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

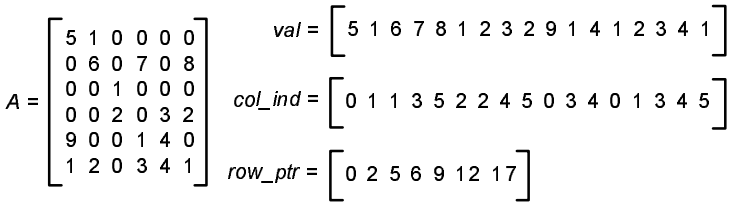
Συγκριτική μελέτη μεθόδων αποθήκευσης αραιών πινάκων σε μπλοκ για την βελτιστοποίηση του αριθμητικού υπολογιστικού πυρήνα του Πολλαπλασιασμού Αραιού Πίνακα με Διάνυσμα

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

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 [1]. Το πρόβλημα SpMV είναι εξαιρετικά απαιτητικό σε εύρος ζώνης μνήμης, καθώς ο λόγος των υπολογισμών προς τις προσβάσεις στην κύρια μνήμη είναι $O(n^2)/O(n^2)$, σε αντίθεση με άλλα αριθμητικά προβλήματα, όπως είναι ο πολλαπλασιασμός πινάκων, όπου ο λόγος αυτός είναι $O(n^3)/O(n^2)$. Επίσης, το SpMV παρουσιάζει και μία πληθώρα εγγενών προβλημάτων επίδοσης στις σύγχρονες αρχιτεκτονικές υπολογιστών [2,5], τα οποία οδηγούν σε πολύ μέτριες επιδόσεις.

Η πιο διαδεδομένη μορφή αποθήκευσης ενός αραιού πίνακα είναι η μορφή CSR. Για την αποθήκευση ενός αραιού πίνακα σε αυτή την μορφή απαιτούνται τρεις μονοδιάστατοι πίνακες: ένας πίνακας για την αποθήκευση των μη μηδενικών στοιχείων του αρχικού πίνακα, ένας πίνακας που αποθηκεύει τους αριθμούς των στηλών κάθε μη



Σχήμα 1: Παράδειγμα αποθήκευσης ενός αραιού πίνακα σε μορφή CSR.

μηδενικού στοιχείου στον αρχικό πίνακα και, τέλος, ένας πίνακας που αποθηκεύει δείκτες στον πίνακα των μη μηδενικών στοιχείων στις θέσεις όπου ξεκινάει η κάθε γραμμή του αρχικού πίνακα. Η μορφή CSR (Compressed Sparse Row) φαίνεται στο Σχήμα 1. Παρόλο που η συγκεκριμένη μορφή εξοικονομεί σημαντικό χώρο στην κύρια μνήμη, δεν αποτελεί την βέλτιστη μορφή αποθήκευσης γενικού σκοπού[1] για κάθε αραιό πίνακα.

Πολλοί αραιοί πίνακες που έχουν προκύψει από συγκεκριμένες εφαρμογές παρουσιάζουν μία τοπικότητα στην κατανομή των μη μηδενικών στοιχείων τους, εμφανίζοντας πολλά μικρότερα ή μεγαλύτερα πυκνά μπλοκ από μη μηδενικά στοιχεία. Για τον λόγο αυτό αυτό έχει προταθεί μία πληθώρα από μορφές αποθήκευσης σε μπλοκ για αραιούς πίνακες, οι οποίες εκμεταλλεύονται τέτοιου είδους σχέδια (patterns), ώστε να βελτιώσουν την επίδοση του SpMV. Ωστόσο, στην πράξη τα πυκνά μπλόκ ενός αραιού πίνακα δεν είναι πλήρη, επομένως, είναι αναγκαίο να αποθηκεύονται μηδενικά στοιχεία, ώστε να σχηματιστούν πλήρη μπλοκ. Αυτό έχει ως αποτέλεσμα, το όποιο όφελος από την αποθήκευση με μπλοκ να υπερκεραστεί με περιττούς υπολογισμούς πάνω σε μηδενικά στοιχεία. Έτσι, έχουν επίσης προταθεί μορφές αποθήκευσης, στις οποίες ο αραιός πίνακας χωρίζεται σε δύο ή περισσότερους πίνακες, εκ των οποίων ο ένας είναι αποθηκευμένος στην κλασική μορφή CSR και οι υπόλοιποι σε κάποια μορφή μπλοκ. Ακόμα και αυτοί οι τρόποι αποθήκευσης έχουν σημαντικά μειονεκτήματα, γιατί για τον τελικό υπολογισμό του SpMV θα πρέπει να εκτελεστούν επιπλέον πρόσθεσεις διανυσμάτων. Στην πράξη, η βέλτιστη μορφή αποθήκευσης εξαρτάται άμεσα τόσο από τον αρχικό αραιό πίνακα όσο και από την υποκείμενη αρχιτεκτονική, στην οποία θα εκτελεστεί το SpMV. Η επιλογή κάθε φορά της βέλτιστης μορφής αποθήκευσης αποτελεί ακόμα ένα αρκετά ενεργό ερευνητικό πεδίο.

