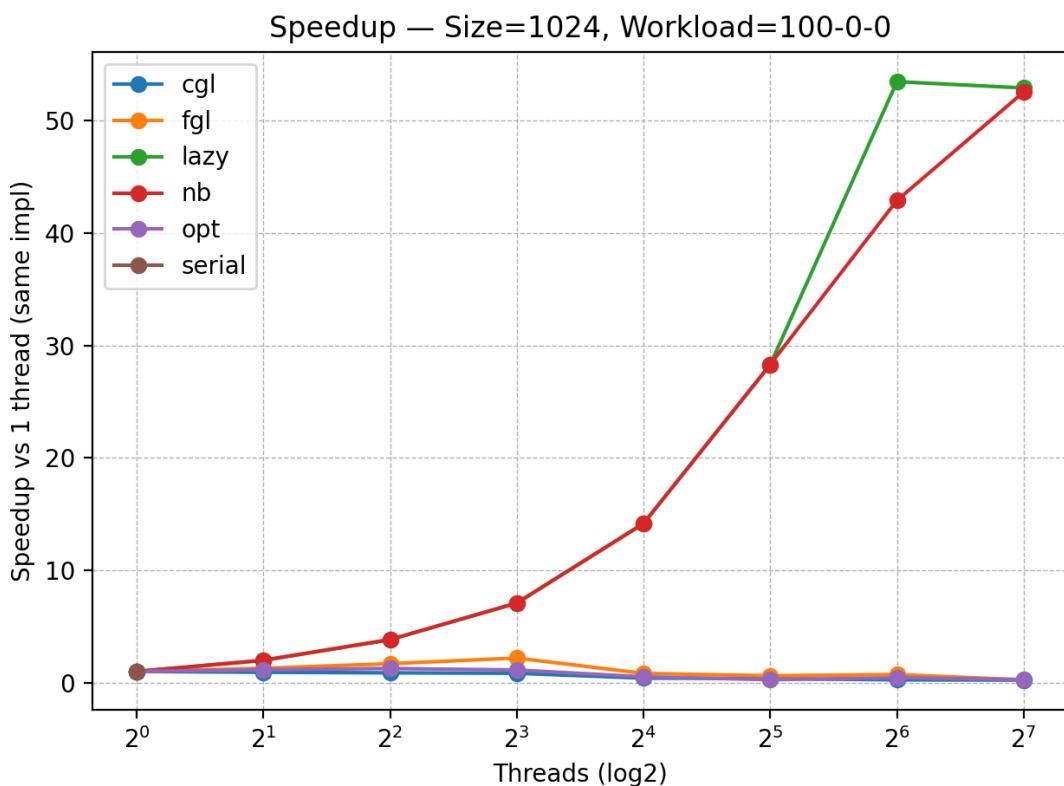
Για μεγαλύτερο μέγεθος λίστας, η ποιοτική εικόνα παραμένει ίδια, αλλά όλες οι υλοποιήσεις παρουσιάζουν σημαντικά χαμηλότερο throughput. Η αύξηση του

μήκους της λίστας συνεπάγεται μεγαλύτερο κόστος διάσχισης και αυξημένα cache misses, γεγονός που περιορίζει τον ρυθμό εκτέλεσης των αναζητήσεων.

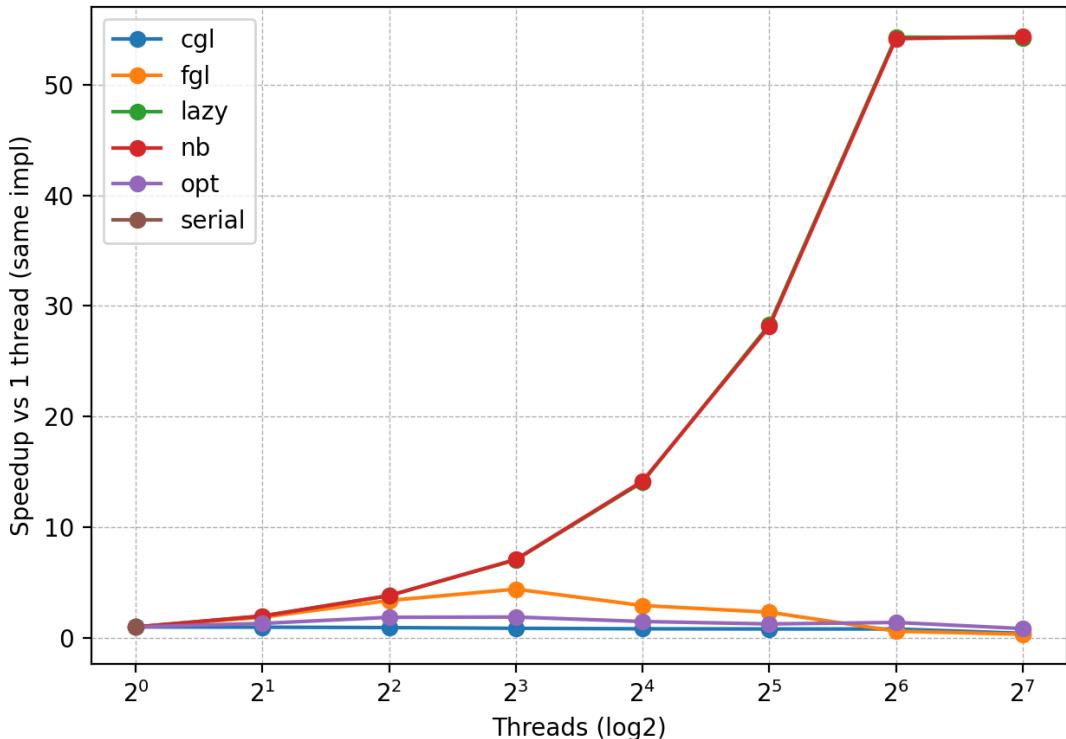
Παρόλα αυτά, οι lazy και non-blocking υλοποιήσεις εξακολουθούν να υπερέχουν σημαντικά έναντι των locking-based προσεγγίσεων, επιβεβαιώνοντας ότι η αποφυγή blocking είναι καθοριστικός παράγοντας απόδοσης σε read-only workloads.

Speedup ως προς τον αριθμό νημάτων

Το speedup υπολογίζεται σε σχέση με την απόδοση ενός νήματος της ίδιας υλοποίησης και αποτυπώνει την ικανότητα κλιμάκωσης ανεξάρτητα από τις απόλυτες τιμές throughput.



Speedup — Size=8192, Workload=100-0-0



Τόσο για $S=1024$ όσο και για $S=8192$, οι υλοποιήσεις *lazy* και *non-blocking* παρουσιάζουν εντυπωσιακό speedup, φτάνοντας περίπου έως $50\times$ σε 64 νήματα. Πέρα από αυτό το σημείο, και ειδικότερα στα 128 νήματα, η επιτάχυνση σταθεροποιείται, γεγονός που αποδίδεται στη χρήση hyperthreading και oversubscription, όπου τα threads μοιράζονται τους ίδιους φυσικούς πόρους.

Οι υλοποιήσεις *coarse-grain*, *fine-grain* και *optimistic* εμφανίζουν περιορισμένο speedup, με τις καμπύλες τους να παραμένουν κοντά στη μονάδα ή να παρουσιάζουν μικρή μόνο βελτίωση. Αυτό υποδηλώνει ότι το κόστος συγχρονισμού και το contention υπερισχύουν των ωφελειών από την παράλληλη εκτέλεση.

Επίδραση του μεγέθους της λίστας

Η σύγκριση μεταξύ $S=1024$ και $S=8192$ δείχνει ότι το μεγαλύτερο μέγεθος λίστας επηρεάζει αρνητικά το throughput σε όλες τις υλοποιήσεις, λόγω αυξημένου κόστους διάσχισης και μνήμης. Ωστόσο, το σχήμα των καμπυλών speedup παραμένει παρόμοιο, γεγονός που υποδηλώνει ότι οι βασικοί περιορισμοί κλιμάκωσης κάθε μηχανισμού συγχρονισμού δεν αλλάζουν με το μέγεθος της λίστας, αλλά σχετίζονται κυρίως με τον τρόπο συγχρονισμού.

Συμπεράσματα για το workload 100-0-0

Στο αμιγώς read-only workload προκύπτουν τα εξής:

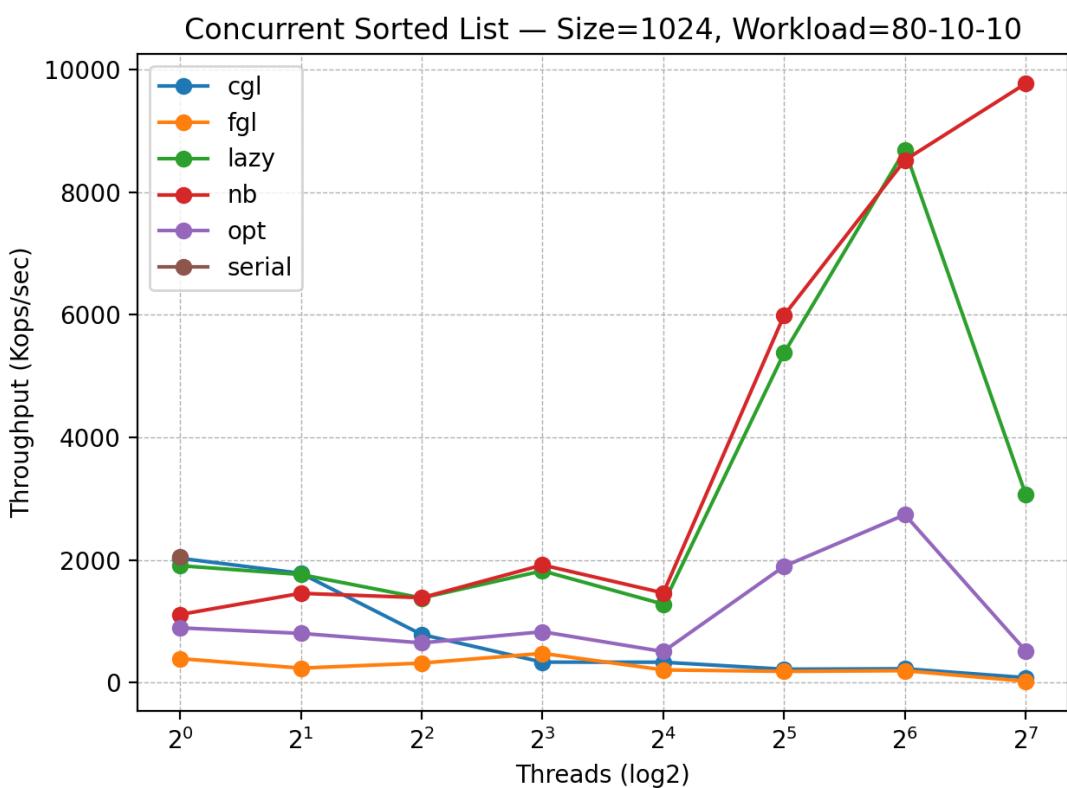
- Οι lazy synchronization και non-blocking υλοποιήσεις είναι οι πλέον κατάλληλες, καθώς οι αναζητήσεις εκτελούνται με ελάχιστο ή μηδενικό blocking.
- Η coarse-grain locking υλοποίηση δεν κλιμακώνεται, λόγω της πλήρους σειριοποίησης των λειτουργιών.
- Οι fine-grain και optimistic υλοποιήσεις παρουσιάζουν καλύτερη συμπεριφορά από την coarse-grain, αλλά παραμένουν κατώτερες σε σχέση με lazy και non-blocking λόγω αυξημένου κόστους συγχρονισμού.
- Η κλιμάκωση περιορίζεται μετά τα 64 νήματα, κυρίως λόγω αρχιτεκτονικών περιορισμών (hyperthreading και oversubscription) και όχι λόγω αλγορίθμικής αστοχίας.

B. Workload 80-10-10 (κυρίως αναζητήσεις)

Στο workload 80-10-10, το 80% των λειτουργιών είναι `contains()`, ενώ το υπόλοιπο 20% κατανέμεται ισομερώς σε `add()` και `remove()`. Το σενάριο αυτό προσομοιώνει ένα ρεαλιστικό `read-mostly` workload, όπου συνυπάρχουν αναζητήσεις και περιορισμένος αριθμός ενημερώσεων.

Σε αντίθεση με το 100-0-0, εδώ αρχίζουν να εμφανίζονται συγκρούσεις μεταξύ νημάτων λόγω των updates, γεγονός που επιτρέπει να αξιολογηθεί η συμπεριφορά των διαφορετικών μηχανισμών συγχρονισμού υπό μέτριο contention.

Throughput ως προς τον αριθμό νημάτων

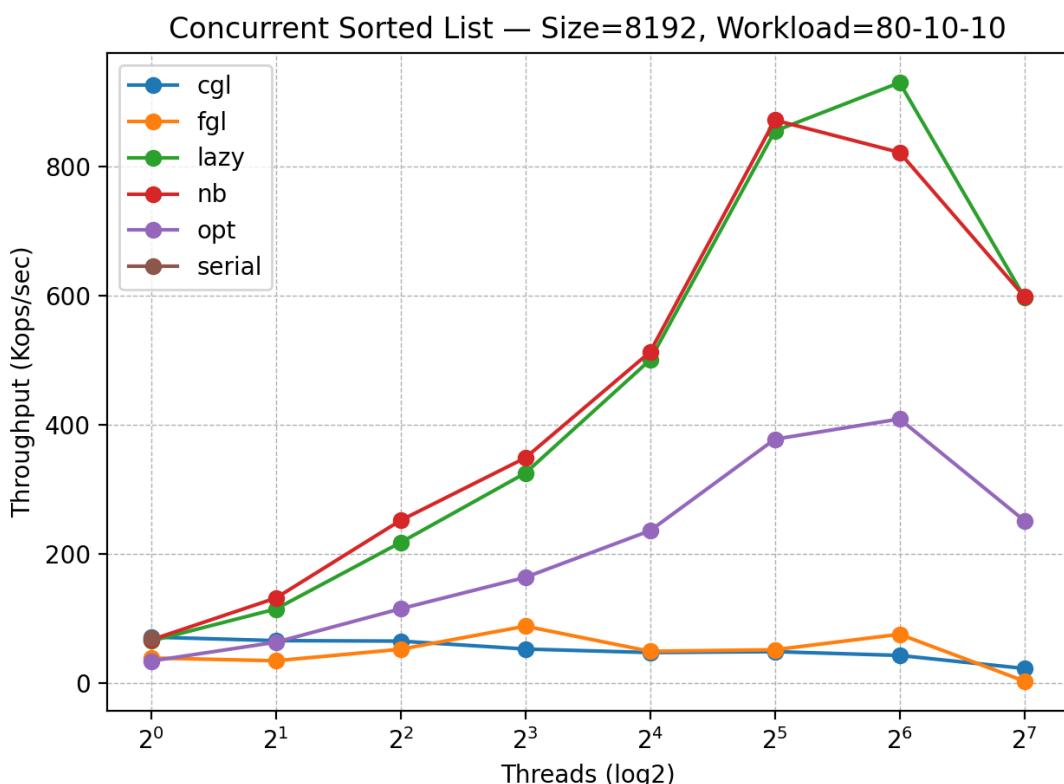


Για μικρό μέγεθος λίστας, η `lazy synchronization` εμφανίζει την καλύτερη συνολική απόδοση. Το throughput αυξάνεται σταθερά μέχρι τα 32–64 νήματα και στη συνέχεια παρουσιάζει ελαφρά σταθεροποίηση. Η καλή αυτή συμπεριφορά

οφείλεται στο γεγονός ότι οι `contains()` εκτελούνται χωρίς locks, ενώ οι ενημερώσεις υλοποιούνται με σύντομο τοπικό locking και λογική διαγραφή.

Η optimistic synchronization ακολουθεί σε απόδοση, όμως η αύξηση του αριθμού των νημάτων οδηγεί σε συχνότερα αποτυχημένα validations, με αποτέλεσμα η κλιμάκωση να περιορίζεται νωρίτερα σε σχέση με το lazy. Η non-blocking υλοποίηση παρουσιάζει καλή απόδοση σε χαμηλό και μεσαίο αριθμό νημάτων, αλλά σε υψηλότερο concurrency αρχίζει να επηρεάζεται από αποτυχημένες CAS και αυξημένο contention σε κοινά σημεία της λίστας.

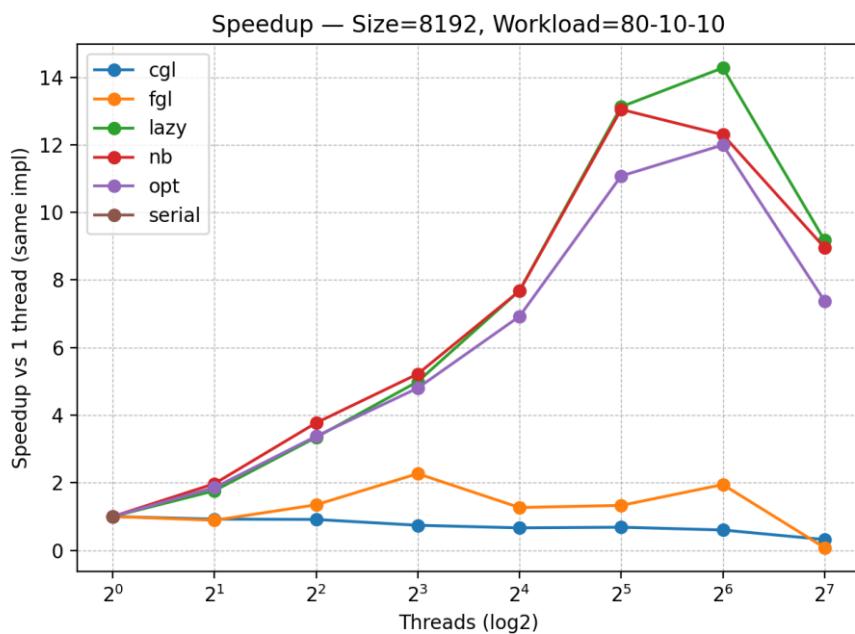
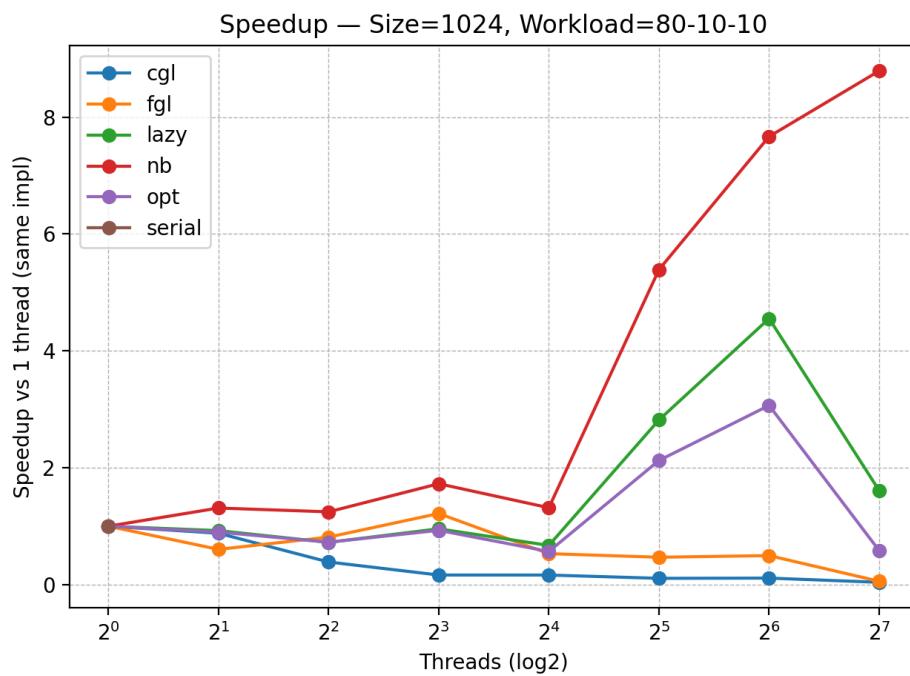
Οι υλοποιήσεις fine-grain και coarse-grain locking εμφανίζουν σαφώς χαμηλότερο throughput. Στην coarse-grain, όλες οι λειτουργίες σειριαποιούνται μέσω ενός καθολικού lock, ενώ στη fine-grain το κόστος των πολλαπλών lock/unlock γίνεται αισθητό ακόμη και με σχετικά περιορισμένο αριθμό ενημερώσεων.



Με τη μεγαλύτερη λίστα, το συνολικό throughput μειώνεται για όλες τις υλοποιήσεις, λόγω αυξημένου κόστους διάσχισης και χειρότερης χωρικής τοπικότητας μνήμης. Παρόλα αυτά, η lazy synchronization διατηρεί ξεκάθαρο προβάδισμα, παρουσιάζοντας την πιο σταθερή συμπεριφορά σε όλο το εύρος των νημάτων.

Η optimistic synchronization επηρεάζεται περισσότερο από το μεγαλύτερο μέγεθος λίστας, καθώς κάθε αποτυχημένο validation συνεπάγεται επαναδιάσχιση μεγαλύτερου τμήματος της δομής. Η non-blocking υλοποίηση συνεχίζει να κλιμακώνει έως ένα σημείο, αλλά ο κορεσμός εμφανίζεται νωρίτερα σε σχέση με το S=1024.

Speedup ως προς τον αριθμό νημάτων



Η ανάλυση του speedup δείχνει ότι:

- Η lazy synchronization παρουσιάζει την καλύτερη κλιμάκωση και στα δύο μεγέθη λίστας, με σχεδόν γραμμική αύξηση έως τα 32–64 νήματα.
- Η optimistic και η non-blocking υλοποίηση εμφανίζουν μέτριο speedup, το οποίο περιορίζεται καθώς αυξάνονται τα retries (λόγω validation ή CAS αποτυχιών).
- Οι locking-based υλοποιήσεις (coarse και fine-grain) παρουσιάζουν περιορισμένη επιτάχυνση, με τις καμπύλες speedup να παραμένουν χαμηλά.
- Σε όλα τα σχήματα, η μετάβαση από 64 σε 128 νήματα δεν οδηγεί σε ουσιαστική επιπλέον επιτάχυνση, γεγονός που αποδίδεται στη χρήση hyperthreading και oversubscription.

Επίδραση του μεγέθους της λίστας

Η αύξηση του μεγέθους της λίστας από 1024 σε 8192 στοιχεία:

- Μειώνει το απόλυτο throughput σε όλες τις υλοποιήσεις.
- Επηρεάζει δυσανάλογα τις optimistic και fine-grain υλοποιήσεις, όπου το κόστος αποτυχημένων προσπαθειών (validation ή locks) αυξάνεται με το μήκος της διάσχισης.

Παρόλα αυτά, η σχετική κατάταξη των υλοποιήσεων παραμένει σταθερή, με τη lazy synchronization να αποτελεί την πιο ανθεκτική επιλογή ως προς την αύξηση του μεγέθους της λίστας.

