ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ

Λογισμικό και Προγραμματισμός Συστημάτων Υψηλής Επίδοσης

Άσκηση στο προγραμματιστικό μοντέλο CUDA

Αλεξάντερ Σκβορτσόβ

ΑΜ:5892

skvortsov@ceid.upatras.gr

Δημήτριος Μονοπάτης

ΑΜ:1040546

(Παλαιός ΑΜ:4776) - Επί πτυχίω

monopatis@ceid.upatras.gr

Ακαδημαϊκό Έτος 2016/2017

Εργαλεία ανάπτυξης

Συγγραφή αναφοράς: LibreOffice 4.3

Compiler: nvcc NVIDIA (R) Cuda compiler driver

Copyright (c) 2005-2015 NVIDIA Corporation

Built on Tue_Aug_11_14:27:32_CDT_2015

Cuda compilation tools, release 7.5, V7.5.17

Οι κώδικες έτρεξαν στην κάρτα γραφικών GeForce 820M

--- General Information for device 0 ---
Name: GeForce 820M
Compute capability: 2.1
Clock rate: 1250000
Device copy overlap: Enabled
Kernel execution timeout : Enabled
 --- Memory Information for device 0 ---
Total global mem: 2081095680
Total constant Mem: 65536
Max mem pitch: 2147483647
Texture Alignment: 512
 --- MP Information for device 0 ---
Multiprocessor count: 2
Shared mem per mp: 49152
Registers per mp: 32768
Threads in warp: 32
Max threads per block: 1024
Max thread dimensions: (1024, 1024, 64)
Max grid dimensions: (65535, 65535, 65535)

Για την μεταγλώττιση των πηγαίων αρχείων μπορεί να γίνει χρήση των Makefile με την εντολή « make » .

Για την υπολογισμό των χρόνων έγινε χρήση bash scripts.
Για το τρέξιμο των εκτελέσιμων είναι απαραίτητο να προστεθούν τα ορίσματα Rows Cols Loops Print, δηλαδή

Rows > 0 οι γραμμές του μητρώου A
Cols > 0 οι στήλες του μητρώου A
Loops: πόσες φορές να τρέξει ο κώδικας υπολογισμού
Print = 1 αν θέλουμε να εμφανίζονται οι πίνακες A και C ή 0 αν όχι

πχ για πίνακα 4x5 για το ερώτημα 1 για μια φορά υπολογισμού και εμφάνιση των μητρώων θα γράφαμε:

```
./ex1 4 5 1 1
```

Έχουμε αποθηκεύσει τα μητρώα A,C σε μονοδιάστατους πίνακες θεωρώντας τους σε column-major format.

