ΛΟΓΙΣΜΙΚΟ & ΠΡΟΓΡΑΜΜΑΤΙΣΜΟΣ ΣΥΣΤΗΜΑΤΩΝ ΥΨΗΛΗΣ ΕΠΙΔΟΣΗΣ

ΑΚΑΔΗΜΑΪΚΟ ΕΤΟΣ 2018/2019

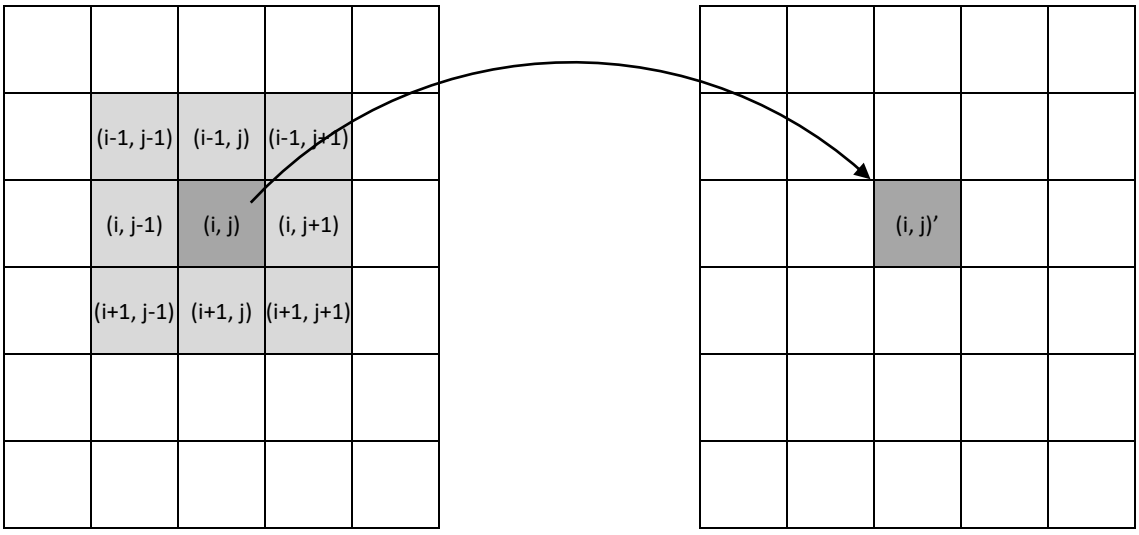
Εισαγωγή

Οι υπολογιστικοί πυρήνες (computational kernels) είναι τμήματα υπολογισμών τα οποία εμφανίζονται συχνά στα πλαίσια μεγαλύτερων εφαρμογών και συνήθως καταναλώνουν σημαντικό ποσοστό του συνολικού χρόνου εκτέλεσης της εφαρμογής. Κατά συνέπεια, είναι σημαντικό ο χρόνος που καταναλώνεται σε έναν υπολογιστικό πυρήνα να συρρικνώνεται όσο περισσότερο γίνεται.

Στα πλαίσια της εργασίας καλείστε να παραλληλοποιήσετε με χρήση του προγραμματιστικού μοντέλου της CUDA τρεις υπολογιστικούς πυρήνες, ο καθένας με ιδιαίτερα χαρακτηριστικά που τον διαφοροποιεί από τους υπόλοιπους.

2-D συνέλιξη (2-D Convolution)

(20% της βαθμολογίας της εργασίας) Ο πρώτος υπολογιστικός πυρήνας υπολογίζει την 2-D συνέλιξη. Συγκεκριμένα, δοθέντος ενός 2-D μητρώου *A*, για τον υπολογισμό της νέας τιμής του στην θέση (i,j) χρησιμοποιείται τόσο η τρέχουσα τιμή της συγκεκριμένης θέσης, όσο και οι τιμές όλων των γειτονικών στοιχείων του (9 τιμές στο σύνολο), καθεμία εκ των οποίων πολλαπλασιάζεται με ένα «βάρος». Ο υπολογισμός φαίνεται στο παρακάτω σχήμα:



$$A'(i, j) = c_{11} \cdot A(i - 1, j - 1) + c_{12} \cdot A(i, j - 1) + c_{13} \cdot A(i + 1, j - 1) +$$
$$+ c_{21} \cdot A(i - 1, j) + c_{22} \cdot A(i, j) + c_{23} \cdot A(i + 1, j) +$$
$$+ c_{31} \cdot A(i - 1, j + 1) + c_{32} \cdot A(i, j + 1) + c_{33} \cdot A(i + 1, j + 1);$$