Συμπεράσματα για το workload 80-10-10

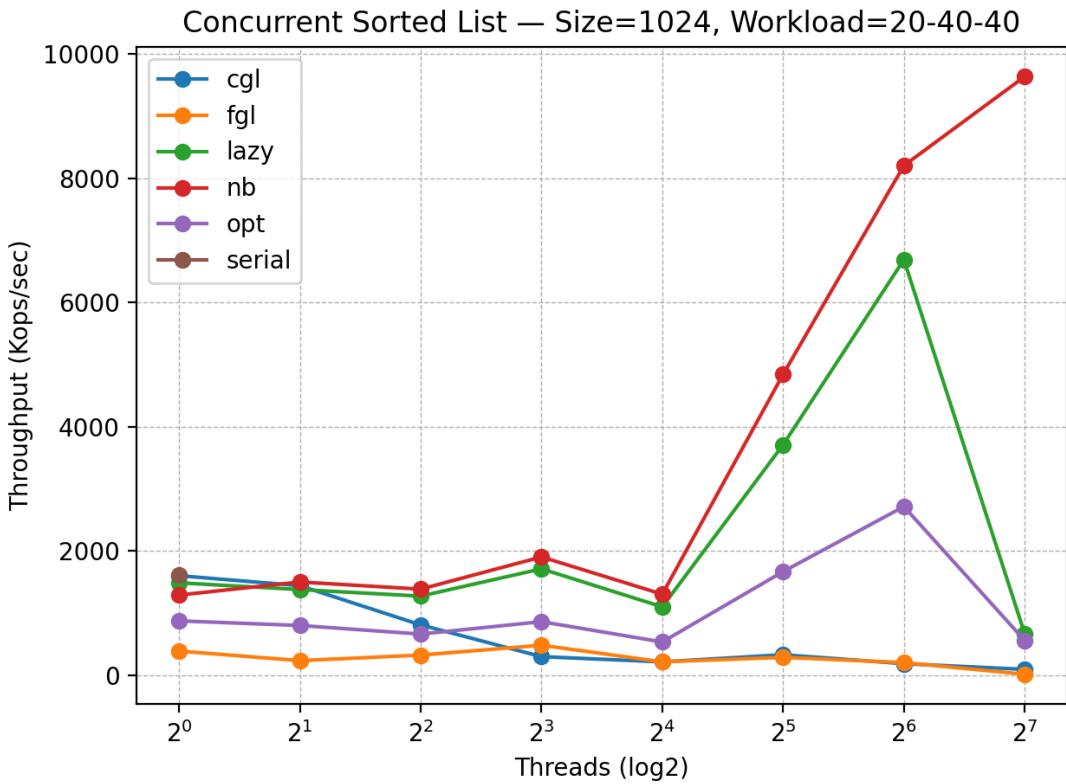
Από την ανάλυση του workload 80-10-10 προκύπτουν τα εξής:

- Η lazy synchronization αποτελεί την πιο αποδοτική και σταθερή υλοποίηση σε read-mostly σενάρια με περιορισμένα updates.
- Η optimistic synchronization λειτουργεί ικανοποιητικά, αλλά η απόδοσή της περιορίζεται από τα αποτυχημένα validations όσο αυξάνεται το concurrency.
- Η non-blocking υλοποίηση παρουσιάζει καλό scaling σε μέτριο αριθμό νημάτων, αλλά εμφανίζει κορεσμό σε υψηλό contention.
- Οι coarse-grain και fine-grain locking υλοποιήσεις υστερούν σημαντικά, επιβεβαιώνοντας ότι το blocking και το lock overhead επηρεάζουν αρνητικά την απόδοση ακόμη και όταν τα updates είναι σχετικά λίγα.

Γ. Workload 20-40-40 (κυρίως ενημερώσεις)

Στο workload 20-40-40, μόνο το 20% των λειτουργιών είναι contains(), ενώ το 80% αφορά ενημερώσεις (add() και remove()). Πρόκειται για ένα update-dominated σενάριο, στο οποίο το contention στη δομή δεδομένων είναι έντονο και η αποδοτικότητα των μηχανισμών συγχρονισμού παίζει καθοριστικό ρόλο.

Σε αντίθεση με τα read-mostly workloads, εδώ αναδεικνύεται το κόστος των locks, των αποτυχημένων validations και των επαναλαμβανόμενων atomic retries.

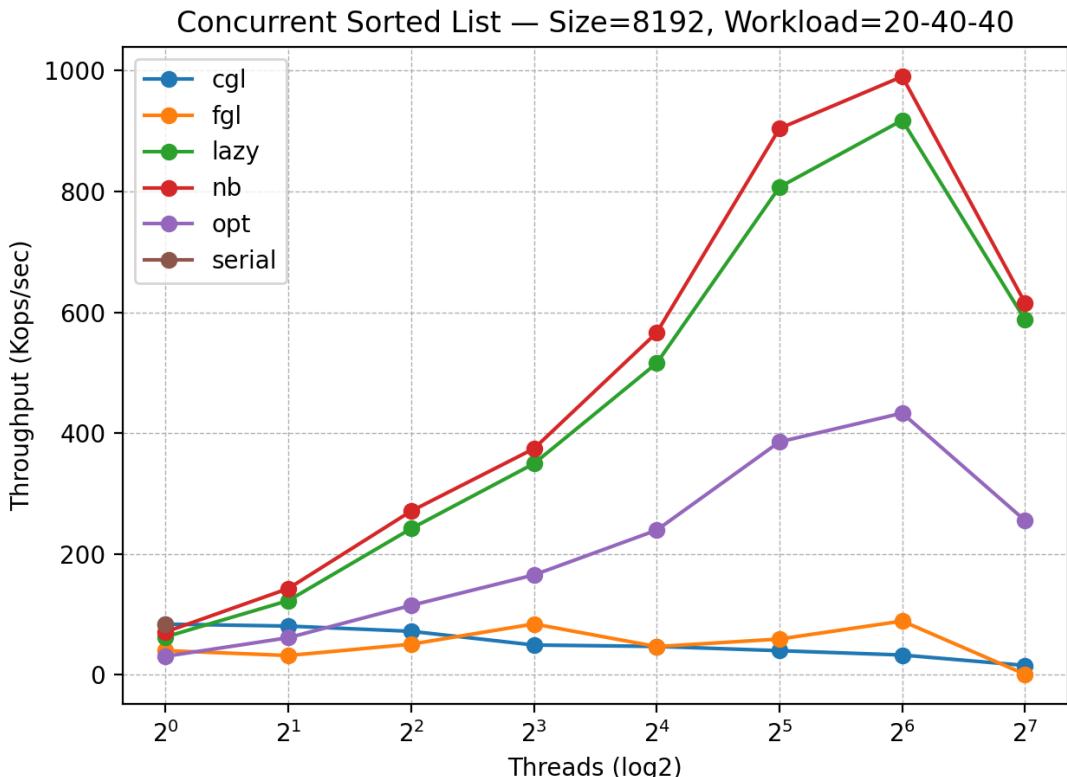


Για μικρό μέγεθος λίστας, η non-blocking (lock-free) υλοποίηση επιτυγχάνει το υψηλότερο throughput, παρουσιάζοντας σαφή υπεροχή σε μεγάλο εύρος αριθμού νημάτων. Το throughput αυξάνεται έντονα έως τα 64 νήματα, γεγονός που υποδηλώνει ότι η αποφυγή locks επιτρέπει μεγαλύτερο βαθμό ταυτόχρονης προόδου ακόμη και υπό υψηλό contention.

Η lazy synchronization ακολουθεί σε απόδοση, με καλή κλιμάκωση έως τα 32–64 νήματα. Παρότι χρησιμοποιεί locks στις ενημερώσεις, ο διαχωρισμός λογικής και φυσικής διαγραφής μειώνει τη διάρκεια κατοχής locks και περιορίζει τις συγκρούσεις.

Η optimistic synchronization παρουσιάζει αισθητά χαμηλότερο throughput. Η συχνότητα των αποτυχημένων validations αυξάνεται λόγω των πολλών updates, με αποτέλεσμα σημαντικό ποσοστό του χρόνου εκτέλεσης να αναλώνεται σε επαναλήψεις (retries).

Οι υλοποιήσεις fine-grain και coarse-grain locking εμφανίζουν τη χαμηλότερη απόδοση. Στην coarse-grain, όλες οι ενημερώσεις σειριοποιούνται, ενώ στη fine-grain το κόστος από τα πολλαπλά locks σε κάθε ενημέρωση επιβαρύνει έντονα το throughput.



Για μεγαλύτερο μέγεθος λίστας, το throughput μειώνεται σε όλες τις υλοποιήσεις, καθώς οι ενημερώσεις απαιτούν μεγαλύτερες διασχίσεις και αυξάνεται το memory contention.

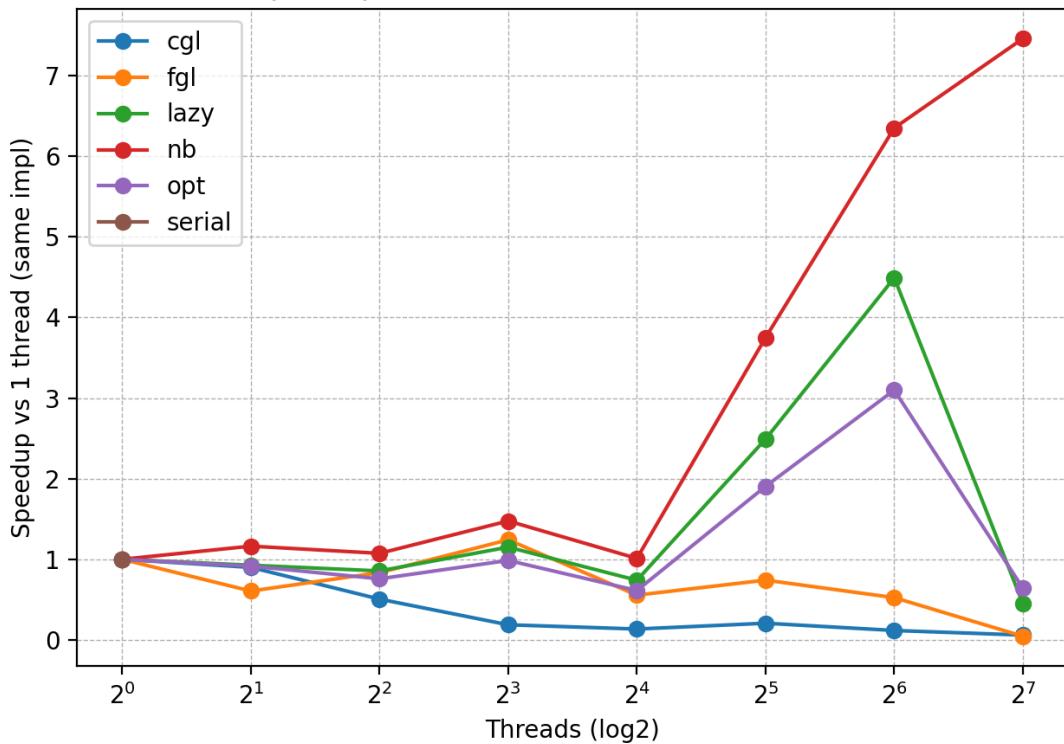
Η non-blocking υλοποίηση παραμένει η καλύτερη σε απόλυτους όρους, φτάνοντας σε μέγιστο throughput γύρω στα 32–64 νήματα, πριν εμφανίσει πτώση στα 128 νήματα. Η πτώση αυτή αποδίδεται στον συνδυασμό αυξημένου CAS contention και hyperthreading/oversubscription.

Η lazy synchronization διατηρεί σταθερά καλή απόδοση, αν και υστερεί ελαφρώς έναντι της non-blocking σε αυτό το έντονα update-heavy σενάριο. Η optimistic synchronization επηρεάζεται ακόμη περισσότερο από το αυξημένο μήκος της λίστας, καθώς κάθε αποτυχημένο validation συνεπάγεται επαναδιάσχιση μεγαλύτερου τμήματος της δομής.

Speedup ως προς τον αριθμό νημάτων

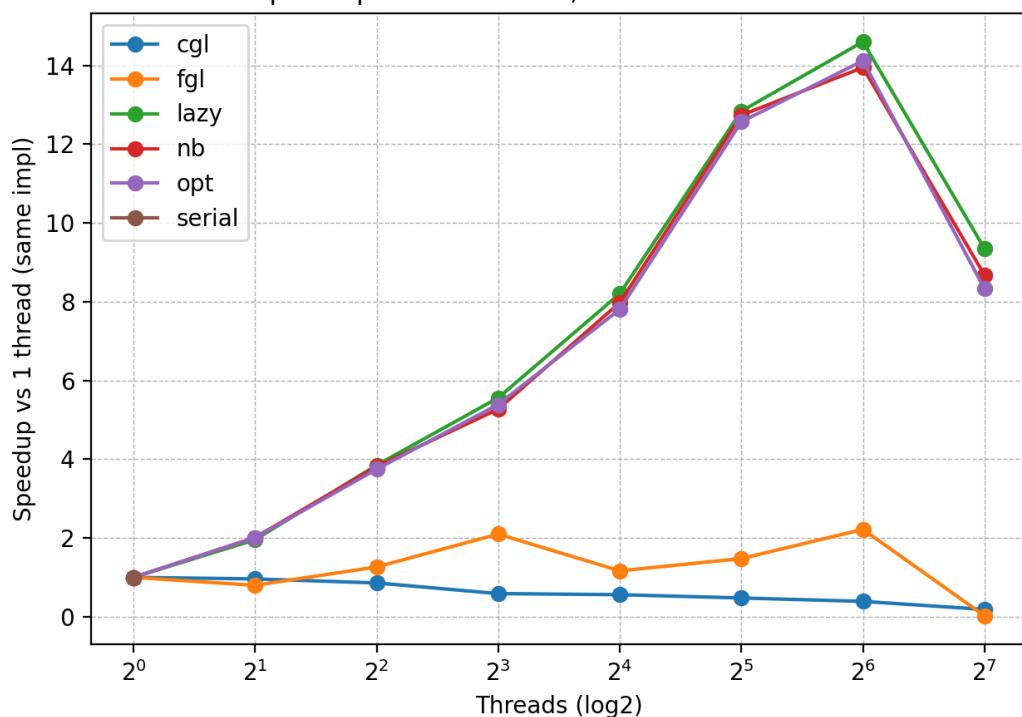
Η ανάλυση του speedup αποκαλύπτει σημαντικές διαφορές στη δυνατότητα κλιμάκωσης

Speedup — Size=1024, Workload=20-40-40



Για $S = 1024$, η non-blocking υλοποίηση επιτυγχάνει τη μεγαλύτερη επιτάχυνση, με speedup που ξεπερνά κατά πολύ τις υπόλοιπες υλοποιήσεις. Η lazy synchronization παρουσιάζει επίσης αξιόλογο speedup, αν και σαφώς χαμηλότερο.

Speedup — Size=8192, Workload=20-40-40



Για $S = 8192$, τόσο η non-blocking όσο και η lazy εμφανίζουν μέγιστο speedup περίπου στα 32–64 νήματα, με αισθητή πτώση στα 128 νήματα.

Οι coarse-grain και fine-grain locking υλοποιήσεις παρουσιάζουν speedup μικρότερο της μονάδας σε αρκετά σημεία, υποδεικνύοντας ότι η προσθήκη νημάτων όχι μόνο δεν επιταχύνει, αλλά επιβαρύνει την εκτέλεση λόγω έντονου contention.

Επίδραση του μεγέθους της λίστας

Η σύγκριση μεταξύ $S=1024$ και $S=8192$ δείχνει ότι το μεγαλύτερο μέγεθος λίστας:

- Μειώνει το throughput σε όλες τις υλοποιήσεις.
- Εντείνει ιδιαίτερα το κόστος των retries σε optimistic και non-blocking προσεγγίσεις.
- Καθιστά πιο εμφανές το πλεονέκτημα της lazy synchronization, η οποία περιορίζει τη διάρκεια κρίσιμων τμημάτων.

Παρότι το absolute throughput μειώνεται, η σχετική κατάταξη των υλοποιήσεων παραμένει παρόμοια, με τις non-blocking και lazy να υπερέχουν σε σχέση με τις locking-based λύσεις.

Συμπεράσματα για το workload 20-40-40

Από την ανάλυση του update-heavy workload προκύπτουν τα εξής:

- Η non-blocking (lock-free) υλοποίηση επιτυγχάνει την υψηλότερη απόδοση, ιδιαίτερα σε μικρό και μεσαίο μέγεθος λίστας, επιβεβαιώνοντας το πλεονέκτημα της απουσίας locks σε περιβάλλον έντονων ενημερώσεων.
- Η lazy synchronization αποτελεί τον καλύτερο συμβιβασμό μεταξύ απόδοσης και σταθερότητας, ειδικά όταν αυξάνεται το μέγεθος της λίστας.
- Η optimistic synchronization υποφέρει σημαντικά λόγω συχνών αποτυχημένων validations και δεν ενδείκνυται για update-dominated workloads.
- Οι coarse-grain και fine-grain locking υλοποιήσεις παρουσιάζουν πολύ περιορισμένη κλιμάκωση και αποτελούν τις λιγότερο αποδοτικές επιλογές στο συγκεκριμένο σενάριο.

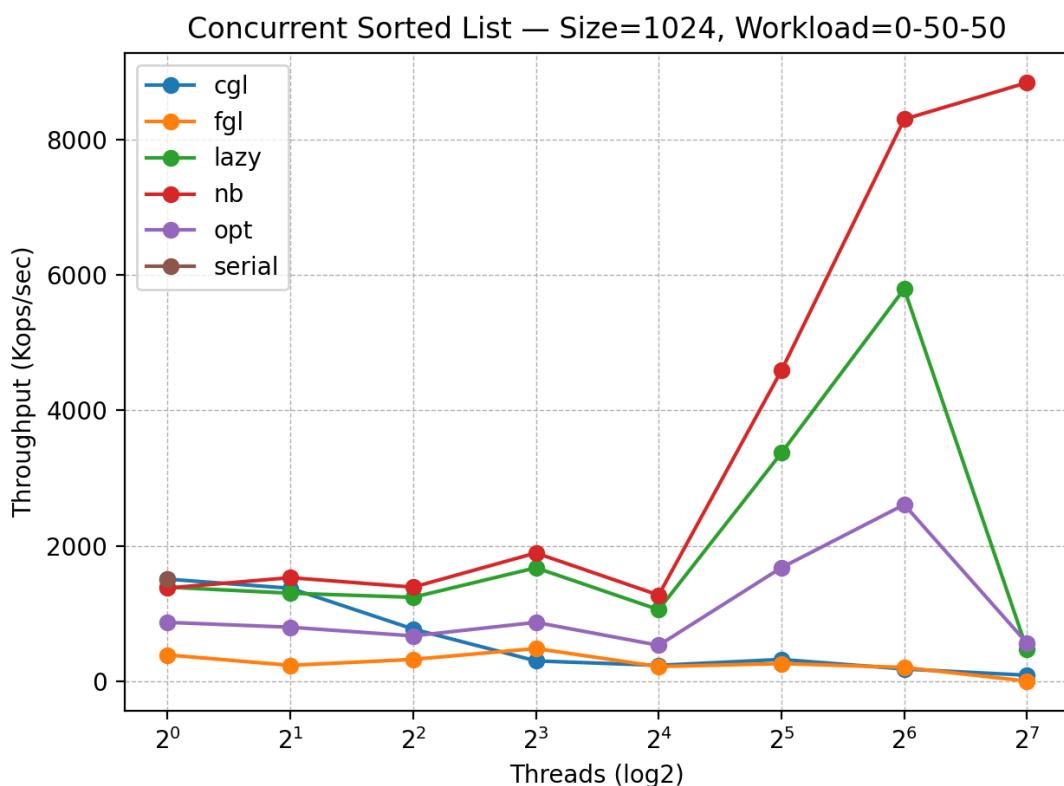
Η πτώση της απόδοσης στα 128 νήματα οφείλεται κυρίως σε αρχιτεκτονικούς περιορισμούς (hyperthreading και oversubscription), σε συνδυασμό με αυξημένο contention σε atomic ή locking μηχανισμούς.

Δ. Workload 0-50-50 (μόνο ενημερώσεις)

Στο workload 0-50-50 όλες οι λειτουργίες είναι ενημερώσεις (add() και remove()), χωρίς καθόλου αναζητήσεις. Πρόκειται για το πιο απαιτητικό σενάριο από πλευράς συγχρονισμού, καθώς κάθε λειτουργία τροποποιεί τη δομή της λίστας και συνεπώς δημιουργείται έντονο contention τόσο σε locks όσο και σε atomic operations.

To workload αυτό αναδεικνύει καθαρά τις διαφορές μεταξύ blocking, optimistic και lock-free προσεγγίσεων.

Throughput ως προς τον αριθμό νημάτων

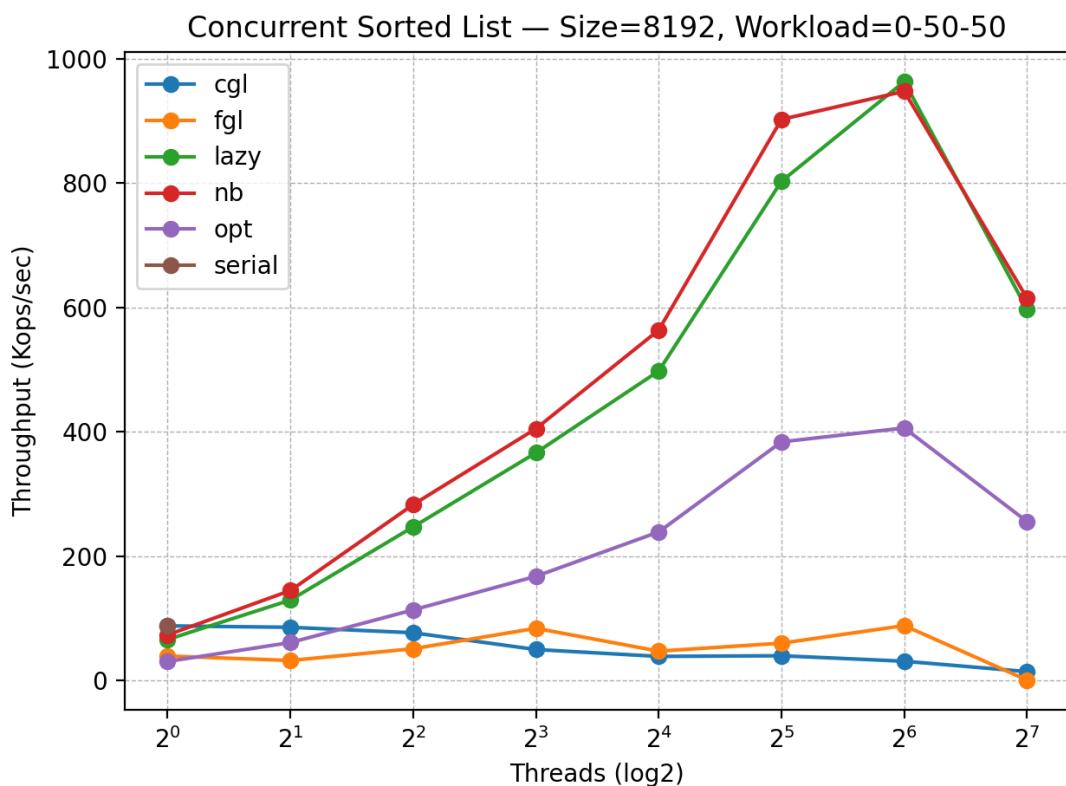


Για μικρό μέγεθος λίστας, η non-blocking (lock-free) υλοποίηση επιτυγχάνει το υψηλότερο throughput σε όλο το εύρος των νημάτων. Το throughput αυξάνεται σημαντικά έως τα 64 νήματα, όπου και παρατηρείται η μέγιστη απόδοση. Η απουσία locks επιτρέπει στα νήματα να προοδεύουν χωρίς να μπλοκάρουν μεταξύ τους, ακόμη και όταν όλες οι λειτουργίες είναι ενημερώσεις.

