Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Η/Υ
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών

Εργαστήριο Υπολογιστικών Συστημάτων
www.cslab.ece.ntua.gr

Διπλωματική εργασία

Μελέτη και υλοποίηση του αριθμητικού υπολογιστικού πυρήνα του Πολλαπλασιασμού Αραιού Πίνακα με Διάνυσμα σε μαξικά πολυνηματικούς επεξεργαστές γραφικών

Καθηγητής: Νεκτάριος Κοζύρης (nkoziris@cslab.ece.ntua.gr)
Επικοινωνία: Βασίλειος Καρακάσης (bkk@cslab.ece.ntua.gr)
Γιώργος Γκούμας (goumas@cslab.ece.ntua.gr)
Άτομα: 1

Εισαγωγή

Τα τελευταία χρόνια παρατηρείται μία έντονη στροφή της τεχνολογίας των υπολογιστών προς αρχιτεκτονικές πολλών πυρήνων. Η συνεχής αύξηση της διαφοράς ταχύτητας μεταξύ της υπολογιστικής ισχύος των σύγχρονων επεξεργαστών και της κύριας μνήμης οδήγησε στην χρησιμοποιήση ογκωδών κρυφών μνημών, ώστε να κρυφτεί κατά το δυνατόν η χαώδης διαφορά στην επίδοση. Παρόλα αυτά, μία τέτοια λύση μόνο ως προσωρινή μπορεί να χαρακτηριστεί, καθότι δεν μπορεί να κλιμακωθεί σωστά όταν το μέγεθος του προς επίλυση προβλήματος –ειδικά όταν αυτό είναι απαιτητικό σε εύρος ζώνης μνήμης– γίνεται πολύ μεγάλο.

Οι επεξεργαστές γραφικών (GPUs) αποτελούσαν ανέκαθεν μία πανίσχυρη υπολογιστική πλατφόρμα, η οποία, όμως, μπορούσε να χρησιμοποιηθεί σχεδόν αποκλειστικά σε προβλήματα γραφικών. Με την στροφή των επεξεργαστών σε πολυπύρηνες αρχιτεκτονικές, οι επεξεργαστές γραφικών με κάποιες μικρές –αλλά καίριες– τροποποιήσεις ήρθαν στο προσκήνιο, καθότι πλέον μπορούν με σχετική ευκολία να αναλάβουν και την επίλυση κοινών υπολογιστικών προβλημάτων. Η θεμελιώδης διαφορά των επεξεργαστών αυτών σε σχέση με τις πιο διαδεδομένες μέχρι τώρα πολυπύρηνες αρχιτεκτονικές

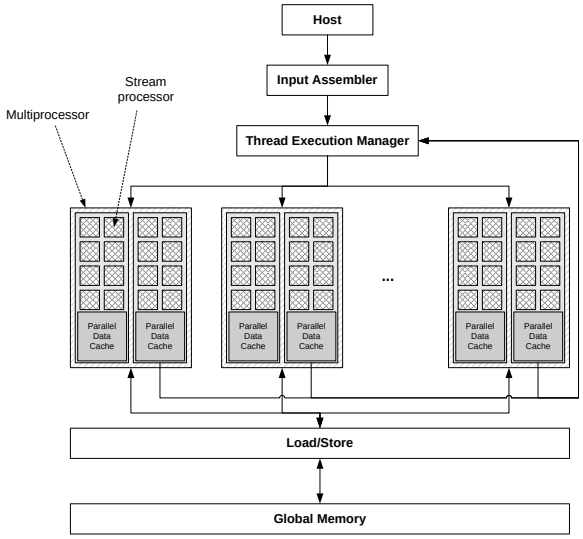
υπολογιστών είναι ότι βασίζονται σχεδόν αποκλειστικά στην παραλληλία σε επίπεδο υλικού, ώστε να καλύψουν την διαφορά επίδοσης με την κύρια μνήμη, σε αντίθεση με τους «κλασικούς» επεξεργαστές πολλαπλών πυρήνων που βασίζονται σε ογκώδεις κρυφές μνήμες. Ένας σύγχρονος επεξεργαστής γραφικών μπορεί να εκτελεί παράλληλα εκατοντάδες νήματα υλικού, επιτυγχάνοντας ασυνήθιστες ταχύτητες επεξεργασίας δεδομένων. Παρόλα αυτά, ο τρόπος προγραμματισμού ενός τέτοιου επεξεργαστή, ώστε να επιτευχθεί η μέγιστη δυνατή επίδοση, δεν είναι μία εύκολη και άμεση διαδικασία. Απαιτείται πολύ καλή εξοικείωση τόσο με το μοντέλο προγραμματισμού, το οποίο συνίσταται στην μέγιστη δυνατή παραλληλία σε επίπεδο υλικού, όσο και με τις λεπτομέρειες τις αρχιτεκτονικής.

