

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

ΕΡΓΑΣΤΗΡΙΟ ΥΠΟΛΟΓΙΣΤΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

www.cslab.ece.ntua.gr

Διπλωματικές Εργασίες

Ακαδημαϊκό έτος 2026-2027

Περιεχόμενα

1 Αρχιτεκτονική Υπολογιστών

1.1 Συστήματα Μνημών και Ασφάλεια Intermittent IoT . . .
1.2 Ανάλυση και αντιμετώπιση της συμφόρησης της IOMMU σε εικονικοποιημένα συστήματα FPGA . . .
1.3 Coarse Grained Reconfigurable Arrays . . .
1.4 Μελέτη, ανάλυση και υλοποίηση βελτιώσεων σε επεξεργαστές RISC-V . . .
1.5 Αρχιτεκτονική Υπολογιστών και Μηχανική Μάθηση . . .
1.6 Μελέτη, Υλοποίηση και σύγκριση Κβαντικών Αλγορίθμων Μηχανικής Μάθησης και Κβαντικών Νευρωνικών Δικτύων (Bayessian) . . .

2 Λειτουργικά Συστήματα

2.1 Απευθείας μεταφορά KV Cache από τη GPU στο storage για LLMs . . .
2.2 Serverless/FaaS Infrastructure Evaluation & Optimization . . .
2.3 Συστήματα Διαμοιραζόμενης Μνήμης . . .
2.4 Confidential Computing . . .
2.5 Page Placement in Tiered-Memory Systems . . .
2.6 Μελέτη και Αξιοποίηση Μηχανισμών Page Fault Handling με χρήση eBPF . . .

3 Παράλληλα Συστήματα

3.1 Διαχείριση πόρων σε Συστήματα Υψηλής Επίδοσης (HPC) . . .
3.1.1 Συγχρονοδρομολόγηση εργασιών (Job Co-Scheduling) . . .
3.1.2 Εργασίες με ευέλικτο τρόπο ανάθεσης πόρων (Moldable Jobs) . . .
3.1.3 Σχεδιασμός και Υλοποίηση Υπηρεσίας AI Inference σε Υποδομές HPC . . .
3.2 Μελέτη μεταφοράς δεδομένων μεταξύ δομών διαφορετικής αρχιτεκτονικής . . .
3.3 Χαρακτηρισμός Εφαρμογών με χρήση micro-benchmarks . . .
3.4 Βελτιστοποίηση του υπολογιστικού πυρήνα πολλαπλασιασμού αραιού πίνακα με διάνυσμα (SpMV) . . .
3.5 Μελέτη της συμπεριφοράς αραιών υπολογιστικών πυρήνων (SpMM, SDDMM) σε αρχιτεκτονικές μηχανικής μάθησης . . .
3.6 Αξιολόγηση επίδοσης και βελτιστοποίηση υπολογιστικών πυρήνων σε υβριδική αρχιτεκτονική με διαμοιρασμό μνήμης μεταξύ CPU-GPU . . .
3.7 Πολυπαραμετρική μελέτη της συμπεριφοράς μεγάλων γλωσσικών μοντέλων μηχανικής μάθησης (LLM) σε σύγχρονες αρχιτεκτονικές . . .

4 Κατανεμημένα Συστήματα - Προχωρημένα θέματα βάσεων δεδομένων

4.1 Ανάπτυξη Chatbot για Βελτίωση Εφαρμογής Χρησιμοποιώντας Προηγμένες Τεχνικές Μηχανικής Μάθησης . . .
4.2 Χρήση Βάσεων Δεδομένων Γράφων για την Παρακολούθηση Γενεαλογίας (Lineage) Ερωτημάτων Ανάλυσης σε Σχεσιακά Δεδομένα . . .
4.3 Αποδεδειγμένα δίκαιη τυχαιότητα για παιχνίδια πολλών παικτών σε blockchain που προστατεύει την ιδιωτικότητα . . .

5 Εργασίες σε συνεπίβλεψη με εξωτερικούς συνεργάτες

5.1 Αυτόματη επιτάχυνση προγραμμάτων κελύφους και ροών εργασίας πρακτόρων τεχνητής νοημοσύνης . . . . . .

5.2 Από Μονολιθικές Εφαρμογές σε Microservices και Serverless, Αυτόματα . . . . . .

5.3 Edge Server Management for Vision-Language-Action Models . . . . . .

5.4 System-Level Performance Optimizations for Physical Artificial Intelligence . . . . . .

5.5 Parallelization Techniques for High-Throughput Retrieval-Augmented Generation . . .

5.6 CXLSim: A Flexible Simulator Framework for Compute Express Link Memory Disaggregation Research . . . . . .

1. Αρχιτεκτονική Υπολογιστών

1.1 Συστήματα Μνημών και Ασφάλεια Intermittent IoT

Τα Intermittent IoT (IIoT, Transient IoT, Ambient IoT, Batteryless IoT κλπ.) είναι low-power microcontrollers που λειτουργούν χωρίς μπαταρία, συλλέγοντας ενέργεια από το περιβάλλον τους (ηλιακή, αιολική, RF). Οι εφαρμογές τους ποικίλουν από αισθητήρες σε δύσβατα περιβάλλοντα, έως σε εφαρμογές στην ιατρική ή το διάστημα. Παραδείγματα τέτοιων συσκευών μπορούν να βρεθούν εδώ. Στα πλαίσια αυτής της διπλωματικής εργασίας προτείνεται η ενασχόληση και εξερεύνηση των παρακάτω θεμάτων:

- Hardware και hardware-software μηχανισμοί για συστήματα μνημών σε IIoT. Efficient caches και ανάλυση λειτουργίας caches σε IoT, Hardware-software codesign για βελτίωση απόδοσης cache. Μελέτη σχημάτων data cache από τη βιβλιογραφία [1] και σχετικής με μνήμες σε IIoT βιβλιογραφίας, π.χ. [2].
- Security and integrity of IIoT. Μελέτη hardware-software μηχανισμών για την ασφάλιση IIoT. Ενασχόληση με το μηχανισμό INCA του εργαστηρίου (παρόμοιος με [3,4]) και προέκτασή του για task-based IIoT.

Παραπομπές:

[1] Data Cache for Intermittent Computing Systems with Non-Volatile Main Memory

[2] DiCA: A Hardware-Software Co-Design for Differential Checkpointing in Intermittently Powered Devices

[3] CASU: Compromise Avoidance via Secure Update for Low-end Embedded Systems

[4] VRASED: A Verified Hardware/Software Co-Design for Remote Attestation

Σχετικά Μαθήματα: Αρχιτεκτονική Υπολογιστών, Προηγμένα Θέματα Αρχιτεκτονικής Υπολογιστών

Επικοινωνία: Φοίβος Ηλιάδης, filiadis@cslab.ece.ntua.gr
Παναγιώτης Μηλιάδης,pmiliad@cslab.ece.ntua.gr
Διονύσης Πνευματικάτος, pnevmati@cslab.ece.ntua.gr

1.2 Ανάλυση και αντιμετώπιση της συμφόρησης της IOMMU σε εικονικοποιημένα συστήματα FPGA

Η εικονικοποίηση επιτρέπει σε πολλαπλές εργασίες να μοιράζονται τους πόρους μιας FPGA, ενώ η IOMMU ελέγχει την πρόσβαση των επιταχυντών στη κύρια μνήμη του υπολογιστή. Όταν οι επιταχυντές προσπελαύνουν ταυτόχρονα διαφορετικές περιοχές της μνήμης, οι μεταφράσεις ενδέχεται να εκτοπίζονται από το IOTLB, προκαλώντας επιπλέον page walks, καθυστέρηση και παρεμβολή μεταξύ των εργασιών. Στόχος της διπλωματικής είναι να αναπαραγάγει αυτή τη συμπεριφορά σε ένα διαμοιραζόμενο περιβάλλον FPGA και να προσδιορίσει υπό ποιες συνθήκες η IOMMU περιορίζει την απόδοση. Μέσα από ελεγχόμενες παράλληλες προσπελάσεις DMA, ο φοιτητής θα μετρήσει την επίδραση του πλήθους των εργασιών και των ενεργών σελίδων μνήμης. Στη συνέχεια, θα αξιολογήσει τρόπους περιορισμού των καθυστερήσεων, όπως η χρήση μεγαλύτερων σελίδων ή η βελτίωση της τοπικότητας των προσπελάσεων.

Παραπομπές:

[1] Rethinking IOMMU for Future IO Devices
[2] Do OS abstractions make sense on FPGAs?
[3] Fast & Safe IO Memory Protection
[4] IOMMU Interference in Mixed Criticality Systems

Σχετικά Μαθήματα: Αρχιτεκτονική Υπολογιστών, Προηγμένα Θέματα Αρχιτεκτονικής Υπολογιστών

Επικοινωνία: Παναγιώτης Μηλιάδης, pmiliad@cslab.ece.ntua.gr
Διονύσης Πνευματικάτος, pnevmati@cslab.ece.ntua.gr

1.3 Coarse Grained Reconfigurable Arrays

Τα Coarse-Grained Reconfigurable Arrays (CGRAs) αποτελούν μια κλάση αναδιατασσόμενων αρχιτεκτονικών (Reconfigurable Architectures) που χαρακτηρίζονται από υψηλή ενεργειακή αποδοτικότητα, γρήγορους χρόνους επαναδιαμόρφωσης και ικανοποιητική ευελιξία [1], πληρώντας έτσι όλα τα απαραίτητα χαρακτηριστικά για να παίξουν σημαντικό ρόλο στην εποχή του AI [3]. Στη συγκεκριμένη θεματολογία θα ασχοληθείτε με μια CGRA [2], με σκοπό την, κατανόηση, μελέτη και επέκταση της.