<u>Πολλαπλασιασμός ανάστροφου μητρώου με το κανονικό μητρώο και διάνυσμα (Multiplication of Transposed with Normal Matrix and a Vector)</u>

(35% της βαθμολογίας της εργασίας) Δοθέντος ενός 2-D μητρώου *A* διαστάσεων *N*x*M* και ενός διανύσματος *x* μεγέθους *M*, ο δεύτερος υπολογιστικός πυρήνας υπολογίζει την έκφραση $A^T \cdot A \cdot x$, δηλαδή το γινόμενο του ανάστροφου του μητρώου επί το αρχικό μητρώο επί το διάνυσμα x. Για τον περιορισμό του πλήθους των πράξεων και του ενδιάμεσου αποθηκευτικού χώρου που απαιτείται για τον υπολογισμό, πραγματοποιείται πρώτα ο υπολογισμός $A \cdot x$.

<u>Συνδιακύμανση (Covariance)</u>

(45% της βαθμολογίας της εργασίας) Ο τρίτος υπολογιστικός πυρήνας υπολογίζει το μητρώο συνδιακύμανσης ενός μητρώου *A*. Συγκεκριμένα:

- Για κάθε στήλη του μητρώου *A* υπολογίζει τον μέσο όρο των στοιχείων της στήλης.
- Από κάθε στοιχείο μιας στήλης του μητρώου *A* αφαιρεί τον μέσο όρο της αντίστοιχης στήλης που υπολογίστηκε στο προηγούμενο βήμα.
- Πολλαπλασιάζει το μητρώο *A* του προηγούμενου βήματος με τον ανάστροφο του μητρώου αυτού. Επειδή το μητρώο που προκύπτει είναι συμμετρικό, αρκεί να υπολογιστεί είτε το άνω είτε το κάτω τριγωνικό μέρος του αποτελέσματος.

Συνίσταται να υλοποιήσετε τους τρεις υπολογιστικούς πυρήνες με την σειρά που παρατίθενται, καθώς η υλοποίηση και κατανόηση ενός πιο εύκολου υπολογιστικού πυρήνα θα σας βοηθήσει να προχωρήσετε στον επόμενο υπολογιστικό πυρήνα.

Ζητούμενα της άσκησης

Στα πλαίσια της άσκησης σας ζητείται να υλοποιήσετε σε CUDA και να πάρετε μετρήσεις απόδοσης για τους υπολογιστικούς πυρήνες που σας δόθηκαν. **Οι υλοποιήσεις θα πρέπει να είναι γενικές, υπό την έννοια ότι θα πρέπει να υποστηρίζεται οποιοδήποτε μέγεθος διανυσμάτων και μητρώων (και όχι, π.χ., μόνο διανύσματα και μητρώα με μεγέθη που είναι δυνάμεις του 2).** Τα στοιχεία των διανυσμάτων και μητρώων θα είναι τύπου "double". Πιο συγκεκριμένα, προσπαθήστε να βελτιστοποιήσετε την απόδοση της υλοποίησης σας αξιοποιώντας την ιεραρχία μνήμης της CUDA. Χρησιμοποιείστε καταχωρητές, κοινή μνήμη (shared memory), streams, εσωτερική αναδιοργάνωση των δεδομένων στην κάρτα γραφικών και ότι άλλο θεωρείτε ότι μπορεί να σας δώσει καλύτερη απόδοση! **Προσπαθήστε να λάβετε υπόψη σας τα μοτίβα προσπέλασης δεδομένων στην μνήμη (memory access patterns) για την όσο το δυνατόν καλύτερη εκμετάλλευση της κοινής μνήμης, όσο και για την αξιοποίηση της προσπέλασης συνεχόμενων στοιχείων στην καθολική μνήμη (coalesced accesses in main memory).** Οποιοδήποτε υλικό βρείτε (βιβλία, GPU Gems, δημοσιεύσεις σε επιστημονικά περιοδικά ή συνέδρια, κλπ) μπορείτε φυσικά να το χρησιμοποιήσετε.