Η lazy synchronization ακολουθεί σε απόδοση, παρουσιάζοντας καλή κλιμάκωση έως τα 32–64 νήματα. Ο διαχωρισμός της λογικής από τη φυσική διαγραφή μειώνει το contention στα κρίσιμα τμήματα, αν και το κόστος των locks παραμένει αισθητό σε σύγκριση με την lock-free προσέγγιση.

Η optimistic synchronization εμφανίζει μέτριο throughput. Τα συχνά updates οδηγούν σε αποτυχημένα validations και retries, τα οποία περιορίζουν την απόδοση, αν και παραμένει σαφώς ανώτερη από τις locking-based υλοποιήσεις.

Οι fine-grain και coarse-grain locking υλοποιήσεις παρουσιάζουν τη χαμηλότερη απόδοση. Στην coarse-grain, όλες οι ενημερώσεις σειριοποιούνται, ενώ στη fine-grain το κόστος από τα πολλαπλά locks σε κάθε ενημέρωση επιβαρύνει έντονα το throughput.



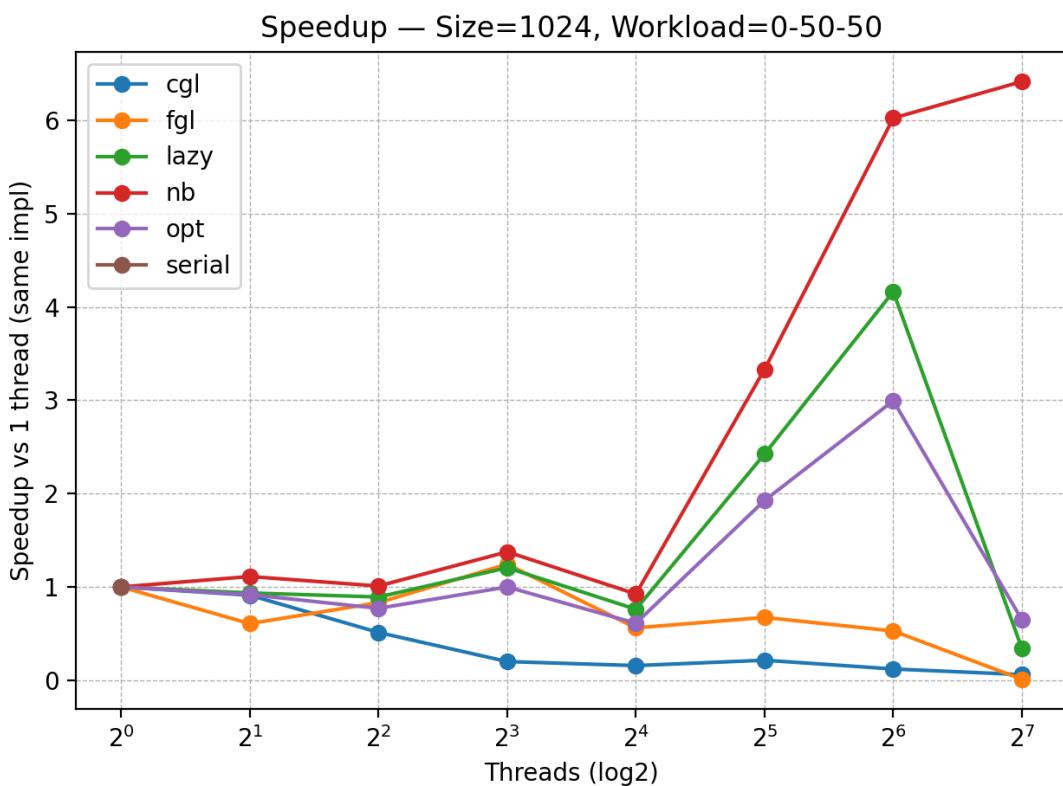
Με μεγαλύτερη λίστα, το throughput μειώνεται για όλες τις υλοποιήσεις, λόγω αυξημένου κόστους διάσχισης και μεγαλύτερης πιθανότητας συγκρούσεων.

Η non-blocking υλοποίηση παραμένει η ταχύτερη, φτάνοντας σε μέγιστο throughput γύρω στα 32–64 νήματα, πριν εμφανίσει πτώση στα 128 νήματα. Η πτώση αυτή αποδίδεται στον συνδυασμό έντονου CAS contention και oversubscription.

Η lazy synchronization παρουσιάζει παρόμοια ποιοτική συμπεριφορά, αλλά με χαμηλότερο απόλυτο throughput σε σχέση με τη non-blocking. Η optimistic synchronization επηρεάζεται ιδιαίτερα από το αυξημένο μέγεθος της λίστας, καθώς κάθε αποτυχημένο validation συνεπάγεται επαναδιάσχιση μεγαλύτερου τμήματος της δομής.

Speedup ως προς τον αριθμό νημάτων

Η ανάλυση του speedup επιβεβαιώνει τα συμπεράσματα από το throughput:

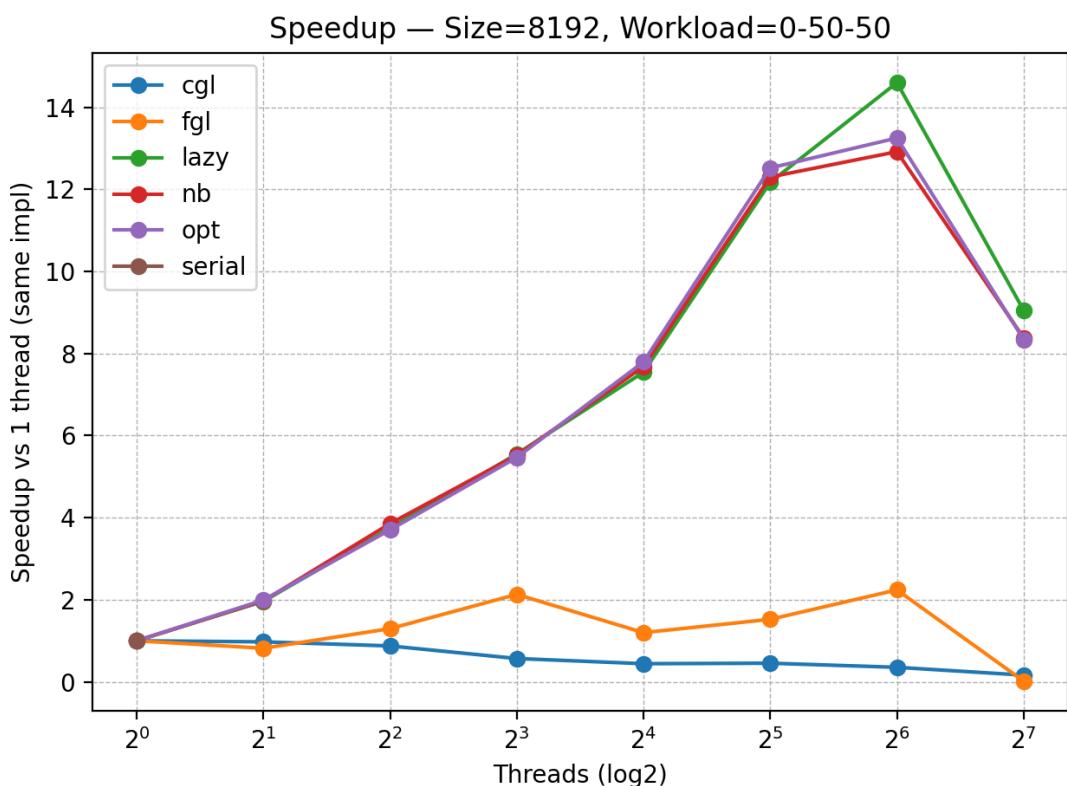


Για $S = 1024$, η non-blocking υλοποίηση παρουσιάζει τη μεγαλύτερη επιτάχυνση, ξεπερνώντας όλες τις άλλες υλοποιήσεις σε όλο το εύρος των νημάτων.

Η lazy synchronization επιτυγχάνει αξιόλογο speedup, αν και χαμηλότερο από τη non-blocking, λόγω του κόστους των locks στις ενημερώσεις.

Η optimistic synchronization εμφανίζει περιορισμένο speedup, καθώς τα retries ακυρώνουν μέρος του οφέλους από την παράλληλη εκτέλεση.

Οι coarse-grain και fine-grain locking υλοποιήσεις παρουσιάζουν speedup μικρότερο της μονάδας σε αρκετές περιπτώσεις, υποδεικνύοντας ότι η αύξηση του αριθμού νημάτων επιβαρύνει την εκτέλεση.



Για $S = 8192$, όλες οι καμπύλες speedup εμφανίζουν κορεσμό ή πτώση στα 128 νήματα, γεγονός που αποδίδεται στη χρήση hyperthreading και oversubscription σε συνδυασμό με έντονο contention.

Επίδραση του μεγέθους της λίστας

Η αύξηση του μεγέθους της λίστας από 1024 σε 8192 στοιχεία:

- Μειώνει σημαντικά το throughput σε όλες τις υλοποιήσεις.
- Εντείνει το κόστος των retries τόσο στις optimistic όσο και στις non-blocking υλοποιήσεις.
- Καθιστά πιο εμφανές το πλεονέκτημα της lazy synchronization έναντι των locking-based λύσεων, λόγω της μείωσης του χρόνου κατοχής locks.

Παρά τη μείωση των απόλυτων τιμών, η σχετική κατάταξη των υλοποιήσεων παραμένει σταθερή, με τις non-blocking και lazy να υπερέχουν.

Συμπεράσματα για το workload 0-50-50

Από την ανάλυση του αμιγώς update-heavy workload προκύπτουν τα εξής:

- Η non-blocking (lock-free) υλοποίηση αποτελεί την πιο αποδοτική επιλογή, επιτυγχάνοντας το υψηλότερο throughput και speedup.
- Η lazy synchronization προσφέρει τον καλύτερο συμβιβασμό μεταξύ απόδοσης και σταθερότητας, ειδικά σε μεγαλύτερα μεγέθη λίστας.
- Η optimistic synchronization δεν ενδείκνυται για workloads με αποκλειστικά ενημερώσεις, λόγω συχνών αποτυχημένων validations.
- Οι coarse-grain και fine-grain locking υλοποιήσεις αποτυγχάνουν να κλιμακώσουν και εμφανίζουν έντονη υποβάθμιση της απόδοσης.

4. Γενικό Συμπέρασμα

Από τη μελέτη των διαφορετικών workloads προκύπτει ότι η απόδοση και η κλιμάκωση μιας ταυτόχρονης ταξινομημένης λίστας εξαρτώνται άμεσα από το ποσοστό ενημερώσεων και τον μηχανισμό συγχρονισμού. Οι υλοποιήσεις με coarse-grain και fine-grain locking παρουσιάζουν περιορισμένη κλιμάκωση λόγω blocking και αυξημένου lock overhead, ανεξάρτητα από το workload. Η optimistic synchronization αποδίδει καλά μόνο σε read-heavy σενάρια, αλλά υποβαθμίζεται σημαντικά όταν αυξάνονται οι ενημερώσεις λόγω συχνών retries.

Η lazy synchronization εμφανίζει τη πιο σταθερή και ισορροπημένη συμπεριφορά σε όλα τα workloads, ενώ η non-blocking (lock-free) υλοποίηση επιτυγχάνει την υψηλότερη απόδοση σε update-heavy σενάρια, με κόστος αυξημένο CAS contention σε υψηλό αριθμό νημάτων. Τέλος, η εμφάνιση κορεσμού ή πτώσης της απόδοσης στα 64–128 νήματα αποδίδεται κυρίως σε αρχιτεκτονικούς περιορισμούς (hyperthreading και oversubscription) και όχι στους ίδιους τους αλγορίθμους.

**Σ.Η.Μ.Μ.Υ. Ε.Μ.Π.
Δεκέμβριος 2025**

ΣΥΣΤΗΜΑΤΑ ΠΑΡΑΛΛΗΛΗΣ ΕΠΕΞΕΡΓΑΣΙΑΣ

ΑΝΑΦΟΡΑ 4^{ης} ΑΣΚΗΣΗΣ



Στοιχεία Ομάδας

- Αναγνωριστικό: parlab05
- Μέλος 1^ο: Πέππας Μιχαήλ – Αθανάσιος, Α.Μ: 03121026
- Μέλος 2^ο: Σαουνάτσος Ανδρέας, Α.Μ: 03121197
- Ημερομηνία Παράδοσης Αναφοράς: 20.10.2025

▪ Εισαγωγή

Σκοπός της άσκησης είναι η παραλληλοποίηση και η βελτιστοποίηση του αλγορίθμου K-means σε επεξεργαστές γραφικών NVIDIA, μέσω CUDA. Η υλοποίηση βασίζεται στο μοντέλο CPU-GPU: η CPU (host) προετοιμάζει τα δεδομένα, εκκινεί τα kernels στην GPU (device) και (ανάλογα το πρόγραμμα) εκτελεί μέρος του αλγορίθμου και διαχειρίζεται μεταφορές μνήμης. Η άσκηση συγκρίνει 4 διαδοχικές εκδόσεις του αλγορίθμου, οι οποίες στοχεύουν στην ανάδειξη και αξιολόγηση κλασικών παραγόντων επίδοσης GPU: προσπελάσεις global memory/coalescing, αξιοποίηση shared memory, κόστος atomic operations και overhead επικοινωνίας host-device.

Σύμφωνα με τα ζητούμενα της άσκησης, δουλέψαμε και παραγάγαμε 4 εκδόσεις του αλγορίθμου K-means:

1. Naive (cuda_kmeans_naive.cu): η GPU υπολογίζει μόνο την ανάθεση στο κοντινότερο cluster ανά αντικείμενο. Η ενημέρωση των κέντρων (update_centroids) γίνεται στην CPU, με μεταφορές host \leftrightarrow device ανά επανάληψη.
2. Transpose (cuda_kmeans_transpose.cu): αναδιάταξη δεδομένων σε column-major transpose μορφή για βελτιωμένο memory coalescing στις προσπελάσεις της GPU (κοντινά threads διαβάζουν γειτονικές διευθύνσεις).
3. Shared (cuda_kmeans_shared.cu): επιπλέον φόρτωση των cluster centers στη shared memory ανά block, ώστε οι επαναλαμβανόμενες αναγνώσεις των clusters κατά τον υπολογισμό αποστάσεων να γίνονται από ταχύτερη on-chip μνήμη.
4. All-GPU (cuda_kmeans_all_gpu.cu): πλήρες offload και του update_centroids στη GPU. Η συσσώρευση (sums/counts) υλοποιείται με atomics, αναδεικνύοντας το κόστος συγχρονισμού/contention ως πιθανό bottleneck.

Οι παραπάνω αυτές εκδοχές είναι αυτές που θα αναλύσουμε και θα συγκρίνουμε στη συνέχεια.

▪ Ενότητα 3.1 – Naive Version

A. Εισαγωγή

Η «naive» έκδοση μεταφέρει στη GPU μόνο το πιο υπολογιστικά βαρύ βήμα του K-means: την ανάθεση κάθε αντικειμένου στο κοντινότερο κέντρο ενός cluster. Η ενημέρωση των κέντρων (update_centroids: αθροίσματα/πλήθη/μέσοι όροι) παραμένει στην CPU. Έτσι, σε κάθε επανάληψη εκτελούνται:

- Host → Device: αντιγραφή των τρεχόντων cluster centers στη GPU,
- GPU kernel: υπολογισμός nearest cluster για κάθε object και ενημέρωση membership/delta,
- Device → Host: αντιγραφή membership και delta πίσω στην CPU,
- CPU: update_centroids, παραγωγή νέων cluster centers για το επόμενο loop.

Η αρχική αντιγραφή του συνόλου των objects (dataset) προς τη GPU γίνεται μία φορά πριν το while-loop (initialization) και δεν αποτελεί μέρος του «per-loop» breakdown.

Ο κώδικας του αρχείου μας (cuda_kmeans_naive.cu) παρατίθεται ακολούθως:

a5/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 blockIdx.x * blockDim.x + threadIdx.x;
25 }
26
27 /* square of Euclid distance between two multi-dimensional points */
28 __host__ __device__ inline static
29 double euclid_dist_2(int numCoords,
30                      int numObjs,
31                      int numClusters,
32                      double *objects,      // [numObjs][numCoords]
33                      double *clusters,     // [numClusters][numCoords]
34                      int objectId,
35                      int clusterId) {
36     int i;
37     double ans = 0.0;
38
39     /* TODO: Calculate the euclid_dist of elem=objectId of objects from elem=clusterId from
clusters*/
40     for (i = 0; i < numCoords; i++) {
41         double objectVal = objects[objectId * numCoords + i];
42         double clusterVal = clusters[clusterId * numCoords + i];
43
44         double diff = objectVal - clusterVal;
45         ans += diff * diff;
46     }
47
48     return (ans);
49 }
50
51 __global__ static
```

```
52 void find_nearest_cluster(int numCoords,
53                           int numObjs,
54                           int numClusters,
55                           double *objects,           // [numObjs][numCoords]
56                           double *deviceClusters,    // [numClusters][numCoords]
57                           int *deviceMembership,     // [numObjs]
58                           double *devdelta) {
59
60     /* Get the global ID of the thread. */
61     int tid = get_tid();
62
63     if (tid < numObjs) {
64         int index, i;
65         double dist, min_dist;
66
67         /* find the cluster id that has min distance to object */
68         index = 0;
69
70         min_dist = euclid_dist_2(numCoords, numObjs, numClusters,
71                               objects, deviceClusters,
72                               tid, index);
73
74         for (i = 1; i < numClusters; i++) {
75
76             dist = euclid_dist_2(numCoords, numObjs, numClusters,
77                               objects, deviceClusters,
78                               tid, i);
79             /* no need square root */
80             if (dist < min_dist) { /* find the min and its array index */
81                 min_dist = dist;
82                 index = i;
83             }
84         }
85
86         if (deviceMembership[tid] != index) {
87
88             atomicAdd(devdelta, 1.0);
89         }
90
91         /* assign the deviceMembership to object objectId */
92         deviceMembership[tid] = index;
93     }
94 }
95
96 //
97 // -----
98 // DATA LAYOUT
99 //
100 // objects      [numObjs][numCoords]
101 // clusters     [numClusters][numCoords]
102 // newClusters   [numClusters][numCoords]
103 // deviceObjects  [numObjs][numCoords]
104 // deviceClusters [numClusters][numCoords]
105 // -----
```

```

106 //                                                 */
107 /* return an array of cluster centers of size [numClusters][numCoords]      */
108 void kmeans_gpu(double *objects,          /* in: [numObjs][numCoords] */
109                  int numCoords,        /* no. features */
110                  int numObjs,         /* no. objects */
111                  int numClusters,      /* no. clusters */
112                  double threshold,     /* % objects change membership */
113                  long loop_threshold,   /* maximum number of iterations */
114                  int *membership,      /* out: [numObjs] */
115                  double *clusters,       /* out: [numClusters][numCoords] */
116                  int blockSize) {
117
118     double timing = wtime(), timing_internal, timer_min = 1e42, timer_max = 0;
119     double timing_gpu, timing_cpu, timing_transfers, transfers_time = 0.0, cpu_time = 0.0,
120     gpu_time = 0.0;
121
122     int loop_iterations = 0;
123     int i, j, index, loop = 0;
124     int *newClusterSize; /* [numClusters]: no. objects assigned in each
125                           new cluster */
126
127     double delta = 0, *dev_delta_ptr;           /* % of objects change their clusters */
128     double **newClusters = (double **) calloc_2d(numClusters, numCoords, sizeof(double));
129
130     double *deviceObjects;
131     double *deviceClusters;
132     int *deviceMembership;
133
134     printf("\n|-----Naive GPU Kmeans-----|\n\n");
135
136
137     /* initialize membership[] */
138     for (i = 0; i < numObjs; i++) membership[i] = -1;
139
140     /* need to initialize newClusterSize and newClusters[0] to all 0 */
141     newClusterSize = (int *) calloc(numClusters, sizeof(int));
142     assert(newClusterSize != NULL);
143
144     timing = wtime() - timing;
145     printf("t_alloc: %lf ms\n\n", 1000 * timing);
146     timing = wtime();
147
148     const unsigned int numThreadsPerClusterBlock = (numObjs > blockSize) ? blockSize :
149     numObjs;
150
151     const unsigned int numClusterBlocks = (numObjs + numThreadsPerClusterBlock - 1) /
152     numThreadsPerClusterBlock; /* TODO: Calculate Grid size, e.g. number of blocks. */
153
154     const unsigned int clusterBlockSharedDataSize = 0;
155
156     checkCuda(cudaMalloc(&deviceObjects, numObjs * numCoords * sizeof(double)));
157     checkCuda(cudaMalloc(&deviceClusters, numClusters * numCoords * sizeof(double)));
158     checkCuda(cudaMalloc(&deviceMembership, numObjs * sizeof(int)));
159     checkCuda(cudaMalloc(&dev_delta_ptr, sizeof(double)));
160
161     timing = wtime() - timing;
162     printf("t_alloc_gpu: %lf ms\n\n", 1000 * timing);
163     timing = wtime();
164
165

```