Παραπομπές:
[1] A Survey of Coarse Grained Reconfigurable Architecture and Design: Taxonomy, Challenges, and Applications
[2] Neura: A Unified Framework for Hierarchical and Adaptive CGRAs
[3] Plasticine: A Reconfigurable Architecture For Parallel Paterns
[4] PICACHU: Plug-In CGRA Handling Upcoming Nonlinear Operations in LLMs
[5] Hardware abstractions and hardware mechanisms to support multi-task execution on Coarse Grained Reconfigurable Arrays

Σχετικά Μαθήματα: Αρχιτεκτονική Υπολογιστών, Προηγμένα Θέματα Αρχιτεκτονικής Υπολογιστών

Επικοινωνία: Αγαμέμνων Κυριαζής, akyriazis@cslab.ece.ntua.gr
Διονύσιος Πνευματικάτος, pnevmati@cslab.ece.ntua.gr

1.4 Μελέτη, ανάλυση και υλοποίηση βελτιώσεων σε επεξεργαστές RISC-V

Η RISC-V[1] αρχιτεκτονική είναι μια ανοικτή και επεκτάσιμη αρχιτεκτονική συνόλου εντολών που ξεκίνησε να αναπτύσσεται στο Πανεπιστήμιο του Berkeley το 2010 και από το 2016 λαμβάνει διεθνή προσοχή τόσο από τον ακαδημαϊκό χώρο όσο και από τον χώρο της βιομηχανίας, με κατάλληλη υποστήριξη σε όλα τα επίπεδα της υπολογιστικής στοίβας (υλικό, λειτουργικό σύστημα, βιβλιοθήκες, μεταγλωττιστές, κτλ). Ως ανοικτή και επεκτάσιμη αρχιτεκτονική προσφέρεται για την έρευνα σε λειτουργικές επεκτάσεις, ενώ πολλές υλοποιήσεις ανοικτού κώδικα είναι άμεσα διαθέσιμες, άλλες απλούστερες με γραμμική in-order pipeline και άλλες μεγαλύτερων επιδόσεων με πυρήνα εκτέλεσης εντολών εκτός σειράς (out-of-order).
Στην συγκεκριμένη θεματολογία θα ασχοληθείτε με αρχιτεκτονικό προσομοιωτή της επιλογής σας όπως ο gem5[2] η ο Firesim[3], με σκοπό την μελέτη συμπεριφοράς/επίδοσης ενός επεξεργαστή RISC-V. Τα θέματα τα οποία θα κληθείτε να μελετήσετε θα αφορούν την ιεραρχία κρυφών μνημών, το σύστημα διαχείρισης εικονικής μνήμης, την υλοποίηση/ανάλυση επεκτάσεων της αρχιτεκτονικής και άλλα.

Παραπομπές:
[1] RISC-V
[2] gem5

[3] FireSim

Σχετικά Μαθήματα: Προηγμένα Θέματα Αρχιτεκτονικής Υπολογιστών, Εργαστήριο Υπολογιστικών Συστημάτων

Επικοινωνία: Νίκος Χ. Παπαδόπουλος, ncpapad@cslab.ece.ntua.gr

1.5 Αρχιτεκτονική Υπολογιστών και Μηχανική Μάθηση

Αλγόριθμοι μηχανικής μάθησης ταξινόμησης (classification) και πρόβλεψης (prediction) εφαρμόζονται κατά κανόνα σε τομείς όπως η όραση υπολογιστών, η επεξεργασία φυσικής γλώσσας κ.α, πετυχαίνοντας εντυπωσιακά αποτελέσματα. Πλέον, έχουν αρχίσει και αυξάνονται οι περιπτώσεις εφαρμογής/χρήσης τους για τη βελτίωση της ίδιας της επίδοσης ενός υπολογιστικού συστήματος.

Πιθανές/Ενδεικτικές εργασίες:

- ***Εφαρμογή αλγορίθμων μηχανικής μάθησης για τη βελτιστοποίηση υπολογιστικών συστημάτων.*** Παραδείγματα αποτελούν βελτιστοποιήσεις στη χρήση των κρυφών μνημών (caches [1], prefetching [2]), στο μηχανισμό πρόβλεψης διακλαδώσεων (branch prediction), κ.α. αλλά πρόσφατα και στην ασφάλεια των συστημάτων.

  [1] Applying Deep Learning to the Cache Replacement Problem

  [2] TransforMAP: Transformer for Memory Access Prediction

  [3] Branch Prediction as a Reinforcement Learning Problem: Why, How and Case Studies

  [4] AutoCAT: Reinforcement Learning for Automated Exploration of Cache-Timing Attacks

- ***Εκτέλεση ML μοντέλων σε IoT devices.*** Τα τελευταία χρόνια έχει δημιουργηθεί ένα πλούσιο οικοσύστημα από μεγάλα, πολύπλοκα μοντέλα, τα οποία όμως έχουν τεράστιες απαιτήσεις τόσο σε υπολογιστικούς πόρους όσο και σε μνήμη. Στόχο της εργασίας αποτελεί η προσαρμογή και εκτέλεση τέτοιων μοντέλων σε επεξεργαστικά συστήματα περιορισμένων πόρων (ARM/RISC-V based IoT devices) [5, 6, 7, 8].

  [1] KWT-Tiny: RISC-V Accelerated, Embedded Keyword Spotting Transformer

  [2] TinyLlama: An Open-Source Small Language Model

  [3] bert-small

  [4] MLPerf: Tiny Deep Learning Benchmarks for Embedded Devices

Σχετικά Μαθήματα: Προηγμένα Θέματα Αρχιτεκτονικής Υπολογιστών

Επικοινωνία: Κωνσταντίνος Νίκας, knikas@cslab.ece.ntua.gr

1.6 Μελέτη, Υλοποίηση και σύγκριση Κβαντικών Αλγορίθμων Μηχανικής Μάθησης και Κβαντικών Νευρωνικών Δικτύων (Bayessian)

Το Quantum Computing[1] είναι εδώ για να μείνει. Ενώ η ενασχόληση με τον κλάδο για χρόνια περιοριζόταν σε θεωρητικό επίπεδο πλέον υλοποιήσεις Quantum Computers από κολοσσούς όπως η Google[2] και η IBM[3], δίνουν τη δυνατότητα πρακτικής εξέτασης κβαντικών αλγορίθμων σε πραγματικό χρόνο και τα αποτελέσματα αποδεικνύονται ολοένα και πιο υποσχόμενα.

Το Quantum Computing εκμεταλλεύεται αρχές της Κβαντομηχανικής (quantum superposition, interference, and entanglement)[4] και την πιθανοτικής της φύση (ένα σωματίδιο στους Κβαντικούς υπολογιστές πριν μετρηθεί δεν έχει μόνο δύο καταστάσεις που μπορεί να βρίσκεται αλλά άπειρες πάνω σε μια σφαίρα πιθανοτήτων -Bloch Sphere)) προκειμένου να παρουσιάσει εκθετική βελτίωση σε προβλήματα που μέχρι πρόσφατα βρίσκονταν στην NP κλάση, παρουσιάζοντας μια νέα κλάση πολυπλοκότητας, την BQP[5].

Στο εργαστήριο έχουμε μελετήσει Κβαντικούς αλγορίθμους μηχανικής μάθησης, ενώ έχουμε υλοποιήσει μια προσέγγιση ενός υβριδικού kmeans σε πραγματικό Κβαντικό υλικό, παρεχόμενο στο cloud της IBM. Συγκρίναμε τον υβριδικό kmeans με τον κλασσικό kmeans και βγάλαμε συμπεράσματα σχετικά με την αποδοτικότητα του ανοιχτού Quantum hardware για την ώρα.

Στόχος αυτής της εργασίας είναι η μελέτη των Κβταντικών Νευρωνικών Δικτύων[6]. Συγκεκριμένα θα ασχοληθούμε με τη μελέτη και την κατασκευή ενός Quantum Bayessian Network[7], όπου τα βάρη σε κάθε νευρώνα δεν είναι ντετερμινιστικά αλλά πιθανοτικά, για να εκμεταλλευτούμε την πιθανοτική φύση του Κβαντικού υπολογισμού.

Θα συγκρίνουμε τα Quantum νευρωνικά δίκτυα με τα αντίστοιχα κλασικά, θα προτείνουμε βελτιώσεις και θα βγάλουμε συμπεράσματα σχετικά με την υπάρχουσα αποδοτικότητα των ανοιχτών Κβαντικών cloud systems, προτείνοντας αλλαγές και βελτιώσεις.

