

Θέματα Διπλωματικών Εργασιών

Ακαδημαϊκό Έτος 2026–2027

ΘΕΜΑ
ΔΙΠΛΩΜΑΤΙΚΗΣ

1

Χαρτογράφηση Μικροαρχιτεκτονικών Υπογραφών και Ευαισθησίας Παραμέτρων με τη Σουίτα Bringup-Bench στον gem5

Mapping Microarchitectural Signatures and Parameter Sensitivity with Bringup-Bench on gem5

Λέξεις κλειδιά: gem5 · Bringup-Bench · Μικροαρχιτεκτονική · Ανάλυση Ευαισθησίας

ΠΕΡΙΓΡΑΦΗ / ΑΝΤΙΚΕΙΜΕΝΟ

Η απόδοση ενός σύγχρονου επεξεργαστή προκύπτει από την αλληλεπίδραση πολλών μικροαρχιτεκτονικών δομών, όπως το μέτωπο προσκόμισης και αποκωδικοποίησης (front-end), η εκτέλεση εκτός σειράς (out-of-order execution), η πρόβλεψη διακλαδώσεων (branch prediction) και η ιεραρχία μνήμης. Επομένως, μία συνολική μετρική, όπως οι εντολές ανά κύκλο (instructions per cycle, IPC), δεν αρκεί για να εξηγήσει την αιτία μιας μεταβολής στην απόδοση. Η εργασία θα αξιοποιήσει τη σουίτα Bringup-Bench [1] και τον μικροαρχιτεκτονικό προσομοιωτή gem5 [2] για τη δημιουργία ελεγχόμενων μικροαρχιτεκτονικών πειραμάτων. Θα πραγματοποιηθούν συστηματικές σαρώσεις παραμέτρων (parameter sweeps) σε επιλεγμένες δομές, όπως οι κρυφές μνήμες, το πλάτος της διοχέτευσης (pipeline width), το παράθυρο εκτέλεσης και η πρόβλεψη διακλαδώσεων. Για κάθε διαμόρφωση θα συλλέγονται μετρικές απόδοσης και αξιοποίησης πόρων, με στόχο να απομονώνεται η δομή που περιορίζει κάθε μικροπρόγραμμα (microbenchmark). Με βάση τις μετρήσεις θα κατασκευαστεί μια ποσοτική «μικροαρχιτεκτονική υπογραφή» (microarchitectural signature) για διαφορετικά μοτίβα εκτέλεσης και θα μελετηθεί η ευαισθησία της στις βασικές παραμέτρους. Το αποτέλεσμα θα είναι μια αναπαραγώγιμη μεθοδολογία χαρακτηρισμού που θα μπορεί να χρησιμοποιηθεί για επιλογή αντιπροσωπευτικών microbenchmarks και για στοχευμένες μελέτες σχεδιασμού επεξεργαστών.

ΣΤΟΧΟΙ

1. Εγκατάσταση και αναπαραγώγιμη εκτέλεση της Bringup-Bench στον gem5 με ελεγχόμενες μικροαρχιτεκτονικές διαμορφώσεις.
2. Ορισμός συνόλου μετρικών και κατασκευή μικροαρχιτεκτονικών υπογραφών (microarchitectural signatures) για διαφορετικά μοτίβα εκτέλεσης.
3. Πραγματοποίηση σαρώσεων παραμέτρων και ποσοτικοποίηση της ευαισθησίας κάθε microbenchmark στις επιλεγμένες δομές του επεξεργαστή.
4. Ομαδοποίηση των microbenchmarks με βάση τη συμπεριφορά τους και τεκμηρίωση μιας επαναχρησιμοποιήσιμης μεθοδολογίας χαρακτηρισμού.

ΕΝΔΕΙΚΤΙΚΗ ΒΙΒΛΙΟΓΡΑΦΙΑ

- [1] Bringup-Bench Benchmark Suite, <https://github.com/toddmaustin/bringup-bench>
- [2] J. Lowe-Power et al., “The gem5 Simulator: Version 20.0+,” arXiv:2007.03152, 2020, <https://arxiv.org/abs/2007.03152>.
- [3] L. Eeckhout, *Computer Architecture Performance Evaluation Methods*, Synthesis Lectures on Computer Architecture, Morgan & Claypool, 2010.
- [4] J. L. Hennessy, D. A. Patterson, and C. Kozyrakis, *Computer Architecture: A Quantitative Approach*, 7th ed., Morgan Kaufmann, 2025.

ΣΧΕΤΙΚΑ ΜΑΘΗΜΑΤΑ

- Αρχιτεκτονική Υπολογιστών (CEID_23Y203)
- Εργαστήριο Αρχιτεκτονικής Υπολογιστών (CEID_23Y212)
- Προχωρημένα Θέματα Αρχιτεκτονικής Υπολογιστών (CEID_NE4617)
- Προγραμματισμός Συστημάτων (CEID_25Y401)
- Λειτουργικά Συστήματα (CEID_24Y330)

ΕΠΙΒΛΕΠΩΝ

Γιώργος Παπαδημητρίου, Επίκουρος Καθηγητής

ΘΕΜΑ
ΔΙΠΛΩΜΑΤΙΚΗΣ

2

**Χαρακτηρισμός της Διατήρησης Μικροαρχιτεκτονικής Κατάστασης
μεταξύ Serverless Κλήσεων και της Επίδρασής της στην Απόδοση**Characterizing Microarchitectural State Retention Across Serverless Invocations and Its
Performance Impact

Λέξεις κλειδιά: Serverless Computing · gem5 · Μικροαρχιτεκτονική Κατάσταση · Invocation Interference

ΠΕΡΙΓΡΑΦΗ / ΑΝΤΙΚΕΙΜΕΝΟ

Οι serverless εφαρμογές εκτελούνται ως σύντομες και ανεξάρτητες κλήσεις συναρτήσεων (function invocations), οι οποίες μπορεί να διαδέχονται η μια την άλλη με πολύ διαφορετικά χρονικά διαστήματα και μοτίβα συνύπαρξης [1]. Η συνηθισμένη διάκριση μεταξύ «ψυχρής» και «θερμής» εκτέλεσης δεν περιγράφει πλήρως τη μικροαρχιτεκτονική κατάσταση του επεξεργαστή. Μια κλήση μπορεί να βρει μέρος της κρυφής μνήμης εντολών, της προσωρινής μνήμης μετάφρασης διευθύνσεων (Translation Lookaside Buffers, TLBs) ή της κατάστασης της μονάδας πρόβλεψης διακλαδώσεων διαθέσιμο, ενώ άλλες δομές έχουν ήδη εκτοπιστεί από ενδιάμεσες κλήσεις διαφορετικών συναρτήσεων [2,3]. Η παρούσα εργασία θα μελετήσει συστηματικά αυτή τη μερική διατήρηση μικροαρχιτεκτονικής κατάστασης (microarchitectural state retention) χρησιμοποιώντας τη σουίτα vSwarm-u [4,5] και τον gem5 [6]. Θα σχεδιαστούν ελεγχόμενες ακολουθίες κλήσεων της ίδιας ή διαφορετικών serverless συναρτήσεων, με μεταβαλλόμενα χρονικά κενά, διαφορετικά μεγέθη αποτυπώματος (footprint) και διαφορετικό βαθμό παρεμβολής (interleaving). Για κάθε σενάριο θα μετρώνται η καθυστέρηση εκτέλεσης, το CPI/IPC, οι αστοχίες στις κρυφές μνήμες εντολών και δεδομένων, οι αστοχίες TLB, οι λανθασμένες προβλέψεις διακλαδώσεων και οι καθυστερήσεις του frontend. Η ανάλυση θα επιδιώξει να προσδιορίσει για πόσο χρόνο και υπό ποιες συνθήκες παραμένει χρήσιμη η μικροαρχιτεκτονική κατάσταση μιας προηγούμενης κλήσης, καθώς και ποιες δομές είναι περισσότερο ευαίσθητες στην παρεμβολή άλλων συναρτήσεων. Στο τελευταίο στάδιο θα αξιολογηθούν απλές πολιτικές, όπως η ομαδοποίηση παρόμοιων κλήσεων, η ελεγχόμενη προθέρμανση (warm-up) επιλεγμένων δομών ή η μεταβολή συγκεκριμένων μικροαρχιτεκτονικών παραμέτρων. Στόχος είναι να εξαχθούν τεκμηριωμένα συμπεράσματα για το πότε η επαναχρησιμοποίηση μικροαρχιτεκτονικής κατάστασης μπορεί να μειώσει ουσιαστικά το κόστος εκτέλεσης των serverless εφαρμογών.

