ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ
ΕΡΓΑΣΤΗΡΙΟ ΥΠΟΛΟΓΙΣΤΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ
http://www.cslab.ece.ntua.gr

Διπλωματικές Εργασίες
Ακαδημαϊκό έτος 2024-2025

I. Αρχιτεκτονική Υπολογιστών

1 Confidential Computing σε IoT και FPGA

Το Confidential Computing (ή Trust Computing) είναι ένα σύνολο μηχανισμών που εγγυόνται την ασφάλεια και την ακεραιότητα του software και του hardware. Στα πλαίσια αυτής της διπλωματικής προτείνεται η ενασχόληση με το CC σε 2 συγγενικά περιβάλλοντα:

- Trust σε FPGAs [1, 2]: Υλοποίηση custom block που ενισχύουν την ασφάλεια σε FPGAs. Εξερεύνηση και μελέτη υπάρχουσων λύσεων. Χρήση επιταχυντών κρυπτογραφικών αλγορίθμων σε FPGAs, security co-processors για ασφαλείς υπολογισμούς, συστήματα μνήμης κλπ. Integration με πιό περίπλοκα συστήματα. Εκτίμηση designs επιταχυντών πρωτοκόλλων Remote Attestation (RA) ή Swarm Remote Attestation (SRA). Trust / Confidential Computing σε περιβάλλον cloud, σε mutli-tenant περιβάλλοντα κλπ.

- Swarm Remote Attestation (SRA) σε IoT [3, 4] : Στο Internet of Things το Remote Attestation ανάγεται σε τεράστια δίκτυα συσκευών που εισάγουν νέες προκλήσεις στη διαδικασία του RA. Εξερεύνηση, μελέτη και σύγκριση υπάρχουσων λύσεων, υλοποίηση SRA σε πραγματικό υλικό, υλοποίηση SRA για συσκευές αρχιτεκτονικής RISC-V, εφαρμογή SRA λύσεων για FPGAs στο cloud.

Παραπομπές:
[1] F. Turan and I. Verbauwhede. “Trust in FPGA-accelerated cloud computing.”
[2] P. Rosero-Montalvo et al. “A survey of trusted computing solutions using FPGAS.”
[3] A. Sprogø Banks, M. Kisiel, and P. Korsholm. “Remote attestation: A literature review.”
[4] T. Yagawa et al. “Delegating Verification for Remote Attestation Using TEE.”

Σχετικά Μαθήματα: Αρχιτεκτονική Υπολογιστών, Προηγμένα Θέματα Αρχιτεκτονικής Υπολογιστών

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

2 Εικονικοποίηση και διαμοιρασμός πόρων συστημάτων FPGA

Οι FPGAs κερδίζουν όλο και περισσότερο έδαφος στο υπολογιστικό νέφος, χάρη στην επαναπρογραμματιζόμενη φύση τους και την ικανότητά τους να υποστηρίζουν επιταχυντές υψηλής απόδοσης. Ωστόσο, η εικονικοποίηση και ο διαμοιρασμός των πόρων μιας FPGA δεν έχουν μελετηθεί εκτενώς, γεγονός που οδηγεί στη μερική αξιοποίηση των πόρων, καθώς συνήθως μόνο ένας χρήστης έχει πρόσβαση σε αυτή. Στόχος της διπλωματικής είναι η διερεύνηση τρόπων διαμοιρασμού των πόρων μεταξύ χρηστών, η διασφάλιση της απομόνωσης και ασφάλειας μεταξύ τους, καθώς και η ανάπτυξη τεχνικών εικονικοποίησης που θα επιτρέπουν στους χρήστες να θεωρούν ότι είναι οι μόνοι που χρησιμοποιούν την πλατφόρμα.

Σχετικά Μαθήματα: Αρχιτεκτονική Υπολογιστών, Προηγμένα Θέματα Αρχιτεκτονικής Υπολογιστών

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

3 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

4 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

5 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

II. Λειτουργικά Συστήματα

6 Τεχνικές βελτιστοποίησης εικονικής μνήμης

Στο πλαίσιο της προτεινόμενης διπλωματικής θα μελετηθούν οι μηχανισμοί εικονικής μνήμης (virtual memory) και σελιδοποίησης (paging), με σκοπό την βελτιστοποίηση της αλληλεπίδρασης του ΛΣ (Linux) με το υλικό εικονικής μνήμης – το MMU (Memory Management Unit) και το TLB (Translation Lookaside Buffers). Συγκεκριμένα, θα δοθεί έμφαση σε νέες δυνατότητες του υλικού εικονικής μνήμης που παρέχουν οι αρχιτεκτονικές ARM καI RISC-V, καθώς και στο πώς μπορούν να αξιοποιηθούν πολυεπίπεδες (multi-tiered) αρχιτεκτονικές μνημών (DRAM, PMEM, CXL) ώστε να ελαχιστοποιηθεί το κόστος της εικονικής μνήμης.