```
157 checkCuda(cudaMemcpy(deviceObjects, objects,
158                     numObjs * numCoords * sizeof(double), cudaMemcpyHostToDevice));
159 checkCuda(cudaMemcpy(deviceMembership, membership,
160                     numObjs * sizeof(int), cudaMemcpyHostToDevice));
161 timing = wtime() - timing;
162 printf("t_get_gpu: %lf ms\n\n", 1000 * timing);
163 timing = wtime();
164
165 do {
166     timing_internal = wtime();
167
168     /* GPU part: calculate new memberships */
169
170     timing_transfers = wtime();
171
172     checkCuda(cudaMemcpy(deviceClusters, clusters,
173                         numClusters * numCoords * sizeof(double),
174                         cudaMemcpyHostToDevice));
175
176     transfers_time += wtime() - timing_transfers;
177
178     checkCuda(cudaMemset(dev_delta_ptr, 0, sizeof(double)));
179
180     //printf("Launching find_nearest_cluster Kernel with grid_size = %d, block_size = %d,
181     //shared_mem = %d KB\n", numClusterBlocks, numThreadsPerClusterBlock, clusterBlockSharedDa-
182     taSize/1000);
183     timing_gpu = wtime();
184     find_nearest_cluster
185     <<< numClusterBlocks, numThreadsPerClusterBlock, clusterBlockSharedDataSize >>>
186         (numCoords, numObjs, numClusters,
187          deviceObjects, deviceClusters, deviceMembership, dev_delta_ptr);
188
189     cudaDeviceSynchronize();
190     checkLastCudaError();
191     gpu_time += wtime() - timing_gpu;
192     //printf("Kernels complete for itter %d, updating data in CPU\n", loop);
193
194     timing_transfers = wtime();
195
196     checkCuda(cudaMemcpy(membership, deviceMembership,
197                         numObjs * sizeof(int),
198                         cudaMemcpyDeviceToHost));
199
200     checkCuda(cudaMemcpy(&delta, dev_delta_ptr,
201                         sizeof(double),
202                         cudaMemcpyDeviceToHost));
203
204     transfers_time += wtime() - timing_transfers;
205
206     /* CPU part: Update cluster centers*/
207     timing_cpu = wtime();
208     for (i = 0; i < numObjs; i++) {
209         /* find the array index of nestest cluster center */
210         index = membership[i];
```

```

209
210     /* update new cluster centers : sum of objects located within */
211     newClusterSize[index]++;
212     for (j = 0; j < numCoords; j++)
213         newClusters[index][j] += objects[i * numCoords + j];
214     }
215
216     /* average the sum and replace old cluster centers with newClusters */
217     for (i = 0; i < numClusters; i++) {
218         for (j = 0; j < numCoords; j++) {
219             if (newClusterSize[i] > 0)
220                 clusters[i * numCoords + j] = newClusters[i][j] / newClusterSize[i];
221             newClusters[i][j] = 0.0; /* set back to 0 */
222         }
223         newClusterSize[i] = 0; /* set back to 0 */
224     }
225
226     delta /= numObjs;
227     //printf("delta is %f - ", delta);
228     loop++;
229     //printf("completed loop %d\n", loop);
230     cpu_time += wtime() - timing_cpu;
231
232     timing_internal = wtime() - timing_internal;
233     if (timing_internal < timer_min) timer_min = timing_internal;
234     if (timing_internal > timer_max) timer_max = timing_internal;
235 } while (delta > threshold && loop < loop_threshold);
236
237 timing = wtime() - timing;
238 printf("\nloops = %d : total = %lf ms\n\t-> t_loop_avg = %lf ms\n\t-> t_loop_min = %lf
ms\n\t-> t_loop_max = %lf ms\n\t"
239         "-> t_cpu_avg = %lf ms\n\t-> t_gpu_avg = %lf ms\n\t-> t_transfers_avg = %lf
ms\n-----|\n",
240         loop, 1000 * timing, 1000 * timing / loop, 1000 * timer_min, 1000 * timer_max,
241         1000 * cpu_time / loop, 1000 * gpu_time / loop, 1000 * transfers_time / loop);
242
243 char outfile_name[1024] = {0};
244 sprintf(outfile_name, "Execution_logs/silver1-V100_Sz-%lu_Coo-%d_C1-%d.csv",
245         numObjs * numCoords * sizeof(double) / (1024 * 1024), numCoords, numClusters);
246 FILE *fp = fopen(outfile_name, "a+");
247 if (!fp) error("Filename %s did not open successfully, no logging performed\n",
outfile_name);
248 fprintf(fp, "%s,%d,%lf,%lf,%lf\n", "Naive", blockSize, timing / loop, timer_min,
timer_max);
249 fclose(fp);
250 checkCuda(cudaFree(deviceObjects));
251 checkCuda(cudaFree(deviceClusters));
252 checkCuda(cudaFree(deviceMembership));
253
254 free(newClusters[0]);
255 free(newClusters);
256 free(newClusterSize);
257
258 return;

```

259 | }

260 |

261 |

B. Υλοποίηση και Ορθότητα

(α) Υπολογισμός απόστασης (euclid dist 2)

Υλοποιείται η τετραγωνική Ευκλείδεια απόσταση σε μορφή row-major (naive έκδοση):

$$d^2(x, c) = \sum_{i=1}^n (x_i - c_i)^2$$

με indexing objects[objectId*numCoords + i], clusters[clusterId*numCoords + i]. Αυτή η προσέγγιση ακολουθήθηκε για τη naive μορφή δεδομένων.

(β) Kernel find_nearest_cluster: αντιστοίχιση threads σε objects

Κάθε thread αντιστοιχεί σε ένα object μέσω global thread id:

$$tid = blockIdx.x * blockDim.x + threadIdx.x.$$

Ο αριθμός blocks ορίζεται ως $\text{ceil}(\text{numObjs} / \text{block_size})$, ώστε να καλύπτονται όλα τα objects, και γίνεται έλεγχος ορίων ($tid < \text{numObjs}$).

(γ) Υπολογισμός delta με atomics

Η μεταβλητή delta μετρά πόσα objects άλλαξαν cluster σε μία επανάληψη. Στο kernel, αν το νέο clusterId διαφέρει από το παλιό membership[tid], γίνεται atomicAdd(devdelta, 1).

Η επιλογή atomics είναι σωστή για αποφυγή race conditions, αλλά αποτελεί και κλασικό σημείο bottleneck (contention) όταν πολλά threads ενημερώνουν την ίδια global μεταβλητή. Δηλαδή, τα atomics επιτυγχάνουν ορθότητα, αλλά όχι απαραίτητα επίδοση, και συχνά αντικαθίστανται από reduction patterns.

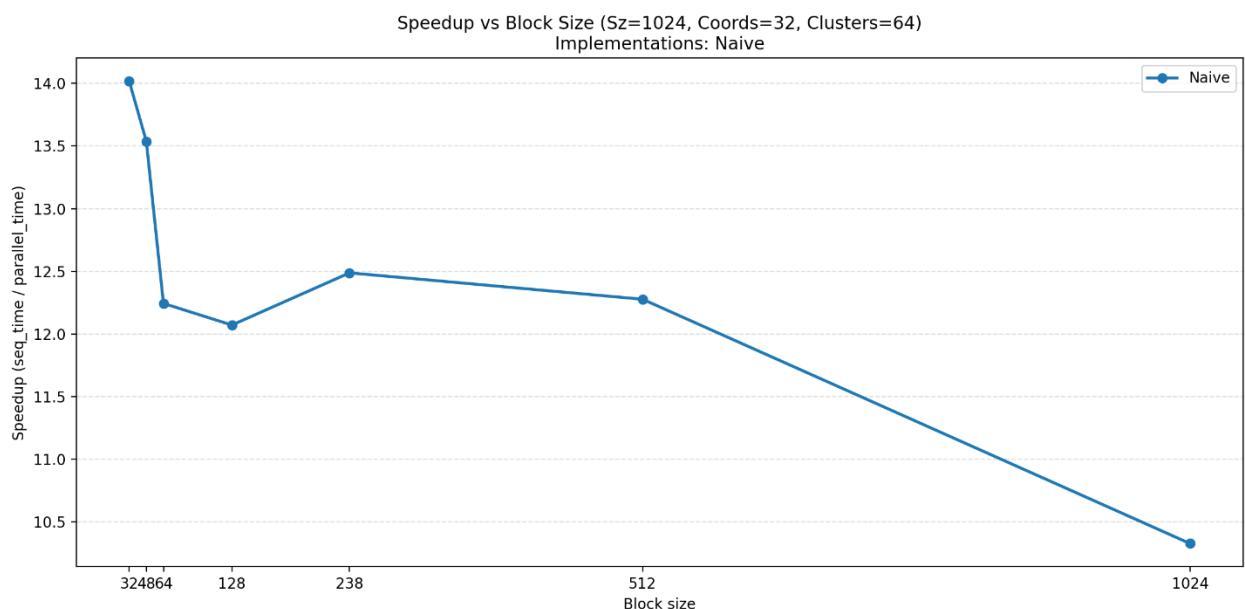
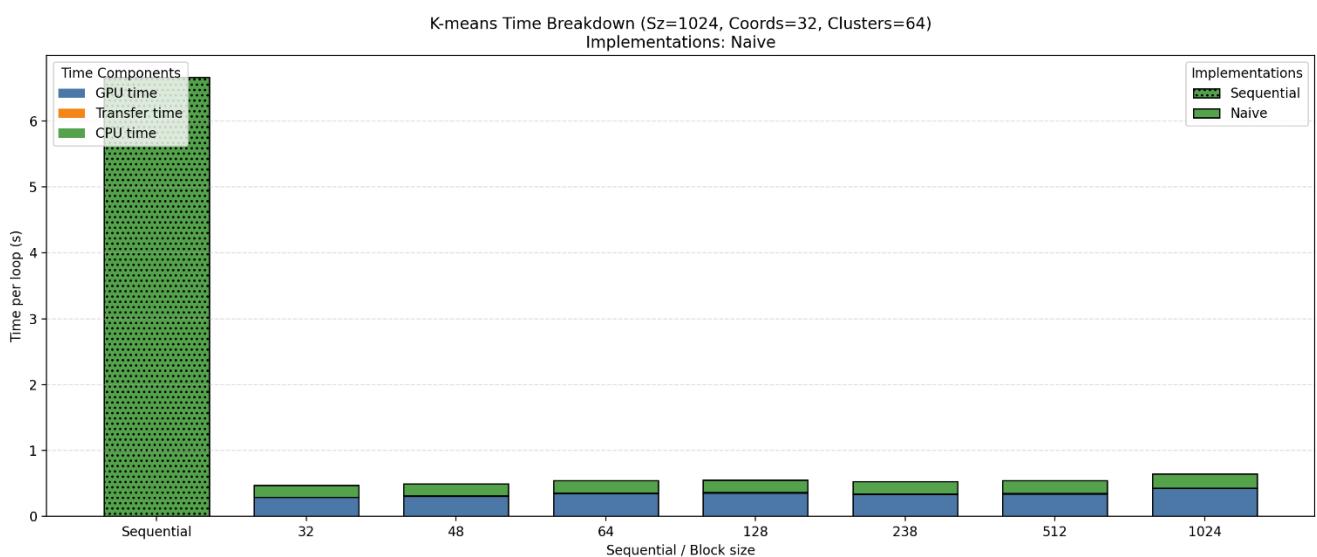
(δ) Timers (breakdown CPU / GPU / Transfers)

Στη naïve έκδοση καταγράφονται τρεις συνιστώσες χρόνου ανά loop:

- GPU time: χρόνος εκτέλεσης του kernel,
- Transfers time: χρόνος αντιγραφών host \leftrightarrow device που γίνονται μέσα στο loop,
- CPU time: χρόνος update_centroids στην CPU.

Επισημαίνεται ότι το «μεγάλο» H \rightarrow D copy του dataset (objects) γίνεται πριν την επανάληψη και μετριέται ξεχωριστά (άρα δεν εμφανίζεται στο per-loop transfers_avg).

Γ. Παρουσίαση Διαγραμμάτων



Δ. Ερμηνεία Διαγραμμάτων

(1) Speedup vs Block Size

Παρατηρείται μέγιστο speedup για μικρά block sizes (π.χ. 32–48) και σταδιακή υποβάθμιση για πολύ μεγάλα block sizes (έως 1024). Η συμπεριφορά αυτή είναι αναμενόμενη βάσει θεωρίας:

- Με μικρότερα blocks, ο scheduler μπορεί να διατηρεί περισσότερα resident blocks/warps ανά SM, αυξάνοντας την ικανότητα απόκρυψης latency (occupancy/latency hiding).
- Με πολύ μεγάλα blocks, μειώνεται ο αριθμός blocks που χωρούν ταυτόχρονα σε ένα SM (λόγω ορίων threads/SM ή πόρων όπως registers), άρα μειώνονται τα ενεργά warps και η GPU δυσκολεύεται να κρύψει memory latency. Επιπλέον, η naïve πρόσβαση σε global μνήμη (objects/clusters) κάνει την επίδοση πιο ευαίσθητη σε occupancy.

(2) Time Breakdown

To breakdown δείχνει ότι:

- Ο συνολικός χρόνος ανά loop της naïve έκδοσης είναι πολύ μικρότερος από το sequential baseline, άρα επιτυγχάνεται σημαντικό speedup.
- To GPU time είναι η κυρίαρχη συνιστώσα (όπως αναμενόταν, αφού το assignment είναι το κύριο υπολογιστικό μέρος).
- Τα transfer times φαίνονται μηδαμηνά για Coords=32 και αυτό είναι λογικό: μέσα στο loop μεταφέρονται κυρίως (i) τα clusters (πολύ μικρά, ~KB) και (ii) το membership (μεγαλύτερο, αλλά όχι συγκρίσιμο με το 1GB dataset). Αντίθετα, η αρχική αντιγραφή του dataset προς τη GPU (1GB) γίνεται εκτός loop και δεν συμπεριλαμβάνεται στο transfers_avg του breakdown. Ωστόσο, το membership είναι $O(N)$ ανά επανάληψη (Device \rightarrow Host) και μπορεί να γίνει σημαντικό όταν το πλήθος objects N μεγαλώνει, ιδιαίτερα στο Coords=2 όπου για ίδιο Size προκύπτει πολύ μεγαλύτερο N. Άρα, το «μικρά transfers» ισχύει εδώ, για Coords=32, και όχι γενικά.

E. Συμπεράσματα

Η παίνε παραλληλοποίηση επιβεβαιώνει ότι το «assignment step» είναι κατάλληλο για GPU (data-parallel, ανεξάρτητος υπολογισμός ανά object), προσφέροντας υψηλό speedup. Ωστόσο, παραμένουν δύο εγγενή όρια:

- Επικοινωνία και CPU work ανά iteration (clusters/membership transfers + update_centroids στην CPU),
- Atomics για το delta (πιθανό contention).

Το K-means δεν είναι «ιδανικός» πυρήνας GPU ως συνολικός αλγόριθμος, αλλά περιέχει ένα τμήμα που είναι ιδιαίτερα κατάλληλο. Συγκεκριμένα, το βήμα ανάθεσης (assignment: για κάθε object υπολογισμός απόστασης από όλα τα clusters και επιλογή του ελάχιστου) είναι έντονα data-parallel, με ανεξάρτητη εργασία ανά object και μεγάλη παραλληλία, άρα ταιριάζει πολύ καλά στο SIMT μοντέλο των GPUs. Ωστόσο, η συνολική δομή του K-means είναι επαναληπτική και απαιτεί συγχρονισμό μεταξύ επαναλήψεων, ενώ το update των κέντρων είναι reduction/accumulation (sums & counts) και συχνά επιβαρύνεται από atomics και μη ευνοϊκές προσπελάσεις μνήμης. Στη παίνε υλοποίησή μας, επιπλέον, μέρος του κόστους παραμένει εκτός GPU (CPU update + μεταφορές membership/centroids ανά loop), άρα η επίδοση δεν εξαρτάται μόνο από το kernel αλλά και από επικοινωνία/overhead.

Αυτά αποτελούν και το κίνητρο για τις επόμενες εκδόσεις: βελτίωση προσπελάσεων global μνήμης (transpose/coalescing), επαναχρησιμοποίηση δεδομένων μέσω shared memory, και στη συνέχεια πλήρες offload (all-gpu) για μείωση CPU/transfer overhead.

▪ Ενότητα 3.2 – Transpose Version

A. Εισαγωγή

Η έκδοση Transpose στοχεύει αποκλειστικά στη βελτιστοποίηση των προσπελάσεων της global μνήμης στην GPU. Στην naïve έκδοση, τα threads ενός warp που επεξεργάζονται διαδοχικά objects προσπελαύνουν τα δεδομένα με «stride» ως προς τις συντεταγμένες (row-major layout), οδηγώντας σε μη coalesced accesses και αυξημένο αριθμό memory transactions. Η Transpose έκδοση αλλάζει τη διάταξη των δεδομένων σε column-based (transpose) μορφή, έτσι ώστε για κάθε συντεταγμένη i, τα 32 threads ενός warp να διαβάζουν συνεχόμενες διευθύνσεις μνήμης (coalescing), μειώνοντας δραστικά το κόστος πρόσβασης στη global memory και άρα τον χρόνο του kernel.

Ο κώδικας του αρχείου μας (cuda_kmeans_transpose.cu) παρατίθεται ακολούθως:

a5/cuda_kmeans_transpose.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 blockIdx.x * blockDim.x + threadIdx.x;
25 }
26
27 /* square of Euclid distance between two multi-dimensional points using column-base format */
28 __host__ __device__ inline static
29 double euclid_dist_2_transpose(int numCoords,
30                                int numObjs,
31                                int numClusters,
32                                double *objects,      // [numCoords][numObjs]
33                                double *clusters,     // [numCoords][numClusters]
34                                int objectId,
35                                int clusterId) {
36     int i;
37     double ans = 0.0;
38
39     /* TODO: Calculate the euclid_dist of elem=objectId of objects from elem=clusterId from
clusters, but for column-base format!!! */
40     for (i = 0; i < numCoords; i++) {
41         double objectVal = objects[i * numObjs + objectId];
42         double clusterVal = clusters[i * numClusters + clusterId];
43
44         double diff = objectVal - clusterVal;
45         ans += diff * diff;
46     }
47
48     return (ans);
49 }
50
```

```
51 __global__ static
52 void find_nearest_cluster(int numCoords,
53                           int numObjs,
54                           int numClusters,
55                           double *objects,           // [numCoords][numObjs]
56                           double *deviceClusters,    // [numCoords][numClusters]
57                           int *membership,           // [numObjs]
58                           double *devdelta) {
59
60     /* Get the global ID of the thread. */
61     int tid = get_tid();
62
63     if (tid < numObjs) {
64         int index, i;
65         double dist, min_dist;
66
67         /* find the cluster id that has min distance to object */
68         index = 0;
69
70         min_dist = euclid_dist_2_transpose(numCoords, numObjs, numClusters,
71                                           objects, deviceClusters,
72                                           tid, index);
73
74         for (i = 1; i < numClusters; i++) {
75
76             dist = euclid_dist_2_transpose(numCoords, numObjs, numClusters,
77                                             objects, deviceClusters,
78                                             tid, i);
79
80             /* no need square root */
81             if (dist < min_dist) { /* find the min and its array index */
82                 min_dist = dist;
83                 index = i;
84             }
85         }
86
87         if (membership[tid] != index) {
88
89             atomicAdd(devdelta, 1.0);
90
91         /* assign the deviceMembership to object objectId */
92         membership[tid] = index;
93     }
94
95     //
96     // -----
97     // DATA LAYOUT
98     //
99     // objects      [numObjs][numCoords]
100    // clusters     [numClusters][numCoords]
101    // dimObjects   [numCoords][numObjs]
102    // dimClusters  [numCoords][numClusters]
103    // newClusters  [numCoords][numClusters]
104    // deviceObjects [numCoords][numObjs]
```

```

105 // deviceClusters [numCoords][numClusters]
106 // -----
107 //
108 /* return an array of cluster centers of size [numClusters][numCoords] */
109 void kmeans_gpu(double *objects,      /* in: [numObjs][numCoords] */
110                 int numCoords,    /* no. features */
111                 int numObjs,     /* no. objects */
112                 int numClusters, /* no. clusters */
113                 double threshold, /* % objects change membership */
114                 long loop_threshold, /* maximum number of iterations */
115                 int *membership,  /* out: [numObjs] */
116                 double *clusters, /* out: [numClusters][numCoords] */
117                 int blockSize) {
118     double timing = wtime(), timing_internal, timer_min = 1e42, timer_max = 0;
119     double timing_gpu, timing_cpu, timing_transfers, transfers_time = 0.0, cpu_time = 0.0,
gpu_time = 0.0;
120     int loop_iterations = 0;
121     int i, j, index, loop = 0;
122     int *newClusterSize; /* [numClusters]: no. objects assigned in each
new cluster */
123     double delta = 0, *dev_delta_ptr; /* % of objects change their clusters */
124
125     /* TODO: Transpose dims */
126     double **dimObjects = (double **) calloc_2d(numCoords, numObjs, sizeof(double));
//calloc_2d(..., numCoords, numObjs)
127     double **dimClusters = (double **) calloc_2d(numCoords, numClusters, sizeof(double));
//calloc_2d(..., numCoords, numClusters)
128     double **newClusters = (double **) calloc_2d(numCoords, numClusters, sizeof(double));
//calloc_2d(..., numCoords, numClusters)
129
130     double *deviceObjects;
131     double *deviceClusters;
132     int *deviceMembership;
133
134     printf("\n|-----Transpose GPU Kmeans-----|\n\n");
135
136     // TODO: Copy objects given in [numObjs][numCoords] layout to new
137     // [numCoords][numObjs] layout
138     for (i=0 ; i < numObjs; i++){
139         for (j=0; j<numCoords; j++){
140             dimObjects[j][i]=objects[i*numCoords + j];
141         }
142     }
143
144     /* pick first numClusters elements of objects[] as initial cluster centers*/
145     for (i = 0; i < numCoords; i++) {
146         for (j = 0; j < numClusters; j++) {
147             dimClusters[i][j] = dimObjects[i][j];
148         }
149     }
150
151     /* initialize membership[] */
152     for (i = 0; i < numObjs; i++) membership[i] = -1;
153
154