ΣΤΟΧΟΙ

1. Δημιουργία αναπαραγώγιμων σεναρίων διαδοχικών και παρεμβαλλόμενων serverless κλήσεων με τη vSwarm-u στον gem5.
2. Ποσοτικοποίηση της διατήρησης κατάστασης στις κρυφές μνήμες, στα TLBs και στη μονάδα πρόβλεψης διακλαδώσεων.
3. Συσχέτιση του βαθμού μικροαρχιτεκτονικής επαναχρησιμοποίησης με την καθυστέρηση, το CPI και τα frontend stalls.
4. Μελέτη της επίδρασης διαφορετικών λειτουργιών και διαφορετικών χρονικών αποστάσεων μεταξύ κλήσεων.
5. Αξιολόγηση απλών πολιτικών διατήρησης ή επαναχρησιμοποίησης κατάστασης και διατύπωση αρχιτεκτονικών κατευθύνσεων για serverless επεξεργασία.

ΕΝΔΕΙΚΤΙΚΗ ΒΙΒΛΙΟΓΡΑΦΙΑ

- [1] D. Schall, A. Margaritov, D. Ustiugov, A. Sandberg and B. Grot, “Lukewarm Serverless Functions: Characterization and Optimization,” in *Proc. 49th Annual International Symposium on Computer Architecture (ISCA)*, 2022, pp. 757–770, doi: 10.1145/3470496.3527390.
- [2] M. Golec, G. Kaur Walia, M. Kumar, F. Cuadrado, S. Singh Gill and S. Uhlig, “Cold Start Latency in Serverless Computing: A Systematic Review, Taxonomy, and Future Directions,” *ACM Computing Surveys*, vol. 57, no. 3, Article 65, 2025, doi: 10.1145/3700875.
- [3] G. Pournaras, V. Karakostas, G. Papadimitriou and D. Gizopoulos, “Benchmarking Support for RISC-V CPUs in Serverless Computing,” in *IEEE International Symposium on Workload Characterization (IISWC)*, 2025.
- [4] vSwarm – Serverless Benchmarking Suite, <https://github.com/vhive-serverless/vSwarm>
- [5] vSwarm-μ – Microarchitectural Research for Serverless, <https://github.com/vhive-serverless/vSwarm-u>
- [6] The gem5 Simulator, <https://www.gem5.org>

ΣΧΕΤΙΚΑ ΜΑΘΗΜΑΤΑ

- Αρχιτεκτονική Υπολογιστών (CEID_23Y203)
- Εργαστήριο Αρχιτεκτονικής Υπολογιστών (CEID_23Y212)
- Προχωρημένα Θέματα Αρχιτεκτονικής Υπολογιστών (CEID_NE4617)
- Προγραμματισμός Συστημάτων (CEID_25Y401)
- Λειτουργικά Συστήματα (CEID_24Y330)

ΕΠΙΒΛΕΠΩΝ

Γιώργος Παπαδημητρίου, Επίκουρος Καθηγητής

ΘΕΜΑ
ΔΙΠΛΩΜΑΤΙΚΗΣ

3

Μικροαρχιτεκτονικά Καθοδηγούμενη Προθέρμανση για Μείωση των Ψυχρών Εκκινήσεων σε Serverless Εφαρμογές

Microarchitecture-Guided Warm-Up for Reducing Cold Starts in Serverless Workloads

Λέξεις κλειδιά: Serverless Computing · Cold Start · Warm-Up · Prefetching

ΠΕΡΙΓΡΑΦΗ / ΑΝΤΙΚΕΙΜΕΝΟ

Η ψυχρή εκκίνηση (cold start) αποτελεί μία από τις βασικές πηγές απρόβλεπτης καθυστέρησης στις serverless εφαρμογές. Συνήθως μελετάται σε επίπεδο περιβάλλοντος εκτέλεσης, κοντέινερ ή φόρτωσης βιβλιοθηκών [1-3]. Ωστόσο, ακόμη και όταν το λογισμικό είναι διαθέσιμο, οι πρώτες φάσεις μιας κλήσης εκτελούνται με «ψυχρή» μικροαρχιτεκτονική κατάσταση: οι κρυφές μνήμες (caches), οι προσωρινές μνήμες μετάφρασης διευθύνσεων (TLBs) και η μονάδα πρόβλεψης διακλαδώσεων δεν έχουν ακόμη προσαρμοστεί στη συγκεκριμένη συνάρτηση. Η εργασία θα εξετάσει κατά πόσο το κόστος αυτό μπορεί να μειωθεί με στοχευμένη και μικροαρχιτεκτονικά καθοδηγούμενη προθέρμανση (microarchitecture-guided warm-up), αντί για μια πλήρη δοκιμαστική εκτέλεση της συνάρτησης. Με τη χρήση της σουίτας vSwarm-u [4-6] και του gem5 [7] θα δημιουργηθούν τρεις βασικές κατηγορίες πειραμάτων: πλήρως ψυχρή εκτέλεση, συμβατική πλήρης προθέρμανση και επιλεκτική προθέρμανση. Η τελευταία μπορεί να περιλαμβάνει περιορισμένη εκτέλεση επιλεγμένων τμημάτων κώδικα, εκ των προτέρων προσκόμιση (prefetching) εντολών ή δεδομένων, και προετοιμασία μεταφράσεων διευθύνσεων όταν αυτό είναι τεχνικά εφικτό στο μοντέλο. Ο φοιτητής/Η φοιτήτρια θα προσδιορίσει ποιες μικροαρχιτεκτονικές δομές ευθύνονται για το μεγαλύτερο μέρος της αρχικής καθυστέρησης ανά συνάρτηση και θα σχεδιάσει έναν απλό μηχανισμό επιλογής της κατάλληλης στρατηγικής προθέρμανσης. Η αξιολόγηση θα λαμβάνει υπόψη όχι μόνο τη μείωση του χρόνου απόκρισης, αλλά και το ίδιο το κόστος της προθέρμανσης σε αριθμό εντολών, προσπελάσεις μνήμης και χρόνο εκτέλεσης. Με αυτόν τον τρόπο, το ζητούμενο δεν είναι η μέγιστη δυνατή προθέρμανση, αλλά η εύρεση του μικρότερου και αποδοτικότερου συνόλου ενεργειών που μειώνει ουσιαστικά την ποινή της ψυχρής εκκίνησης.