Σχετικά Μαθήματα: Μηχανική μάθηση, Γραμμική Άλγεβρα, Αρχιτεκτονική Υπολογιστών

Επικοινωνία: Κωνσταντίνος Μπιτσάκος, kbitsak@cslab.ece.ntua.gr
Κωνσταντίνος Νίκας , knikas@cslab.ece.ntua.gr

2. Λειτουργικά Συστήματα

2.1 Απευθείας μεταφορά KV Cache από τη GPU στο storage για LLMs

Το long-context inference σε Μεγάλα Γλωσσικά Μοντέλα (LLMs) περιορίζεται συχνά από τη διαθέσιμη μνήμη για την αποθήκευση της KV cache [1]. Όταν οι απαιτήσεις σε μνήμη ξεπερνούν τα όρια της τοπικής μνήμης της GPU (HBM), το σύστημα αναγκάζεται να μεταφέρει τα KV blocks σε χαμηλότερα επίπεδα (Tiers) της ιεραρχίας μνήμης, στη μνήμη (Tier-1) της CPU ή στο δίσκο (Tier-2).
Στις αρχιτεκτονικές που στηρίζονται στο PCIe bus, η μεταφορά σε χαμηλότερα (π.χ. στο δίσκο) Tiers περνάει συνήθως από τα ενδιάμεσα (π.χ. DRAM της CPU) και αντιστρόφως, εισάγοντας επιβάρυνση (latency) και σπαταλώντας υπολογιστικούς πόρους. Για το λόγο αυτό έχουν αναπτυχθεί τεχνολογίες [2] οι οποίες επιτρέπουν την απευθείας μεταφορά δεδομένων από την HBM στο δίσκο, μετριάζοντας τις επιβαρύνσεις των ενδιάμεσων μεταφορών.
Η αρχιτεκτονική Nvidia GraceHopper (GH200) προσφέρει ένα μοντέλο ενοποιημένης μνήμης (unified memory) με hardware coherence, όπου η CPU και η GPU μοιράζονται κοινό page table και αντιμετωπίζονται από το λειτουργικό σύστημα ως ξεχωριστά NUMA nodes. Στόχος της διπλωματικής είναι ο σχεδιασμός και η υλοποίηση ενός συστήματος διαχείρισης της KV cache, το οποίο θα επιτρέπει την απευθείας μεταφορά των δεδομένων στο δίσκο αξιοποιώντας τα ιδιαίτερα χαρακτηριστικά της εν λόγω αρχιτεκτονικής, παρακάμπτοντας πλήρως τη DRAM της CPU.
Για τον σκοπό αυτό, θα διερευνηθούν τεχνολογίες όπως το Nvidia GDS (GPUDirect Storage) [2] και το io_uring [3] του Linux, αξιοποιώντας παράλληλα το LMCache [4,5] ως το framework που θα αναλαμβάνει τη μετακίνηση των KV cache blocks μεταξύ του vLLM [6] και των storage tiers.

Παραπομπές:

[1] https://huggingface.co/blog/not-lain/kv-caching
[2] https://docs.nvidia.com/gpudirect-storage/
[3] https://unixism.net/loti/
[4] https://docs.lmcache.ai/kv_cache/gds.html#gds-backend
[5] https://docs.lmcache.ai/mp/l2_storage/raw_block.html
[6] https://vllm.ai/blog/2023-06-20-vllm

Σχετικά Μαθήματα: Λειτουργικά Συστήματα, Εργαστήριο Λειτουργικών Συστημάτων

Επικοινωνία: Ορέστης Λάγκας Νικολός, olagkas@cslab.ece.ntua.gr
Παναγιώτης Μπάκος, pmpakos@cslab.ece.ntua.gr

2.2 Serverless/FaaS Infrastructure Evaluation & Optimization

Ενδεικτικές εργασίες:

- Optimized device-mapper snapshotter[3] implementation as Rust[4] crate
- Comparative study (design & performance evaluation): firecracker-containerd[5] vs Kata Containers[6]
- Performance evaluation & optimizations of Kubernetes-based FaaS stacks[1][2][6][7]

Παραπομπές:

[1] FaaSRail: Employing Real Workloads to Generate Representative Load for Serverless Research
[2] KNative
[3] Device Mapper storage Driver
[4] Rust Language
[5] Firecracker containerd
[6] Kata Containers
[7] vHive ecosystem

Σχετικά Μαθήματα: Λειτουργικά Συστήματα, Εργαστήριο Λειτουργικών Συστημάτων

Επικοινωνία: Χρήστος Κατσακιώρης, ckatsak@cslab.ece.ntua.gr
Κωνσταντίνος Νίκας, knikas@cslab.ece.ntua.gr

2.3 Συστήματα Διαμοιραζόμενης Μνήμης

Το CXL (Compute Express Link) είναι μια νέα τεχνολογία διασύνδεσης που επιτρέπει σε επεξεργαστές και επιταχυντές να μοιράζονται έναν ενιαίο χώρο μνήμης, με συναφείς εντολές προσπέλασης μνήμης (load/store), επεκτείνοντας την παραδοσιακή έννοια διαμοιρασμού πέρα από τα όρια ενός μόνο υπολογιστικού κόμβου σε υπολογιστικές πλέον συστοιχίες. Τα κατανεμημένα αυτά συστήματα αποκτούν ιδιότητες που θυμίζουν πολυεπεξεργαστές αλλά κάθε κόμβος συνεχίζει να τρέχει το δικό του λειτουργικό σύστημα και συνεπώς κατ αρχήν εφαρμόζονται προγραμματιστικά μοντέλα κατακερματισμού (π.χ. MPI). Απουσιάζουν οι διεπαφες λογισμικού που θα επιτρέψουν τον αποδοτικό προγραμματισμό τους. Στο πλαίσιο των διπλωματικών εργασιών που προτείνονται, οι φοιτητές θα μελετήσουν τον ρόλο του λογισμικού συστήματος (λειτουργικού, διαχειριστών μνήμης, μηχανισμών επικοινωνίας) και των διεπαφών του με προγραμματιστικά μοντέλα παραλληλισμού κοινής μνήμης (όπως OpenMP) σε κατανεμημένα περιβάλλοντα με αποσυζευγμένη, διαμοιραζόμενη μνήμη αμεσης προσπέλασης. Οι κύριες προκλήσεις περιλαμβάνουν την επεκταση βασικών μηχανισμών του ΛΣ, π.χ. συγχρονισμός, χρονοδρομολογηση, διαχείριση αρχειων κ.α., ωστε να αποκτήσουν επίγνωση της κατανεμημένης φύσης του νέου συστήματος (σε επιπεδο rack).

Παραπομπές:

1) https://dl.acm.org/doi/10.1145/3676641.3715988
2) https://arxiv.org/abs/2607.15569

Σχετικά Μαθήματα: Λειτουργικά Συστήματα, Εργαστήριο Λειτουργικών Συστημάτων

Επικοινωνία: Χλόη Αλβέρτη, xalverti@cslab.ece.ntua.gr
Στράτος Ψωμαδάκης, psomas@cslab.ece.ntua.gr

2.4 Confidential Computing

Οι σύγχρονες εφαρμογές νέφους, από συνεργαζόμενες μικροϋπηρεσίες (microservices) έως agents τεχνητής νοημοσύνης (AI agents) που χρησιμοποιούν εργαλεία και εκτελούν κώδικα, διαχειρίζονται συχνά ευαίσθητα δεδομένα σε κοινόχρηστες υποδομές. Οι εμπιστευτικές εικονικές μηχανές (Confidential VMs) αξιοποιούν μηχανισμούς υλικού για την προστασία του κώδικα και των δεδομένων κατά την εκτέλεση, ακόμη και από το priviledged λογισμικό διαχείρισης της υποδομής (πχ hypervisor). Στο πλαίσιο της διπλωματικής εργασίας θα εξερευνήσουμε την αποδοτική εκτέλεση microservices ή/και agentic workloads σε Confidential VMs, με έμφαση στον χρόνο εκκίνησης, στη διαχείριση μνήμης και στην ασφαλή μετανάστευση μεταξύ υπολογιστικών κόμβων. Ιδιαίτερο ενδιαφέρον παρουσιάζει η μεταφορά της κατάστασης εκτελούμενων υπηρεσιών, για παράδειγμα για συντήρηση της υποδομής ή

προσαρμογή σε νέες απαιτήσεις ασφάλειας, διατηρώντας την προστασία των δεδομένων και περιορίζοντας τη διακοπή λειτουργίας. Στόχος είναι η μελέτη και ανάπτυξη μηχανισμών λειτουργικού συστήματος και εικονικοποίησης για ασφαλείς και αποδοτικές υποδομές νέφους.

Παραπομπές:

1) https://arxiv.org/pdf/2605.03213 (Θα εστιάσουμε στο tool execution και όχι το GPU confidentiality)
2) https://dl.acm.org/doi/10.1145/3700418
3) https://arxiv.org/html/2609.04478v1
4) https://www.usenix.org/conference/nsdi26/presentation/sabanic

Επικοινωνία: Χλόη Αλβέρτη, xalverti@cslab.ece.ntua.gr

2.5 Page Placement in Tiered-Memory Systems

Τα σύγχρονα υπολογιστικά συστήματα υιοθετούν ολοένα και περισσότερο ετερογενείς αρχιτεκτονικές μνήμης (tiered-memory systems), οι οποίες συνδυάζουν διαφορετικές τεχνολογίες μνήμης, όπως DRAM και CXL (Compute Express Link), με διαφορετικά χαρακτηριστικά χωρητικότητας, latency και bandwidth. Η αξιοποίηση αυτών των αρχιτεκτονικών επιτρέπει την επέκταση της διαθέσιμης μνήμης με χαμηλότερο κόστος, αλλά εισάγει νέες προκλήσεις για το λειτουργικό σύστημα, καθώς η τοποθέτηση των δεδομένων στα διαφορετικά επίπεδα της ιεραρχίας επηρεάζει σημαντικά την επίδοση των εφαρμογών.

Οι υπάρχουσες πολιτικές τοποθέτησης σελίδων μνήμης βασίζονται συχνά στη συχνότητα προσπέλασής τους, θεωρώντας ότι οι συχνότερα προσπελάσιμες σελίδες πρέπει να βρίσκονται στην ταχύτερη μνήμη [4, 5, 6]. Ωστόσο, διαφορετικές προσπελάσεις μπορεί να έχουν διαφορετική επίδραση στην επίδοση, ανάλογα με τη δυνατότητα επικάλυψης των καθυστερήσεων μνήμης με την εκτέλεση άλλων εντολών [1, 2]. Επιπλέον, οι μετακινήσεις σελίδων μεταξύ επιπέδων μνήμης επιβαρύνουν το σύστημα και ενδέχεται να προκαλέσουν παρεμβολές στην εκτέλεση των εφαρμογών. Συνεπώς, η βέλτιστη τοποθέτηση δεδομένων απαιτεί την εκτίμηση τόσο του οφέλους από την πρόσβαση σε ταχύτερη μνήμη όσο και του κόστους μετακίνησής τους.