Ο υπολογιστικός πυρήνας του Πολλαπλασιασμού Αραιού Πίνακα με Διάνυσμα (Sparse Matrix-Vector Multiplication – SpMV) συναντάται σε πληθώρα υπολογιστικών προβλημάτων και αποτελεί ένα από τα επτά σημαντικότερα υπολογιστικά προβλήματα των επόμενων δεκαετιών [3]. Το πρόβλημα SpMV είναι εξαιρετικά απαιτητικό σε εύρος ζώνης μνήμης και παρουσιάζει μία πληθώρα εγγενών προβλημάτων επίδοσης στις σύγχρονες αρχιτεκτονικές υπολογιστών [6,9]. Επομένως, η μεταφορά του σε πιο σύγχρονες ή εξεζητημένες αρχιτεκτονικές υπολογιστών αποτελεί μία μεγάλη πρόκληση. Το ενδιαφέρον της ερευνητικής κοινότητας για την μεταφορά του συγκεκριμένου πυρήνα στους επεξεργαστές γραφικών είναι ήδη αρκετά έντονο, ενώ έχουν ήδη παρουσιαστεί κάποιες πρώτες υλοποιήσεις [4,5,8].

Σκοπός

Ο σκοπός της συγκεκριμένης διπλωματικής εργασίας είναι αφενός η εξοικείωση του/της ενδιαφερομένου/ης με μία πραγματική αρχιτεκτονική πολλών πυρήνων, όπως είναι οι επεξεργαστές γραφικών της σειράς 8 της NVIDIA, και αφετέρου η μελέτη και αξιολόγηση των προβλημάτων που παρουσιάζονται κατά την υλοποίηση του SpMV σε μία τέτοια αρχιτεκτονική. Κάποια από τα ερωτήματα που πρόκειται να απαντηθούν από την εκπόνηση της συγκεκριμένης εργασίας είναι τα ακόλουθα:

- Ποια από τα προβλήματα που παρουσιάζει το SpMV σε μία «κλασική» πολυπύρηνη αρχιτεκτονική εξακολουθούν να υφίστανται σε έναν επεξεργαστή γραφικών; Ποια εξαλείφονται;
- Πώς πρέπει να υλοποιηθεί ο πυρήνας του SpMV, ώστε να εξαχθεί η μέγιστη δυνατή παραλληλία και πώς επηρεάζεται η επίδοση από τις διάφορες αρχιτεκτονικές παραμέτρους;
- Έχουν νόημα χρήσιμες και αποδοτικές βελτιστοποιήσεις του SpMV, όπως το blocking, σε μία τόσο παράλληλη αρχιτεκτονική, όπως είναι ένας επεξεργαστής γραφικών;



Σχήμα 1: Η αρχιτεκτονική G80 της NVIDIA. Κάθε επεξεργαστής γραφικών είναι μία ομάδα πολυ-επεξεργαστών, καθένας εκ των οποίων αποτελείται από οκτώ επεξεργαστές σε ροές. Κάθε πολυ-επεξεργαστής μπορεί να αναλάβει οποιουδήποτε τύπου υπολογιστικά καθήκοντα, γι΄ αυτό και η αρχιτεκτονική αυτή των επεξεργαστών γραφικών ονομάζεται *ενοποιημένη (unified)*.

- Τελικά, πόσο μπορεί να κερδίσει κανείς σε σχέση με ένα βελτιστοποιημένο κώδικα SpMV για «κλασικές» αρχιτεκτονικές, αναλαμβάνοντας το κόστος της υλοποίησης του συγκεκριμένου υπολογιστικού πυρήνα σε μία τέτοια εξεζητημένη αρχιτεκτονική;

Τα πειράματα θα διεξαχθούν σε έναν ή δύο επεξεργαστές γραφικών της NVIDIA, συγκεκριμένα στους επεξεργαστές GeForce 8800 Ultra [7] της αρχιτεκτονικής G80. Κάθε τέτοιος επεξεργαστής γραφικών αποτελείται από 16 πολυ-επεξεργαστικά στοιχεία (multi-processors), καθένα εκ των οποίων αποτελείται από οκτώ επεξεργαστές υπολογισμών σε ροές (streaming processors). Συνολικά, επομένως, η αρχιτεκτονική αυτή αποτελείται από 128 επεξεργαστικούς πυρήνες (Σχήμα 1). Η υλοποίηση θα γίνει με χρήση του προγραμματιστικού μοντέλου CUDA [2] μέσω του αντίστοιχου κιτ ανάπτυξης λογισμικού (SDK) σε περιβάλλον λειτουργικού συστήματος GNU/Linux.

