6ο Εργαστήριο Αρχιτεκτονικής Η/Υ: Κρυφές μνήμες (cache)

Α. Ευθυμίου

Παραδοτέο: Παρασκευή 16 Δεκέμβρη 2022, 23:59

Ο σκοπός αυτής της άσκησης είναι η εμβάθυνση της κατανόησης της λειτουργίας των κρυφών μνημών. Θα πρέπει να έχετε μελετήσει τα μαθήματα για το υποσύστημα μνήμης που αντιστοιχούν στις ενότητες 5.1-5.3 και τμήμα του 5.5 του συγγράμματος των Patterson, Hennessy.

Οι ερωτήσεις είναι σχεδιασμένες για να απαντηθούν με τη σειρά που βρίσκονται στο κείμενο γιατί ακολουθούν τη σειρά των «πειραμάτων». Αν δεν μπορείτε να απαντήσετε σε κάποια, συνεχίστε στην επόμενη. Μη δοκιμάσετε όμως να εξετάσετε μια ερώτηση χωρίς τουλάχιστον να προσπαθήσετε την προηγούμενή της.

Οι απαντήσεις θα δοθούν σε quiz στο ecourse και δεν θα χρειαστεί να ανεβάσετε αρχεία στο Github. Επομένως, σε αυτή την άσκηση θα χρειαστεί μόνο να πάρετε τον σκελετό της από το Github: https://github.com/UoI-CSE-MYY505/starter_lab06. Δεν θα έχετε το δικό σας ξεχωριστό αποθετήριο.

1 Προσομοιωτής κρυφής μνήμης του Ripes

Θα χρησιμοποιήσετε τον Ripes και κυρίως το "Cache" tab, την τρίτη από τις επιλογές στα αριστερά, που είναι ένας προσομοιωτής κρυφής μνήμης. Αν και να προσομοιώσει και την κρυφή μνήμη εντολών, στην άσκηση θα περιοριστούμε στη κρυφή μνήμη δεδομένων (L1 Data Cache). Μπορεί κανείς να ορίσει τον τρόπο οργάνωσης και τις βασικές παραμέτρους μιας κρυφής μνήμης και να δεί πως γίνονται οι προσπελάσεις στις γραμμές της καθώς και να πάρει πληροφορίες σχετικά με τον αριθμό και το ποσοστό ευστοχιών και άλλων μετρικών επίδοσης κρυφής μνήμης. **Γι'αυτή την άσκηση είναι απαραίτητο να ορίσετε τον επεξεργαστή ώστε να μην γίνεται διοχέτευση (single-cycle processor)**. Διαβάστε το https://github.com/mortbopet/Ripes/blob/master/docs/cache_sim.md για πληροφορίες.

Οι παράμετροι οργάνωσης του cache simulator βρίσκονται στο επάνω αριστερά μέρος του παραθύρου. Από κάτω υπάρχουν επιλογές για παρουσίαση γραφημάτων (plot configutation) και ένα πλαίσιο που εμφανίζονται μετρήσεις (ονομάζεται statistics). Στο μεγάλο τμήμα στα δεξιά εμφανίζεται σχηματικά η κρυφή μνήμη. Στον παραπάνω σύνδεσμο περιγράφεται πως εμφανίζονται direct mapped και set-associative οργανώσεις.

Παρατηρήστε ιδιαίτερα τον τρόπο που ορίζονται το πλήθος των γραμμών, η συσχετιστικότητα (αριθμός ways), και το μέγεθος κάθε γραμμής: δίνονται ως ο εκθέτης μιας δύναμης του 2. Δείτε τον παραπάνω σύνδεσμο για λεπτομέρειες. **Προσοχή, αυτό που αναφέρεται ως γραμμές στο πλαίσιο cache configuration, είναι ο αριθμός των σετ!** Για παράδειγμα, αν βάλετε ως 2^N Lines το 4 και 2^N Ways 1, θα δείτε ότι σχηματίζονται 16 σετ των 2 γραμμών το καθένα. Επίσης προσέξτε ότι το μέγεθος γραμμής είναι σε λέξεις όχι bytes.