Στο πλαίσιο της διπλωματικής εργασίας, θα μελετηθούν και θα αξιολογηθούν σύγχρονες πολιτικές διαχείρισης ιεραρχικής μνήμης. Ιδιαίτερη έμφαση θα δοθεί στη διερεύνηση και αξιοποίηση μετρητών υλικού (hardware performance counters) για την παρακολούθηση της συμπεριφοράς των εφαρμογών και την εκτίμηση της επίδρασης των προσπελάσεων μνήμης στην επίδοση. Μέσω πειραματικής αξιολόγησης διαφορετικών κατηγοριών εφαρμογών και σεναρίων χρήσης, θα μελετηθεί η αποτελεσματικότητα πολιτικών τοποθέτησης και μετακίνησης σελίδων μεταξύ των επιπέδων μνήμης, με στόχο την αναγνώριση των περιορισμών τους και τη διερεύνηση πιθανών βελτιώσεων στο Linux.
Παραπομπές:

1) Performance Predictability in Heterogeneous Memory
2) PACT: A Criticality-First Design for Tiered Memory
3) M5
4) MEMTIS
5) HeMem

Σχετικά Μαθήματα: Λειτουργικά Συστήματα, Εργαστήριο Λειτουργικών Συστημάτων, Αρχιτεκτονική Υπολογιστών

Επικοινωνία: Κωνσταντίνος Μορές kmores@cslab.ece.ntua.gr

2.6 Μελέτη και Αξιοποίηση Μηχανισμών Page Fault Handling με χρήση eBPF

Η διαχείριση της εικονικής μνήμης αποτελεί βασική λειτουργία των σύγχρονων λειτουργικών συστημάτων, επηρεάζοντας σημαντικά την απόδοση των εφαρμογών. Ειδικότερα, η διαχείριση των page faults πραγματοποιείται παραδοσιακά από τον πυρήνα του λειτουργικού συστήματος, ενώ μηχανισμοί όπως το userfaultfd [4] επιτρέπουν στις εφαρμογές να υλοποιούν εξειδικευμένους page fault handlers στο userspace, εισάγοντας ωστόσο επιπλέον επιβαρύνσεις στην απόδοση.
Πρόσφατες ερευνητικές προσεγγίσεις, όπως το bpf_fault [1, 2], αξιοποιούν την τεχνολογία eBPF για την εκτέλεση εξειδικευμένων page fault handlers απευθείας στον Linux kernel, μειώνοντας το overhead ενώ παράλληλα, αναδεικνύουν τις δυνατότητες του προσαρμοσμένου χειρισμού page faults μέσα από εφαρμογές όπως η δημιουργία live snapshots εικονικών μηχανών (VM live snapshotting).
Στόχος της διπλωματικής εργασίας είναι η μελέτη και πειραματική αξιολόγηση σύγχρονων μηχανισμών διαχείρισης page faults στον Linux kernel, με έμφαση στις δυνατότητες που προσφέρει το eBPF. Στο πλαίσιο αυτό, θα εξοικειωθείτε με το framework bpf_fault, θα μελετήσετε την αρχιτεκτονική και τη λειτουργία του και θα αξιολογήσετε την απόδοσή του σε διαφορετικά workloads.
Με βάση τα αποτελέσματα της αρχικής μελέτης, θα διερευνηθούν πιθανές βελτιστοποιήσεις, επεκτάσεις ή νέες εφαρμογές των παραπάνω μηχανισμών, με έμφαση σε περιβάλλοντα εικονικοποίησης και σύγχρονες υπολογιστικές υποδομές, όπως το serverless computing.

Παραπομπές:

[1] Zussman et al., It's the Kernel's Fault! Custom Page Fault Handling With bpf_fault
[2] eBPF
[3] bpf_fault Github
[4] Userfaultfd documentation
[5] Benchmarking, Analysis, and Optimization of Serverless Function Snapshots
[6] Cloud Programming Simplified: A Berkeley View on Serverless Computing

Σχετικά Μαθήματα: Λειτουργικά Συστήματα, Εργαστήριο Λειτουργικών Συστημάτων

Επικοινωνία: Βασιλική Κωστούλα, vkostoula@cslab.ece.ntua.gr

3. Παράλληλα Συστήματα

3.1 Διαχείριση πόρων σε Συστήματα Υψηλής Επίδοσης (HPC)

Τα υπολογιστικά συστήματα υψηλής επίδοσης (High Performance Computing clusters) συχνά χρησιμοποιούνται για την επίλυση πολύπλοκων προβλημάτων από ποικίλες ερευνητικές περιοχές όπως η μοντελοποίηση των καιρικών φαινομένων καθώς και η διερεύνηση της ακολουθίας του ανθρώπινου γονιδιώματος. Η κατάλληλη χρονοδρομολόγηση των εργασιών σ' αυτά, αποτελεί καθοριστικό παράγοντα για την αποτελεσματική χρήση και την εξοικονόμηση ενέργειας. Το βασικό λογισμικό που συνθέτει ένα τέτοιο σύστημα είναι ο διαχειριστής πόρων (resource manager) με τον χρονοδρομολογητή εργασιών (job scheduler). Ο πρώτος κατανέμει τους πόρους, ενώ ο δεύτερος αποφασίζει πότε και πού θα εκτελούνται οι εργασίες. Κοινός παρονομαστής των ερευνητικών κατευθύνσεων σε αυτόν τον τομέα είναι η βελτίωση της λήψης αποφάσεων από τον χρονοδρομολογητή εργασιών.

3.1.1 Συγχρονοδρομολόγηση εργασιών (Job Co-Scheduling)

Μελέτες δείχνουν πως το co-execution, δηλαδή η εκτέλεση διαφορετικών εφαρμογών ταυτόχρονα στον ίδιο κόμβο, οδηγεί σε αποτελεσματικότερη χρήση των υπολογιστικών πόρων. Η επιλογή των εφαρμογών που θα εκτελεστούν μαζί παίζει σημαντικό ρόλο για την επίδοση της εκτέλεσης λόγω των race conditions που θα αναπτυχθούν ανάλογα με τους πόρους που ζητά η εκάστοτε εφαρμογή. Ο σκοπός των διπλωματικών εργασιών σε αυτόν τον τομέα είναι η μελέτη, προσομοίωση, υλοποίηση και αξιολόγηση αλγορίθμων συγχρονοδρομολόγησης (co-scheduling) με στόχο τη βελτιστοποίηση δεικτών όπως τη ρυθμαπόδοση του συστήματος (system throughput) και την ενεργειακή αποδοτικότητα. Στο πλαίσιο αυτό προτείνονται οι ακόλουθες ερευνητικές κατευθύνσεις.

Διπλωματικές εργασίες:

1. **Μελέτη της συγχρονοδρομολόγησης n MPI εφαρμογών (n-way co-scheduling) και των φαινομένων που παρατηρούνται στα υπερυπολογιστικά συστήματα ARIS και ICARUS**: Πειραματική μελέτη του n-way co-scheduling σε multi-node MPI εφαρμογές, με έμφαση στο *Straggler Effect*, όπου η επίδοση περιορίζεται από τα πιο επιβραδυμένα MPI ranks, και στο *Hammering Effect*, όπου η συνδυασμένη παρεμβολή πολλαπλών εφαρμογών προκαλεί μεγαλύτερη υποβάθμιση από αυτή που προβλέπεται από τις επιμέρους κατά ζεύγη συνεκτελέσεις.

2. **Μηχανισμοί απομόνωσης κρυφής μνήμης (cache) και εύρους ζώνης μνήμης (memory bandwidth) σε συνθήκες συγχρονοδρομολόγησης (co-scheduling) στην υπολογιστική υποδομή Grid5000**: Μελέτη τεχνικών απομόνωσης cache και memory bandwidth σε συνδυασμό με CPU co-scheduling, με στόχο τον περιορισμό των παρεμβολών χωρίς αποκλειστική ανάθεση πόρων (*compact*). Οι πολιτικές θα αξιολογηθούν πειραματικά σε εφαρμογές και workloads με χρήση του Flurm (`https://github.com/cslab-ntua/flurm`).

3. **Δημιουργία μοντέλου πρόβλεψης της απαιτούμενης απομόνωσης κρυφής μνήμης και εύρους ζώνης σε συνθήκες συγχρονοδρομολόγησης**: Πειραματική μελέτη της συμπεριφοράς της απομόνωσης της κρυφής μνήμης (cache) και του εύρους ζώνης μνήμης (memory bandwidth) σε συνθήκες συγχρονοδρομολόγησης (co-scheduling), με σκοπό την ανάπτυξη ενός αναλυτικού μοντέλου ή μοντέλου βασισμένου σε τεχνικές μηχανικής μάθησης, το οποίο θα μπορεί να

καθορίζει το κατάλληλο ποσοστό των διαθέσιμων πόρων που θα πρέπει να διατίθεται σε κάθε εργασία.

4. **Δημιουργία μοντέλου πρόβλεψης της επίδοσης εφαρμογών σε καθεστώς συγχρονοδρομολόγησης**: Αξιοποίηση αναλυτικών μεθόδων, τεχνικών μηχανικής μάθησης και υφιστάμενων μοντέλων επίδοσης (π.χ. Roofline Model, Cache-Aware Roofline Model) για τον χαρακτηρισμό και την ομαδοποίηση εφαρμογών, με στόχο την πρόβλεψη της συμπεριφοράς τους σε περιβάλλον συγχρονοδρομολόγησης.