ΣΤΟΧΟΙ

1. Ποσοτικοποίηση της μικροαρχιτεκτονικής συνιστώσας της ψυχρής εκκίνησης σε αντιπροσωπευτικές serverless συναρτήσεις.
2. Σύγκριση πλήρως ψυχρής εκτέλεσης, πλήρους προθέρμανσης και επιλεκτικών στρατηγικών προθέρμανσης.
3. Εντοπισμός των δομών που κυριαρχούν στο αρχικό κόστος, όπως κρυφές μνήμες, TLBs και μονάδα πρόβλεψης διακλαδώσεων.
4. Σχεδίαση και αξιολόγηση μικροαρχιτεκτονικά καθοδηγούμενης πολιτικής προθέρμανσης με μετρικές κόστους/οφέλους.
5. Διατύπωση προτάσεων για runtime μηχανισμούς που μπορούν να αξιοποιήσουν τα αποτελέσματα της ανάλυσης.

ΕΝΔΕΙΚΤΙΚΗ ΒΙΒΛΙΟΓΡΑΦΙΑ

- [1] M. Golec, G. Kaur Walia, M. Kumar, F. Cuadrado, S. Singh Gill and S. Uhlig, “Cold Start Latency in Serverless Computing: A Systematic Review, Taxonomy, and Future Directions,” *ACM Computing Surveys*, vol. 57, no. 3, Article 65, 2025, doi: 10.1145/3700875.
- [2] D. Schall, A. Margaritov, D. Ustiugov, A. Sandberg and B. Grot, “Lukewarm Serverless Functions: Characterization and Optimization,” in *Proc. 49th Annual International Symposium on Computer Architecture (ISCA)*, 2022, pp. 757–770, doi: 10.1145/3470496.3527390.
- [3] A. Bhattacharya and T. Wen, “Understanding and Remediating Cold Starts: An AWS Lambda Perspective,” <https://aws.amazon.com/blogs/compute/understanding-and-remediating-cold-starts-an-aws-lambda-perspective/>
- [4] G. Pournaras, V. Karakostas, G. Papadimitriou and D. Gizopoulos, “Benchmarking Support for RISC-V CPUs in Serverless Computing,” in *IEEE International Symposium on Workload Characterization (IISWC)*, 2025.
- [5] vSwarm – Serverless Benchmarking Suite, <https://github.com/vhive-serverless/vSwarm>
- [6] vSwarm-μ – Microarchitectural Research for Serverless, <https://github.com/vhive-serverless/vSwarm-u>
- [7] The gem5 Simulator, <https://www.gem5.org>

ΣΧΕΤΙΚΑ ΜΑΘΗΜΑΤΑ

- Αρχιτεκτονική Υπολογιστών (CEID_23Y203)
- Εργαστήριο Αρχιτεκτονικής Υπολογιστών (CEID_23Y212)
- Προχωρημένα Θέματα Αρχιτεκτονικής Υπολογιστών (CEID_NE4617)
- Προγραμματισμός Συστημάτων (CEID_25Y401)
- Λειτουργικά Συστήματα (CEID_24Y330)

ΕΠΙΒΛΕΠΩΝ

Γιώργος Παπαδημητρίου, Επίκουρος Καθηγητής

ΘΕΜΑ
ΔΙΠΛΩΜΑΤΙΚΗΣ

4

Ανάλυση Διάδοσης Σφαλμάτων και Επιλεκτικής Προστασίας σε Επιταχυντές NVDLA με gem5

Fault Propagation Analysis and Selective Protection in NVDLA Accelerators with gem5

Λέξεις κλειδιά: NVDLA · Fault Injection · Fault Propagation · Selective Protection

ΠΕΡΙΓΡΑΦΗ / ΑΝΤΙΚΕΙΜΕΝΟ

Οι επιταχυντές βαθέων νευρωνικών δικτύων (Deep Neural Network accelerators, DNN accelerators) αποτελούν πλέον βασικό υποσύστημα σε συστήματα τεχνητής νοημοσύνης, όμως η μεγάλη πυκνότητα υπολογιστικών και αποθηκευτικών δομών τους αυξάνει τη σημασία της συστηματικής αξιολόγησης αξιοπιστίας. Ένα παροδικό σφάλμα (transient fault) σε έναν καταχωρητή, έναν ενδιάμεσο αποθηκευτικό χώρο (buffer) ή σε δεδομένα που τροφοδοτούν τις μονάδες υπολογισμού μπορεί να καλυφθεί, να οδηγήσει σε εμφανή αποτυχία ή να διαδοθεί μέχρι την τελική έξοδο του νευρωνικού δικτύου προκαλώντας σιωπηλή αλλοίωση δεδομένων (silent data corruption, SDC). Η εργασία θα χρησιμοποιήσει το gem5-NVDLA [1,4,5] για να μελετήσει τη διαδρομή ενός σφάλματος από επιλεγμένες και προσβάσιμες καταστάσεις του μοντέλου του επιταχυντή μέχρι την παρατηρούμενη έξοδο της εφαρμογής. Θα αναπτυχθεί μηχανισμός ελεγχόμενης στατιστικής εισαγωγής σφαλμάτων (fault injection) και θα εκτελεστούν αντιπροσωπευτικά νευρωνικά δίκτυα που υποστηρίζονται από την εργαλειοθήκη του NVDLA. Τα πειράματα θα διαφοροποιούνται ως προς τη θέση, τον χρόνο και τον τύπο της αλλοίωσης, ώστε να καταγραφεί αν το σφάλμα καλύπτεται, προκαλεί αποτυχία της εκτέλεσης ή μεταβάλλει το τελικό αποτέλεσμα [2,3]. Η βασική διαφοροποίηση από μια απλή διαδικασία εισαγωγής σφαλμάτων είναι ότι θα επιδιωχθεί η συσχέτιση της προέλευσης του σφάλματος με τον τρόπο διάδοσης και με τη σημασιολογική επίπτωσή του στην έξοδο. Με βάση αυτή την ανάλυση θα εντοπιστούν οι πλέον ευάλωτες δομές ή κατηγορίες δεδομένων και θα αξιολογηθούν αφαιρετικά σενάρια επιλεκτικής προστασίας (selective protection), όπως προστασία μόνο συγκεκριμένων buffers ή κρίσιμων bits. Η αξιολόγηση θα συγκρίνει το ποσοστό κάλυψης σφαλμάτων (coverage) με το κόστος της προστατευόμενης κατάστασης και τυχόν επίπτωση στην απόδοση, χωρίς να απαιτείται πλήρης υλοποίηση τεχνικών ανοχής σφαλμάτων σε επίπεδο RTL.

ΣΤΟΧΟΙ

1. Κατανόηση της οργάνωσης του NVDLA και των δομών που είναι διαθέσιμες για αξιόπιστη εισαγωγή σφαλμάτων στο gem5-NVDLA.
2. Υλοποίηση αναπαραγωγίμου μηχανισμού εισαγωγής σφαλμάτων σε επιλεγμένες καταστάσεις και δομές του μοντέλου.
3. Κατηγοριοποίηση των αποτελεσμάτων σε κάλυψη (masking), εμφανή αποτυχία και σιωπηλή αλλοίωση δεδομένων (SDC), μαζί με ποσοτικοποίηση της επίπτωσης στην έξοδο του DNN.
4. Συσχέτιση της θέσης και του χρόνου του σφάλματος με τη διαδρομή διάδοσης και την ευπάθεια διαφορετικών δομών.
5. Αξιολόγηση επιλεκτικών σεναρίων προστασίας και προσδιορισμός του καλύτερου συμβιβασμού μεταξύ κάλυψης και κόστους.