Αυτό που ζητείται είναι να δείξετε ότι έχετε κατανοήσει την αρχιτεκτονική των καρτών γραφικών και πως μπορείτε να απεικονίσετε αποδοτικά έναν αλγόριθμο στην αρχιτεκτονική αυτή, λαμβάνοντας υπ' όψη σας όλες τις παραμέτρους του αλγόριθμου και της αρχιτεκτονικής. Μας ενδιαφέρει η απόδοση της λύσης σας!

Διαδικαστικά

Η εργασία θα πρέπει να γίνει σε ομάδες των 2 ή 3 ατόμων. Η διαχείριση των ομάδων θα γίνει μέσω της ηλεκτρονικής πλατφόρμας "Open eClass" του Πανεπιστημίου Πατρών (http://eclass.upatras.gr). Για τον σκοπό αυτό θα πρέπει όλοι οι φοιτητές που επιθυμούν να παραδώσουν εργασία να εγγραφούν πρώτα στην παραπάνω πλατφόρμα. Στην συνέχεια, ένα άτομο από κάθε ομάδα θα αναλάβει να δηλώσει την ομάδα του μέχρι την **Τρίτη, 06/11/2018 και ώρα 23:59:59**. Το άτομο αυτό θα είναι επίσης υπεύθυνο για όλη την επικοινωνία της ομάδας μαζί μας, καθ' όλη την διάρκεια του εξαμήνου και μέχρι την παράδοση της άσκησης. Η ομάδα θα δηλωθεί μέσω e-mail στην διεύθυνση venetis@ceid.upatras.gr. Για την ευκολότερη ταξινόμηση από την μεριά μας και την δυνατότητα αυτόματης προώθησης, το e-mail θα πρέπει να έχει τον εξής τίτλο (subject):

`[HPC18-19] Δήλωση ομάδας`

Το περιεχόμενο του e-mail θα πρέπει να είναι ο Α.Μ. και το ονοματεπώνυμο του φοιτητή που κάνει την δήλωση της ομάδας. Στην συνέχεια θα αναλάβουμε να φτιάξουμε μια ομάδα στο "Open eClass" και θα σας ενημερώσουμε για τον αριθμό της ομάδας σας.

Για όσες ομάδες έχουν ήδη ζητήσει μέσω e-mail πρόσβαση στο σύστημα GPU του εργαστηρίου, μαζί με το λογαριασμό τους στο σύστημα δημιουργήθηκε και η αντίστοιχη ομάδα στο eClass οπότε **δεν** χρειάζεται να ξαναστείλουν e-mail.

Σε περίπτωση που χρειαστεί επιπλέον επικοινωνία μαζί μας μέσω e-mail, αυτή θα πρέπει να γίνει με τον κ. Ιωάννη Βενέτη (venetis@ceid.upatras.gr). **Για την ευκολότερη ταξινόμηση από την μεριά μας και την δυνατότητα αυτόματης προώθησης, ο τίτλος (subject) κάθε e-mail θα πρέπει να ξεκινάει με [HPC18-19].**

Παραδοτέα

Τα παραδοτέα για την εργασία σας είναι μια γραπτή αναφορά και ο κώδικας της άσκησης που θα αναπτύξετε. **Η προθεσμία παράδοσης της εργασίας ορίζεται η Κυριακή 13/01/2019 και ώρα 23:59:59. Η εργασία θα πρέπει να παραδωθεί αποκλειστικά μέσω της ηλεκτρονικής πλατφόρμας "Open eClass" (εργασίες που θα αποσταλούν μέσω e-mail δεν θα βαθμολογηθούν).** Μετά την είσοδο σας στο σύστημα θα πρέπει να μεταβείτε στο μάθημα "Λογισμικό & Προγραμματισμός Συστημάτων Υψηλής Επίδοσης " και στο μενού αριστερά να μεταβείτε στο "Εργασίες". **Κάθε ομάδα θα παραδώσει μια φορά μόνο την εργασία (όχι κάθε φοιτητής ξεχωριστά). Βεβαιωθείτε πως στο εξώφυλλο της γραπτής αναφοράς αναφέρονται τα ονόματα και οι ΑΜ όλων των συμμετεχόντων της ομάδας.**