3.1.2 Εργασίες με ευέλικτο τρόπο ανάθεσης πόρων (Moldable Jobs)

Οι εργασίες με ευέλικτη ανάθεση πόρων (moldable jobs) μπορούν να εκτελεστούν χρησιμοποιώντας διαφορετικούς πιθανούς αριθμούς διεργασιών, επιλεγόμενους από ένα σύνολο δυνατών τιμών που καθορίζει ο χρήστης. Έπειτα, ο χρονοδρομολογητής επιλέγει μία από αυτές τις πιθανές τιμές αριθμού διεργασιών για κάθε εργασία στην ουρά. Οι βασικές συνιστώσες του προβλήματος περιλαμβάνουν την πρόβλεψη της κλιμακωσιμότητας (scalability) των εργασιών και την ανάπτυξη αλγορίθμων βελτιστοποίησης για την επιλογή του κατάλληλου αριθμού διεργασιών ανά εργασία, με στόχο την ικανοποίηση ενός ή περισσότερων αντικειμενικών στόχων. Στο πλαίσιο αυτό προτείνονται οι ακόλουθες ερευνητικές κατευθύνσεις.

1. **Ανάπτυξη πολυκριτηριακού αλγορίθμου χρονοδρομολόγησης εργασιών με ευέλικτο τρόπο ανάθεσης πόρων (moldable jobs)**: Ο αλγόριθμος πρέπει να λαμβάνει υπόψη τόσο την κλιμακωσιμότητα (scalability) όσο και την ενεργειακή κατανάλωση των εργασιών. Στόχος είναι η βελτίωση της συνολικής επίδοσης του συστήματος, επιτυγχάνοντας παράλληλα αυξημένη ενεργειακή αποδοτικότητα.

2. **Μελέτη και πρόβλεψη της κλιμακωσιμότητας (scalability) εφαρμογών HPC**: Η ερευνητική αυτή κατεύθυνση εστιάζει στην ανάπτυξη ενός αναλυτικού ή βασισμένου σε τεχνικές μηχανικής μάθησης μοντέλου, το οποίο θα μπορεί να προβλέπει την κλιμακωσιμότητα HPC εργασιών αξιοποιώντας μετρικές επίδοσης που συλλέγονται από την εκτέλεση της εργασίας σε έναν ή, κατ' ανώτατο όριο, δύο διαφορετικούς αριθμούς διεργασιών.

Σχετικά Μαθήματα: Συστήματα Παράλληλης Επεξεργασίας

Επικοινωνία: Νίκος Τριανταφύλλης, ntriantafyl@cslab.ece.ntua.gr
Θάνος Τσουκλείδης - Καρυδάκης, ttsoukl@cslab.ece.ntua.gr
Αλέξανδρος Χαριτάτος, aharit@cslab.ece.ntua.gr
Γιώργος Γκούμας, goumas@cslab.ece.ntua.gr

3.1.3 Σχεδιασμός και Υλοποίηση Υπηρεσίας AI Inference σε Υποδομές HPC

Η αυξανόμενη χρήση μοντέλων Τεχνητής Νοημοσύνης (AI), και ιδιαίτερα Μεγάλων Γλωσσικών Μοντέλων (LLMs), δημιουργεί την ανάγκη για υπολογιστικές υποδομές που υποστηρίζουν τη συνεχή εκτέλεση μοντέλων και την παροχή υπηρεσιών inference. Σε αντίθεση με τις παραδοσιακές εργασίες HPC, οι υπηρεσίες αυτές απαιτούν μόνιμα περιβάλλοντα εκτέλεσης, πρόσβαση σε GPUs και δικτυακή συνδεσιμότητα.

Στόχος της διπλωματικής εργασίας είναι ο σχεδιασμός και η υλοποίηση μιας πλατφόρμας AI inference, αξιοποιώντας το Incus [1] για τη διαχείριση containers και το Waldur [2] για την αυτοματοποιημένη παροχή υπολογιστικών πόρων.

Στο πλαίσιο της εργασίας θα αναπτυχθεί λογισμικό ανοικτού κώδικα (open source) ως μηχανισμός διασύνδεσης του Waldur Site Agent με το Incus, επιτρέποντας τη δημιουργία και διαχείριση GPU-enabled containers, την κατανομή CPU, μνήμης και αποθηκευτικού χώρου, καθώς και την ασφαλή πρόσβαση σε υπηρεσίες inference μέσω APIs, με χρήση εργαλείων όπως το vLLM [3].

Παραπομπές:

[1] Incus Documentation
[2] Waldur Site Agent
[3] vLLM – Efficient LLM Inference and Serving

Σχετικά Μαθήματα: Συστήματα Παράλληλης Επεξεργασίας

Επικοινωνία: Νίκος Τριανταφύλλης, ntriantafyl@cslab.ece.ntua.gr
Διονύσης Πνευματικάτος, pnevmati@cslab.ece.ntua.gr
Γιώργος Γκούμας, goumas@cslab.ece.ntua.gr

3.2 Μελέτη μεταφοράς δεδομένων μεταξύ δομών διαφορετικής αρχιτεκτονικής

Στο πλαίσιο αυτής την διπλωματικής εργασίας, θα ελέγξουμε μηχανισμούς δομών δεδομένων ως προς την απόδοσή τους και την απόδοση μεταφοράς δεδομένων από και προς αυτούς σε μεγάλες κλίμακες μεγέθους. Πιο συγκεκριμένα, εξετάζουμε το πόσο κοστίζει μια αλλαγή δομής δεδομένων κατά την διάρκεια εκτέλεσης μιας εφαρμογής, ανάλογα με το μέγεθος δεδομένων και το πρότυπο αλληλεπιδράσεων με την ιεραρχία μνήμης. Κάποιες δομές που θα εξεταστούν πρώτα είναι δομές δέντρων τύπου AVL, B-Tree και πιθανώς άλλες, και θα δοκιμαστούν τεχνικές άμεσης μεταφοράς και έμμεσης μεταφοράς χρησιμοποιώντας αρχεία καταγραφής. Οι οποιεσδήποτε υλοποιήσεις αλγορίθμων θα γίνουν με την χρήση C/C++ σε περιβάλλον linux.

Σχετικά Μαθήματα: Συστήματα Παράλληλης Επεξεργασίας

Επικοινωνία: Δημήτρης Γιαννόπουλος, dimian@cslab.ece.ntua.gr

3.3 Χαρακτηρισμός Εφαρμογών με χρήση micro-benchmarks

Καθώς η εκτέλεση πολλών τύπων υπηρεσιών μεταφέρεται σε συστήματα μεγάλης κλίμακας, η πρόκληση της διατήρησης υψηλής ποιότητας υπηρεσίας συνεχώς μεγαλώνει. Η απουσία αποδοτικών λύσεων διαμοιρασμού των κοινόχρηστων πόρων οδηγεί τους Cloud Service Providers στην απομόνωση ολόκληρων servers για την εκτέλεση εφαρμογών με αυστηρούς περιορισμούς για την επίδοση τους. Αυτό οδηγεί στην υποχρησιμοποίηση αυτών των πόρων και την αύξηση του λειτουργικού κόστους. Για την αντιμετώπιση των ζητημάτων αυτών προτείνονται τεχνικές χαρακτηρισμού των εφαρμογών ως προς τους κρίσιμους πόρους με σκοπό τη συνεκτέλεση εφαρμογών με συμπληρωματικές απαιτήσεις για πόρους. Σκοπός της διπλωματικής είναι η ανάπτυξη ενός μηχανισμού που θα προβλέπει τις ανάγκες των εφαρμογών από άποψης πόρων με χρήση micro-benchmarks που, στερώντας πόρους από την κάθε εφαρμογή, θα αποκαλύπτουν τις απαιτήσεις της.

Σχετικά Μαθήματα: Συστήματα Παράλληλης Επεξεργασίας

Επικοινωνία: Γιάννης Παπαδάκης, ypap@cslab.ece.ntua.gr

3.4 Βελτιστοποίηση του υπολογιστικού πυρήνα πολλαπλασιασμού αραιού πίνακα με διάνυσμα (SpMV)

Μελέτη παράλληλων υλοποιήσεων του SpMV πυρήνα σε αρχιτεκτονικές CPU (Intel, AMD, ARM) ή GPU (Nvidia). Τα προγραμματιστικά μοντέλα που μπορούν να χρησιμοποιηθούν είναι το OpenMP για τις CPUs και η CUDA για τις GPUs, αλλά και όποια άλλη προτίμηση υπάρχει. Η πραγματοποίηση της διπλωματικής μπορεί να έχει διάφορες προσεγγίσεις, για παράδειγμα:

- βελτίωση ήδη υπάρχουσας υλοποίησης του πυρήνα
- μελέτη της αναπαράστασης της δομής του αραιού πίνακα (συντεταγμένες των μη μηδενικών τιμών)
- συμπίεση των τιμών του πίνακα και μελέτη της επίδρασης lossy μεθόδων συμπίεσης

Σχετικά Μαθήματα: Συστήματα Παράλληλης Επεξεργασίας

Επικοινωνία: Γιώργος Γκούμας, goumas@cslab.ece.ntua.gr
Δημήτριος Γαλανόπουλος, dgal@cslab.ece.ntua.gr

3.5 Μελέτη της συμπεριφοράς αραιών υπολογιστικών πυρήνων (SpMM, SDDMM) σε αρχιτεκτονικές μηχανικής μάθησης

Τα συστήματα μηχανικής μάθησης και ειδικότερα οι τεχνικές βελτιστοποίησης μηχανικης μαθησης έχουν γνωρίσει ραγδαία ανάπτυξη τα τελευταία χρόνια. Στον πυρήνα των εφαρμογών βαθιάς μάθησης βρίσκονται οι πολλαπλασιασμοί πινάκων οι οποίοι λόγω του όγκου των περιττών βαρών μπορούν να καταστούν αραιοί. Στόχος της προτεινόμενης διπλωματικής εργασίας είναι η μελέτη των χαρακτηριστικών αραιών πινάκων οι οποίοι προέρχονται από εφαρμογές βαθιάς μηχανικής μάθησης και η αξιολόγηση της επίδοσης βασικών αραιών υπολογιστικών πυρήνων (SpMM, SDDMM) με στόχο τον εντοπισμό των κύριων προβλημάτων υπάρχουσων υλοποιήσεων και την πρόταση βελτιώσεων.