Στάδια Υλοποίησης

Η προτεινόμενη διπλωματική εργασία μπορεί να χωριστεί ενδεικτικά στα εξής επιμέρους στάδια:

1. Μελέτη και εξοικείωση με την αρχιτεκτονική πολλών πυρήνων του επεξεργαστή

γραφικών της NVIDIA και αρχική εξοικείωση με το περιβάλλον και το μοντέλο προγραμματισμού CUDA.

2. Μελέτη της υλοποίησης του SpMV σε «κλασικές» αρχιτεκτονικές υπολογιστών και μελέτη της σχετικής βιβλιογραφίας, τόσο για το SpMV όσο και για την μεταφορά του σε επεξεργαστές γραφικών.

3. Υλοποίηση του SpMV για τον επεξεργαστή γραφικών.

4. Εκτέλεση πειραμάτων.

5. Αξιολόγηση και ανάλυση της επίδοσης.

6. (*Προαιρετικό*) Μελέτη μεθόδων blocking για την υλοποίηση του SpMV στον επεξεργαστή γραφικών. Έχουν νόημα;

7. (*Προαιρετικό*) Μελέτη και ανάλυση επίδοσης της εκτέλεσης του SpMV σε δύο επεξεργαστές γραφικών παράλληλα.

8. Συγγραφή της διπλωματικής εργασίας.

Προαπαιτούμενες Γνώσεις

Για την εκπόνηση της προτεινόμενης διπλωματικής εργασίας απαιτούνται ή είναι επιθυμητά τα εξής:

- Καλή γνώση της γλώσσας προγραμματισμού C ή C++.
- Γνώση αρχιτεκτονικής πολυπύρηνων επεξεργαστών.

Γνώση που θα αποκτηθεί

Με το πέρας της συγκεκριμένης διπλωματικής εργασίας, ο/η ενδιαφερόμενος/η θα έχει αποκτήσει πολύ καλή γνώση μιας σύγχρονης μαζικά παράλληλης πολυνηματικής αρχιτεκτονικής καθώς και μεγάλη εξοικείωση με τον τρόπο προγραμματισμού της. Επίσης, θα αποκτηθεί πολύτιμη εμπειρία στο πεδίο μελέτης της επίδοσης του υπολογιστικού πυρήνα SpMV, καθώς και στις μεθόδους που χρησιμοποιούνται για την βελτίωση της επίδοσής του, τόσο σε πιο «κλασικές» αρχιτεκτονικές υπολογιστών όσο και σε πιο εξεζητημένες, όπως είναι η αρχιτεκτονική του αντικειμένου μελέτης της προτεινόμενης εργασίας.

Αναφορές

[1] NVIDIA CUDA Forum. http://forums.nvidia.com/index.php?showforum=62.

[2] NVIDIA CUDA Zone. http://www.nvidia.com/object/cuda_home.html.

[3] K. Asanovic and et al. The Landscape of Parallel Computing Research: A View from Berkeley. Technical Report UCB/EECS-2006-183, EECS Department, University of California, Berkeley, December 2006.

[4] M. Christen, O. Schenk, and H. Burkhart. General-purpose sparse matrix building blocks using NVIDIA CUDA technology platform. In *Workshop on General Processing on Graphics Processing Units*, Boston, MA, USA, October 2007.

[5] M. Garland. Sparse matrix computations on manycore GPU's. *Design Automation Conference, 2008. DAC 2008. 45th ACM/IEEE*, pages 2–6, June 2008.

[6] G. Goumas, K. Kourtis, N. Anastopoulos, V. Karakasis, and N. Koziris. Performance Evaluation of the Sparse Matrix-Vector Multiplication on Modern Architectures. *The Journal of Supercomputing*, (to appear).

[7] NVIDIA Corp. NVIDIA GeForce 8800 GPU Architecture Overview. Technical Report TB-02787-001_v01, NVIDIA Corp., November 2006.

[8] V. Volkov and J. Demmel. LU, QR and Cholesky Factorizations using Vector Capabilities of GPUs. Technical Report UCB/EECS-2008-49, EECS Department, University of California, Berkeley, May 2008.

[9] S. Williams, L. Oilker, R. Vuduc, J. Shalf, K. Yelick, and J. Demmel. Optimization of sparse matrix-vector multiplication on emerging multicore platforms. In *Supercomputing'07*, Reno, NV, November 2007.

[10] General Purpose Computation Using Graphics Hardware. http://www.gpgpu.org.