ΕΝΔΕΙΚΤΙΚΗ ΒΙΒΛΙΟΓΡΑΦΙΑ

- [1] C. Lai and W. Zhang, “gem5-NVDLA: A Simulation Framework for Compiling, Scheduling, and Architecture Evaluation on AI System-on-Chips,” *ACM Transactions on Design Automation of Electronic Systems*, vol. 29, no. 5, Article 84, pp. 1–20, 2024, doi: 10.1145/3661997.
- [2] G. Li, S. K. S. Hari, M. B. Sullivan, T. Tsai, K. Pattabiraman, J. S. Emer and S. W. Keckler, “Understanding Error Propagation in Deep Learning Neural Network (DNN) Accelerators and Applications,” in *Proc. SC17: International Conference for High Performance Computing, Networking, Storage and Analysis*, 2017, pp. 1–12, doi: 10.1145/3126908.3126964.
- [3] G. Papadimitriou and D. Gizopoulos, “Demystifying the System Vulnerability Stack: Transient Fault Effects Across the Layers,” in *Proc. 48th Annual International Symposium on Computer Architecture (ISCA)*, 2021, pp. 902–915, doi: 10.1109/ISCA52012.2021.00075.
- [4] gem5-NVDLA source repository, <https://github.com/suchandler96/gem5-NVDLA>
- [5] NVIDIA Deep Learning Accelerator (NVDLA), <https://nvdla.org>

ΣΧΕΤΙΚΑ ΜΑΘΗΜΑΤΑ

- Αρχιτεκτονική Υπολογιστών (CEID_23Y203)
- Εργαστήριο Λογικού Σχεδιασμού (CEID_23Y211)
- Εργαστήριο Αρχιτεκτονικής Υπολογιστών (CEID_23Y212)
- Προχωρημένα Θέματα Αρχιτεκτονικής Υπολογιστών (CEID_NE4617)
- Προγραμματισμός Συστημάτων (CEID_25Y401)
- Λειτουργικά Συστήματα (CEID_24Y330)

ΕΠΙΒΛΕΠΩΝ

Γιώργος Παπαδημητρίου, Επίκουρος Καθηγητής

ΘΕΜΑ
ΔΙΠΛΩΜΑΤΙΚΗΣ

5

Ανάλυση Διάδοσης Σφαλμάτων και Αξιοπιστίας σε Πολυπύρηνους Επεξεργαστές

Fault Propagation and Reliability Analysis in Multicore Processors

Λέξεις κλειδιά: Multicore Processors · Fault Propagation · Fault Injection · Reliability

ΠΕΡΙΓΡΑΦΗ / ΑΝΤΙΚΕΙΜΕΝΟ

Η αξιολόγηση της αξιοπιστίας ενός πολυπύρηνου επεξεργαστή δεν περιορίζεται στον εντοπισμό του σημείου στο οποίο εμφανίζεται ένα σφάλμα. Ένα παροδικό σφάλμα (transient fault) μπορεί να καλυφθεί από τη μικροαρχιτεκτονική, να επηρεάσει την κατάσταση ενός μόνο πυρήνα ή να διαδοθεί μέσω κοινόχρηστων δεδομένων, της ιεραρχίας κρυφών μνημών και των μηχανισμών συνοχής (cache coherence), μέχρι να γίνει ορατό στην έξοδο μιας παράλληλης εφαρμογής. Η ύπαρξη αλληλεπιδράσεων μεταξύ πυρήνων καθιστά τη μελέτη της διάδοσης σημαντικά πιο σύνθετη από την αντίστοιχη ανάλυση σε μονοπύρηνια συστήματα. Στην παρούσα εργασία θα αναπτυχθεί μια ελεγχόμενη μεθοδολογία εισαγωγής σφαλμάτων (fault injection) για επιλεγμένες μικροαρχιτεκτονικές δομές, όπως αρχεία καταχωρητών (register files), ιδιωτικές ή κοινόχρηστες κρυφές μνήμες και επιλεγμένη κατάσταση που συμμετέχει στην επικοινωνία μεταξύ πυρήνων. Η μελέτη θα επικεντρωθεί σε σαφώς ορισμένα σενάρια και σε αντιπροσωπευτικά παράλληλα φορτία εργασίας (parallel workloads), ώστε να είναι εφικτή η εκτέλεση στατιστικά επαρκούς αριθμού πειραμάτων χωρίς να απαιτείται εξαντλητική προσομοίωση ολόκληρου του χώρου σφαλμάτων. Για κάθε πείραμα θα καταγράφεται η πορεία του σφάλματος από την αρχική αλλοίωση μέχρι το παρατηρούμενο αποτέλεσμα: μικροαρχιτεκτονική κάλυψη (masking), σφάλμα που παραμένει τοπικό σε έναν πυρήνα, διάδοση σε κοινόχρηστη κατάσταση, ανιχνεύσιμη αποτυχία ή σιωπηλή αλλοίωση δεδομένων (silent data corruption, SDC). Ιδιαίτερο ενδιαφέρον θα δοθεί στη σύγκριση ιδιωτικών και κοινόχρηστων δομών, καθώς και στη σχέση μεταξύ μοτίβων κοινής χρήσης δεδομένων και πιθανότητας διάδοσης σε άλλο νήμα ή πυρήνα. Το τελικό αποτέλεσμα θα είναι ένας ποσοτικός χάρτης ευπάθειας (vulnerability map) που θα συνδέει τη θέση και τον τύπο του σφάλματος με τον τρόπο διάδοσης και την τελική επίπτωση στην εφαρμογή. Με βάση τα ευρήματα θα εξεταστούν στοχευμένες στρατηγικές προστασίας, ώστε να αναδειχθεί ποιες δομές ή κατηγορίες κατάστασης αξίζει να προστατεύονται κατά προτεραιότητα σε έναν πολυπύρηνιο επεξεργαστή.

ΣΤΟΧΟΙ

1. Σχεδίαση και υλοποίηση ελεγχόμενου μηχανισμού εισαγωγής παροδικών σφαλμάτων στον gem5 για επιλεγμένες δομές πολυπύρηνων επεξεργαστών.
2. Εκτέλεση στατιστικής εκστρατείας εισαγωγής σφαλμάτων σε αντιπροσωπευτικούς παράλληλους φόρτους εργασίας.
3. Κατηγοριοποίηση της διάδοσης σε τοπική επίδραση, διάδοση μέσω κοινόχρηστης κατάστασης, εμφανή αποτυχία και σιωπηλή αλλοίωση δεδομένων (SDC).
4. Μελέτη της επίδρασης της κοινής χρήσης δεδομένων, της ιεραρχίας κρυφών μνημών και της συνοχής (coherence) στην πιθανότητα διαπυρηνικής διάδοσης σφαλμάτων.
5. Δημιουργία χάρτη ευπάθειας και αξιολόγηση στοχευμένων επιλογών προστασίας με βάση τα πειραματικά αποτελέσματα.