Σχετικά Μαθήματα: Συστήματα Παράλληλης Επεξεργασίας

Επικοινωνία: Ιωάννα Τάσου, itasou@cslab.ece.ntua.gr
Παναγιώτης Μπάκος, pmpakos@cslab.ece.ntua.gr
Γιώργος Γκούμας, goumas@cslab.ece.ntua.gr

3.6 Αξιολόγηση επίδοσης και βελτιστοποίηση υπολογιστικών πυρήνων σε υβριδική αρχιτεκτονική με διαμοιρασμό μνήμης μεταξύ CPU-GPU

Το σύστημα Grace Hopper της NVIDIA αποτελεί μία καινοτόμα αρχιτεκτονική προσέγγιση για την εκτέλεση υπολογιστικά απαιτητικών εφαρμογών από το πεδίο της τεχνητής νοημοσύνης. Περιλαμβάνει CPU και GPU για πρώτη φορά συνδεδεμένα σε κοινή μνήμη υποστηρίζοντας ταυτόχρονα και συνάφεια κρυφής μνήμης. Στο πλαίσιο της προτεινόμενης διπλωματικής εργασίας θα πραγματοποιηθούν λεπτομερείς μετρήσεις επίδοσης του συγκεριμένου συστήματος προκειμένου να γίνουν κατανοητά τα ιδιαίτερα χαρακτηριστικά της αρχιτεκτονικής και να αναδειχθούν τα δυνατά και αδύναμα σημεία της.

Στη συνέχεια, θα επιλεγούν κρίσιμοι υπολογιστικοί πυρήνες, θα σχεδιαστεί και θα υλοποιηθεί η υβριδική απεικόνισή τους στη συγκεκριμένη αρχιτεκτονική με στόχο της βελτιστοποίηση της επίδοσης και της ενεργειακής κατανάλωσης.

Σχετικά Μαθήματα: Συστήματα Παράλληλης Επεξεργασίας, Προηγμένα Θέματα Αρχιτεκτονικής Υπολογιστών

Επικοινωνία: Γιώργος Γκούμας, goumas@cslab.ece.ntua.gr
Παναγιώτης Μπάκος, pmpakos@cslab.ece.ntua.gr

3.7 Πολυπαραμετρική μελέτη της συμπεριφοράς μεγάλων γλωσσικών μοντέλων μηχανικής μάθησης (LLM) σε σύγχρονες αρχιτεκτονικές

Η εκτέλεση μεγάλων γλωσσικών μοντέλων (LLM) σε σύγχρονες αρχιτεκτονικές (CPUs και GPUs) αποτελεί μία από τις πιο σημαντικές εφαρμογές στα σύγχρονα υπολογιστικά συστήματα μεγάλης κλίμακας. Η συγκεκριμένη διπλωματική εργασία εστιάζει στο υπολογιστικά απαιτητικό στάδιο του prefil και σκοπεύει να μελετήσει ολιστικά τη συμπεριφορά της εκτέλεσης στοχεύοντας σε τρεις κρίσιμες παραμέτρους: επίδοση (π.χ. latency ή Time To First Token - TTFT), κατανάλωση ενέργειας (π.χ. Energy To First Token, GFLOPS/Watt) και ακρίβεια. Στο πλαίσιο της εργασίας θα μελετηθούν διαφορετικοί τρόποι εκτέλεσης των μοντέλων με διαφορετικά επίπεδα κβαντισμού (quantization) και αραιοποίησης (sparsification) και θα εξαχθούν συμπεράσματα σχετικά με τη βέλτιστη λειτουργία τους ανάλογα με το περιβάλλον εκτέλεσης και τη μετρική στόχου.

Σχετικά Μαθήματα: Συστήματα Παράλληλης Επεξεργασίας

Επικοινωνία: Αρετή Μέη, amei@cslab.ece.ntua.gr
Γιώργος Γκούμας, goumas@cslab.ece.ntua.gr

4. Κατανεμημένα Συστήματα - Προχωρημένα θέματα βάσεων δεδομένων

4.1 Ανάπτυξη Chatbot για Βελτίωση Εφαρμογής Χρησιμοποιώντας Προηγμένες Τεχνικές Μηχανικής Μάθησης

Με τη θεαματική πρόοδο στον τομέα της τεχνητής νοημοσύνης και της μηχανικής μάθησης, τα chatbots έχουν γίνει απαραίτητο κομμάτι των σύγχρονων εφαρμογών για τη διευκόλυνση της διάδρασης με τους χρήστες. Η εργασία αυτή στοχεύει στο σχεδιασμό και την ανάπτυξη ενός εξελιγμένου βοηθού chatbot στα ελληνικά, που βελτιώνει την εμπειρία χρήστη μιας εφαρμογής παραγωγής γραφημάτων, ενσωματώνοντας προηγμένες τεχνικές μηχανικής μάθησης όπως μεγάλα γλωσσικά μοντέλα (LLMs) [1,5] ή μεθόδους ανάκτησης (πχ word embeddings [2], transformers [3,6], deep averaging networks [4] κλπ).
Στο πλαίσιο της διπλωματικής: (α) θα μελετηθούν σε επίπεδο βιβλιογραφικό σύγχρονες μεθοδολογίες για την ανάπτυξη chatbots, εστιάζοντας σε LLMs ή/και μεθόδους ανάκτησης και θα γίνει η επιλογή της πιο κατάλληλης, (β) θα σχεδιαστεί και θα υλοποιηθεί ml pipeline που θα περιλαμβάνει όλα τα στάδια από την προετοιμασία των δεδομένων μέχρι την εκπαίδευση του ML μοντέλου και την παραγωγή αποτελεσμάτων, (γ) θα υλοποιηθεί απλό interface αλληλεπίδρασης με τους χρήστες και (δ) θα αποτιμηθεί η επίδοση του συστήματος ως προς μετρικές όπως η ακρίβεια των αποτελεσμάτων, ο χρόνος απόκρισης, η ικανοποίηση των χρηστών κ.α.

Παραπομπές:

[1] Meltemi
[2] Understanding Word2Vec: A Beginner's Guide to Word Embeddings
[3] Sentence Transformers
[4] Deep Averaging network in Universal sentence encoder
[5] Meltemi-7B-Instruct-v1.5
[6] bert-base-greek-uncased-v1

Σχετικά Μαθήματα: Κατανεμημένα Συστήματα

Επικοινωνία: Κατερίνα Δόκα, katerina@cslab.ece.ntua.gr

4.2 Χρήση Βάσεων Δεδομένων Γράφων για την Παρακολούθηση Γενεαλογίας (Lineage) Ερωτημάτων Ανάλυσης σε Σχεσιακά Δεδομένα

Το query lineage, που αφορά την ιχνηλάτηση της ιστορίας των δεδομένων που παράγονται από την εκτέλεση ερωτημάτων, είναι κρίσιμη για τη βελτίωση της κατανόησης των ροών ανάλυσης δεδομένων, την εύρεση σφαλμάτων σε αυτές και τη βελτιστοποίησή τους. Ο κύριος στόχος αυτής της διπλωματικής είναι η αξιοποίηση των δομικών πλεονεκτημάτων των graph databases, όπως το Neo4j [1], για την αναπαράσταση και ανάλυση των περίπλοκων σχέσεων και εξαρτήσεων που εμπεριέχονται στις διαδικασίες εκτέλεσης ερωτημάτων.

Στο πλαίσιο της διπλωματικής θα: (α) σχεδιαστεί και υλοποιηθεί ένα σύστημα αποθήκευσης δεδομένων lineage βασισμένο πάνω σε τεχνολογίες βάσεων γράφων (graph databases), (β) θα αναπτυχθούν αλγόριθμοι για την εξαγωγή, τη μετατροπή και τη φόρτωση πληροφοριών lineage στην graph database, (γ) θα αξιολογηθεί η απόδοση του συστήματος με χρήση δεδομένων και ερωτημάτων ανάλυσης που παρέχονται από γνωστά benchmarks (πχ, TPC-H [2]).

Παραπομπές:

[1] Neo4j
[2] TPC-H

Σχετικά Μαθήματα: Βάσεις Δεδομένων

Επικοινωνία: Κατερίνα Δόκα, katerina@cslab.ece.ntua.gr

4.3 Αποδεδειγμένα δίκαιη τυχαιότητα για παιχνίδια πολλών παικτών σε blockchain που προστατεύει την ιδιωτικότητα