```

```
155 /* need to initialize newClusterSize and newClusters[0] to all 0 */
156 newClusterSize = (int *) calloc(numClusters, sizeof(int));
157 assert(newClusterSize != NULL);
158
159 timing = wtime() - timing;
160 printf("t_alloc: %lf ms\n\n", 1000 * timing);
161 timing = wtime();
162
163 const unsigned int numThreadsPerClusterBlock = (numObjs > blockSize) ? blockSize :
164 numObjs;
165 const unsigned int numClusterBlocks = (numObjs + numThreadsPerClusterBlock - 1) /
166 numThreadsPerClusterBlock;
167 const unsigned int clusterBlockSharedDataSize = 0;
168
169 checkCuda(cudaMalloc(&deviceObjects, numObjs * numCoords * sizeof(double)));
170 checkCuda(cudaMalloc(&deviceClusters, numClusters * numCoords * sizeof(double)));
171 checkCuda(cudaMalloc(&deviceMembership, numObjs * sizeof(int)));
172 checkCuda(cudaMalloc(&dev_delta_ptr, sizeof(double)));
173 timing = wtime() - timing;
174 printf("t_alloc_gpu: %lf ms\n\n", 1000 * timing);
175 timing = wtime();
176
177 checkCuda(cudaMemcpy(deviceObjects, dimObjects[0],
178                      numObjs * numCoords * sizeof(double), cudaMemcpyHostToDevice));
179 checkCuda(cudaMemcpy(deviceMembership, membership,
180                      numObjs * sizeof(int), cudaMemcpyHostToDevice));
181 timing = wtime() - timing;
182 printf("t_get_gpu: %lf ms\n\n", 1000 * timing);
183 timing = wtime();
184
185 do {
186     timing_internal = wtime();
187
188     /* GPU part: calculate new memberships */
189
190     timing_transfers = wtime();
191     // TODO: Copy clusters to deviceClusters
192     checkCuda(cudaMemcpy(deviceClusters, dimClusters[0],
193                          numClusters * numCoords * sizeof(double),
194                          cudaMemcpyHostToDevice));
195
196     transfers_time += wtime() - timing_transfers;
197
198     checkCuda(cudaMemset(dev_delta_ptr, 0, sizeof(double)));
199
200     //printf("Launching find_nearest_cluster Kernel with grid_size = %d, block_size = %d,
201     //shared_mem = %d KB\n", numClusterBlocks, numThreadsPerClusterBlock, clusterBlockSharedDa-
202     taSize/1000);
203     timing_gpu = wtime();
204     find_nearest_cluster
205     <<< numClusterBlocks, numThreadsPerClusterBlock, clusterBlockSharedDataSize >>>
206         (numCoords, numObjs, numClusters,
207          deviceObjects, deviceClusters, deviceMembership, dev_delta_ptr);
```

```
205     cudaDeviceSynchronize();
206     checkLastCudaError();
207     gpu_time += wtime() - timing_gpu;
208     //printf("Kernels complete for itter %d, updating data in CPU\n", loop);
209
210     timing_transfers = wtime();
211
212     checkCuda(cudaMemcpy(membership, deviceMembership,
213                         numObjs * sizeof(int),
214                         cudaMemcpyDeviceToHost));
215
216     checkCuda(cudaMemcpy(&delta, dev_delta_ptr,
217                         sizeof(double),
218                         cudaMemcpyDeviceToHost));
219     transfers_time += wtime() - timing_transfers;
220
221     /* CPU part: Update cluster centers*/
222
223     timing_cpu = wtime();
224     for (i = 0; i < numObjs; i++) {
225         /* find the array index of nestest cluster center */
226         index = membership[i];
227
228         /* update new cluster centers : sum of objects located within */
229         newClusterSize[index]++;
230         for (j = 0; j < numCoords; j++)
231             newClusters[j][index] += objects[i * numCoords + j];
232     }
233
234     /* average the sum and replace old cluster centers with newClusters */
235     for (i = 0; i < numClusters; i++) {
236         for (j = 0; j < numCoords; j++) {
237             if (newClusterSize[i] > 0)
238                 dimClusters[j][i] = newClusters[j][i] / newClusterSize[i];
239             newClusters[j][i] = 0.0; /* set back to 0 */
240         }
241         newClusterSize[i] = 0; /* set back to 0 */
242     }
243
244     delta /= numObjs;
245     //printf("delta is %f - ", delta);
246     loop++;
247     //printf("completed loop %d\n", loop);
248     cpu_time += wtime() - timing_cpu;
249
250     timing_internal = wtime() - timing_internal;
251     if (timing_internal < timer_min) timer_min = timing_internal;
252     if (timing_internal > timer_max) timer_max = timing_internal;
253 } while (delta > threshold && loop < loop_threshold);
254
255 /*TODO: Update clusters using dimClusters. Be carefull of layout!!!
clusters[numClusters][numCoords] vs dimClusters[numCoords][numClusters] */
256 for (i = 0; i < numClusters; i++) {
257     for (j = 0; j < numCoords; j++) {
```

```
258         clusters[i * numCoords + j] = dimClusters[j][i];
259     }
260 }
261
262 timing = wtime() - timing;
263 printf("nloops = %d : total = %lf ms\n\t-> t_loop_avg = %lf ms\n\t-> t_loop_min = %lf
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
ms\n\n|-----|\n",
264     loop, 1000 * timing, 1000 * timing / loop, 1000 * timer_min, 1000 * timer_max,
265     1000 * cpu_time / loop, 1000 * gpu_time / loop, 1000 * transfers_time / loop);
266
267
268 char outfile_name[1024] = {0};
269 sprintf(outfile_name, "Execution_logs/silver1-V100_Sz-%lu_Coo-%d_C1-%d.csv",
270         numObjs * numCoords * sizeof(double) / (1024 * 1024), numCoords, numClusters);
271 FILE *fp = fopen(outfile_name, "a+");
272 if (!fp) error("Filename %s did not open successfully, no logging performed\n",
outfile_name);
273 fprintf(fp, "%s,%d,%lf,%lf,%lf\n", "Transpose", blockSize, timing / loop, timer_min,
timer_max);
274 fclose(fp);
275
276 checkCuda(cudaFree(deviceObjects));
277 checkCuda(cudaFree(deviceClusters));
278 checkCuda(cudaFree(deviceMembership));
279
280 free(dimObjects[0]);
281 free(dimObjects);
282 free(dimClusters[0]);
283 free(dimClusters);
284 free(newClusters[0]);
285 free(newClusters);
286 free(newClusterSize);
287
288 return;
289 }
290
291 }
```

B. Υλοποίηση και Ορθότητα

Η λογική του αλγορίθμου παραμένει ίδια: η GPU εκτελεί το assignment (membership) και η CPU εκτελεί το update_centroids. Η αλλαγή είναι καθαρά στη δομή δεδομένων:

- Αντί για `objects[object][coord]`, δημιουργείται `dimObjects[coord][object]`.
- Αντί για `clusters[cluster][coord]`, δημιουργείται `dimClusters[coord][cluster]`.

Επισημαίνουμε τα εξής σημεία:

(α) Νέα συνάρτηση απόστασης euclid dist 2 transpose

Υπολογίζεται η ίδια Ευκλείδεια απόσταση, αλλά με indexing που ευνοεί coalescing:

`objects[i*numObjs + objectId]` και `clusters[i*numClusters + clusterId]`.

Έτσι, για σταθερό i , τα threads του warp διαβάζουν συνεχόμενα `objects` (`objectId` διαδοχικά), άρα οι αναγνώσεις είναι coalesced.

(β) Μετασχηματισμός των δεδομένων (transpose) πριν την επανάληψη

Πριν ξεκινήσει το loop, ο host κατασκευάζει τον `dimObjects` πίνακα από το αρχικό row-major `objects`. Αυτό είναι κόστος προεπεξεργασίας (εκτός loop) και δεν επηρεάζει το per-loop breakdown.

(γ) Ενημέρωση clusters με transpose μορφή

Στο CPU `update_centroids`, τα αθροίσματα/μέσοι όροι ενημερώνονται στη μορφή `dimClusters[coord][cluster]` ώστε το επόμενο H \rightarrow D copy να διατηρεί τη coalesced διάταξη. Στο τέλος γίνεται back-transform σε `clusters[cluster][coord]` μόνο για λόγους συμβατότητας/εκτύπωσης.

Γενικά, η Transpose έκδοση είναι αριθμητικά ισοδύναμη με τη Naive (ίδια μετρική απόστασης, ίδια διαδικασία ανάθεσης/ενημέρωσης), αλλά με διαφορετική διάταξη στη μνήμη. Επομένως, αναμένουμε τα ίδια clusters (εντός floating-point διαφορών) και το ίδιο κριτήριο σύγκλισης· η διαφορά αφορά αποκλειστικά την επίδοση λόγω memory access pattern.

Γ. Παρουσίαση Διαγραμμάτων



Δ. Ερμηνεία Διαγραμμάτων

(1) Speedup vs Block Size

Η Transpose έκδοση παρουσιάζει σημαντικά υψηλότερο speedup από τη Naive (περίπου 23–25 έναντι ~10–14), και μάλιστα με σχετικά «επίπεδη» συμπεριφορά ως προς το block size. Αυτό είναι αναμενόμενο, καθώς:

- Με coalesced προσπελάσεις, μειώνονται τα global memory transactions ανά warp, άρα αυξάνεται το effective bandwidth και μειώνεται ο χρόνος kernel.
- Όταν το κύριο bottleneck είναι οι προσπελάσεις μνήμης, η βελτίωση στο memory access pattern έχει μεγαλύτερη επίδραση από μικρο-βελτιστοποιήσεις scheduling/occupancy μέσω block size, με αποτέλεσμα πιο σταθερή καμπύλη.

Στη Transpose έκδοση το block_size παίζει σαφώς μικρότερο ρόλο σε σχέση με τη Naive, όπως φαίνεται από τη σχεδόν επίπεδη καμπύλη speedup. Ο λόγος είναι ότι με το transpose πετυχαίνουμε coalesced προσπελάσεις στη global μνήμη (τα threads ενός warp διαβάζουν συνεχόμενες διευθύνσεις για κάθε συντεταγμένη), άρα μειώνεται δραστικά το κόστος memory transactions και το kernel γίνεται λιγότερο ευαίσθητο σε αλλαγές occupancy/scheduling που προκαλεί το block_size. Εφόσον το block_size είναι πολλαπλάσιο του 32 (warp size) και διατηρεί επαρκή ενεργά warps ανά SM, η απόδοση παραμένει σχεδόν σταθερή. Μόνο σε ακραία μεγέθη blocks ενδέχεται να εμφανιστεί μικρή πτώση (π.χ. λόγω μειωμένων resident blocks/warps ανά SM ή αυξημένων απαιτήσεων πόρων), αλλά συνολικά το κυρίαρχο κέρδος στη Transpose προέρχεται από το βελτιωμένο memory access pattern και όχι από την επιλογή block_size.

(2) Time Breakdown (GPU / Transfers / CPU)

To breakdown δείχνει ότι η κύρια μείωση χρόνου προέρχεται από το GPU time (kernel). Αυτό είναι ακριβώς το αναμενόμενο αποτέλεσμα της βελτιστοποίησης coalescing: δεν αλλάζουμε τις μεταφορές ανά loop ούτε το CPU update_centroids, αλλά μειώνουμε δραστικά τον χρόνο της φάσης assignment στη GPU.

Τα transfer times παραμένουν χαμηλά στο συγκεκριμένο σενάριο (Coords=32), διότι εντός loop μεταφέρεται:

- Host→Device: clusters (μικρό μέγεθος),
- Device→Host: membership + delta.

Η αρχική αντιγραφή των objects (1GB) γίνεται εκτός loop και δεν περιλαμβάνεται στο per-loop transfer χρόνο.

E. Σύγκριση Αποτελεσμάτων (με Naive)

1. Κύριο εύρημα: Η Transpose έκδοση επιτυγχάνει $\sim 1.7\times\text{--}2\times$ καλύτερο speedup από τη Naive, παρότι ο αλγόριθμος παραμένει ο ίδιος.
2. Η βελτίωση δεν οφείλεται σε περισσότερους υπολογισμούς στη GPU, αλλά σε καθαρά αρχιτεκτονικό λόγο: καλύτερη αξιοποίηση του memory subsystem μέσω coalescing (32-thread warps → συνεχόμενες διευθύνσεις → λιγότερα transactions).
3. Η συνιστώσα CPU (update_centroids) παραμένει πρακτικά η ίδια, άρα το συνολικό κέρδος έρχεται από τη μείωση του kernel time.
4. Η εξάρτηση από block size είναι μικρότερη σε σχέση με τη Naive, επειδή η Transpose μειώνει το memory overhead και σταθεροποιεί την απόδοση.

ΣΤ. Συμπεράσματα

Η Transpose έκδοση επιβεβαιώνει ότι η διάταξη των δεδομένων στη μνήμη μπορεί να είναι καθοριστική για την απόδοση σε GPU. Με την αναδιάταξη σε column-based μορφή πετυχαίνουμε coalesced global memory accesses κατά τον υπολογισμό αποστάσεων, μειώνοντας αισθητά τον χρόνο του kernel και αυξάνοντας το speedup. Αυτό αποτελεί το φυσικό επόμενο βήμα μετά τη Naive προσέγγιση και δημιουργεί τη βάση για την επόμενη βελτιστοποίηση (Shared), όπου στοχεύουμε επιπλέον στη μείωση των επαναλαμβανόμενων αναγνώσεων clusters μέσω shared memory (on-chip reuse).

- **Ενότητα 3.3 – Shared Version**

A. Εισαγωγή

Η έκδοση Shared επεκτείνει την βελτιστοποίηση που εισαγάγαμε στην Transpose και στοχεύει στη μείωση των επαναλαμβανόμενων αναγνώσεων των cluster centers από την global μνήμη. Στο K-means, για κάθε object υπολογίζονται αποστάσεις από όλα τα clusters. Άρα, τα ίδια cluster centers επαναχρησιμοποιούνται πολλές φορές από τα threads ενός block. Με τη φόρτωσή τους στη shared memory (on-chip), μειώνουμε σημαντικά το global memory traffic και επιταχύνουμε το assignment kernel.

Ο κώδικας του αρχείου μας (cuda_kmeans_shared.cu) παρατίθεται ακολούθως:

a5/cuda_kmeans_shared.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 blockIdx.x * blockDim.x + threadIdx.x;
25 }
26
27 /* square of Euclid distance between two multi-dimensional points using column-base format */
28 __host__ __device__ inline static
29 double euclid_dist_2_transpose(int numCoords,
30                                int numObjs,
31                                int numClusters,
32                                double *objects,      // [numCoords][numObjs]
33                                double *clusters,     // [numCoords][numClusters]
34                                int objectId,
35                                int clusterId) {
36     int i;
37     double ans = 0.0;
38
39     /* TODO: Calculate the euclid_dist of elem=objectId of objects from elem=clusterId from
clusters, but for column-base format!!! */
40     for (i = 0; i < numCoords; i++) {
41         double objectVal = objects[i * numObjs + objectId];
42         double clusterVal = clusters[i * numClusters + clusterId];
43
44         double diff = objectVal - clusterVal;
45         ans += diff * diff;
46     }
47
48     return (ans);
49 }
50
```

```

51 __global__ static
52 void find_nearest_cluster(int numCoords,
53                           int numObjs,
54                           int numClusters,
55                           double *objects,           // [numCoords][numObjs]
56                           double *deviceClusters,    // [numCoords][numClusters]
57                           int *deviceMembership,     // [numObjs]
58                           double *devdelta) {
59     extern __shared__ double shmemClusters[];
60
61     // TODO: Copy deviceClusters to shmemClusters so they can be accessed faster.
62     int tid_in_block = threadIdx.x;      // Το ID του νήματος μέσα στο Block
63     int block_size = blockDim.x;         // Πόσα νήματα έχει το Block
64     int total_cluster_doubles = numClusters * numCoords; // Συνολικά νούμερα προς αντιγραφή
65
66     // Κάθε νήμα αντιγράφει όσα στοιχεία του αναλογούν (με βήμα block_size)
67     for (int k = tid_in_block; k < total_cluster_doubles; k += block_size) {
68         shmemClusters[k] = deviceClusters[k];
69     }
70
71     /* Συγχρονισμός (BARRIER) */
72
73     __syncthreads();
74
75     /* Get the global ID of the thread. */
76     int tid = get_tid();
77
78     /* TODO: Maybe something is missing here... should all threads run this? */
79     if (tid < numObjs) {
80         int index, i;
81         double dist, min_dist;
82
83         /* find the cluster id that has min distance to object */
84         index = 0;
85         /* TODO: call min_dist = euclid_dist_2(...) with correct objectId/clusterId using
clusters in shmem*/
86
87
88         min_dist = euclid_dist_2_transpose(numCoords, numObjs, numClusters,
89                                           objects, shmemClusters,
90                                           tid, index);
91
92         for (i = 1; i < numClusters; i++) {
93             dist = euclid_dist_2_transpose(numCoords, numObjs, numClusters,
94                                           objects, shmemClusters,
95                                           tid, i);
96
97             /* no need square root */
98             if (dist < min_dist) { /* find the min and its array index */
99                 min_dist = dist;
100                index = i;
101            }
102        }
103    }

```

```

104     if (deviceMembership[tid] != index) {
105         /* TODO: Maybe something is missing here... is this write safe? */
106         atomicAdd(devdelta, 1.0);
107     }
108
109     /* assign the deviceMembership to object objectId */
110     deviceMembership[tid] = index;
111 }
112 }
113
114 //
115 // -----
116 // DATA LAYOUT
117 //
118 // objects      [numObjs][numCoords]
119 // clusters     [numClusters][numCoords]
120 // dimObjects   [numCoords][numObjs]
121 // dimClusters  [numCoords][numClusters]
122 // newClusters  [numCoords][numClusters]
123 // deviceObjects [numCoords][numObjs]
124 // deviceClusters [numCoords][numClusters]
125 //
126 //
127 /* return an array of cluster centers of size [numClusters][numCoords]      */
128 void kmeans_gpu(double *objects,          /* in: [numObjs][numCoords] */
129                  int numCoords,    /* no. features */
130                  int numObjs,     /* no. objects */
131                  int numClusters, /* no. clusters */
132                  double threshold, /* % objects change membership */
133                  long loop_threshold, /* maximum number of iterations */
134                  int *membership, /* out: [numObjs] */
135                  double *clusters, /* out: [numClusters][numCoords] */
136                  int blockSize) {
137     double timing = wtime(), timing_internal, timer_min = 1e42, timer_max = 0;
138     double timing_gpu, timing_cpu, timing_transfers, transfers_time = 0.0, cpu_time = 0.0,
gpu_time = 0.0;
139     int loop_iterations = 0;
140     int i, j, index, loop = 0;
141     int *newClusterSize; /* [numClusters]: no. objects assigned in each
new cluster */
142     double delta = 0, *dev_delta_ptr; /* % of objects change their clusters */
143     /* TODO: Copy me from transpose version*/
144     double **dimObjects = (double **) calloc_2d(numCoords, numObjs, sizeof(double));
//calloc_2d(...)->[numCoords][numObjs]
145     double **dimClusters = (double **) calloc_2d(numCoords, numClusters, sizeof(double));
//calloc_2d(...)->[numCoords][numClusters]
146     double **newClusters = (double **) calloc_2d(numCoords, numClusters, sizeof(double));
//calloc_2d(...)->[numCoords][numClusters]
147
148     double *deviceObjects;
149     double *deviceClusters;
150     int *deviceMembership;
151
152     printf("\n|-----Shared GPU Kmeans-----|\n\n");

```

```

154
155     /* TODO: Copy me from transpose version*/
156     for (i=0 ; i < numObjs; i++){
157         for (j=0; j<numCoords; j++){
158             dimObjects[j][i]=objects[i*numCoords + j];
159         }
160     }
161
162     /* pick first numClusters elements of objects[] as initial cluster centers*/
163     for (i = 0; i < numCoords; i++) {
164         for (j = 0; j < numClusters; j++) {
165             dimClusters[i][j] = dimObjects[i][j];
166         }
167     }
168
169     /* initialize membership[] */
170     for (i = 0; i < numObjs; i++) membership[i] = -1;
171
172     /* need to initialize newClusterSize and newClusters[0] to all 0 */
173     newClusterSize = (int *) calloc(numClusters, sizeof(int));
174     assert(newClusterSize != NULL);
175
176     timing = wtime() - timing;
177     printf("t_alloc: %lf ms\n\n", 1000 * timing);
178     timing = wtime();
179     const unsigned int numThreadsPerClusterBlock = (numObjs > blockSize) ? blockSize :
180     numObjs;
181     const unsigned int numClusterBlocks = (numObjs + numThreadsPerClusterBlock - 1) /
182     numThreadsPerClusterBlock; /* TODO: Calculate Grid size, e.g. number of blocks. */
183
184     /* Define the shared memory needed per block.
185      - BEWARE: We can overrun our shared memory here if there are too many
186      clusters or too many coordinates!
187      - This can lead to occupancy problems or even inability to run.
188      - Your exercise implementation is not requested to account for that (e.g. always
189      assume deviceClusters fit in shmemClusters */
190     const unsigned int clusterBlockSharedDataSize = numClusters*numCoords*sizeof(double);
191
192     cudaDeviceProp deviceProp;
193     int deviceNum;
194     cudaGetDevice(&deviceNum);
195     cudaGetDeviceProperties(&deviceProp, deviceNum);
196
197     if (clusterBlockSharedDataSize > deviceProp.sharedMemPerBlock) {
198         error("Your CUDA hardware has insufficient block shared memory to hold all cluster
199         centroids\n");
200     }
201
202     checkCuda(cudaMalloc(&deviceObjects, numObjs * numCoords * sizeof(double)));
203     checkCuda(cudaMalloc(&deviceClusters, numClusters * numCoords * sizeof(double)));
204     checkCuda(cudaMalloc(&deviceMembership, numObjs * sizeof(int)));
205     checkCuda(cudaMalloc(&dev_delta_ptr, sizeof(double)));
206
207     timing = wtime() - timing;

```

```
204     printf("t_alloc_gpu: %lf ms\n\n", 1000 * timing);
205     timing = wtime();
206
207     checkCuda(cudaMemcpy(deviceObjects, dimObjects[0],
208                          numObjs * numCoords * sizeof(double), cudaMemcpyHostToDevice));
209     checkCuda(cudaMemcpy(deviceMembership, membership,
210                         numObjs * sizeof(int), cudaMemcpyHostToDevice));
211     timing = wtime() - timing;
212     printf("t_get_gpu: %lf ms\n\n", 1000 * timing);
213     timing = wtime();
214
215 do {
216     timing_internal = wtime();
217
218     /* GPU part: calculate new memberships */
219
220     timing_transfers = wtime();
221     // TODO: Copy clusters to deviceClusters
222     checkCuda(cudaMemcpy(deviceClusters, dimClusters[0],
223                          numClusters * numCoords * sizeof(double),
224                          cudaMemcpyHostToDevice));
225
226     transfers_time += wtime() - timing_transfers;
227
228     checkCuda(cudaMemset(dev_delta_ptr, 0, sizeof(double)));
229
230     timing_gpu = wtime();
231     //printf("Launching find_nearest_cluster Kernel with grid_size = %d, block_size = %d,
232     shared_mem = %d KB\n", numClusterBlocks, numThreadsPerClusterBlock, clusterBlockSharedDataSize/1000);
233     find_nearest_cluster
234     <<< numClusterBlocks, numThreadsPerClusterBlock, clusterBlockSharedDataSize >>>
235         (numCoords, numObjs, numClusters,
236          deviceObjects, deviceClusters, deviceMembership, dev_delta_ptr);
237
238     cudaDeviceSynchronize();
239     checkLastCudaError();
240     gpu_time += wtime() - timing_gpu;
241     //printf("Kernels complete for itter %d, updating data in CPU\n", loop);
242
243     timing_transfers = wtime();
244
245     checkCuda(cudaMemcpy(membership, deviceMembership,
246                          numObjs * sizeof(int),
247                          cudaMemcpyDeviceToHost));
248
249     checkCuda(cudaMemcpy(&delta, dev_delta_ptr,
250                         sizeof(double),
251                         cudaMemcpyDeviceToHost));
252
253     transfers_time += wtime() - timing_transfers;
254
255     /* CPU part: Update cluster centers*/
```