ΕΝΔΕΙΚΤΙΚΗ ΒΙΒΛΙΟΓΡΑΦΙΑ

- [1] G. Papadimitriou and D. Gizopoulos, “Demystifying the System Vulnerability Stack: Transient Fault Effects Across the Layers,” in *Proc. 48th Annual International Symposium on Computer Architecture (ISCA)*, 2021, pp. 902–915, doi: 10.1109/ISCA52012.2021.00075.
- [2] S. S. Mukherjee, C. Weaver, J. Emer, S. K. Reinhardt and T. Austin, “A Systematic Methodology to Compute the Architectural Vulnerability Factors for a High-Performance Microprocessor,” in *Proc. 36th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 2003, pp. 29–40, doi: 10.1109/MICRO.2003.1253181.
- [3] S. K. S. Hari, S. V. Adve, H. Naeimi and P. Ramachandran, “Relyzer: Exploiting Application-Level Fault Equivalence to Analyze Application Resiliency to Transient Faults,” in *Proc. 17th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, 2012, pp. 123–134, doi: 10.1145/2150976.2150990.
- [4] J. L. Hennessy, D. A. Patterson, and C. Kozyrakis, *Computer Architecture: A Quantitative Approach*, 7th ed., Morgan Kaufmann, 2025.

ΣΧΕΤΙΚΑ ΜΑΘΗΜΑΤΑ

- Αρχιτεκτονική Υπολογιστών (CEID_23Y203)
- Εργαστήριο Αρχιτεκτονικής Υπολογιστών (CEID_23Y212)
- Προχωρημένα Θέματα Αρχιτεκτονικής Υπολογιστών (CEID_NE4617)
- Προγραμματισμός Συστημάτων (CEID_25Y401)
- Παράλληλη Επεξεργασία (CEID_25EE607)
- Λειτουργικά Συστήματα (CEID_24Y330)

ΕΠΙΒΛΕΠΩΝ

Γιώργος Παπαδημητρίου, Επίκουρος Καθηγητής

ΘΕΜΑ
ΔΙΠΛΩΜΑΤΙΚΗΣ

6

**Μοντελοποίηση Υποστήριξης Δυνατοτήτων CHERI-RISC-V στον gem5:
Αρχιτεκτονικές Επεκτάσεις και Μικροαρχιτεκτονική Ανάλυση**Modeling CHERI-RISC-V Capability Support in gem5: Architectural Extensions and
Microarchitectural Implications

Λέξεις κλειδιά: CHERI-RISC-V · Capabilities · gem5 · Memory Safety

ΠΕΡΙΓΡΑΦΗ / ΑΝΤΙΚΕΙΜΕΝΟ

Η αρχιτεκτονική CHERI (Capability Hardware Enhanced RISC Instructions) επεκτείνει τη συμβατική έννοια του δείκτη μνήμης με δυνατότητες (capabilities) που περιλαμβάνουν πληροφορία ορίων, δικαιωμάτων και εγκυρότητας. Στόχος είναι η παροχή ισχυρότερων εγγυήσεων ασφάλειας μνήμης (memory safety) και απομόνωσης (compartmentalization) με υποστήριξη από το υλικό. Η τρέχουσα προδιαγραφή CHERI για RISC-V ορίζει τις επεκτάσεις της οικογένειας RVY και περιγράφει νέες εντολές, κανόνες ελέγχου δυνατοτήτων, σημασιολογία εξαιρέσεων και απαιτήσεις για μνήμη με ετικέτες (tagged memory) [1,4,5]. Η εργασία θα εξετάσει τη μοντελοποίηση ενός σαφώς οριοθετημένου υποσυνόλου της CHERI-RISC-V στον μικροαρχιτεκτονικό προσομοιωτή gem5 [6]. Θα επιλεγεί ένα συνεκτικό σύνολο λειτουργιών που επιτρέπει την εκτέλεση ελεγχόμενων μικροπρογραμμάτων (microbenchmarks), όπως αναπαράσταση καταχωρητών δυνατοτήτων, βασικές εντολές δημιουργίας/περιορισμού δυνατοτήτων, έλεγχος δικαιωμάτων και ορίων, μεταφορά της ετικέτας εγκυρότητας και χειρισμός επιλεγμένων εξαιρέσεων. Η υλοποίηση θα γίνει με τρόπο που να διαχωρίζει καθαρά τις αλλαγές στην αρχιτεκτονική συνόλου εντολών (Instruction Set Architecture, ISA) από τις μικροαρχιτεκτονικές συνέπειες. Η ορθότητα του πρωτοτύπου θα ελεγχθεί με μικροπρογράμματα που παράγονται χειροκίνητα ή με χρήση του CHERI LLVM/Clang [2,3], και όπου είναι πρακτικό με σύγκριση της συμπεριφοράς με διαθέσιμες αναφορές της CHERI-RISC-V. Στη συνέχεια θα μελετηθούν επιπτώσεις όπως η πρόσθετη κατάσταση στους καταχωρητές, η αποθήκευση και μεταφορά των tags, οι επιπλέον έλεγχοι στην κρίσιμη διαδρομή των προσπελάσεων μνήμης και η μεταβολή του μείγματος εντολών (instruction mix). Τελικός στόχος είναι ένα τεχνικά τεκμηριωμένο πρωτότυπο που θα δείχνει τι απαιτείται για την ενσωμάτωση CHERI-RISC-V στον gem5 και θα ποσοτικοποιεί τις βασικές μικροαρχιτεκτονικές επιπτώσεις ενός συγκεκριμένου και επαληθεύσιμου υποσυνόλου της προδιαγραφής. Η εργασία θα καταγράψει επίσης με σαφήνεια ποια στοιχεία παραμένουν εκτός πεδίου για μια μελλοντική πλήρη υλοποίηση.

ΣΤΟΧΟΙ

1. Μελέτη της τρέχουσας προδιαγραφής RISC-V CHERI και του μοντέλου δυνατοτήτων (capability model), με έμφαση στις λειτουργίες που απαιτούν αλλαγές στον επεξεργαστή και στη μνήμη.
2. Επιλογή και υλοποίηση ενός συνεκτικού υποσυνόλου εντολών και κατάστασης CHERI-RISC-V στον gem5.
3. Υλοποίηση βασικής υποστήριξης για ετικέτες εγκυρότητας, ελέγχους ορίων/δικαιωμάτων και επιλεγμένες εξαιρέσεις.
4. Δημιουργία μικροπρογραμμάτων ελέγχου και επαλήθευση της συμπεριφοράς με τη βοήθεια CHERI LLVM/Clang και διαθέσιμων μοντέλων αναφοράς.
5. Ποσοτικοποίηση των μικροαρχιτεκτονικών επιπτώσεων και τεκμηρίωση των απαιτήσεων για μελλοντική επέκταση προς πληρέστερη υποστήριξη.