2 Το πρόγραμμα cache.s

Στην άσκηση θα χρησιμοποιηθεί το cache.s που εκτελεί τον παρακάτω ψευτοκώδικα.

```
int array[arraySize];  // sizeof(int) == 4 bytes == 1 word
j = repCount;
do {
    // Step through the selected array with the given step size.
    i = 0;
    do {
        if (option == 0)
            // Option 0: One cache access - write
            array[i] = 0;
```

```
        else
            // Option 1: Two cache accesses - read AND write
            array[i] = array[i] + 1;
        i += stepSize;
    } while (i < arraySize)
    j--;
} while (j > 0) { // repeat the inner loop repCount times
```

Στις εκτελέσεις του προγράμματος θα γίνονται αλλαγές στις εξής παραμέτρους[1] του.

arraySize - το μέγεθος του πίνακα array σε λέξεις. Αντιστοιχεί στον καταχωρητή s0.

option - Ελέγχει τί θα κάνει ο εσωτερικός βρόχος: 0 μηδενίζει τα στοιχεία του πίνακα και 1 τα αυξάνει κατά 1. Αντιστοιχεί στον καταχωρητή s1.

stepSize - το βήμα με το οποίο προσπελαύνεται ο πίνακας array. Αντιστοιχεί στον καταχωρητή s2.

repCount - ο αριθμός επαναλήψεων του εσωτερικού βρόχου. Αντιστοιχεί στον καταχωρητή s3.

Στην αρχή του τμήματος δεδομένων (.data), υπάρχει το **label padding** που έχει ως σκοπό να αλλάξει την διεύθυνση από όπου αρχίζει ο πίνακας array. Η οδηγία .zero X, που χρησιμοποιείται σε αυτό το label, δεσμεύει και μηδενίζει X bytes μνήμης. Αρχικά είναι μηδέν, αλλά σε κάποια ερωτήματα θα το αλλάξετε σε 4 ή 8. Με αυτό τον τρόπο το array θα ξεκινά από την διεύθυνση 0x10000000, 0x10000004, 0x10000008, και αυτό θα προκαλεί κάποιες διαφορές στην αντιστοίχιση λέξεων του array σε γραμμές της κρυφής μνήμης.

Το σημαντικό στο cache.s είναι ο ψευτοκώδικας, οι παράμετροι και το padding, και το γεγονός ότι υπάρχουν δύο επίπεδα επαναλήψεων (φωλιασμένη επανάληψη). Επίσης προσέξτε ότι όταν το option είναι 1 γίνονται **δύο** προσπελάσεις μνήμης ανά (εσωτερική) επανάληψη: ανάγνωση και εγγραφή στην μνήμη.

Τέλος, υπενθυμίζω ότι πρέπει να έχετε επιλέξει τον επεξεργαστή με εκτέλεση σε έναν κύκλο ρολογιού (single-cycle processor) στις ρυθμίσεις του Ripes.

2.1 Παρατηρήσεις

Στα παρακάτω πειράματα προσπαθήσετε να κάνετε μια εκτίμηση για τα αποτελέσματα **πριν** τρέξετε την προσομοίωση. Μετά την προσομοίωση, σιγουρευτείτε ότι καταλαβαίνετε **γιατί** βλέπετε ό,τι βλέπετε.

Αναρωτηθείτε τα παρακάτω σε κάθε πείραμα - αλλαγή παραμέτρων προγράμματος, κρυφής μνήμης:

- Ποιό είναι το μέγεθος γραμμής (cache block size); Προσοχή μην μπερδεύετε το μέγεθος γραμμής με τον συνολικό αριθμό γραμμών της cache.
- Πόσες συνεχόμενες προσπελάσεις "χωράνε" σε μιά γραμμή (παίρνοντας υπόψη το stepSize);
- Σε ποιά θέση στην κρυφή μνήμη τοποθετείται μια (ή κάθε) συγκεκριμένη γραμμή;
- Πόσα δεδομένα του προγράμματος χωράνε σε **ολόκληρη** την κρυφή μνήμη;
- Πόσο απέχουν στην μνήμη οι γραμμές που αντιστοιχίζονται στο ίδιο set (και που μπορούν να προκαλέσουν συγκρούσεις);
- Όταν εξετάζετε αν μια προσπέλαση θα ευστοχίσει ή όχι, σκεφτείτε αν έχει ξαναγίνει προσπέλαση σε αυτήν την γραμμή στο παρελθόν. Αν έχει ήδη προσπελαστεί, υπάρχει ακόμα στην κρυφή μνήμη, ή έχει αντικατσταθεί;

[1] Προσοχή να μην συγχέονται με τις παραμέτρους της κρυφής μνήμης!

3 Μέρος 1: Αντιστοίχιση γραμμών σε θέσεις της cache

> Φορτώστε το cache.s στον Ripes τρέξτε assemble και ανοίξτε το cache tab από τα εικονίδια στα αριστερά. Οι παράμετροι του cache.s θα πρέπει να είναι: arraySize=32, option=0, stepSize=1, repCount=1. Ο αριθμός μετά το .zero στο label padding θα πρέπει να είναι 0 (η αρχική τιμή). Στο cache configuration, επάνω αριστερά στο tab Cache, επιλέξτε τις κατάλληλες τιμές ώστε να καθορίσετε μια κρυφή μνήμη direct mapping, με 8 γραμμές (lines/blocks), μέγεθος γραμμής (cache block/line size) 2 λέξεων. Θα χρησιμοποιηθεί μόνο η κρυφή μνήμη δεδομένων: L1 Data Cache.

Σημαντικό. Αν τρέξετε όλο το πρόγραμμα θα δείτε την τελική κατάσταση της κρυφής μνήμης και τα τελικά αποτελέσματα ευστοχιών - αστοχιών. Στις ερωτήσεις του quiz όμως θα χρειαστούν και ενδιάμεσα αποτελέσματα. Αντί να τρέχετε τις εντολές μία-μία, είναι καλύτερο να θέσετε breakpoints αμέσως μετά (ή πριν, αν σας βολεύει περισσότερο) από κάθε εντολή προσπέλασης μνήμης για να βλέπετε τις αλλαγές που προκαλεί στην cache κάθε μία από αυτές. Αυτό γίνεται στο Editor tab του Ripes, κάνοντας κλικ στο μπλε πλαίσιο (στήλη) στα αριστερά της εντολής που σας ενδιαφέρει. Η εκτέλεση πλέον θα σταματά πρίν εκτελεστεί η "σημαδεμένη" εντολή, όταν τρέχετε το πρόγραμμα με την επιλογή "Execute simulation (F8)..." πατώντας το εικονίδιο που μοιάζει με "»", που κανονικά εκτελεί ολόκληρο το πρόγραμμα. Έχω παρατηρήσει ότι αυτό **δεν συμβαίνει** όταν τρέχει κανείς το πρόγραμμα με το "Clock the circuit with the selected freuency (F6)" (το πράσινο "play" εικονίδιο). Για να προχωρήσει η εκτέλεση μετά το breakpoint, δυστυχώς χρειάζεται μια φορά να γίνει το "Clock the Circuit (F5)", πατώντας το εικονίδιο που μοιάζει με ">". Πατώντας το "»" δεν προχωράει στο επόμενο breakpoint, αν είναι ήδη σταματημένο σε ένα breakpoint Επειδή θα χρειαστεί αυτό να γίνει πολλές φορές, χρησιμοποιείστε τις συντμήσεις πληκτρολογίου, F8 και F5.