Τα παιχνίδια πολλών παικτών με ποντάρισμα (staked) χρειάζονται τυχαιότητα, την οποία κανένας συμμετέχων δεν μπορεί να προβλέψει ή να επηρεάσει ούτε καν ο διαχειριστής που φιλοξενεί το παιχνίδι. Το Midnight είναι ένα blockchain του οποίου τα έξυπνα συμβόλαια (smart contracts) εκτελούνται με αποδείξεις μηδενικής γνώσης (zero-knowledge proofs), καθιστώντας εφικτούς διάφορους τρόπους επίτευξης αυτού του στόχου. Ένας είναι μέσω τυφλής επαληθεύσιμη τυχαίας συνάρτησης (Blinded VRF), όπου κάθε παίκτης υποβάλλει ένα ερώτημα που μόνο ο ίδιος μπορεί να διαμορφώσει, και ο διαχειριστής απαντά σε αυτό με ένα δεσμευμένο κλειδί (committed key) και μια απόδειξη. Ένας δεύτερος τρόπος μπορεί να είναι η πλήρη κατάργηση του διαχειριστή, όσον αφορά την συμμετοχή του στην τυχαιότητα, με το κάθε τυχαίο αποτέλεσμα να προκύπτει από μυστικά random seeds που οι παίκτες έχουν δεσμεύσει εκ των προτέρων να αποκαλύψουν σε ένα καθορισμένο χρονικό σημείο όπου γίνεται και ο έλεγχός τους με μηδενική γνώση (zero-knowledge proof).
Σε οποιοδήποτε από τα δύο σχήματα, ένας διαχειριστής μπορεί να εισπράξει τα τέλη συναλλαγών (transaction fees) των παικτών χωρίς να μπορεί να τροποποιήσει τις κινήσεις τους, ενώ κάθε ολοκληρωμένο παιχνίδι μπορεί να επανεπαληθευτεί αποκλειστικά από το δημόσιο καθολικό (public ledger).
Κανένα από τα δύο μοντέλα δεν είναι τέλειο. Το καθένα αφήνει περιθώρια για κάποια μορφή συμπαιγνίας (collusion) ή στρατηγικής απόκρυψης πληροφοριών (strategic withholding). Επιπλέον, κάθε επιπλέον δικλείδα ασφαλείας αυξάνει το μέγεθος της απόδειξης και τον χώρο on-chain, τον οποίο το δίκτυο περιορίζει αυστηρά.
Η διπλωματική εργασία θα εξερευνήσει αυτό το πεδίο ελεύθερα: θα πειραματιστεί με τα υπάρχοντα σχήματα, θα εντοπίσει τις αδυναμίες τους και θα προτείνει, θα κατασκευάσει και θα αξιολογήσει βελτιώσεις ή νέες προσεγγίσεις σχεδιασμένες από τον ίδιο τον φοιτητή. Ένα λειτουργικό σύνολο συμβολαίων, μια υπηρεσία διαχειριστή (operator service) και ένας ελεγκτής (verifier) είναι διαθέσιμα ως σημείο εκκίνησης.

Σχετικά Μαθήματα: Κατανεμημένα Συστήματα

Επικοινωνία: Τάσος Κατσιγιάννης, tkats@cslab.ece.ntua.gr
Κατερίνα Δόκα, katerina@cslab.ece.ntua.gr

5. Εργασίες σε συνεπίβλεψη με εξωτερικούς συνεργάτες

5.1 Αυτόματη επιτάχυνση προγραμμάτων κελύφους και ροών εργασίας πρακτόρων τεχνητής νοημοσύνης

Οι πράκτορες τεχνητής νοημοσύνης (AI agents) χρησιμοποιούνται ολοένα και περισσότερο για την αυτοματοποίηση εργασιών, αξιοποιώντας κυρίως το κέλυφος (shell) του λειτουργικού συστήματος για να εκτελούν εντολές και να αλληλεπιδρούν με το υπολογιστικό τους περιβάλλον. Πρόσφατες μελέτες χρήσης εργαλείων όπως το Claude και Codex δείχνουν ότι το κέλυφος είναι το συχνότερα χρησιμοποιούμενο εργαλείο, αντιπροσωπεύοντας περίπου το 57% τόσο των κλήσεων εργαλείων όσο και του συνολικού χρόνου χρήσης τους. Η σύνθεση εργαλείων δημιουργεί ευκαιρίες επιτάχυνσης. Ο στόχος διπλωματικών εργασιών σε αυτή την περιοχή είναι (1) η αυτόματη επιτάχυνση με εκτέλεση εργασιών παράλληλα, κατανεμημένα και εκτός σειράς, (2) η επαναχρησιμοποίηση προηγούμενων αποτελεσμάτων και η αποφυγή περιττών επανεκτελέσεων, (3) η προσαρμογή της εκτέλεσης με βάση τις διαθέσιμες πληροφορίες και τους υπολογιστικούς πόρους, (4) η βελτιστοποίηση του χρονοπρογραμματισμού και της διαχείρισης δεδομένων μεταξύ εργαλείων, και (5) η αξιολόγηση αυτών των τεχνικών σε πραγματικές εφαρμογές και περιβάλλοντα νέφους, με στόχο τη μείωση του χρόνου εκτέλεσης και του υπολογιστικού κόστους, διατηρώντας την ορθότητα των αποτελεσμάτων.

Σχετικές Δημοσιεύσεις:

[1] Speculative Script Reordering at Subprocess Granularity (OSDI'26)

[2] Faster Re-Execution via Bolt-On Incrementalization (OSDI'26)

[3] Fractal: Fault-Tolerant Shell-Script Distribution (NSDI'26)

Σχετικά Μαθήματα: Λειτουργικά Συστήματα, Συστήματα Παράλληλης Επεξεργασίας, Κατανεμημένα Συστήματα, Μεταγλωττιστές

Επικοινωνία: Γιώργος Γκούμας, goumas@cslab.ece.ntua.gr

Νίκος Βασιλάκης, nikos@vasilak.is, Brown University, Μετασχηματιστικά Συστήματα

5.2 Από Μονολιθικές Εφαρμογές σε Microservices και Serverless, Αυτόματα

Νέες μορφές παράλληλων και κατανεμημένων εφαρμογών, όπως ο υπολογισμός με microservices και serverless, έχουν γίνει εξαιρετικά δημοφιλείς διότι επιτρέπουν διαφορετικά τμήματα ή υπηρεσίες της εφαρμογής να αναπτύσσονται, να συντηρούνται, και κυρίως να κλιμακώνονται ανεξάρτητα από την υπόλοιπη εφαρμογή. Δυστυχώς όμως μια παραδοσιακή "μονολιθική" εφαρμογή γραμμένη π.χ. σε Python ή JavaScript, ειδικά όταν αυτή χρησιμοποιεί εκτενώς βιβλιοθήκες τρίτων, απαιτεί εξαιρετικό κόπο και χρόνο για να μετασχηματιστεί με τρόπο οποίος να αξιοποιεί τέτοια περιβάλλοντα — και ακόμη και τότε μπορεί να καταλήξει να υπολειτουργεί σε επιδόσεις ή ορθότητα.

Ο στόχος διπλωματικών εργασιών σε αυτή την περιοχή είναι (1) η αυτόματη ανάλυση παραδοσιακών εφαρμογών και ο αυτόματος μετασχηματισμός τους σε microservices ή serverless (2) η βελτιστοποίηση επιδόσεων τους αξιοποιώντας ευριστικές, στατιστικές, και αναλυτικές μεθόδους, (3) η χρήση αναλυτικών μεθόδων και μαθηματικών μοντέλων για περεταίρω βελτιστοποίηση, (4) η δημιουργία βιβλιοθη-

κών για την υποστήριξη της ορθής τους εκτέλεσης (π.χ. recoverability logic), παρά τους περιορισμούς του εκάστοτε περιβάλλοντος, και (5) η αξιολόγηση αυτών τεχνικών και συστημάτων σε πραγματικά περιβάλλοντα microservices ή serverless (π.χ., AWS Lambda).

Σχετικά Μαθήματα: Λειτουργικά Συστήματα, Συστήματα Παράλληλης Επεξεργασίας

Επικοινωνία: Γιώργος Γκούμας, goumas@cslab.ece.ntua.gr
Νίκος Βασιλάκης, nikos@vasilak.is, Brown University

5.3 Edge Server Management for Vision-Language-Action Models

Vision-Language-Action (VLA) models are at the forefront of physical artificial intelligence and have emerged as state-of-the-art for robot manipulation tasks, including object relocation, sorting, packing, and assembling. A VLA model takes as input camera images and a text prompt, and outputs an action chunk for the robot to execute, making VLA models foundational to robot AI factories.

However, strict weight, cost, and power constraints prevent robots from integrating high-end GPUs onboard (e.g., A100, H100), and robots instead rely on low-area, low-cost embedded GPUs such as the NVIDIA Jetson series. Recent works show that these embedded GPUs cannot meet the strict latency guarantees VLA inference requires: increased inference latency delays the action output, causing task failures, unsafe robot navigation, or broader disruption such as halted assembly lines. This motivates a local edge server within the robot AI factory that serves VLA inference for many robots by integrating one or more high-end GPUs (e.g., A100, H100), hosting multiple VLA models, and communicating with robots over low-latency wireless networks to receive requests and return action chunks.

The goal of this thesis is to develop an efficient management system and architecture for such a local edge server. The thesis will cover three directions: (i) a deep characterization of VLA models and their runtime characteristics, (ii) an analysis of the needs and constraints of local edge servers in a robot AI factory setting, and (iii) architectural guidelines and a management system for allocating, replicating, and/or disaggregating VLA models across the GPUs of a local edge server. The proposed system will be evaluated across multiple VLA models and edge server architectures, targeting performance, energy efficiency, and cost.

This diploma thesis has the possibility of being conducted as part of an internship at the Max Planck Institute for Software Systems (MPI-SWS) in Germany.

References:

[1] Dionysios Adamopoulos, Nattapol Chanpaisit, Basel Fakhri, Christina Giannoula, “Efficient Vision-Language-Action Management and Serving for Robot Factories”, arXiv 2026

[2] Wenqi Jiang, Jason Clemons, Rowland O’Flaherty, Hugo Hadfield, Alperen Degirmenci, Shuran Song, Yashraj Narang, Christos Kozyrakis, “ROSA: A Robotics Foundation Model Serving System for Robot Factories”, arXiv 2026.

[3] Yinwei Dai, Ganesh Ananthanarayanan, Landon Cox, Xenofon Foukas, Bozidar Radunovic, and Ravi Netravali, “Kairos: A Scalable Serving System for Physical AI”, arXiv 2026.

[4] Dionysios Adamopoulos, Anastasia Poulopoulou, Georgios Goumas, Christina Giannoula, “Spira: Exploiting Voxel Data Structural Properties for Efficient Sparse Convolution in Point Cloud Networks”, MLSys 2026.