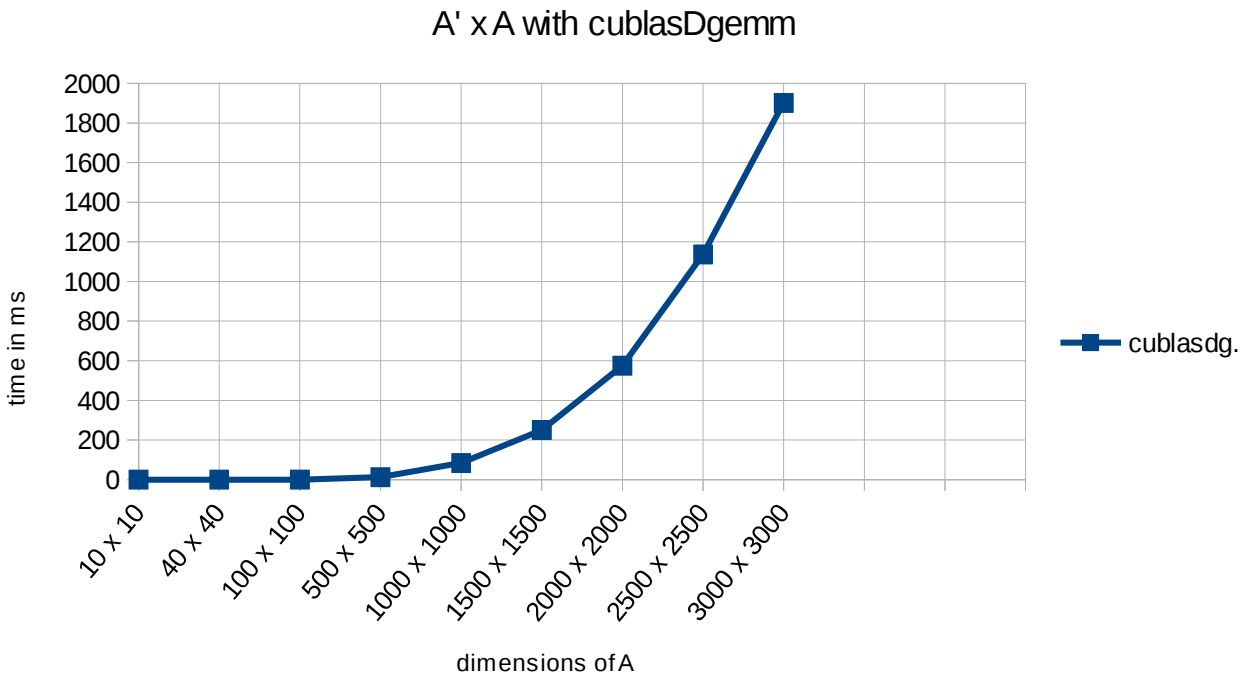
Υπάρχουν και κάποια σχόλια στους κώδικες.

Ερώτημα 1

er1.cu

Στο πρώτο ερώτημα χρησιμοποιήσαμε την έτοιμη συνάρτηση cublasDgemm από την βιβλιοθήκη cublas_v2 και την τρέξαμε 20 φορές για κάθε μέγεθος, χρονομετρώντας μόνο την συνάρτηση υπολογισμού (kernel). Το πρόγραμμα ονομάζεται ex1.cu.

Παρακάτω φαίνεται το γράφημα με τους χρόνους, όπου πήραμε το μέσο όρο των χρονομετρήσεων πλην της πρώτης.