```

256
257     timing_cpu = wtime();
258     for (i = 0; i < numObjs; i++) {
259         /* find the array index of nestest cluster center */
260         index = membership[i];
261
262         /* update new cluster centers : sum of objects located within */
263         newClusterSize[index]++;
264         for (j = 0; j < numCoords; j++)
265             newClusters[j][index] += objects[i * numCoords + j];
266     }
267
268     /* average the sum and replace old cluster centers with newClusters */
269     for (i = 0; i < numClusters; i++) {
270         for (j = 0; j < numCoords; j++) {
271             if (newClusterSize[i] > 0)
272                 dimClusters[j][i] = newClusters[j][i] / newClusterSize[i];
273             newClusters[j][i] = 0.0; /* set back to 0 */
274         }
275         newClusterSize[i] = 0; /* set back to 0 */
276     }
277
278     delta /= numObjs;
279     //printf("delta is %f - ", delta);
280     loop++;
281     //printf("completed loop %d\n", loop);
282     cpu_time += wtime() - timing_cpu;
283
284     timing_internal = wtime() - timing_internal;
285     if (timing_internal < timer_min) timer_min = timing_internal;
286     if (timing_internal > timer_max) timer_max = timing_internal;
287 } while (delta > threshold && loop < loop_threshold);
288
289 /*TODO: Update clusters using dimClusters. Be carefull of layout!!!
clusters[numClusters][numCoords] vs dimClusters[numCoords][numClusters] */
290 for (i = 0; i < numClusters; i++) {
291     for (j = 0; j < numCoords; j++) {
292         clusters[i * numCoords + j] = dimClusters[j][i];
293     }
294 }
295
296 timing = wtime() - timing;
297 printf("nloops = %d : total = %lf ms\n\t-> t_loop_avg = %lf ms\n\t-> t_loop_min = %lf
ms\n\t-> t_loop_max = %lf ms\n\t"
298         "-> t_cpu_avg = %lf ms\n\t-> t_gpu_avg = %lf ms\n\t-> t_transfers_avg = %lf
ms\n\n|-----|\n",
299         loop, 1000 * timing, 1000 * timing / loop, 1000 * timer_min, 1000 * timer_max,
300         1000 * cpu_time / loop, 1000 * gpu_time / loop, 1000 * transfers_time / loop);
301
302 char outfile_name[1024] = {0};
303 sprintf(outfile_name, "Execution_logs/silver1-V100_Sz-%lu_Coo-%d_Cl-%d.csv",
304         numObjs * numCoords * sizeof(double) / (1024 * 1024), numCoords, numClusters);
305 FILE *fp = fopen(outfile_name, "a+");

```

```
306 if (!fp) error("Filename %s did not open successfully, no logging performed\n",
307     outfile_name);
308     fprintf(fp, "%s,%d,%lf,%lf,%lf\n", "Shmem", blockSize, timing / loop, timer_min,
309     timer_max);
310     fclose(fp);
311
312     checkCuda(cudaFree(deviceObjects));
313     checkCuda(cudaFree(deviceClusters));
314     checkCuda(cudaFree(deviceMembership));
315
316     free(dimObjects[0]);
317     free(dimObjects);
318     free(dimClusters[0]);
319     free(dimClusters);
320     free(newClusters[0]);
321     free(newClusters);
322     free(newClusterSize);
323
324     return;
325 }
```

B. Υλοποίηση και Ορθότητα

Η δομή δεδομένων παραμένει transpose (dimObjects[coord][obj], dimClusters[coord][cluster]) για coalescing. Η βασική αλλαγή είναι ότι στον kernel:

- Τα cluster centers αντιγράφονται μια φορά ανά block από global σε shared memory.
- Όλοι οι υπολογισμοί απόστασης χρησιμοποιούν πλέον τη shared μνήμη για τα clusters.

Επισημαίνουμε τα εξής σημεία:

(α) Δυναμική shared memory και αντιγραφή clusters

Χρησιμοποιείται extern `__shared__ double shmemClusters[]` και αντιγράφεται ολόκληρος ο πίνακας dimClusters (numCoords*numClusters στοιχεία) στη shared memory, με «striding» ως προς threadIdx (k += blockDim.x). Έτσι η φόρτωση μοιράζεται σε threads και γίνεται μία φορά ανά block.

(β) Συγχρονισμός (`__syncthreads`)

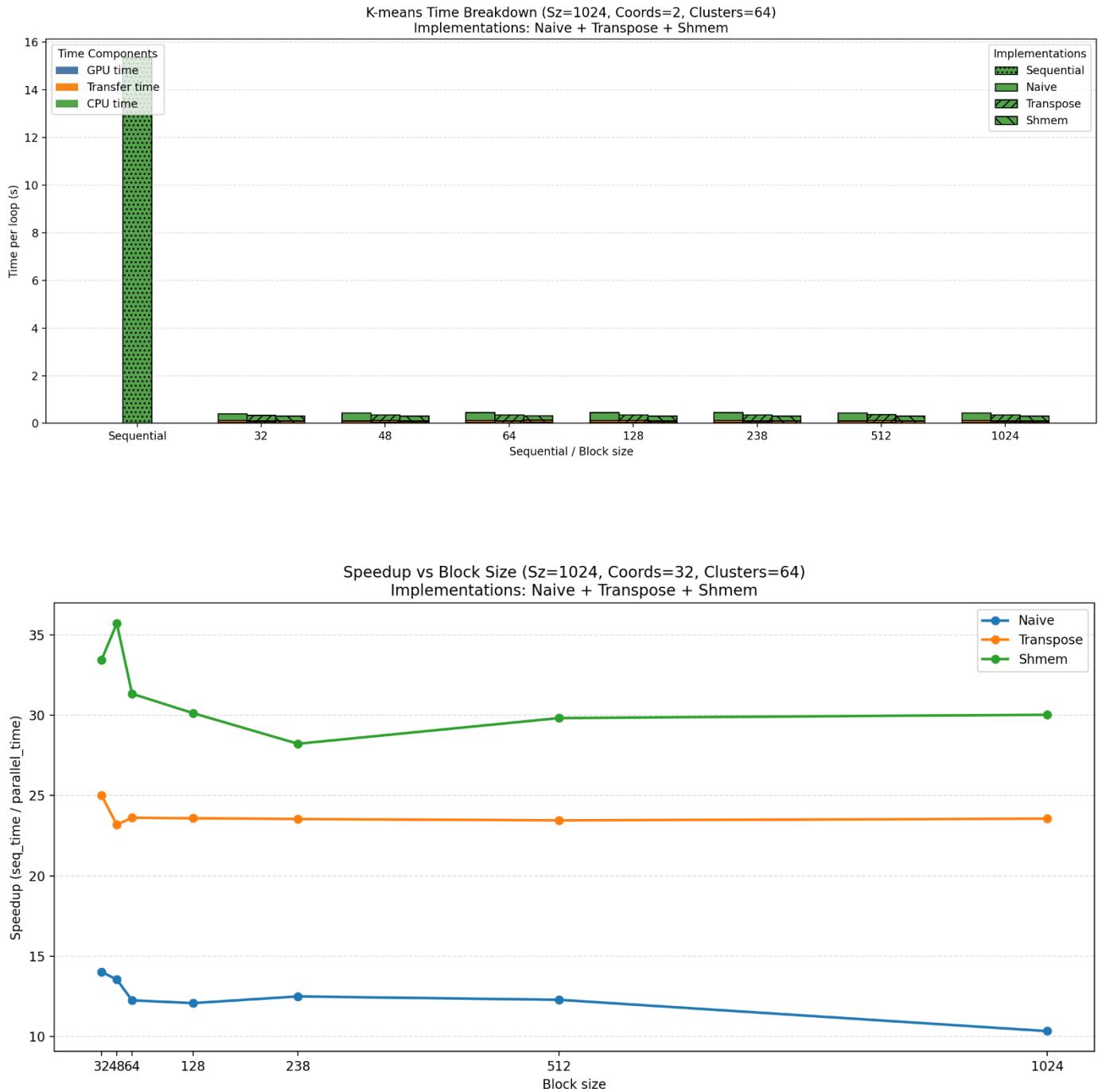
Μετά τη φόρτωση στη shared, γίνεται `__syncthreads()` ώστε να εξασφαλιστεί ότι όλα τα threads του block βλέπουν πλήρως γραμμένα τα δεδομένα πριν ξεκινήσουν τους υπολογισμούς αποστάσεων. Αυτό είναι απαραίτητο για ορθότητα (διαφορετικά κάποια threads θα διάβαζαν μη αρχικοποιημένες τιμές).

(γ) Έλεγχος διαθέσιμης shared μνήμης ανά block

Στον host υπολογίζεται το απαιτούμενο shared size = numClusters * numCoords * sizeof(double) και ελέγχεται έναντι της ιδιότητας sharedMemPerBlock της συσκευής. Αν το όριο ξεπεραστεί, η έκδοση δεν μπορεί να τρέξει (σωστό safeguard, καθώς το shared memory είναι περιορισμένος πόρος).

Η έκδοση Shared είναι αριθμητικά ισοδύναμη με Transpose/Naive: δεν αλλάζει ο ορισμός απόστασης ούτε το κριτήριο ανάθεσης. Αλλάζει μόνο η τοποθέτηση των cluster centers (shared αντί global), άρα αναμένουμε ίδια αποτελέσματα σύγκλισης (εντός floating-point διαφορών), με χαμηλότερο χρόνο kernel.

Γ. Παρουσίαση Διαγραμμάτων



Δ. Ερμηνεία Διαγραμμάτων

(1) Speedup vs Block Size

Η Shared υπερέχει σαφώς, με speedup περίπου 28–36, έναντι ~23–25 της Transpose και ~10–14 της Naive. Η βελτίωση είναι αναμενόμενη: ενώ η Transpose μειώνει τα global transactions μέσω coalescing, η Shared μειώνει και το πλήθος των global loads για τα clusters, αφού κάθε block φέρνει τα clusters μία φορά και τα επαναχρησιμοποιεί σε όλους τους distance υπολογισμούς.

(2) Time Breakdown (GPU / Transfers / CPU)

Το breakdown δείχνει ότι το κύριο κέρδος της Shared προέρχεται από περαιτέρω μείωση του GPU time (kernel). Οι μεταφορές (clusters H→D, membership+delta D→H) και ο CPU χρόνος (update_centroids) παραμένουν ουσιαστικά παρόμοια με Transpose/Naive, άρα η επιτάχυνση οφείλεται σχεδόν αποκλειστικά στη βελτίωση του memory access/reuse εντός του kernel.

Ρόλος του block_size στη Shared

Σε αντίθεση με τη Transpose (όπου η εξάρτηση από block_size ήταν μικρή), στη Shared το block_size μπορεί να επηρεάζει περισσότερο την απόδοση, επειδή η shared memory εισάγει πρόσθετους περιορισμούς στους resident πόρους ανά SM:

- Κάθε block δεσμεύει σταθερό shared size (εδώ: numClusters*numCoords*sizeof(double)), άρα ο μέγιστος αριθμός blocks/SM μπορεί να περιοριστεί από τη διαθέσιμη shared μνήμη, μειώνοντας occupancy (active warps/SM).
- Με πολύ μεγάλα blocks, περιοριζόμαστε επιπλέον από το όριο threads/SM, άρα μπορεί να μειωθούν ταυτόχρονα resident blocks και warps, και να αυξηθεί η ευαισθησία σε latency.

Συνεπώς, παρατηρείται συνήθως ένα ιδανικό σημείο (small block sizes) όπου συνδυάζονται αρκετά active warps και χαμηλό global traffic, ενώ σε ακραία μεγέθη blocks η απόδοση μπορεί να σταθεροποιείται ή να πέφτει.

E. Σύγκριση Αποτελεσμάτων (με Naive + Transpose)

1. Από Naive → Transpose: μεγάλο κέρδος λόγω coalescing (μείωση global memory transactions).
2. Από Transpose → Shared: επιπλέον μεγάλο κέρδος, διότι μειώνουμε τις επαναλαμβανόμενες αναγνώσεις clusters από global (on-chip reuse). Το κέρδος εμφανίζεται κυρίως ως περαιτέρω μείωση του GPU time, ενώ CPU και transfers παραμένουν περίπου σταθερά.
3. Η Shared εμφανίζει μεγαλύτερη (αλλά λογική) εξάρτηση από block_size σε σχέση με Transpose, λόγω των πόρων shared memory/occupancy.

ΣΤ. Συμπεράσματα

Η Shared έκδοση επιβεβαιώνει τη βασική αρχή βελτιστοποίησης GPU: πέρα από το coalescing, η επαναχρησιμοποίηση «hot» δεδομένων στη shared memory μπορεί να μειώσει δραστικά το global memory traffic και να επιταχύνει σημαντικά memory-bound kernels όπως το assignment του K-means. Για το συγκεκριμένο σενάριο (Coords=32, Clusters=64), η Shared είναι η καλύτερη από τις τρεις εκδόσεις, με το κέρδος να προέρχεται κυρίως από τη μείωση του χρόνου kernel και δευτερευόντως από επιλογές block_size που επηρεάζουν occupancy.

■ Ενότητα 3.4 – Σύγκριση Υλοποιήσεων/Bottleneck Analysis

A. Εισαγωγή

Στο σημείο αυτό συγκρίνουμε τις τρεις υλοποιήσεις (Naive, Transpose, Shared) και εντοπίζουμε το bottleneck χρησιμοποιώντας τα δεδομένα χρόνου ανά επανάληψη (GPU kernel / transfers CPU↔GPU / CPU update). Υπενθυμίζουμε ότι τα transfers που μετράμε εδώ αφορούν τις αντιγραφές μέσα στο loop (π.χ. clusters H→D και membership+delta D→H) και όχι την αρχική μεταφορά του dataset προς τη GPU, η οποία γίνεται μία φορά πριν ξεκινήσουν οι επαναλήψεις.

B. Σύγκριση Υλοποιήσεων/Bottleneck Analysis

1. Ποιο bottleneck περιορίζει την επίδοση (Sz=1024MB, Coords=32, Clusters=64);

Από τα διαγράμματα για Coords=32, το αποτέλεσμα είναι ξεκάθαρο:

- Naive → Transpose: η κύρια μείωση χρόνου έρχεται από το GPU time, επειδή το transpose βελτιώνει το memory coalescing (λιγότερα global memory transactions ανά warp).
- Transpose → Shared: το GPU time μειώνεται περαιτέρω, επειδή τα cluster centers επαναχρησιμοποιούνται από shared memory (on-chip) αντί για επαναλαμβανόμενες αναγνώσεις από global.

Μετά τις δύο αυτές βελτιστοποιήσεις, όμως, παρατηρείται ότι το συνολικό κέρδος δεν αυξάνεται αναλογικά (ιδανικά, δηλαδή δεν κλιμακώνει τέλεια) με τη μείωση του GPU time. Ο λόγος είναι ότι πλέον αρχίζουν να κυριαρχούν/να γίνονται συγκρίσιμα τα μη-επιταχυνόμενα τμήματα:

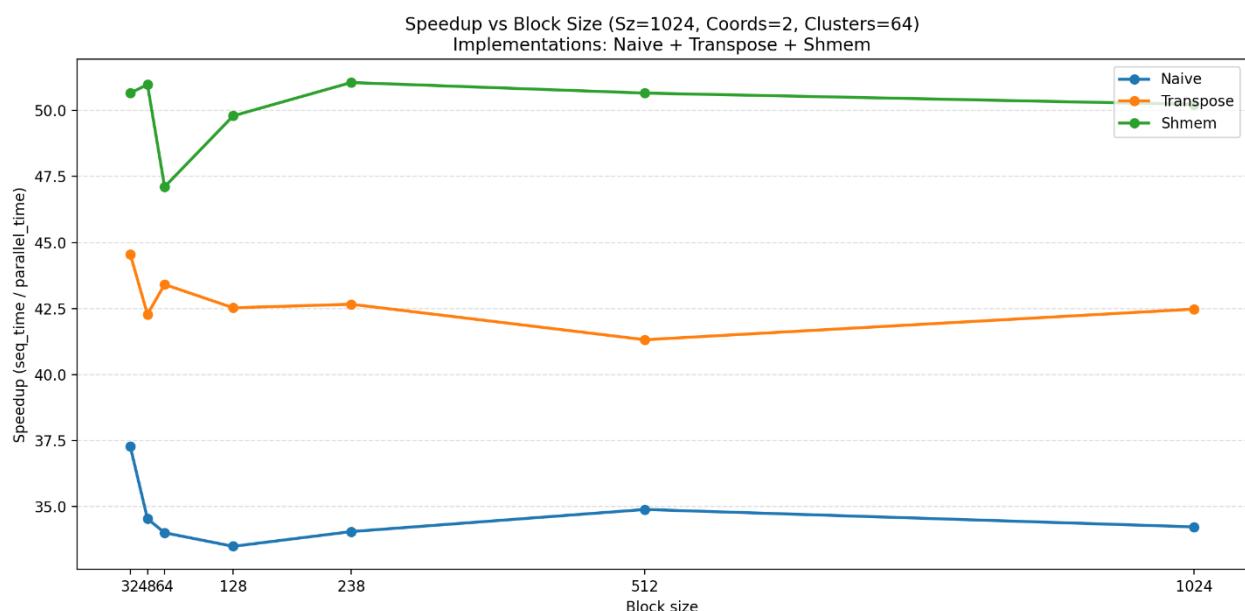
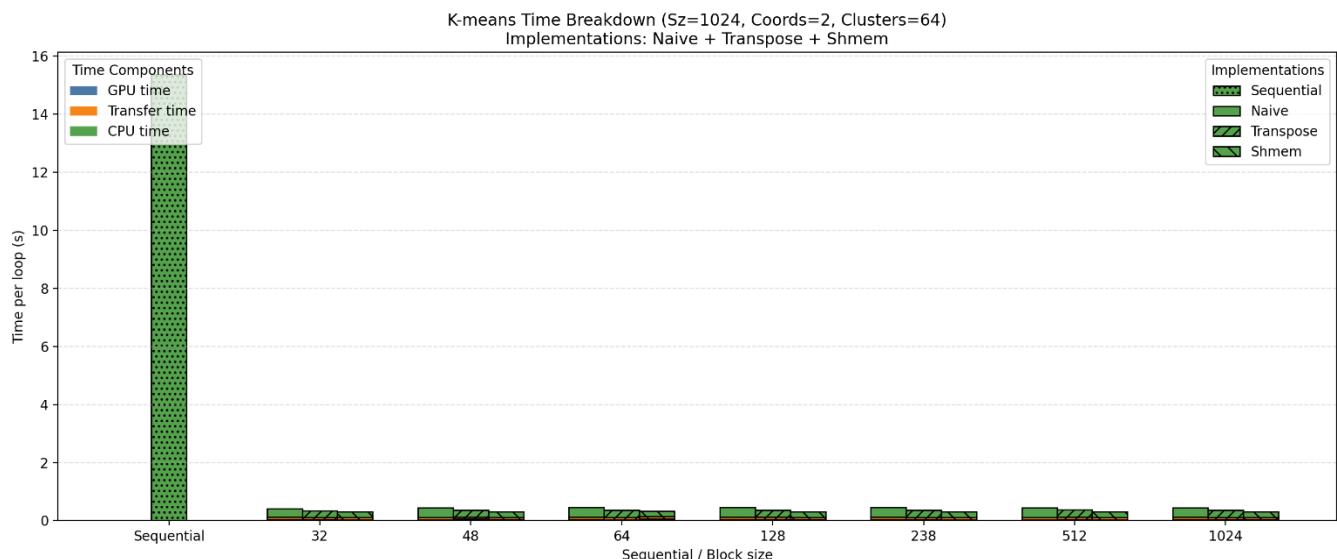
- CPU time (update_centroids): παραμένει σειριακό στην CPU στις τρεις εκδόσεις και θέτει όριο βάσει του Amdahl (όσο μικραίνει το GPU kernel, τόσο μεγαλύτερο ποσοστό του συνολικού χρόνου καταλαμβάνει το CPU update).

- Transfers time: για Coords=32 είναι σχετικά μικρό, αλλά είναι σταθερό overhead ανά επανάληψη (ιδίως η επιστροφή του membership), το οποίο δεν μειώνεται από τις βελτιστοποιήσεις μέσα στον kernel.

Συνεπώς, για Coords=32 το bottleneck «μετατοπίζεται» σταδιακά: αρχικά είναι κυρίως η απόδοση του GPU kernel (Naive), ενώ στις βελτιστοποιημένες εκδόσεις (Transpose/Shared) το όριο τίθεται ολοένα περισσότερο από το CPU update και το σταθερό κόστος επικοινωνίας ανά loop.

2. Τι αλλάζει για Coords=2 και είναι η προσέγγιση Shared κατάλληλη για arbitrary configs;

Αρχικά, παρουσιάζουμε τα διαγράμματα με 2 συντεταγμένες ακολούθως:



Με σταθερό dataset size (1024MB), όταν μειώνουμε τις διαστάσεις από 32 σε 2, ο αριθμός objects αυξάνεται περίπου κατά 16 \times (numObjs \propto 1/numCoords). Αυτό έχει δύο κρίσιμες συνέπειες:

1. Αυξάνεται δραστικά το μέγεθος του membership που πρέπει να επιστρέψει στη CPU σε κάθε επανάληψη ($D \rightarrow H$), άρα ο χρόνος transfers γίνεται πολύ πιο σημαντικός σε σχέση με το Coords=32.
2. Ταυτόχρονα, ο υπολογισμός απόστασης ανά object γίνεται ελαφρύτερος (μόνο 2 συντεταγμένες), άρα το GPU kernel έχει μικρότερο arithmetic work ανά element και η συνολική εκτέλεση τείνει να γίνεται λιγότερο compute-bound και πιο overhead/communication sensitive.

Τα διαγράμματα για Coords=2 επιβεβαιώνουν αυτή τη μετατόπιση: ενώ η Shared παραμένει η ταχύτερη υλοποίηση (Shared > Transpose > Naive), η διαφορά μεταξύ των GPU εκδόσεων προκύπτει πλέον κυρίως από σχετικά μικρότερες βελτιώσεις στο GPU time, επειδή ένα μεγαλύτερο ποσοστό του συνολικού χρόνου ανά loop ανήκει στις μεταφορές (και στο CPU update). Με άλλα λόγια, όταν το bottleneck είναι η επικοινωνία (membership transfer) και το σειριακό update, οι βελτιστοποιήσεις εντός του kernel έχουν περιορισμένο χώρο να αποδώσουν.

Γ. Συμπέρασμα για arbitrary configs

Η τεχνική shared memory για τα clusters είναι γενικά αποδοτική όταν:

- το (numClusters \times numCoords) χωράει σε shared ανά block, και
- υπάρχει αρκετή επαναχρησιμοποίηση/υπολογιστικό έργο ανά φόρτωση (ώστε το κόστος φόρτωσης + `__syncthreads` να αποσβεστεί).

Ωστόσο, δεν είναι καθολική αλήθεια για όλα τα configs: σε περιπτώσεις όπως Coords=2, όπου αυξάνεται έντονα το communication overhead (membership $D \rightarrow H$) και μειώνεται το arithmetic intensity του distance computation, το συνολικό bottleneck μετακινείται εκτός kernel. Τότε η Shared εξακολουθεί να βοηθά (μειώνει το GPU time), αλλά το συνολικό speedup περιορίζεται κυρίως από transfers και CPU update, δηλαδή από τμήματα που η Shared δεν μπορεί να βελτιώσει.

- **Full-Offload (All-GPU) Version**

A. Εισαγωγή

Στην έκδοση Full-Offload (All-GPU) μεταφέρουμε ολόκληρο το iterative μέρος του K-means στη GPU: όχι μόνο το assignment (εύρεση κοντινότερου cluster για κάθε object), αλλά και το update των centroids (συσσώρευση sums/counts και υπολογισμός νέων κέντρων). Στόχος είναι να εξαλειφθούν (i) το CPU load ανά επανάληψη (update_centroids στην CPU) και (ii) οι μεγάλες μεταφορές CPU↔GPU μέσα στο loop (ιδίως το D2H membership), ώστε το bottleneck να περιοριστεί στον καθαρό GPU υπολογισμό.

Ο κώδικας του αρχείου μας (cuda_kmeans_all_gpu.cu) παρατίθεται ακολούθως:

a5/cuda_kmeans_all_gpu.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 {
11     if (e != cudaSuccess)
12     {
13         // cudaGetErrorString() isn't always very helpful. Look up the error
14         // number in the cudaError enum in driver_types.h in the CUDA includes
15         // directory for a better explanation.
16         error("CUDA Error %d: %s\n", e, cudaGetErrorString(e));
17     }
18 }
19
20 inline void checkLastCudaError()
21 {
22     checkCuda(cudaGetLastError());
23 }
24 #endif
25
26 __device__ int get_tid()
27 {
28     return blockIdx.x * blockDim.x + threadIdx.x;
29 }
30
31 /* square of Euclid distance between two multi-dimensional points using column-base format
 */
32 __host__ __device__ inline static double euclid_dist_2_transpose(int numCoords,
33                                                               int numObjs,
34                                                               int numClusters,
35                                                               double *objects, // [numCoords][numObjs]
36                                                               double *clusters, // [numCoords][numClusters]
37                                                               int objectId,
38                                                               int clusterId)
39 {
40     int i;
41     double ans = 0.0;
42
43     /* TODO: Calculate the euclid_dist of elem=objectId of objects from elem=clusterId from
        clusters, but for column-base format!!! */
44     for (i = 0; i < numCoords; i++)
45     {
46         double objectVal = objects[i * numObjs + objectId];
47         double clusterVal = clusters[i * numClusters + clusterId];
48     }
```

```

49     double diff = objectVal - clusterVal;
50     ans += diff * diff;
51 }
52
53 return (ans);
54 }
55
56 __global__ static void find_nearest_cluster(int numCoords,
57                                             int numObjs,
58                                             int numClusters,
59                                             double *deviceObjects, // [numCoords]
60                                             [numObjs]
61                                             /*
62 TODO: If you choose to do (some of) the new centroid calculation here, you will need some
63 extra parameters here (from "update_centroids").
64 */
65                                             int *devicenewClusterSize,
66                                             double *devicenewClusters, // [numCoords]
67                                             [numClusters]
68                                             double *deviceClusters, // [numCoords]
69                                             int *deviceMembership, // [numObjs]
70                                             double *devdelta)
71 {
72     extern __shared__ double shmemClusters[];
73     // TODO: Copy deviceClusters to shmemClusters so they can be accessed faster.
74     int tid_in_block = threadIdx.x; // To ID του νήματος μέσα στο Block
75     int block_size = blockDim.x; // Πόσα νήματα έχει το Block
76     int total_cluster_doubles = numClusters * numCoords; // Συνολικά νούμερα προς αντιγραφή
77
78     // Κάθε νήμα αντιγράφει όσα στοιχεία του αναλογούν (με βήμα block_size)
79     for (int k = tid_in_block; k < total_cluster_doubles; k += block_size)
80     {
81         shmemClusters[k] = deviceClusters[k];
82     }
83
84     /* Συγχρονισμός (BARRIER) */
85
86     __syncthreads();
87
88     /* Get the global ID of the thread. */
89     int tid = get_tid();
90
91     /* TODO: Maybe something is missing here... should all threads run this? */
92     if (tid < numObjs)
93     {
94         int index, i;
95         double dist, min_dist;
96
97         /* find the cluster id that has min distance to object */
98         index = 0;
99         /* TODO: call min_dist = euclid_dist_2(...) with correct objectId/clusterId using
100            clusters in shmem*/