Επικοινωνία: Χριστίνα Γιαννούλα, christina.giann@gmail.com

5.4 System-Level Performance Optimizations for Physical Artificial Intelligence

Physical AI applications in robotics and autonomous driving and real-world perception systems rely on sophisticated Machine Learning (ML) models that must process complex sensory data in real-time with stringent latency and safety requirements. These systems employ diverse ML model architectures specifically designed for real-world environments, including 3D point cloud networks, vision transformers, semantic segmentation networks, simultaneous localization and mapping (SLAM) algorithms, and perception-action transformers that directly map sensory inputs to robotic control outputs. However, the unique computational characteristics of these models, including irregular memory access patterns, multi-modal data processing, temporal dependencies, and variable-sized inputs, combined with the real-time constraints of physical environments, create distinct performance bottlenecks that differ significantly from traditional ML workloads. While extensive research has focused on optimizing dense transformers and Large Language Models (LLMs), there has been limited exploration of systems optimizations specifically tailored for physical AI models used in robotics and autonomous driving executed on modern GPUs, thereby leaving significant performance potential untapped.

The goal of this thesis project is to develop effective systems optimization techniques specifically tailored for the unique computational characteristics of physical AI workloads. Such optimization techniques include algorithmic optimizations, intelligent pipelining strategies, software-based prefetching mechanisms, adaptive caching policies, computation-communication overlap techniques, fine-grained parallelization strategies, and automated tuning frameworks that adapt to specific hardware configurations and real-time characteristics. The proposed runtime optimization engine will be comprehensively evaluated across diverse physical AI workloads, using representative datasets and benchmarks and running on various modern GPU architectures, to demonstrate the broad applicability and effectiveness of systems-level optimizations for next-generation physical AI deployments.

This diploma thesis has the possibility of being conducted as part of an internship at the Max Planck Institute for Software Systems (MPI-SWS) in Germany.

References:

[1] Dionysios Adamopoulos, Anastasia Poulopoulou, Georgios Goumas, Christina Giannoula, “Spira: Exploiting Voxel Data Structural Properties for Efficient Sparse Convolution in Point Cloud Networks”, MLSys 2026.

[2] Jiacheng Yang, Christina Giannoula, Jun Wu, Mostafa Elhoushi, James Gleeson, Gennady Pekhimenko, “Minuet: Accelerating 3D Sparse Convolutions on GPUs”, EuroSys 2024.

[3] Jiacheng Yang, Jun Wu, Zhen Zhang, Xinwei Fu, Zhiying Xu, Zhen Jia, Yida Wang, Gennady Pekhimenko, ”ScaleFusion: Scalable Inference of Spatial-Temporal Diffusion Transformers for High-Resolution Long Video Generation”, MLSys 2025.

[4] Qidong Su, Wei Zhao, Xin Li, Muralidhar Andoorveedu, Chenhao Jiang, Zhanda Zhu, Kevin Song, Christina Giannoula, Gennady Pekhimenko, ”Seesaw: High-throughput LLM Inference via Model Re-Sharding”, MLSys 2025.

[5] Ke Hong, Zhongming Yu, Guohao Dai, Xinhao Yang, Yaoxiu Lian, Zehao Liu, Ningyi Xu, Yuhan Dong, Yu Wang, “Exploiting Hardware Utilization and Adaptive Dataflow for Efficient Sparse Convolution in 3D Point Clouds”, MLSys 2024.

Επικοινωνία: Χριστίνα Γιαννούλα, christina.giann@gmail.com

5.5 Parallelization Techniques for High-Throughput Retrieval-Augmented Generation

Large Language Models (LLMs) have made important advances in natural language processing and generation. However, LLMs rely on pretrained data and face data staleness: they generate factually incorrect information because that data grows outdated. Retrieval-Augmented Generation (RAG) addresses this limitation by combining an external knowledge base of up-to-date documents with the LLM, retrieving relevant documents and passages to improve the credibility of generation. Specifically, a RAG system retrieves the data most relevant to a user query from the knowledge base and passes it to the LLM, which leverages this data to produce high-quality output.

Existing systems run retrieval on CPUs and dedicate GPUs to LLM inference for final output generation. This design has two limitations: (i) it assumes the external knowledge base fits in CPU DRAM, and (ii) it restricts retrieval to CPU cores alone. In practice, knowledge bases can be far too large for CPU DRAM and must reside on external storage (i.e., disk). Moreover, GPUs can achieve much faster retrieval than CPU cores when the RAG algorithm is tailored to GPU architectural characteristics.

The goal of this thesis is to explore efficient parallelization techniques for RAG with large knowledge databases, leveraging both CPU and GPU accelerators in the end-to-end RAG pipeline, with the aim of achieving high-throughput generation across many concurrent user queries. These techniques include CPU-GPU co-execution and pipelining, software prefetching, parallel RAG algorithms, and query batching within the RAG inference pipeline. The proposed techniques will be evaluated on large knowledge databases across diverse query distributions and benchmarks, demonstrating robustness under realistic execution settings.

References:

[1] Wenqi Jiang, Marco Zeller, Roger Waleffe, Torsten Hoefler, Gustavo Alonso, “Chameleon: a Heterogeneous and Disaggregated Accelerator System for Retrieval-Augmented Language Models”, VLDB 2025.

[2] Chien-Yu Lin, Keisuke Kamahori, Yiyu Liu, Xiaoxiang Shi, Madhav Kashyap, Yile Gu, Rulin Shao, Zihao Ye, Kan Zhu, Rohan Kadekodi, Stephanie Wang, Arvind Krishnamurthy, Luis Ceze, Baris Kasikci, “TeleRAG: Efficient Retrieval-Augmented Generation Inference with Lookahead Retrieval”, MLSys 2026.

[3] Chao Liu, Haifeng Liu, Dan Chen, Yu Huang, Yi Zhang, Wen-Jing Xiao, Xiaofei Liao, Hai Jin, ”HeterRAG: Heterogeneous Processing-in-Memory Acceleration for Retrieval-augmented Generation”, ISCA 2025.

[4] Qidong Su, Wei Zhao, Xin Li, Muralidhar Andoorveedu, Chenhao Jiang, Zhanda Zhu, Kevin Song, Christina Giannoula, Gennady Pekhimenko, ”Seesaw: High-throughput LLM Inference via Model Re-Sharding”, MLSys 2025.

Επικοινωνία: Χριστίνα Γιαννούλα, christina.giann@gmail.com

5.6 CXLSim: A Flexible Simulator Framework for Compute Express Link Memory Disaggregation Research

Compute Express Link (CXL) technology can tackle the memory capacity problem by enabling large pools of memory to be accessed remotely at high speed. As CXL is expected to be commercialized soon and achieve widespread adoption, recent research has increasingly focused on optimizing performance to leverage the benefits of CXL networks and memory disaggregation. Various prior works have designed application-specific accelerators and near-data processing solutions upon CXL-based disaggregated systems, demonstrating the technology’s potential for enhancing computational efficiency and resource utilization.

These research efforts highlight the critical need for a configurable, easy-to-use CXL simulator that can enable further research and improvements not only in CXL characteristics themselves, but also in accelerators based on CXL for important applications as well as CPU/GPU architecture designs that leverage CXL infrastructure.

Currently, existing approaches to CXL evaluation rely primarily on real-system emulation. These approaches include either (i) using a CPU NUMA system where applications run locally on one NUMA node while other nodes serve as remote memory pools with increased access latency, or (ii) FPGA-based emulation, where there is an FPGA connected to a CPU and the FPGA memory serves as the remote memory pool, while the hardware logic functions as a controller/accelerator. Although real-system emulation provides valuable insights, a dedicated cycle-accurate simulator (similar to Ramulator) would be significantly more beneficial for collecting low-level statistics, enabling detailed research and exploring additional CXL designs, particularly at the hardware level. In contrast to CXL emulations on real systems that have limited configuration options and constrained experimental parameters, a cycle-accurate CXL simulator would provide extensive configurability and enable deeper hardware-level exploration of design trade-offs and optimizations. The goal of this thesis is to design a flexible and configurable CXL simulator, potentially based on the Ramulator, that can be seamlessly integrated with existing processor-centric simulators such as gem5, Sniper, and GPGPU-Sim, thereby providing the research community with a powerful tool for advancing CXL-based system design and optimization. The interface and configuration can allow cycle-accurate simulation of NUMA CPU systems and CXL and Near-Data-Processing (CXL+NDP) systems.

This diploma thesis has the possibility of being conducted as part of an internship at the Max Planck Institute for Software Systems (MPI-SWS) in Germany.

References:

[1] Seoyoung Ko, Hyunjeong Shim, Wanju Doh, Sungmin Yun, Jinin So, Yongsuk Kwon, Sang-Soo Park, Si-Dong Roh, Minyong Yoon, Taeksang Song, Jung Ho Ahn, "COSMOS: A CXL-Based Full In-Memory System for Approximate Nearest Neighbor Search", IEEE CAL 2025.

[2] Yan Sun, Yifan Yuan, Zeduo Yu, Reese Kuper, Chihun Song, Jinghan Huang, Houxiang Ji, Siddharth Agarwal, Jiaqi Lou, Ipoom Jeong, Ren Wang, Jung Ho Ahn, Tianyin Xu, Nam Sung Kim, "Demystifying CXL Memory with Genuine CXL-Ready Systems and Devices", MICRO 2023.

[3] Yoongu Kim; Weikun Yang; Onur Mutlu, "Ramulator: A Fast and Extensible DRAM Simulator", IEEE CAL 2016.

[4] Yujie Yang, Lingfeng Xiang, Peiran Du, Zhen Lin, Weishu Deng, Ren Wang, Andrey Kudryavtsev, Louis Ko, Hui Lu, Jia Rao, "Architectural and System Implications of CXL-enabled Tiered Memory", arXiv 2025.

Επικοινωνία: Χριστίνα Γιαννούλα, christina.giann@gmail.com