Σκοπός

Ο σκοπός της συγκεκριμένης διπλωματικής εργασίας είναι η μελέτη, υλοποίηση και πειραματική σύγκριση σε μία σειρά από σύγχρονες υπολογιστικές πλατφόρμες των δια-

[1]Ο όρος *μορφή αποθήκευσης γενικού σκοπού* ενός αραιού πίνακα σημαίνει ότι οποιοσδήποτε αραιός πίνακας μπορεί να αποθηκευθεί σ΄ αυτή την μορφή, σε αντίθεση με εξειδικευμένες μορφές αποθήκευσης, που βασίζονται σε συγκεκριμένα σχέδια (patterns) των μη μηδενικών στοιχείων του αραιού πίνακα.

φόρων μορφών αποθήκευσης αραιών πινάκων σε μπλοκ. Συγκεκριμένα, θα πρέπει να μελετηθούν οι παρακάτω μορφές αποθήκευσης:

Blocked CSR – BCSR Ουσιαστικά πρόκειται για την μορφή αποθήκευσης CSR, αλλά με χρήση μπλοκ. Είναι η πιο διαδεδομένη μορφή αποθήκευσης αραιών πινάκων σε μπλοκ, αλλά μπορεί να οδηγήσει σε υπερβολική χρήση μηδενικών στοιχείων για την δημιουργία πλήρων μπλοκ.

Unaligned BCSR – UBCSR Ίδια με την μορφή BCSR, μόνο που πλέον τα μπλοκ δεν είναι αυστήρα στοιχισμένα ως προς τις γραμμές ή/και τις στήλες του αρχικού αραιού πίνακα. Αυτό έχει ως αποτέλεσμα να χρειάζονται να αποθηκευτούν λιγότερα μηδενικά στοιχεία για την δημιουργία πλήρων μπλοκ. Απαιτεί, όμως, επιπλέον δομές δεικτοδότησης.

Blocked Compressed Sparse Diagonal – BCSD Αντίστοιχη μορφή με την BCSR, αλλά αναζητά μονοδιάστατα μπλοκ στις διεύθυνσεις των διαγωνίων του πίνακα.

Cache-blocking Δημιουργία μπλοκ που χωρούν στην L1 κρυφή μνήμη.

One-Dimensional Variable Length Blocking – 1DVL Μορφή αποθήκευσης μονοδιάστατων μπλοκ (κατά γραμμές ή στήλες), αλλά μεταβλητού μεγέθους.

Variable Block Row – VBR Αντίστοιχο του 1DVL, αλλά σε δύο διαστάσεις.