```

```
97
98     min_dist = euclid_dist_2_transpose(numCoords, numObjs, numClusters,
99                                         deviceObjects, shmemClusters,
100                                        tid, index);
101
102    for (i = 1; i < numClusters; i++)
103    {
104        dist = euclid_dist_2_transpose(numCoords, numObjs, numClusters,
105                                         deviceObjects, shmemClusters,
106                                         tid, i);
107
108        /* no need square root */
109        if (dist < min_dist)
110        { /* find the min and its array index */
111            min_dist = dist;
112            index = i;
113        }
114    }
115
116    if (deviceMembership[tid] != index)
117    {
118        /* TODO: Maybe something is missing here... is this write safe? */
119        atomicAdd(devdelta, 1.0);
120    }
121
122    /* assign the deviceMembership to object objectId */
123    deviceMembership[tid] = index;
124
125    /* TODO: additional steps for calculating new centroids in GPU? */
126
127    atomicAdd(&devicenewClusterSize[index], 1);
128
129    for (int j = 0; j < numCoords; j++)
130    {
131        // Διαβάζουμε την τιμή του αντικειμένου (Coordinate j, Object tid)
132        double objVal = deviceObjects[j * numObjs + tid];
133
134        // Προσθέτουμε στο άθροισμα (Coordinate j, Cluster index)
135        atomicAdd(&devicenewClusters[j * numClusters + index], objVal);
136    }
137 }
138 }
139
140 __global__ static void update_centroids(int numCoords,
141                                         int numClusters,
142                                         int *devicenewClusterSize, // [numClusters]
143                                         double *devicenewClusters, // [numCoords]
144                                         [numClusters]
145                                         [numClusters])
146 {
147     /* Κάθε νήμα αναλαμβάνει ΜΙΑ τιμή (double) του πίνακα clusters.
148     Συνολικά νήματα = numCoords * numClusters
149     */
150 }
```

```

149 int tid = get_tid();
150 int total_elements = numCoords * numClusters;
151
152 if (tid < total_elements)
153 {
154     // Αποκαδικοποίηση του 1D tid σε 2D (Coordinate, Cluster)
155     // Layout: [numCoords][numClusters] --> index = coord * numClusters + cluster
156     int clusterId = tid % numClusters;
157     // int coordId = tid / numClusters; // Δεν το χρειαζόμαστε άμεσα για τον υπολογισμό,
αλλά για το reset
158
159     int count = devicenewClusterSize[clusterId];
160
161     // Υπολόγισε το νέο κέντρο (Average)
162     if (count > 0)
163     {
164         double sum = devicenewClusters[tid];
165         deviceClusters[tid] = sum / count;
166     }
167     // Av count == 0, κρατάμε την παλιά τιμή (ή δεν κάνουμε τίποτα), όπως και στον CPU
κώδικα
168
169     // RESET για τον επόμενο γύρο (Πολύ σημαντικό!)
170     // Μηδενίζουμε το άθροισμα που μόλις χρησιμοποιήσαμε
171     devicenewClusters[tid] = 0.0;
172 }
173 }
174
175 //
176 // -----
177 // DATA LAYOUT
178 //
179 // objects      [numObjs][numCoords]
180 // clusters     [numClusters][numCoords]
181 // dimObjects   [numCoords][numObjs]
182 // dimClusters  [numCoords][numClusters]
183 // newClusters  [numCoords][numClusters]
184 // deviceObjects [numCoords][numObjs]
185 // deviceClusters [numCoords][numClusters]
186 //
187 //
188 /* return an array of cluster centers of size [numClusters][numCoords] */
189 void kmeans_gpu(double *objects,    /* in: [numObjs][numCoords] */
190                 int numCoords,    /* no. features */
191                 int numObjs,     /* no. objects */
192                 int numClusters, /* no. clusters */
193                 double threshold,/* % objects change membership */
194                 long loop_threshold,/* maximum number of iterations */
195                 int *membership, /* out: [numObjs] */
196                 double *clusters, /* out: [numClusters][numCoords] */
197                 int blockSize)
198 {
199     double timing = wtime(), timing_internal, timer_min = 1e42, timer_max = 0;

```

```

200  double timing_gpu, timing_cpu, timing_transfers, transfers_time = 0.0, cpu_time = 0.0,
201  gpu_time = 0.0;
202  int loop_iterations = 0;
203  int i, j, index, loop = 0;
204  double delta = 0, *dev_delta_ptr; /* % of objects change their clusters */
205  /* TODO: Copy me from transpose version*/
206  double **dimObjects = (double **)calloc_2d(numCoords, numObjs, sizeof(double));      // 
207  calloc_2d(...) -> [numCoords][numObjs]
208  double **dimClusters = (double **)calloc_2d(numCoords, numClusters, sizeof(double)); // 
209  calloc_2d(...) -> [numCoords][numClusters]
210  double **newClusters = (double **)calloc_2d(numCoords, numClusters, sizeof(double));
211
212  printf("\n|-----Full-offload GPU Kmeans-----|\n\n");
213
214  /* TODO: Copy me from transpose version*/
215  for (i = 0; i < numObjs; i++)
216  {
217      for (j = 0; j < numCoords; j++)
218      {
219          dimObjects[j][i] = objects[i * numCoords + j];
220      }
221  }
222
223  double *deviceObjects;
224  double *deviceClusters, *devicenewClusters;
225  int *deviceMembership;
226  int *devicenewClusterSize; /* [numClusters]: no. objects assigned in each new cluster */
227
228  /* pick first numClusters elements of objects[] as initial cluster centers*/
229  for (i = 0; i < numCoords; i++)
230  {
231      for (j = 0; j < numClusters; j++)
232      {
233          dimClusters[i][j] = dimObjects[i][j];
234      }
235  }
236
237
238  /* initialize membership[] */
239  for (i = 0; i < numObjs; i++)
240      membership[i] = -1;
241
242  timing = wtime() - timing;
243  printf("t_alloc: %lf ms\n\n", 1000 * timing);
244  timing = wtime();
245  const unsigned int numThreadsPerClusterBlock = (numObjs > blockSize) ? blockSize :
246  numObjs;
247  const unsigned int numClusterBlocks = (numObjs + numThreadsPerClusterBlock - 1) /
248  numThreadsPerClusterBlock; /* TODO: Calculate Grid size, e.g. number of blocks. */
249
250  /* Define the shared memory needed per block.
251   - BEWARE: We can overrun our shared memory here if there are too many
252   clusters or too many coordinates!
253   - This can lead to occupancy problems or even inability to run.

```

```
248     - Your exercise implementation is not requested to account for that (e.g. always
249     assume deviceClusters fit in shmemClusters */
250
251     const unsigned int clusterBlockSharedDataSize = numClusters * numCoords *
252     sizeof(double);
253
254     cudaDeviceProp deviceProp;
255     int deviceNum;
256     cudaGetDevice(&deviceNum);
257     cudaGetDeviceProperties(&deviceProp, deviceNum);
258
259     if (clusterBlockSharedDataSize > deviceProp.sharedMemPerBlock)
260     {
261         error("Your CUDA hardware has insufficient block shared memory to hold all cluster
262         centroids\n");
263     }
264
265     checkCuda(cudaMalloc(&deviceObjects, numObjs * numCoords * sizeof(double)));
266     checkCuda(cudaMalloc(&deviceClusters, numClusters * numCoords * sizeof(double)));
267     checkCuda(cudaMalloc(&devicenewClusters, numClusters * numCoords * sizeof(double)));
268     checkCuda(cudaMalloc(&devicenewClusterSize, numClusters * sizeof(int)));
269     checkCuda(cudaMalloc(&deviceMembership, numObjs * sizeof(int)));
270     checkCuda(cudaMalloc(&dev_delta_ptr, sizeof(double)));
271
272     timing = wtime() - timing;
273     printf("t_alloc_gpu: %lf ms\n\n", 1000 * timing);
274     timing = wtime();
275
276     checkCuda(cudaMemcpy(deviceObjects, dimObjects[0],
277                         numObjs * numCoords * sizeof(double), cudaMemcpyHostToDevice));
278     checkCuda(cudaMemcpy(deviceMembership, membership,
279                         numObjs * sizeof(int), cudaMemcpyHostToDevice));
280     checkCuda(cudaMemcpy(deviceClusters, dimClusters[0],
281                         numClusters * numCoords * sizeof(double), cudaMemcpyHostToDevice));
282     checkCuda(cudaMemset(devicenewClusterSize, 0, numClusters * sizeof(int)));
283     free(dimObjects[0]);
284
285     timing = wtime() - timing;
286     printf("t_get_gpu: %lf ms\n\n", 1000 * timing);
287     timing = wtime();
288
289     do
290     {
291         timing_internal = wtime();
292         checkCuda(cudaMemset(dev_delta_ptr, 0, sizeof(double)));
293         checkCuda(cudaMemset(devicenewClusterSize, 0, numClusters * sizeof(int)));
294         timing_gpu = wtime();
295         // printf("Launching find_nearest_cluster Kernel with grid_size = %d, block_size = %d,
296         // shared_mem = %d KB\n", numClusterBlocks, numThreadsPerClusterBlock, clusterBlockSharedDa-
297         taSize/1000);
298         // TODO: change invocation if extra parameters needed
299         find_nearest_cluster<<<numClusterBlocks, numThreadsPerClusterBlock,
300         clusterBlockSharedDataSize>>>(numCoords, numObjs, numClusters,
301
302         deviceObjects, devicenewClusterSize, devicenewClusters, deviceClusters, deviceMembership,
303         dev_delta_ptr);
```

```
295     cudaDeviceSynchronize();
296     checkLastCudaError();
297
298     gpu_time += wtime() - timing_gpu;
299
300     // printf("Kernels complete for itter %d, updating data in CPU\n", loop);
301
302     timing_transfers = wtime();
303     // TODO: Copy dev_delta_ptr to &delta
304     checkCuda(cudaMemcpy(&delta, dev_delta_ptr, sizeof(double), cudaMemcpyDeviceToHost));
305     transfers_time += wtime() - timing_transfers;
306
307     const unsigned int update_centroids_block_sz = (numCoords * numClusters > blockSize) ?
308 blockSize : numCoords * numClusters;           /* TODO: can use different blocksize here if
309 deemed better */
310     const unsigned int update_centroids_dim_sz = (numCoords * numClusters +
311 update_centroids_block_sz - 1) / update_centroids_block_sz; /* TODO: calculate dim for
312 "update_centroids" */
313     timing_gpu = wtime();
314     // TODO: use dim for "update_centroids" and fire it
315     update_centroids<<<update_centroids_dim_sz, update_centroids_block_sz, 0>>>(numCoords,
316 numClusters, devicenewClusterSize, devicenewClusters, deviceClusters);
317     cudaDeviceSynchronize();
318     checkLastCudaError();
319     gpu_time += wtime() - timing_gpu;
320
321     timing_cpu = wtime();
322     delta /= numObjs;
323     // printf("delta is %f - ", delta);
324     loop++;
325     // printf("completed loop %d\n", loop);
326     cpu_time += wtime() - timing_cpu;
327
328     timing_internal = wtime() - timing_internal;
329     if (timing_internal < timer_min)
330         timer_min = timing_internal;
331     if (timing_internal > timer_max)
332         timer_max = timing_internal;
333 } while (delta > threshold && loop < loop_threshold);
334
335     checkCuda(cudaMemcpy(membership, deviceMembership,
336                           numObjs * sizeof(int), cudaMemcpyDeviceToHost));
337     checkCuda(cudaMemcpy(dimClusters[0], deviceClusters,
338                           numClusters * numCoords * sizeof(double), cudaMemcpyDeviceToHost));
339
340     for (i = 0; i < numClusters; i++)
341     {
342         for (j = 0; j < numCoords; j++)
343         {
344             clusters[i * numCoords + j] = dimClusters[j][i];
345         }
346     }
347 }
```

```
344     timing = wtime() - timing;
345     printf("nloops = %d : total = %lf ms\n\t-> t_loop_avg = %lf ms\n\t-> t_loop_min = %lf
346 ms\n\t-> t_loop_max = %lf ms\n\t-\n"
347         "-> t_cpu_avg = %lf ms\n\t-> t_gpu_avg = %lf ms\n\t-> t_transfers_avg = %lf
348 ms\n\t-\n|-----|\n",
349         loop, 1000 * timing, 1000 * timing / loop, 1000 * timer_min, 1000 * timer_max,
350         1000 * cpu_time / loop, 1000 * gpu_time / loop, 1000 * transfers_time / loop);
351
350     char outfile_name[1024] = {0};
351     sprintf(outfile_name, "Execution_logs/silver1-V100_Sz-%lu_Coo-%d_Cl-%d.csv",
352             numObjs * numCoords * sizeof(double) / (1024 * 1024), numCoords, numClusters);
353     FILE *fp = fopen(outfile_name, "a+");
354     if (!fp)
355         error("Filename %s did not open successfully, no logging performed\n", outfile_name);
356     fprintf(fp, "%s,%d,%lf,%lf,%lf\n", "All_GPU", blockSize, timing / loop, timer_min,
357             timer_max);
358     fclose(fp);
359
360     checkCuda(cudaFree(deviceObjects));
361     checkCuda(cudaFree(deviceClusters));
362     checkCuda(cudaFree(devicenewClusters));
363     checkCuda(cudaFree(devicenewClusterSize));
364     checkCuda(cudaFree(deviceMembership));
365
365     return;
366 }
367 }
```

B. Υλοποίηση και Ορθότητα

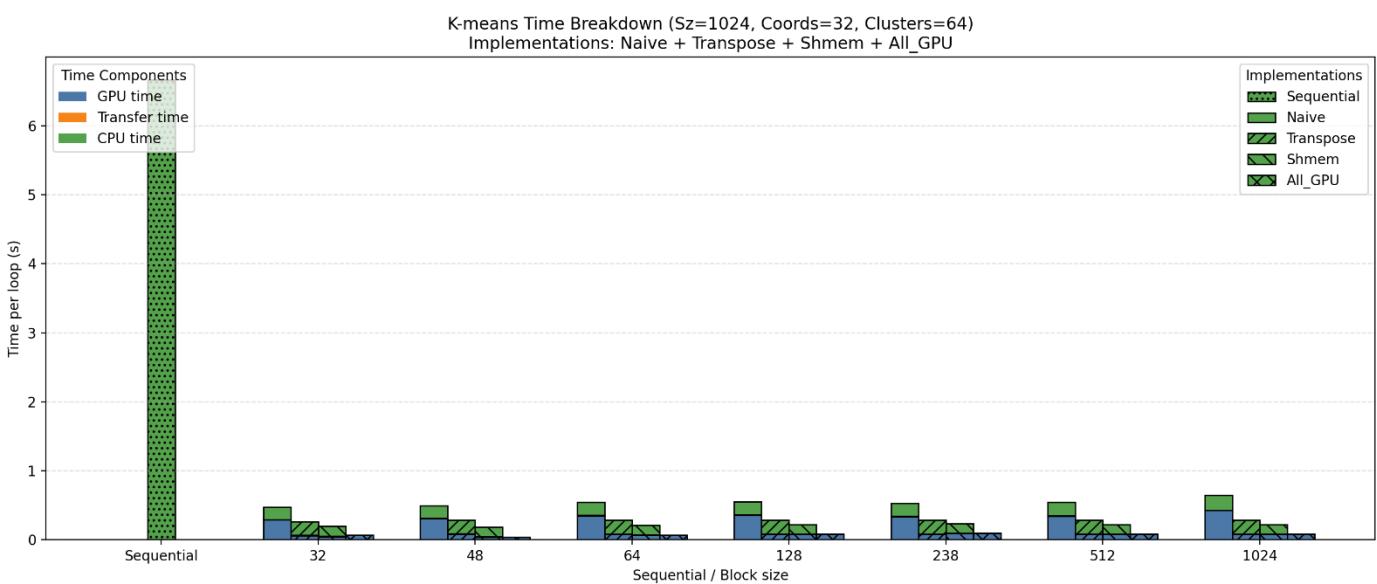
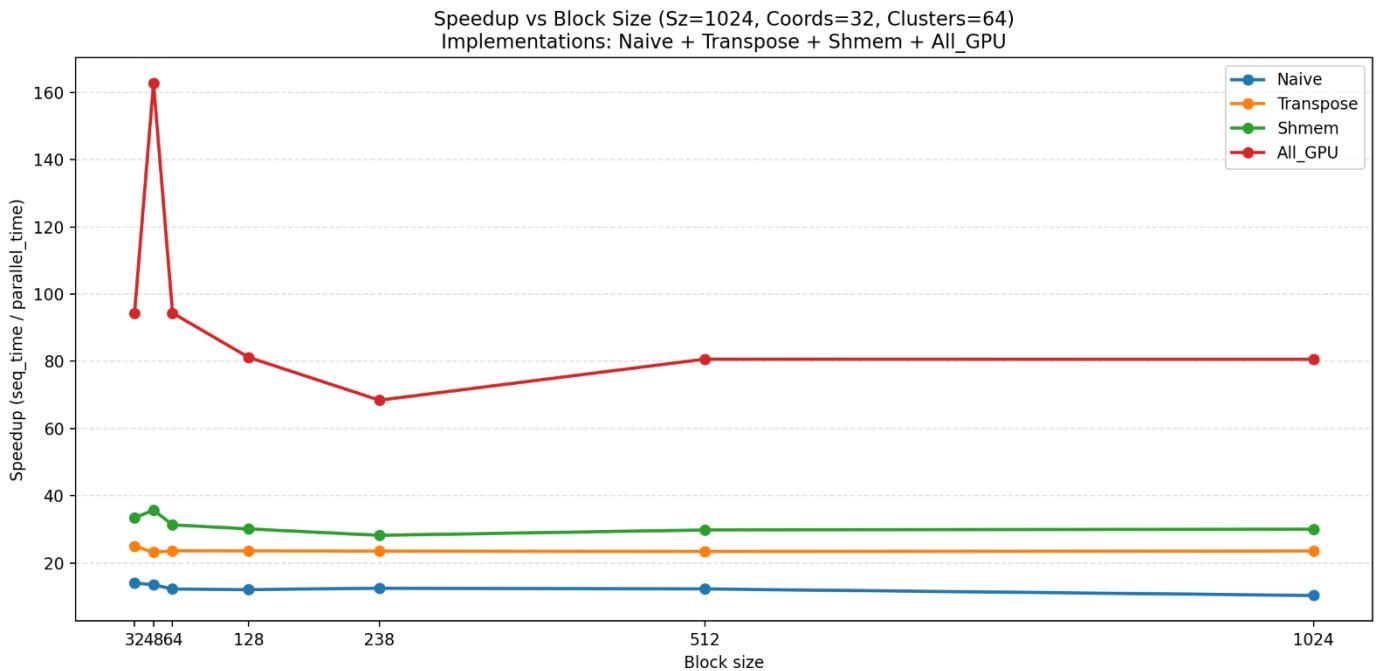
Η λογική της All-GPU υλοποίησης είναι να σπάσει το update_centroids σε βήματα που μπορούν να γίνουν ασφαλώς στη GPU χωρίς καθολικό barrier μέσα σε ένα kernel:

1. Μηδενισμός/αρχικοποίηση device arrays για newClusters_sums και newClusters_counts (και ό,τι άλλο χρειάζεται).
2. Kernel ανάθεσης (find_nearest_cluster): κάθε thread επεξεργάζεται ένα object, υπολογίζει αποστάσεις προς όλα τα clusters, ενημερώνει το membership και ταυτόχρονα συσσωρεύει τη συνεισφορά του object στο cluster που ανήκει.
 - Η συσσώρευση sums/counts γίνεται με atomics σε global μνήμη (atomicAdd σε counts και σε κάθε διάσταση του sum), ώστε να αποφευχθούν race conditions.
 - Τα cluster centers μπορούν να φορτωθούν ανά block στη shared memory (όπως στη shared έκδοση) ώστε οι επαναλαμβανόμενες αναγνώσεις κατά τον υπολογισμό αποστάσεων να γίνονται από on-chip μνήμη.
3. Kernel τελικοποίησης centroids: για κάθε cluster (και διάσταση) υπολογίζεται ο μέσος όρος (sum/count) και παράγονται τα νέα centers για το επόμενο iteration.
4. Για τον τερματισμό του while-loop, στον host επιστρέφει μόνο το delta (ή/και ελάχιστη μετα-πληροφορία). Έτσι, οι μεταφορές μέσα στο loop ελαχιστοποιούνται.