Στην αναφορά **δεν** θα πρέπει να περιλαμβάνεται επεξήγηση των ακολουθιακών αλγορίθμων που σας δόθηκαν. Επικεντρωθείτε στην επεξήγηση της παραλληλοποίησης που κάνατε, πως αξιοποιήσατε τις δυνατότητες της CUDA, πως απεικονίσατε τον αλγόριθμο στην αρχιτεκτονική σας, στις μετρήσεις σας και στα διαγράμματα που θα προσθέσετε.

Ο βαθμός της εργασίας αποτελεί το 60% της τελικής βαθμολογίας. Το υπόλοιπο 40% προκύπτει από την τελική εξέταση. Για να περάσει κάποιος φοιτητής το μάθημα δεν είναι απαραίτητη η συμμετοχή στην τελική εξέταση. **Στην περίπτωση αυτή ωστόσο, θεωρείται πως ο φοιτητής έχει πάρει στην τελική εξέταση βαθμό 0 (μηδέν). Ο τελικός βαθμός τότε προκύπτει μόνο από το 60% της εργασίας και θα πρέπει να είναι προβιβάσιμος (≥ 5).**

Ο βαθμός της εργασίας διατηρείται μέχρι και την εξεταστική Σεπτεμβρίου 2019. Αν κάποιος φοιτητής δεν περάσει το μάθημα μέχρι τότε θα πρέπει να παρακολουθήσει εξ αρχής το μάθημα και να ανταποκριθεί στις υποχρεώσεις του μαθήματος για το ακαδημαϊκό έτος που θα το παρακολουθήσει ξανά.

Σημειώνεται επίσης πως δεν είναι δυνατή η παράδοση της εργασίας σε άλλη ημερομηνία, π.χ. κατά την άτυπη εξεταστική Ιουνίου ή την εξεταστική του Σεπτεμβρίου.

Παράρτημα Α

Για να εγκαταστήσετε την CUDA στο σύστημα σας μεταβείτε στην ιστοσελίδα https://developer.nvidia.com/cuda-zone και επιλέξτε τον σύνδεσμο “Downloads”. Επιλέξτε το πακέτο που θα κατεβάσετε ανάλογα με το σύστημα σας.

Σε περιβάλλον Windows, αν και δεν είναι απαραίτητο, η ύπαρξη του Visual Studio βοηθάει ιδιαίτερα στην ανάπτυξη εφαρμογών. Κατά την εγκατάσταση της CUDA εγκαθίσταται μια επέκταση για το Visual Studio ειδικά για την ανάπτυξη εφαρμογών CUDA. Αν δεν υπάρχει το Visual Studio η ανάπτυξη προγραμμάτων μπορεί να γίνει σε οποιονδήποτε κειμενογράφο (editor) και να χρησιμοποιείται απευθείας ο μεταγλωττιστής της CUDA (nvcc) από την “Γραμμή Εντολών” (“Command Prompt”). Αντίστοιχα, σε περιβάλλον Linux ο μεταγλωττιστής καλείται από το κέλυφος (shell).

Αν η ομάδα σας δεν διαθέτει σύστημα με κάρτα γραφικών της NVidia (ή αυτή δεν υποστηρίζεται από την CUDA) επικοινωνήστε μαζί μας για να σας δώσουμε πρόσβαση σε δικό μας σύστημα. Θα μπορείτε να συνδέεστε με απομακρυσμένη πρόσβαση σε αυτό (ssh).

Σημαντική παρατήρηση: Το σύστημα στο οποίο θα σας δωθεί πρόσβαση έχει δύο κάρτες γραφικών που υποστηρίζουν CUDA. Μόνο η κάρτα με Device ID 1 έχει δυνατότητα επεξεργασίας αριθμών κινητής υποδιαστολής διπλής ακρίβειας (double). Δείτε την συνάρτηση cudaSetDevice() για να θέσετε ποια κάρτα γραφικών θα χρησιμοποιηθεί για την εκτέλεση των υπολογιστικών πυρήνων της εφαρμογής σας.