> Εκτελέστε το πρόγραμμα μέχρι την πρώτη προσπέλαση μνήμης και συμπληρώστε τις πληροφορίες που ζητούνται στο quiz (αριστερό μισό του πρώτου πίνακα απαντήσεων) στην γραμμή για padding 0 και την πρώτη προσπέλαση: Αν η κρυφή μνήμη ευστόχησε (Hit ή Miss), τον αριθμό του block που προσπελάστηκε (0 το πρώτο μέχρι 7 το τελευταίο) και το tag της γραμμής (φαίνεται στο runtime log του cache simulator). Για το tag χρησιμοποιείστε μόνο δεκαεξαδικούς αριθμούς.

Ο αριθμός του block που ζητείται είναι ίδιος με το index στην περίπτωση της direct mapped οργάνωσης. Παρακάτω ζητείται ο αριθμός block και για associative cache. Εκεί συμπληρώνετε την θέση, με το πρώτο, από επάνω, να είναι το 0 και το τελευταίο να είναι το 7.

> Συνεχίστε την εκτέλεση μέχρι και την δεύτερη προσπέλαση μνήμης και συμπληρώστε ξανά τις αντίστοιχες πληροφορίες στο quiz.

Ο σκοπός σας είναι να καταλάβετε σε ποιά γραμμή (cache block) αντιστοιχούν οι λέξεις που προσπελαύνονται, πώς ο προσομοιωτής αποφασίζει σε ποιό block της cache αντιστοιχεί - αποθηκεύεται κάθε γραμμή της μνήμης, πώς υπολογίζεται το tag από την διεύθυνση. Αν δυσκολεύεστε, χρησιμοποιείστε χαρτί και μολύβι, "σπάστε" την διεύθυνση στα τμήματα tag:index:offset, αφού πρώτα υπολογίσετε πόσα bits θα πρέπει να είναι το καθένα από αυτά. Ο αριθμός που σχηματίζει το index θα πρέπει να είναι ο αριθμός του block στον cache simulator και το tag που βρήκατε θα πρέπει να είναι το ίδιο. Στο επάνω μέρος του Cache tab ο Ripes δείχνει πως "σπάει" η διεύθυνση για την τρέχουσα οργάνωση και δείχνει τις λεπτομέρειες όταν εκτελείται μια εντολή προσπέλασης μνήμης.

Συνεχίστε μέχρι να αναγνωρίσετε το μοτίβο των προσπελάσεων, αλλά δεν υπάρχουν άλλες ερωτήσεις να συμπληρωθούν στο quiz γι'αυτό το πείραμα.

> Αλλάξτε το padding σε 4. Δείτε και καταγράψτε στις αντίστοιχες γραμμές του πίνακα του quiz ό,τι ζητείται για τις δύο πρώτες προσπελάσεις του προγράμματος.

Οι διαφορές είναι σχετικά μικρές αλλά θα πρέπει να μπορείτε να καταλαβαίνετε γιατί συμβαίνουν.

Επαναλάβετε με padding 8 αυτή τη φορά και συμπληρώστε το quiz.

Αλλάξτε την οργάνωση σε Fully Associative, με αλγόριθμο αντικατάστασης (Replacement Policy) LRU, και τις υπόλοιπες παραμέτρους ίδιες (συνολικό αριθμό γραμμών 8, μέγεθος γραμμής 2 λέξεις). Επαναλάβετε τις τρεις παραπάνω παραλλαγές του padding (0, 4, 8). Αυτή τη φορά συμπληρώνετε το δεξί μισό του πίνακα απαντήσεων στο quiz.

4 Μέρος 2: Επίδραση μεγέθους γραμμής

Βάλτε τις παρακάτω τιμές παραμέτρων στο cache.s: arraySize 128, option 1, stepSize 1, repCount 1, και padding 0. Ρυθμίστε την κρυφή μνήμη ως εξής: οργάνωση Direct Mapped, αριθμός γραμμών 8, μέγεθος γραμμής 2 λέξεις. Τρέξτε το πρόγραμμα και καταγράψτε το μοτίβο ευστοχιών αστοχιών ως μια ακολουθία από γράμματα M - miss, H - hit. Η απάντησή σας θα πρέπει να είναι το συντομότερο μοτίβο που επαναλαμβάνεται, για παράδειγμα το "HHMHHM" θα πρέπει να γραφεί ως "HHM". Καταγράψτε και το τελικό ποσοστό ευστοχίας του προγράμματος.