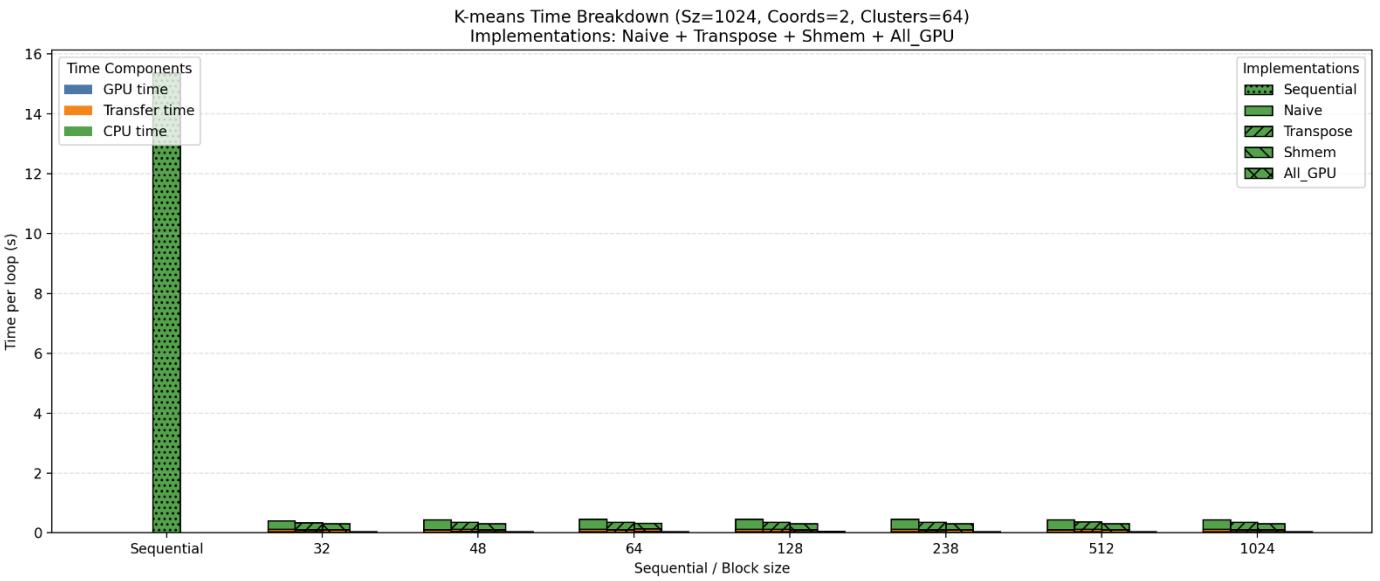
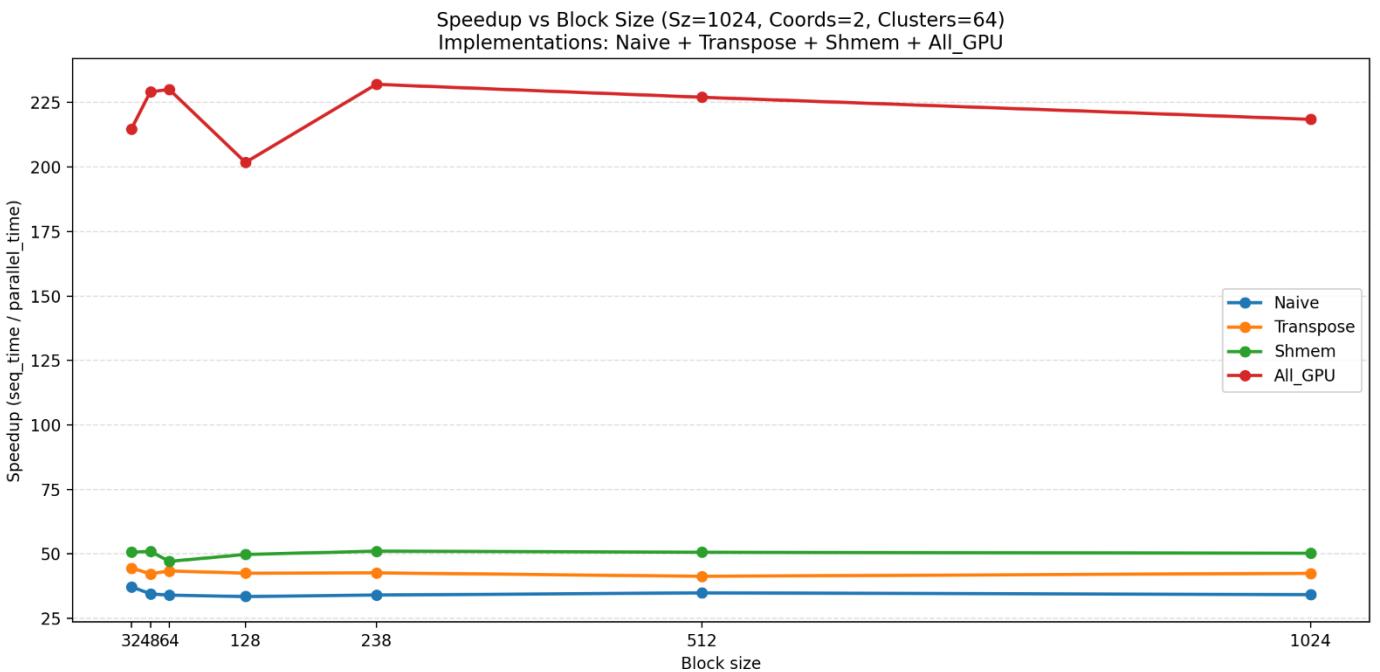
Η χρήση atomics εξασφαλίζει το σωστό αποτέλεσμα στα αθροίσματα, παρά το ταυτόχρονο update από πολλά threads. Ο απαιτούμενος καθολικός συγχρονισμός επιτυγχάνεται φυσικά με τη διάσπαση σε πολλαπλά kernels (τα kernels εκτελούνται σειριακά ως προς τη σειρά κλήσης τους), κάτι που αντικαθιστά την απουσία global barrier μέσα σε έναν kernel.

Γ. Παρουσίαση Διαγραμμάτων

(α) Configuration {1024,32,64,10}



(β) Configuration {1024,2,64,10}



Δ. Ερμηνεία Διαγραμμάτων – Απάντηση Ερωτημάτων (1) και (2)

(1) Επίδοση All-GPU σε σχέση με naive/transpose/shared (και για τα δύο configurations)

Τα διαγράμματα δείχνουν ότι η All-GPU υπερέχει σημαντικά έναντι όλων των προηγούμενων εκδόσεων και στα δύο configurations. Αυτό είναι αναμενόμενο, καθώς:

- Το CPU time μέσα στο loop (update_centroids στην CPU) πρακτικά μηδενίζεται.
- Το transfer time μέσα στο loop μειώνεται δραστικά, επειδή δεν απαιτείται πλέον αντιγραφή του membership πίσω στην CPU σε κάθε επανάληψη. Στον host επιστρέφει μόνο το delta, άρα οι per-iteration μεταφορές περιορίζονται σε πολύ μικρά δεδομένα (σε αντίθεση με τις naive/transpose/shared όπου υπάρχει O(N) Device→Host membership).

Άρα, το iterative μέρος παύει να είναι υβριδικό (GPU assignment + CPU update + transfers) και γίνεται σχεδόν αποκλειστικά GPU workload, το οποίο ταιριάζει καλύτερα στη φιλοσοφία throughput της GPU. Το νέο bottleneck προέρχεται κυρίως στον GPU χρόνο (distance computations + atomics για sums/counts).

Σημείωση: Η εκτέλεση έχει σχετικά μικρό warp divergence (κυρίως bounds checks και η απλή ενημέρωση membership), άρα το bottleneck προέρχεται κυρίως από global memory traffic και atomic contention, όχι από branching

(2) Παίζει διαφορετικό ρόλο το block_size και γιατί;

Ναι, στην All-GPU έκδοση το block_size επηρεάζει έντονα την επίδοση και αυτό φαίνεται καθαρά στα διαγράμματα (ιδίως στο Coords=32, όπου υπάρχει πολύ μεγάλη διακύμανση speedup ανά block size). Ο λόγος είναι ότι, αφού σχεδόν μηδενίζονται τα per-loop transfers και το CPU update, το συνολικό runtime καθορίζεται σχεδόν αποκλειστικά από καθαρά GPU φαινόμενα, τα οποία εξαρτώνται άμεσα από το block_size:

1. Occupancy / latency hiding: Το block_size καθορίζει πόσα blocks/warps μπορούν να είναι resident ανά SM. Με μεγαλύτερα blocks αυξάνονται οι απαιτήσεις σε threads/SM (και σε registers ανά block), άρα συχνά μειώνονται

τα ταυτόχρονα resident blocks/warps. Όταν μειωθούν τα active warps, η GPU κρύβει χειρότερα τη latency της global μνήμης και η επίδοση πέφτει.

2. Πίεση σε registers και shared: Στον assignment kernel κάθε thread κάνει σχετικά βαριά δουλειά (loop σε numClusters και numCoords). Αυτό τείνει να αυξάνει τα registers/thread. Όσο μεγαλώνει το block_size, το συνολικό register footprint/block μεγαλώνει και μπορεί να περιορίσει τα blocks/SM. Αν χρησιμοποιείται και shared caching για τα clusters, το shared ανά block είναι σταθερό, αλλά σε συνδυασμό με τα registers/threads μπορεί να κλειδώσει το occupancy.
3. Atomic contention στο update_centroids: Η All-GPU κάνει συσσώρευση sums/counts με atomics. Το block_size επηρεάζει πόσα threads πηγαίνουν ταυτόχρονα τους ίδιους counters/αθροίσματα (ιδίως όταν πολλά objects καταλήγουν στα ίδια clusters). Μεγαλύτερη ταυτόχρονη πίεση σε atomics οδηγεί σε serialization και απώλεια throughput, άρα μπορεί να εμφανίζεται ισχυρό sweet spot σε συγκεκριμένα block sizes. Θεωρητικά, για να μειωθεί το contention, μια κλασική τεχνική είναι block-level partial sums/counts σε shared memory (με reduction) και στη συνέχεια ένα μόνο atomicAdd ανά (block, cluster, coord) προς global μνήμη
4. Warp efficiency / μη ιδανικά block sizes: Επειδή η εκτέλεση γίνεται σε warps των 32 threads, block sizes που δεν είναι πολλαπλάσια του 32 δημιουργούν μερικώς γεμάτα warps (wasted lanes). Αυτό μπορεί να επιδεινώσει την αποδοτικότητα και να αλλάξει το ισοζύγιο occupancy–contention.

Συμπέρασμα: Σε All-GPU, το block_size δεν είναι δευτερεύον όπως μπορεί να φαινόταν σε Transpose-only σενάρια. Αντίθετα καθορίζει άμεσα το occupancy και το atomic contention (και άρα τον GPU χρόνο), οπότε εμφανίζονται έντονα βέλτιστα σημεία και απότομες μεταβολές στην επίδοση, ειδικά στο Coords=32 όπου αυξάνεται το έργο/νήμα και το πλήθος atomicAdds ανά object. Συνολικά, το block_size καθορίζει ένα trade-off ανάμεσα σε occupancy/latency hiding και σε contention/πόρους (registers/shared), οπότε εμφανίζεται φυσιολογικά sweet spot.

E. Είναι το update_centroids κατάλληλο για GPUs; Και γιατί η All-GPU διαφέρει τόσο σε επίδοση;

To update_centroids δεν είναι ιδανικό GPU kernel με την έννοια του τέλειου, ανεξάρτητου per-thread υπολογισμού: απαιτεί συνάθροιση (reduction) πολλών contributions σε κοινά arrays (sums/counts), άρα:

- Εισάγει συγχρονισμό μέσω atomics και contention (πολλά threads ενημερώνουν τα ίδια clusters), που μπορεί να περιορίσει το scaling.
- Περιλαμβάνει στάδια που απαιτούν καθολικό συγχρονισμό (π.χ. πρώτα να ολοκληρωθούν όλα τα sums/counts πριν γίνει η διαίρεση για τα νέα centroids), κάτι που μας αναγκάζει να το σπάσουμε σε πολλαπλά kernels.

Παρόλα αυτά, η All-GPU είναι πολύ ταχύτερη συνολικά, επειδή αφαιρεί τα προηγούμενα dominant bottlenecks:

- Δεν πληρώνουμε πλέον CPU χρόνο ανά iteration για update_centroids.
- Δεν πληρώνουμε πλέον μεγάλο D2H transfer του membership ανά iteration (ούτε το H2D των clusters σε κάθε γύρο).

Άρα, ακόμη κι αν το update_centroids στη GPU “δεν είναι τέλειο” και έχει atomic overhead, το συνολικό κέρδος από την εξάλειψη CPU+PCIe κόστους είναι πολύ μεγαλύτερο, με αποτέλεσμα την εντυπωσιακή αύξηση speedup έναντι naive transpose/shared.

ΣΤ. Τι διαφέρει μεταξύ των δύο configurations και πώς αιτιολογείται η διαφορά επίδοσης;

Το κρίσιμο σημείο είναι ότι το “Size=1024” αντιστοιχεί σε σταθερό συνολικό μέγεθος dataset, άρα αλλάζει ο αριθμός των objects όταν αλλάζει το Coords:

- Με Coords=2, κάθε object έχει πολύ λιγότερα bytes → έχουμε πολύ περισσότερα objects.
- Με Coords=32, κάθε object είναι “βαρύτερο” → έχουμε πολύ λιγότερα objects.

Αυτό επηρεάζει και τη σειριακή και την παράλληλη εκτέλεση, αλλά και το είδος bottleneck:

- Coords=2: τεράστιος αριθμός objects → πολύ υψηλός παραλληλισμός (η GPU γεμίζει εύκολα), και στις παλιές εκδόσεις υπήρχαν πολύ μεγάλα per-iteration transfers (membership), τα οποία η All-GPU εξαφανίζει. Έτσι βλέπουμε πολύ υψηλό speedup.
- Coords=32: λιγότερα objects → λιγότερος παραλληλισμός και μεγαλύτερη σημασία στα σταθερά/overhead κόστη (kernel launches, reset/finalize kernels). Επιπλέον, στην All-GPU αυξάνεται η δουλειά ανά object (περισσότερες διαστάσεις σε distance + περισσότερα atomicAdds ανά object για sums), οπότε ο GPU χρόνος ανεβαίνει και το speedup περιορίζεται σε σχέση με το Coords=2.

Z. Συμπεράσματα

Η Full-Offload (All-GPU) εκδοχή επιβεβαιώνει ότι η μεγαλύτερη πηγή απώλειας στις προηγούμενες υλοποιήσεις ήταν το CPU work + PCIe transfers μέσα στο iterative loop. Με το πλήρες offload, το πρόγραμμα γίνεται πραγματικά GPU-centric και το bottleneck μεταφέρεται κυρίως στον GPU χρόνο και ειδικά στο κόστος των atomics του update_centroids. Παρ' όλα αυτά, η συνολική επίδοση βελτιώνεται θεαματικά και η All-GPU αποτελεί το φυσικό επόμενο βήμα μετά τις βελτιώσεις πρόσβασης μνήμης (transpose) και επαναχρησιμοποίησης δεδομένων (shared).

▪ Γενικά Συμπεράσματα

Η συνολική εικόνα που προκύπτει είναι ότι η επίδοση δεν καθορίζεται μόνο από το πόσο γρήγορο είναι το kernel, αλλά από το πού βρίσκεται κάθε φορά το bottleneck (PCIe transfers, CPU τμήμα, global memory traffic, atomics). Έτσι, όσον αφορά τις 4 εκδόσεις (προγράμματα) που υλοποιήσαμε:

1. Στη naïve εκδοχή, το βασικό κέρδος προκύπτει από τη μεταφορά του assignment step στη GPU, όμως η συνολική επιτάχυνση περιορίζεται από το ότι παραμένουν σημαντικά κόστη εκτός GPU: (i) οι μεταφορές δεδομένων (ιδίως η μεταφορά του membership προς τον host σε κάθε επανάληψη, που είναι $O(N)$) και (ii) το update_centroids στην CPU. Έτσι, ακόμη κι αν το kernel βελτιωθεί, το speedup περιορίζεται λόγω Amdahl (μη παραλληλοποιημένο/μη offloaded μέρος).
2. Η transpose εκδοχή δείχνει καθαρά τη σημασία της διάταξης δεδομένων και της συν-αξιοποίησης της μνήμης (coalescing). Με το transposed layout, οι προσπελάσεις γίνονται πιο συνεκτικές (coalesced) και μειώνεται η σπατάλη bandwidth, με αποτέλεσμα αισθητή βελτίωση σε σχέση με τη naïve, ειδικά όταν το workload είναι memory-bound. Παράλληλα, ο ρόλος του block_size γίνεται πιο επηρεάζει περισσότερο τη GPU (occupancy/latency hiding), καθώς η διαφορά από transfers/CPU αρχίζει να μειώνεται.
3. Στη shared εκδοχή, η μεταφορά των centroids σε shared memory λειτουργεί ως user-managed cache και μειώνει περαιτέρω τα global reads, οδηγώντας σε επιπλέον επιτάχυνση όταν το configuration το επιτρέπει. Ωστόσο, η τεχνική δεν έρχεται χωρίς κόστος: περιορίζεται από τη διαθέσιμη shared memory και μπορεί να επηρεάσει το occupancy, άρα υπάρχει πρακτικό όριο ως προς τα K-Coords και το block_size. Το συμπέρασμα είναι ότι η shared memory δίνει κέρδος όταν υπάρχει επανάχρηση δεδομένων ανά block, αλλά απαιτεί προσεκτικό διάβασμα των resource constraints.
4. Η all-GPU εκδοχή επιβεβαιώνει ότι το μεγαλύτερο κέρδος έρχεται όταν εξαλειφθούν τα υβριδικά κομμάτια μέσα στο iterative loop. Αφαιρώντας το per-iteration Device→Host membership και μεταφέροντας και το update_centroids στη GPU, μειώνεται δραστικά το transfer/CPU bottleneck, και ο συνολικός χρόνος κυριαρχείται πλέον από καθαρά GPU κόστη. Σε αυτήν τη φάση, το block_size επηρεάζει κυρίως μέσω occupancy/latency hiding, πίεσης σε registers και (κυρίως στο update_centroids) μέσω atomic contention. Ειδικά στο update_centroids, τα atomics μπορούν να αποτελέσουν σημαντικό περιορισμό, κάτι που εξηγεί γιατί το

all-GPU δεν κλιμακώνει πάντα όσο ιδανικά θα περιμέναμε χωρίς πρόσθετες τεχνικές μείωσης contention (π.χ. block-level partial sums σε shared και λιγότερα atomics προς global).

Όσον αφορά τις συντεταγμένες και το block size:

1. Η σύγκριση Coords=32 με Coords=2 δείχνει ότι το ίδιο «Size» δεν συνεπάγεται ίδιο υπολογιστικό κόστος: με μικρότερο Coords προκύπτει πολύ μεγαλύτερο πλήθος points N, άρα αυξάνει έντονα το workload του assignment και το μέγεθος του membership ($O(N)$). Αυτό μεταβάλλει το bottleneck: στο Coords=2 είναι πολύ πιο εύκολο να κυριαρχήσουν bandwidth/atomics ή ακόμη και οι μεταφορές membership (στις υβριδικές εκδοχές), ενώ στο Coords=32 το προφίλ είναι πιο ισορροπημένο και οι per-iteration μεταφορές centroids είναι πράγματι αμελητέες.
2. Τέλος, από τη μελέτη του block_size προκύπτει ότι δεν υπάρχει μία σωστή τιμή: η βέλτιστη επιλογή είναι αποτέλεσμα trade-off ανάμεσα σε occupancy, latency hiding, register/shared pressure και contention. Γι' αυτό βλέπουμε sweet spots και όχι μονοτονικές τάσεις, ενώ οι πολύ μικρές ή πολύ μεγάλες τιμές μπορούν να υποβαθμίσουν την επίδοση (είτε λόγω χαμηλής αξιοποίησης είτε λόγω περιορισμού πόρων).

Σ.Η.Μ.Μ.Υ. Ε.Μ.Π.
Ιανουάριος 2026