ΕΝΔΕΙΚΤΙΚΗ ΒΙΒΛΙΟΓΡΑΦΙΑ

- [1] RISC-V International, “RISC-V Specification for CHERI Extensions,” version v0.9.9, intermediate release 20260810, 2026, <https://riscv.github.io/riscv-cheri/0.9.9-ar20260810/riscv-cheri-full-0.9.9-ar20260810.html>.
- [2] LLVM Project, “LLVM Language Reference Manual – Non-Integral Pointer Types and Pointers with External State,” <https://llvm.org/docs/LangRef.html>.
- [3] CTSRD-CHERI, “CHERI LLVM/Clang Toolchain (llvm-project),” <https://github.com/CTSRD-CHERI/llvm-project>.
- [4] R. N. M. Watson, J. Woodruff, P. G. Neumann, S. W. Moore, J. Anderson, D. Chisnall, N. Dave, B. Davis, K. Gudka, B. Laurie, S. J. Murdoch, R. Norton, M. Roe, S. Son and M. Vadera, “CHERI: A Hybrid Capability-System Architecture for Scalable Software Compartmentalization,” in *2015 IEEE Symposium on Security and Privacy*, 2015, pp. 20–37, doi: 10.1109/SP.2015.9.
- [5] R. N. M. Watson, P. G. Neumann, J. Woodruff, M. Roe, H. Almatary, J. Anderson, J. Baldwin, G. Barnes, D. Chisnall, J. Clarke, B. Davis, L. Eisen, N. W. Filardo, F. A. Fuchs, R. Grisenthwaite, A. Joannou, B. Laurie, A. T. Markettos, S. W. Moore, S. J. Murdoch, K. Nienhuis, R. Norton, A. Richardson, P. Rugg, P. Sewell, S. Son and H. Xia, *Capability Hardware Enhanced RISC Instructions: CHERI Instruction-Set Architecture (Version 9)*, University of Cambridge Computer Laboratory, Tech. Rep. UCAM-CL-TR-987, 2023, doi: 10.48456/tr-987.
- [6] The gem5 Simulator, <https://www.gem5.org>

ΣΧΕΤΙΚΑ ΜΑΘΗΜΑΤΑ

- Αρχιτεκτονική Υπολογιστών (CEID_23Y203)
- Εργαστήριο Αρχιτεκτονικής Υπολογιστών (CEID_23Y212)
- Προχωρημένα Θέματα Αρχιτεκτονικής Υπολογιστών (CEID_NE4617)
- Προγραμματισμός Συστημάτων (CEID_25Y401)
- Ασφάλεια σε Υλικό (CEID_NE591)
- Λειτουργικά Συστήματα (CEID_24Y330)

ΕΠΙΒΛΕΠΩΝ

Γιώργος Παπαδημητρίου, Επίκουρος Καθηγητής

ΘΕΜΑ
ΔΙΠΛΩΜΑΤΙΚΗΣ

7

**Επιτάχυνση Πυρήνων Βιοπληροφορικής μέσω Συσχεδίασης
Αλγορίθμων και Αρχιτεκτονικής**

Accelerating Bioinformatics Kernels through Algorithm/Architecture Co-Design

Λέξεις κλειδιά: Bioinformatics · Sequence Alignment · Hardware Acceleration · Algorithm/Architecture Co-Design

ΠΕΡΙΓΡΑΦΗ / ΑΝΤΙΚΕΙΜΕΝΟ

Η ταχεία αύξηση του όγκου γονιδιωματικών δεδομένων έχει μεταφέρει ένα μεγάλο μέρος του υπολογιστικού κόστους από την ίδια την αλληλούχιση στην επεξεργασία και ανάλυση των δεδομένων. Πυρήνες όπως η αντιστοίχιση αναγνώσεων (read mapping), η στοίχιση ακολουθιών (sequence alignment), ο υπολογισμός απόστασης επεξεργασίας (edit distance) και τα φίλτρα πριν από τη στοίχιση (pre-alignment filters) περιλαμβάνουν συνδυασμό έντονης αριθμητικής επεξεργασίας, ακανόνιστων προσπελάσεων μνήμης και δυναμικού προγραμματισμού [1-4]. Για την ουσιαστική επιτάχυνσή τους δεν αρκεί πάντα η μεταφορά ενός υπάρχοντος αλγορίθμου σε περισσότερο παράλληλο υλικό, αλλά συχνά απαιτείται ταυτόχρονη αναδιαμόρφωση του αλγορίθμου και της οργάνωσης. Η εργασία θα επιλέξει έναν κύριο πυρήνα βιοπληροφορικής, με προτιμώμενη κατεύθυνση τη στοίχιση ακολουθιών ή έναν στενά συνδεδεμένο πυρήνα, και θα ξεκινήσει από μια σωστή υλοποίηση αναφοράς σε λογισμικό. Θα πραγματοποιηθεί χαρακτηρισμός της απόδοσης και της συμπεριφοράς μνήμης, ώστε να εντοπιστούν οι κυρίαρχοι περιορισμοί. Στη συνέχεια θα διερευνηθούν αλγοριθμικές μετασχηματίσεις όπως tiling, περιορισμός του ενεργού παραθύρου, επαναχρησιμοποίηση ενδιάμεσων αποτελεσμάτων ή αναδιοργάνωση της διαδικασίας οπισθοδρόμησης (traceback), με κριτήριο την καλύτερη χαρτογράφηση σε παράλληλες μονάδες και τοπική μνήμη. Στο αρχιτεκτονικό σκέλος θα αναπτυχθεί ένα οριοθετημένο μοντέλο επιτάχυνσης, είτε ως πρωτότυπο υψηλού επιπέδου (high-level synthesis, HLS) είτε ως αναλυτικό μοντέλο, ανάλογα με τον επιλεγμένο πυρήνα και το επίπεδο πολυπλοκότητας. Η αξιολόγηση θα συγκρίνει διαφορετικές επιλογές ως προς τη διεκπεραιωτική ικανότητα (throughput), τον λανθάνοντα χρόνο (latency), κίνηση μνήμης, απαιτούμενη τοπική αποθήκευση και, όπου είναι διαθέσιμα αξιόπιστα δεδομένα, κόστος πόρων ή ενεργειακή αποδοτικότητα. Στόχος είναι η εργασία να καταλήξει σε συγκεκριμένο παράδειγμα συσχεδίασης αλγορίθμου και αρχιτεκτονικής (algorithm/architecture co-design), στο οποίο η προτεινόμενη επιτάχυνση να προκύπτει από σαφή ανάλυση των ιδιοτήτων του προβλήματος και όχι απλώς από αύξηση του παραλληλισμού. Η τελική υλοποίηση θα πρέπει να είναι επαληθεύσιμη ως προς την ορθότητα και να συνοδεύεται από πλήρη πειραματική σύγκριση με τη βασική υλοποίηση λογισμικού.

ΣΤΟΧΟΙ

1. Επιλογή ενός αντιπροσωπευτικού πυρήνα βιοπληροφορικής και δημιουργία επαληθευμένης υλοποίησης αναφοράς.
2. Χαρακτηρισμός των υπολογιστικών και μνημονικών σημείων συμφόρησης και προσδιορισμός των τμημάτων που είναι κατάλληλα για εξειδικευμένη επιτάχυνση.
3. Σχεδίαση αλγοριθμικών μετασχηματισμών που αυξάνουν τον παραλληλισμό ή μειώνουν τις απαιτήσεις μνήμης χωρίς απώλεια ορθότητας.
4. Ανάπτυξη οριοθετημένου αρχιτεκτονικού πρωτοτύπου και μελέτη εναλλακτικών επιλογών οργάνωσης υπολογισμού και τοπικής αποθήκευσης.
5. Ποσοτική σύγκριση με τη βασική υλοποίηση ως προς απόδοση, κίνηση μνήμης, κλιμάκωση και κόστος υλοποίησης.