Υβριδικές μορφές αποθήκευσης Οι μορφές αποθήκευσης, όπου τα μπλοκ έχουν σταθερό μέγεθος, μπορούν να συνδυαστούν με την μορφή CSR σχηματίζοντας υβριδικές μορφές αποθήκευσης, ώστε να μειωθεί η ανάγκη εισαγωγής μηδενικών στοιχείων για τον σχηματισμό πλήρων μπλοκ. Συγκεκριμένα, ο αρχικός πίνακας μπορεί να χωριστεί σε δύο υποπίνακες: ο πρώτος είναι σε μορφή αποθήκευσης με μπλοκ χωρίς να έχουν χρησιμοποιηθεί επιπλέον μηδενικά στοιχεία, ενώ ο δεύτερος περιέχει τα εναπομείναντα μη μηδενικά στοιχεία αποθηκευμένα με την μορφή του CSR.

Στο πλαίσιο της συγκεκριμένης διπλωματικής θα πρέπει να υλοποιηθούν οι μορφές αποθήκευσης UBCSR, Cache-blocking, VBR και κάποιες υβριδικές, οι οποίες θα ενσωματωθούν σε υπάρχουσα βιβλιοθήκη λογισμικού που αναπτύσσεται από το Εργαστήριο Υπολογιστικών Συστημάτων. Η βιβλιοθήκη αυτή υλοποιεί ήδη τις υπόλοιπες μορφές αποθήκευσης. Τα συγκριτικά πειράματα θα εκτελεστούν σε μία σειρά από σύγχρονες πολυπύρηνες ή/και πολυνηματικές μικροαρχιτεκτονικές, όπως είναι οι Intel Core, Intel Netburst και Sun Niagara.

Στάδια Υλοποίησης

Τα στάδια υλοποίησης της προτεινόμενης διπλωματικής εργασίας μπορούν να συνοψιστούν ενδεικτικά στα εξής:

1. Μελέτη της βιβλιογραφίας σχετικά με τις μορφές αποθήκευσης αραιών πινάκων με χρήση μπλοκ.
2. Μελέτη και εξοικείωση με τον υπάρχοντα κώδικα της βιβλιοθήκης αραιών πινάκων, καθώς επίσης και των τρόπων, μεθόδων και εργαλείων ανάπτυξής της (GNU autotools, coding standards, version control, testing, κ.ο.κ.).
3. Υλοποίηση των ζητούμενων μεθόδων αποθήκευσης.
4. Εκτέλεση πειραμάτων, ανάλυση αποτελεσμάτων και εξαγωγή συμπερασμάτων.
5. Συγγραφή διπλωματικής εργασίας.

Προαπαιτούμενες Γνώσεις

Για την εκπόνηση της διπλωματικής εργασίας απαιτείται καλή γνώση της γλώσσας προγραμματισμού C, βασικές γνώσεις ανάπτυξης λογισμικού και δομημένου προγραμματισμού. Επίσης, απαραίτητη είναι η βασική εξοικείωση με συστήματα UNIX/Linux.

Γνώση που θα αποκτηθεί

Στο πλαίσιο της συγκεκριμένης διπλωματικής εργασίας, ο/η ενδιαφερόμενος/η θα αποκτήσει βαθειά γνώση για τις διάφορες μορφές αποθήκευσης αραιών πινάκων τόσο ποιοτικά όσο και πραγματιστικά. Σε τεχνικό επίπεδο, θα αποκτήσει σημαντική εμπειρία στην ανάπτυξη δομημένου κώδικα μεγάλης κλίμακας στην γλώσσα C, θα εξοικειωθεί με κανόνες και πρότυπα προγραμματισμού, ενώ παράλληλα θα εξοικειωθεί με βασικά και ευρέως διαδεδομένα εργαλεία ανάπτυξης λογισμικού για συστήματα Linux, όπως είναι τα GNU automake/autoconf, libtool, svn κ.ά.

Αναφορές

[1] 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.

[2] 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).

[3] E.-J. Im, K. Yelick, and R. Vuduc. SPARSITY: Optimization Framework for Sparse Matrix Kernels. *International Journal of High Performance Computing Applications*, 18:135–158, February 2004.

[4] R. W. Vuduc and H. Moon. Fast sparse matrix-vector multiplication by exploiting variable block structure. In *High Performance Computing and Communications*, volume 3726 of *Lecture Notes in Computer Science*, pages 807–816. Springer, 2005.

[5] 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.