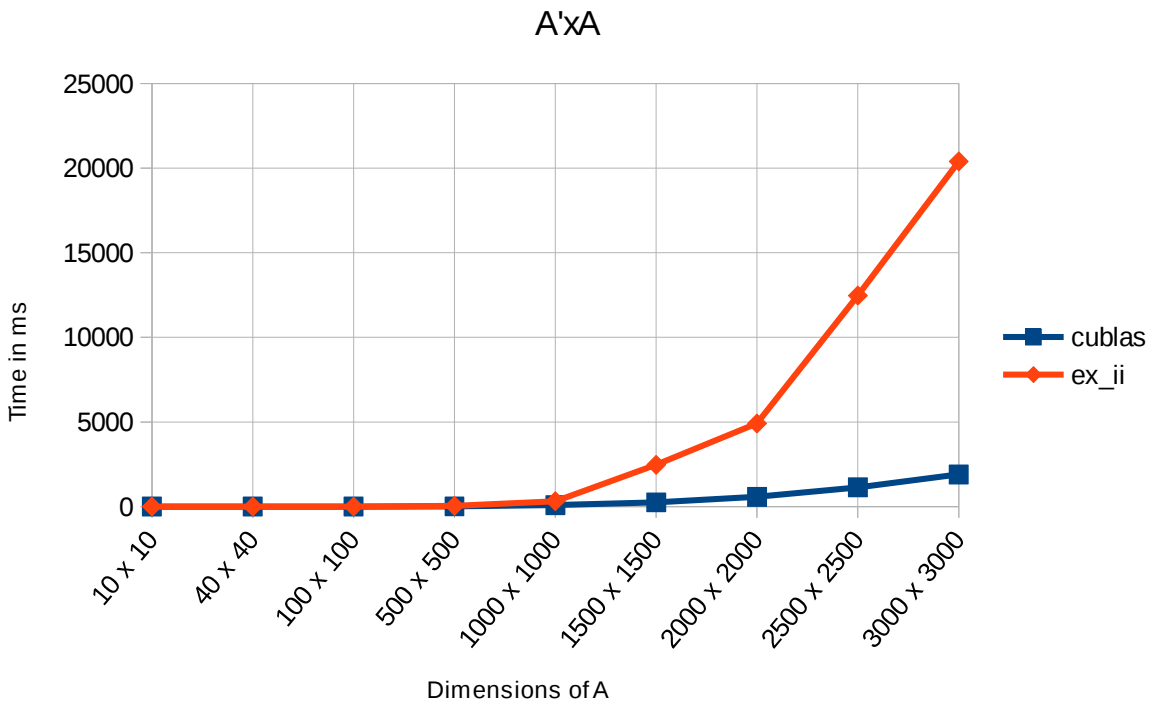


Ερώτημα 2

er2.cu

Στο δεύτερο ερώτημα υλοποιήσαμε ένα πυρήνα που κάθε νήμα υπολογίζει ένα στοιχείο του τελικού πίνακα C. Κάθε νήμα δηλαδή υπολογίζει το εσωτερικό γινόμενο κάποιας γραμμής του πίνακα A^T με την αντίστοιχη στήλη του A. Δεν γίνεται καμία εκμετάλλευση της κοινής μνήμη, ούτε κάποια άλλη τεχνική.

Παρακάτω φαίνεται το γράφημα των χρόνων σε σύγκριση με της συνάρτησης cublas.