Σκεφτείτε πως το μοτίβο μπορεί άμεσα να σας δώσει το τελικό ποστοστό ευστοχίας.

Αλλάξτε το μέγεθος γραμμής σε 4 λέξεις, αλλά για να διατηρηθεί η συνολική χωρητικότητα της cache ίδια, μειώστε τον αριθμό γραμμών σε 4. Βρείτε και γράψτε το νέο μοτίβο και το ποσοστό ευστοχίας.

Σκεφτείτε σε ποιό από τα δύο είδη τοπικότητας αναφορών μνήμης οφείλεται κάθε ευστοχία του μοτίβου. Αντιγράψτε το μοτίβο που βρήκατε στο τελευταίο ερώτημα, αλλάζοντας κάθε H (hit) είτε σε T (temporal), για χρονική τοπικότητα αναφοράς, είτε σε S (spatial), για χωρική τοπικότητα. Αφήστε τα M ως είναι.

Σκεφτείτε αν το ποσοστό ευστοχίας θα άλλαζε αν αυξάνατε το repCount με όλες τις υπόλοιπες παραμέτρους προγράμματος και cache αμετάβλητες. Αλλάξτε το cache.s και τρέξτε το για να επιβεβαιώσετε ότι το βρήκατε σωστά. Αν η οργάνωση ήταν Fully Associative και όλοι οι άλλοι παράμετροι έμεναν ίδιοι[2] θα άλλαζε κάτι στο μοτίβο ή στο ποσοστό ευστοχίας; Οι παραπάνω ερωτήσεις δεν χρειάζονται απάντηση στο quiz.

5 Μέρος 3: Επίδραση repCount

Βάλτε τις παρακάτω τιμές παραμέτρων στο cache.s: arraySize 32, option 1, stepSize 4, repCount 1, και padding 0. Ρυθμίστε τον cache simulator ως εξής: οργάνωση Direct Mapped, αριθμός γραμμών 16, μέγεθος γραμμής 2 λέξεις. Καταγράψτε τον αριθμό προσπελάσεων που συμβαίνουν σε μία επανάληψη του **εξωτερικού βρόχου** και το τελικό ποσοστό ευστοχίας (στο τέλος της εκτέλεσης όλου του προγράμματος).

Αλλάξτε το repCount σε 2 και επαναλάβετε το πείραμα. Καταγράψτε το νέο (τελικό) ποσοστό ευστοχίας.

Σκεφτείτε τί θα συμβεί αν συνεχίσετε να αυξάνετε το repCount. Συμπληρώστε το μικρότερο δυνατό μοτίβο που επαναλαμβάνεται, καταγράφοντας και το είδος τοπικότητας αναφορών μνήμης σε κάθε ευστοχία, **από την δεύτερη επανάληψη και έπειτα**. (Όπως παραπάνω, γράψτε το μοτίβο, αλλά στις ευστοχίες, γράψτε S ή T, ενώ τυχόν αστοχίες συμβολίζονται με M.)

[2] Ο αλγόριθμος αντικατάστασης μπορεί να είναι LRU. Παίζει ρόλο στο συγκεκριμένο πρόγραμμα και cache;

Συμπληρώστε στο quiz τον μαθηματικό τύπο που δίνει το ποσοστό αστοχιών σε σχέση με το repCount. Γράψτε το σαν να το γράφατε π.χ. σε Java. Μπορείτε να το επιβεβαιώσετε με μερικά πειράματα, αλλά προσέξτε ότι ο Ripes παρουσιάζει το hit rate (στο πλαίσιο statistics) ενώ εδώ ζητείται το miss rate.