ΕΝΔΕΙΚΤΙΚΗ ΒΙΒΛΙΟΓΡΑΦΙΑ

- [1] H. Sadasivan, A. Klauser, J. Hench, Y. Turakhia, G. Singh, A. Zeni, S. Beecroft, S. Narayanasamy, J. Nivala, B. Robey, O. Mutlu, K. Denolf and S. Sitaraman, “The Genomic Computing Revolution: Defining the Next Decades of Accelerating Genomics,” in *2024 IEEE High Performance Extreme Computing Conference (HPEC)*, 2024, pp. 1–9, doi: 10.1109/HPEC62836.2024.10938492.
- [2] T. Robinson, J. Harkin and P. Shukla, “Hardware Acceleration of Genomics Data Analysis: Challenges and Opportunities,” *Bioinformatics*, vol. 37, no. 13, pp. 1785–1795, 2021, doi: 10.1093/bioinformatics/btab017.
- [3] S. Walia, C. Ye, A. Bera, D. Lodhavia and Y. Turakhia, “TALCO: Tiling Genome Sequence Alignment Using Convergence of Traceback Pointers,” in *2024 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, 2024, pp. 91–107, doi: 10.1109/HPCA57654.2024.00044.
- [4] D. Senol Cali, G. S. Kalsi, Z. Bingöl, C. Firtina, L. Subramanian, J. S. Kim, R. Ausavarungnirun, M. Alser, J. Gomez-Luna, A. Boroumand, A. Nori, A. Scibisz, S. Subramoney, C. Alkan, S. Ghose and O. Mutlu, “GenASM: A High-Performance, Low-Power Approximate String Matching Acceleration Framework for Genome Sequence Analysis,” in *Proc. 53rd IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 2020, doi: 10.1109/MICRO50266.2020.00081.

ΣΧΕΤΙΚΑ ΜΑΘΗΜΑΤΑ

- Αρχιτεκτονική Υπολογιστών (CEID_23Y203)
- Εργαστήριο Αρχιτεκτονικής Υπολογιστών (CEID_23Y212)
- Προχωρημένα Θέματα Αρχιτεκτονικής Υπολογιστών (CEID_NE4617)
- Προγραμματισμός Συστημάτων (CEID_25Y401)
- Εισαγωγή στη Βιοπληροφορική (CEID_NE548)
- Λειτουργικά Συστήματα (CEID_24Y330)

ΕΠΙΒΛΕΠΩΝ

Γιώργος Παπαδημητρίου, Επίκουρος Καθηγητής

ΘΕΜΑ
ΔΙΠΛΩΜΑΤΙΚΗΣ

8

Χαρακτηρισμός της Κλιμάκωσης και των Στενωμάτων στη Μετάφραση Διευθύνσεων σε GPUs

Characterizing Address-Translation Scaling and Bottlenecks in GPUs

Λέξεις κλειδιά: GPU Virtual Memory · Address Translation · WCPI · Page Divergence · gem5

ΠΕΡΙΓΡΑΦΗ / ΑΝΤΙΚΕΙΜΕΝΟ

Η εικονική μνήμη αποτελεί βασικό μηχανισμό των σύγχρονων μονάδων επεξεργασίας γραφικών (Graphics Processing Units, GPUs), όμως η υψηλή παραλληλία και το μοντέλο εκτέλεσης SIMT (Single Instruction, Multiple Threads) δημιουργούν ιδιαίτερα απαιτητική συμπεριφορά στη μετάφραση διευθύνσεων. Μετά τη συνένωση προσπελάσεων (memory coalescing), μια εντολή μνήμης ενός warp μπορεί να απαιτεί μεταφράσεις για πολλές διαφορετικές εικονικές σελίδες. Η απόκλιση σελίδων (page divergence) αυξάνει έτσι τον αριθμό των αιτημάτων μετάφρασης (translation requests) και μπορεί να προκαλέσει ριπές αστοχιών στις προσωρινές μνήμες μετάφρασης διευθύνσεων (Translation Lookaside Buffers, TLBs), συμφόρηση στους μηχανισμούς διάσχισης πινάκων σελίδων (Page Table Walkers, PTWs) και πίεση στα TLB Miss Status Holding Registers (MSHRs), τα οποία παρακολουθούν εκκρεμείς αστοχίες μετάφρασης [2,3]. Η εργασία θα μελετήσει πως αυτά τα φαινόμενα κλιμακώνονται με το αποτύπωμα μνήμης (memory footprint) και ποια στάδια καθορίζουν το συνολικό κόστος της μετάφρασης διευθύνσεων στην GPU. Η μεθοδολογία θα βασιστεί στον δείκτη Walk Cycles Per Instruction (WCPI) για την ποσοτικοποίηση του κόστους μετάφρασης [1] και θα τον προσαρμόσει στις GPUs. Θα διερευνηθεί ένα μοντέλο πέντε παραγόντων που εισάγει ρητά τον λόγο των αιτημάτων μετάφρασης προς τις προσπελάσεις ενός warp ως μέτρο του page divergence. Στον gem5 θα προστεθούν μετρήσεις για τις προσπελάσεις των warps, τα αιτήματα μετάφρασης μετά το coalescing, τις αστοχίες των TLBs, τη δραστηριότητα των PTWs, τις καθυστερήσεις στις ουρές και τα περιστατικά εξάντλησης των διαθέσιμων θέσεων για εκκρεμή αιτήματα μετάφρασης. Το GPU-WCPI θα συσχετιστεί με το πραγματικό κόστος της μετάφρασης διευθύνσεων και θα εξεταστεί αν αυτό ακολουθεί λογαριθμική κλιμάκωση ως προς το memory footprint σε φορτία εργασίας με κανονικά και μη κανονικά μοτίβα προσπέλασης. Για την απομόνωση των σημείων συμφόρησης θα χρησιμοποιηθούν ελεγχόμενες διαμορφώσεις, όπως μεγαλύτερες σελίδες, αυξημένη χωρητικότητα για εκκρεμή αιτήματα μετάφρασης ή αυξημένο εύρος εξυπηρέτησης των PTWs. Η ανάλυση θα εξετάσει την αλληλεπίδραση μεταξύ της χωρητικότητας των TLB MSHRs και των PTWs, καθώς και τον ανταγωνισμό μεταξύ δεδομένων και μεταδεδομένων μετάφρασης για τους πόρους της DRAM. Στόχος είναι ο χαρακτηρισμός των μικροαρχιτεκτονικών σημείων συμφόρησης της μετάφρασης διευθύνσεων στις GPUs στον gem5 και η εξαγωγή τεκμηριωμένων κατευθύνσεων για μελλοντικές βελτιστοποιήσεις. [3-5].

ΣΤΟΧΟΙ

1. Προσαρμογή του Walk Cycles Per Instruction (WCPI) στις GPUs με ενσωμάτωση του page divergence.
2. Επέκταση των μετρήσεων στον gem5 και συσχέτιση του GPU-WCPI με το πραγματικό address-translation overhead.
3. Μελέτη της κλιμάκωσης με το memory footprint και απομόνωση της συνεισφοράς των TLBs, MSHRs, PTWs και PWCs.
4. Ποσοτικοποίηση της αλληλεπίδρασης MSHR-PTW με κατάλληλες μετρικές ουρών, καθυστέρησης και χρήσης.
5. Διερεύνηση του ανταγωνισμού δεδομένων και μεταδεδομένων στη DRAM και εξαγωγή κατευθύνσεων για μελλοντικές βελτιστοποιήσεις.