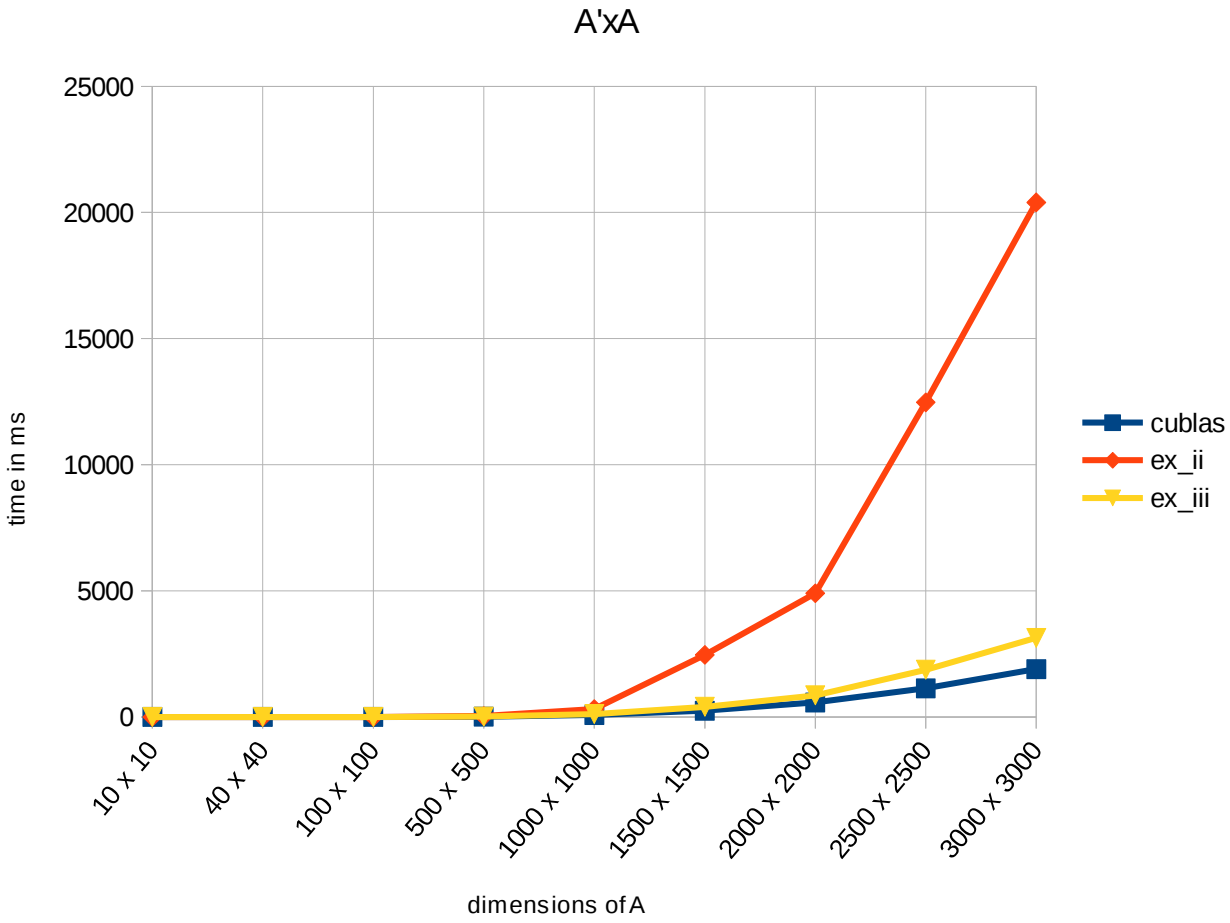
Ερώτημα 3

er3.cu

Στο τρίτο ερώτημα υλοποιήσαμε ένα πυρήνα που δουλεύει δυστυχώς μόνο για τετραγωνικά μητρώα με όνομα « multiply ». Και κάναμε χρήση του πυρήνα από το ερώτημα 2 για τα υπόλοιπα.
Υπολογίζει μόνο τον άνω τριγωνικό πίνακα και μερικά στοιχεία ακόμα.
Αφού ο C είναι από την φύση του συμμετρικός (δηλαδή C[i,j]=C[j,i]) θέτει τις τιμές και του κάτω τριγωνικού πίνακα.
Κάνει χρήση shared memory μεταξύ των block. Συγκεκριμένα ο πίνακας ds_N αποθηκεύει με την βοήθεια πολλών νημάτων για κάθε block ένα κομμάτι του A και ο ds_M του A^T.
Στο Cvalue γίνεται η προσωρινή αποθήκευση των αποτελεσμάτων όπου με την εντολή __syncthreads() το συνολικό εσωτερικό γινόμενο από όλα τα blocks αποθηκεύεται σε ένα (αυτό γίνεται αόρατα).
Οι εντολές if είναι απαραίτητες για να μην γίνει κλήση σε θέσεις του A που δεν υπάρχουν (σε περιπτώσεις που οι διαστάσεις του A δεν είναι ακέραια πολλαπλάσια του TILE_WIDTH (γραμμές 63,67)
Οι εντολές if είναι απαραίτητες για να δίνει σωστά αποτελέσματα σε ακραίες τιμές (γραμμές 61,90)
Επιλέχθηκε TILE_WIDTH=8 διότι έτρεχε σε καλύτερους χρόνους, έτσι έγινε ξετύλιγμα βρόχου στον υπολογισμό του Cvalue που επιτάχυνε λίγο ακόμη τους υπολογισμούς.

Τέλος, έγινε προσπάθεια πλήρους εκμετάλλευσης της συμμετρίας, καλώντας κάθε φορά blocks μόνο για τον υπολογισμό του άνω τριγωνικού μητρώου. Με αυτόν τον τρόπο το σύστημα θα χρειαζόταν σχεδόν τα μισά blocks κάθε φορά εξοικονομώντας έτσι πόρους της κάρτας γραφικών. Δυστυχώς όμως δεν μπορέσαμε να ολοκληρώσουμε πλήρως τον κώδικα.

Παρακάτω φαίνεται το γράφημα των χρόνων όλων των ερωτημάτων



Για τις γραφικές παραστάσεις χρησιμοποιήσαμε αυτές τις μετρήσεις

dimensions	ex1	ex2	ex3
10 x 10	0,0196682	0,0150232	0,014816
40 x 40	0,0262434	0,0415731	0,027392
100 x 100	0,169022	0,376093	0,19008
500 x 500	12,2171	39,8631	20,114
1000 x 1000	83,8438	314,903	122,653
1500 x 1500	249,995	2464,81	407,403
2000 x 2000	575,502	4905,41	861,753
2500 x 2500	1136,72	12473	1873,75