6 Μέρος 4: Επίδραση stepSize και associativity

Βάλτε τις παρακάτω τιμές παραμέτρων στο cache.s: arraySize 128, option 1, stepSize 16, repCount 4, και padding 0. Ρυθμίστε τον cache simulator ως εξής: οργάνωση Direct Mapped, αριθμός γραμμών 16, μέγεθος γραμμής 4 λέξεις. Καταγράψτε τον αριθμό προσπελάσεων **όλου του εσωτερικού βρόχου (δηλαδή της πρώτης επανάληψης του εξωτερικού βρόχου)** και το ποσοστό ευστοχίας. Επιπλέον, γράψτε το μοτίβο προσπελάσεων στην μνήμη της **πρώτης επανάληψης του εξωτερικού βρόχου**, αλλά κάθε σε αστοχία θα γράψετε το είδος της αστοχίας που συμβαίνει: Υ-υποχρεωτική, Σ - σύγκρουσης, Χ - Χωρητικότητας και Ε για ευστοχία. Χρησιμοποιούμε Ελληνικούς χαρακτήρες εδώ γιατί στα Αγγλικά όλα τα είδη αστοχιών ξεκινούν με C! **Προσοχή** ενώ προηγουμένως είχε ζητηθεί το συντομότερο δυνατό μοτίβο, εδώ ζητείται πλήρης καταγραφή - χαρακτηρισμός όλων των προσπελάσεων της πρώτης επανάληψης του εξωτερικού βρόχου.

Συνεχίστε την προσομοίωση και ολοκληρώστε την **δεύτερη** επανάληψη του εξωτερικού βρόχου. Σημειώστε το συνολικό προσοστό ευστοχίας από την αρχή της εκτέλεσης μέχρι αυτό το σημείο της εκτέλεσης. (Όχι μεχρι το τέλος του προγράμματος.) Γράψτε το μοτίβο ολόκληρης της **δεύτερης** επανάληψης, όπως και προηγουμένως.

Αν αλλάζατε το repCount σε μεγαλύτερο αριθμό, θα βελτιωνόταν το ποσοστό ευστοχίας;

Αν είχαμε την παράμετρο option του προγράμματος στην τιμή 0, ποιό θα ήταν το τελικό ποσοστό ευστοχίας;

Επαναφέρετε τις παραμέτρους του cache.s στις αρχικές του μέρους 3: arraySize 128, option 1, stepSize 16, repCount 4, και padding 0. Ρυθμίστε την cache ώστε να ακολουθεί την Fully Associative οργάνωση, με όλες τις υπόλοιπες παραμέτρους ίδιες (αριθμός γραμμών 16, μέγεθος γραμμής 4 λέξεις). Τρέξτε τον **πρώτο εξωτερικό** βρόχο και παρατηρείστε πώς τοποθετούνται οι γραμμές δεδομένων στην cache. Καταγράψτε το πλήθος των θέσεων (γραμμών) της cache που χρησιμοποιούνται, και το ποσοστό ευστοχίας. Το μοτίβο προσπελάσεων (ακολουθία ευστοχιών - αστοχιών και το είδος των αστοχιών) είναι ίδιο με την Direct Mapped οργάνωση για την πρώτη (εξωτερική) επανάληψη;

Συνεχίστε την προσομοίωση και ολοκληρώστε την **δεύτερη** επανάληψη του εξωτερικού βρόχου. Σημειώστε το συνολικό προσοστό ευστοχίας μέχρι αυτό το σημείο της εκτέλεσης. Το μοτίβο προσπελάσεων (ακολουθία ευστοχιών-αστοχιών και το είδος των αστοχιών) είναι ίδιο με την δεύτερη επανάλληψη της Direct Mapped οργάνωσης;

Αν αλλάζατε το repCount σε μεγαλύτερο αριθμό, θα βελτιωνόταν το ποσοστό ευστοχίας;

Βρείτε την μικρότερη associativity (set size στον cache simulator) που να δίνει το ίδιο ποσοστό ευστοχίας με την Fully Associative, χωρίς να αλλάξετε άλλες παραμέτρους ούτε στο πρόγραμμα ούτε στην κρυφή μνήμη.