ΕΝΔΕΙΚΤΙΚΗ ΒΙΒΛΙΟΓΡΑΦΙΑ

- [1] N. Lindsay and A. Bhattacharjee, “Understanding Address Translation Scaling Behaviours Using Hardware Performance Counters,” in *2024 IEEE International Symposium on Workload Characterization (IISWC)*, 2024, pp. 236–246, doi: 10.1109/IISWC63097.2024.00029.
- [2] J. Power, M. D. Hill and D. A. Wood, “Supporting x86-64 Address Translation for 100s of GPU Lanes,” in *2014 IEEE 20th International Symposium on High Performance Computer Architecture (HPCA)*, 2014, pp. 568–578, doi: 10.1109/HPCA.2014.6835965.
- [3] Y. Ha, J. Park, H. Cha, J. Lee, J. Kim, W. W. Ro and Y. Kim, “LATPC: Accelerating GPU Address Translation Using Locality-Aware TLB Prefetching and MSHR Compression,” in *Proc. 58th IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 2025, pp. 418–431, doi: 10.1145/3725843.3756069.
- [4] Y. Feng, Y. Li, J. Lee, W. W. Ro and H. Jeon, “Heliostat: Harnessing Ray Tracing Accelerators for Page Table Walks,” in *Proc. 52nd Annual IEEE/ACM International Symposium on Computer Architecture (ISCA)*, 2025, pp. 122–136, doi: 10.1145/3695053.3731011.
- [5] J. Power, J. Hestness, M. S. Orr, M. D. Hill and D. A. Wood, “gem5-gpu: A Heterogeneous CPU-GPU Simulator,” *IEEE Computer Architecture Letters*, vol. 14, no. 1, pp. 34–36, 2015, doi: 10.1109/LCA.2014.2299539.
- [6] The gem5 Simulator, <https://www.gem5.org>

ΣΧΕΤΙΚΑ ΜΑΘΗΜΑΤΑ

- Αρχιτεκτονική Υπολογιστών (CEID_23Y203)
- Εργαστήριο Αρχιτεκτονικής Υπολογιστών (CEID_23Y212)
- Προχωρημένα Θέματα Αρχιτεκτονικής Υπολογιστών (CEID_NE4617)
- Προγραμματισμός Συστημάτων (CEID_25Y401)
- Λειτουργικά Συστήματα (CEID_24Y330)

ΕΠΙΒΛΕΠΩΝ

Γιώργος Παπαδημητρίου, Επίκουρος Καθηγητής

ΘΕΜΑ
ΔΙΠΛΩΜΑΤΙΚΗΣ

9

Βελτίωση της Διαχείρισης TLB για εκ των Προτέρων Προσκόμιση Πέρα από τα Όρια Σελίδας

Improving TLB Management for Page-Crossing Prefetching

Λέξεις κλειδιά: Data Prefetching · TLB · Address Translation · Page Crossing

ΠΕΡΙΓΡΑΦΗ / ΑΝΤΙΚΕΙΜΕΝΟ

Η εκ των προτέρων προσκόμιση δεδομένων (data prefetching) αποτελεί βασική τεχνική απόκρυψης της καθυστέρησης μνήμης στους σύγχρονους επεξεργαστές. Προσκομιστές όπως ο Berti [1] και ο Best-Offset [2] προβλέπουν μελλοντικές προσπελάσεις από χωρικά ή τοπικά μοτίβα και μεταφέρουν τις αντίστοιχες γραμμές κρυφής μνήμης πριν ζητηθούν από τον επεξεργαστή. Η πρόβλεψη, όμως, γίνεται συνήθως στο επίπεδο διευθύνσεων δεδομένων, ενώ κάθε προσπέλαση πρέπει προηγουμένως να διαθέτει έγκυρη μετάφραση από εικονική σε φυσική διεύθυνση [3]. Το πρόβλημα γίνεται ιδιαίτερα ενδιαφέρον όταν μια ακολουθία προβλεπόμενων προσπελάσεων περνά το όριο μιας εικονικής σελίδας (page crossing). Σε αυτή την περίπτωση, ο prefetcher μπορεί να προβλέπει σωστά το επόμενο μπλοκ δεδομένων αλλά η αντίστοιχη μετάφραση να μην υπάρχει στην προσωρινή μνήμη μετάφρασης διευθύνσεων (Translation Lookaside Buffer, TLB). Η πρόοδος του prefetching μπορεί τότε να προκαλέσει επιπλέον αστοχίες TLB και επανειλημμένες παραπομπές του πινάκων σελίδων (page-table walks), ενώ η συντηρητική διακοπή στο όριο της σελίδας μπορεί να χάνει χρήσιμες ευκαιρίες prefetching. Η εργασία θα μελετήσει συστηματικά αυτόν τον συμβιβασμό στον gem5 [4]. Αρχικά θα υλοποιηθεί ή θα προσαρμοστεί ένας σύγχρονος data prefetcher, με βασικές αναφορές τον Best-Offset και τον Berti, και θα προστεθεί λεπτομερής καταγραφή των περιπτώσεων στις οποίες μια πρόβλεψη διασχίζει το όριο σελίδας. Θα μετρηθούν η ακρίβεια (accuracy), η κάλυψη (coverage) και η έγκαιρη άφιξη (timeliness) των prefetches, μαζί με τις αστοχίες TLB, τις παραπομπές των πινάκων σελίδων, την πίεση της μνήμης και την απόδοση. Η τελική αξιολόγηση θα αναζητήσει τον καλύτερο συμβιβασμό μεταξύ των επιδόσεων που επιτυγχάνονται από την έγκαιρη προσκόμιση και του πρόσθετου κόστους που δημιουργείται στο υποσύστημα μετάφρασης διευθύνσεων.

ΣΤΟΧΟΙ

1. Υλοποίηση ή προσαρμογή σύγχρονου data prefetcher στον gem5 και αξιολόγηση της βασικής του συμπεριφοράς.
2. Καταγραφή και χαρακτηρισμός των προσκομίσεων που ξεπερνούν τα όρια εικονικών σελίδων (page-crossing prefetches).
3. Ποσοτικοποίηση της αλληλεπίδρασης μεταξύ prefetching, TLB misses, page-table walks και τελικής απόδοσης.
4. Σχεδίαση μηχανισμού συντονισμού prefetcher-TLB.
5. Σύγκριση του προτεινόμενου μηχανισμού με πολιτικές που σταματούν ή επιτρέπουν ανεξέλεγκτα το prefetching στα όρια σελίδας.

ΕΝΔΕΙΚΤΙΚΗ ΒΙΒΛΙΟΓΡΑΦΙΑ

- [1] A. Navarro-Torres, B. Panda, J. Alastruey-Benedé, P. Ibáñez, V. Viñals-Yúfera and A. Ros, “Berti: An Accurate Local-Delta Data Prefetcher,” in *Proc. 55th IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 2022, pp. 975–991, doi: 10.1109/MICRO56248.2022.00072.
- [2] P. Michaud, “Best-Offset Hardware Prefetching,” in *2016 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, 2016, pp. 469–480, doi: 10.1109/HPCA.2016.7446087.
- [3] A. Bhattacharjee, “Appendix L: Advanced Concepts on Address Translation,” 2018, <https://www.cs.yale.edu/homes/abhishek/abhishek-appendix-l.pdf>.
- [4] The gem5 Simulator, <https://www.gem5.org>
- [5] J. L. Hennessy, D. A. Patterson, and C. Kozyrakis, *Computer Architecture: A Quantitative Approach*, 7th ed., Morgan Kaufmann, 2025.

ΣΧΕΤΙΚΑ ΜΑΘΗΜΑΤΑ

- Αρχιτεκτονική Υπολογιστών (CEID_23Y203)
- Εργαστήριο Αρχιτεκτονικής Υπολογιστών (CEID_23Y212)
- Προχωρημένα Θέματα Αρχιτεκτονικής Υπολογιστών (CEID_NE4617)
- Προγραμματισμός Συστημάτων (CEID_25Y401)
- Λειτουργικά Συστήματα (CEID_24Y330)

ΕΠΙΒΛΕΠΩΝ

Γιώργος Παπαδημητρίου, Επίκουρος Καθηγητής