Σχετικά Μαθήματα: Λειτουργικά Συστήματα, Εργαστήριο Λειτουργικών Συστημάτων, Αρχιτεκτονική Υπολογιστών, Προηγμένα Θέματα Αρχιτεκτονικής Υπολογιστών

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

7 Operating System principles for Serverless Systems

Το μοντέλο του Serverless Computing [1] αποτελεί μια σχετικά νέα προσέγγιση στο σχεδιασμό των υπολογιστικών υποδομών των σύγχρονων εφαρμογών και την αποδοτική διαχείριση των υπολογιστικών τους πόρων (CPU, μνήμη, δίκτυο). Τα κύρια χαρακτηριστικά της προσέγγισης αυτής είναι η μετακύλιση της ευθύνης διαχείρισης των πόρων των εφαρμογών από το χρήστη προς τον πάροχο των υπολογιστικών υποδομών (Amazon AWS, Microsoft Azure κλπ) και ο αγνωστικισμός για τον "οικοδεσπότη" (host) που φιλοξενεί την εκτέλεση των σχετικών προγραμμάτων. Οι εφαρμογές φαινομενικά - για τον χρήστη - δεν ανήκουν σε κάποιο συγκεκριμένο server (server- less). Το μοντέλο αυτό αφενός διευκολύνει τους χρήστες απλοποιώντας την διαδικασία ανάπτυξης μιας εφαρμογής - ο χρήστης είναι υπεύθυνος μονο για τον κώδικα της εφαρμογής του - και αφετέρου απελευθερώνει τους παρόχους ώστε να διαχειρίζονται τους υπολογιστικούς πόρους που διαθέτουν με ακόμη πιο αποδοτικούς τρόπους.
Στο πλαίσιο αυτό οι παραδοσιακές προσεγγίσεις που τα τελευταία χρόνια είχαν κυριεύσει στο σχεδιασμό των νεφοϋπολογιστικών συστημάτων σταδιακά προσαρμόζονται στις νέες ανάγκες που γεννά το Serverless. Στην περιοχή ενδιαφέροντος που αναφερόμαστε (Λειτουργικά Συστήματα για πλατφόρμες Serverless) περιλαμβάνεται η μελέτη, αξιοποίηση και προσαρμογή βασικών αρχών και μηχανισμών των Λειτουργικών Συστημάτων όπως η εικονικοποίηση (virtualization) μέσω hardware (KVM [3]) και software μηχανισμών (cgroups, namespaces), η διαχείριση μνήμης (memory allocation) καθώς και η αποθήκευση δεδομένων (storage) για συστήματα Serverless [6] [7]. Παράλληλα σε μια πιο ολιστική προσέγγιση, μελετάμε τη συσχέτιση τέτοιων συστημάτων με συγγενείς περιοχές, αναζητώντας και αξιολογώντας την επιρροή των αρχών σχεδίασης Serverless συστημάτων στη διασύνδεση με απομακρυσμένες υπηρεσίες αποθήκευσης (Βάσεις Δεδομένων) και δικτύου [4].
Η εκπόνηση Διπλωματικής Εργασίας σε αυτή την περιοχή ενδιαφέροντος περιλαμβάνει (ενδεικτικά):

- τη μελέτη θεωρητικού υπόβαθρου (cloud computing [5], serverless [1])
- την πρακτική εξοικείωση με τεχνολογίες: hypervisor (πχ QEMU, Firecracker, Cloud Hypervisor), containers (πχ containerd, kata-containers), Linux / kernel (virtio, memory management, sockets, cgroups, namespaces)
- την αξιοποίηση γνώσεων από τα προπτυχιακά μαθήματα "Λειτουργικά Συστήματα" και "Εργαστήριο Λειτουργικών Συστημάτων" (semaphores, sockets, descriptors [2])

- την εξοικείωση με γλώσσες προγραμματισμού (πχ C, Rust, Python)
- την ενδεχόμενη πειραματική αξιολόγηση σε πλατφόρμες παρόχων cloud υπηρεσιών (Amazon AWS).

Παραπομπές:

[1] Cloud Programming Simplified: A Berkeley View on Serverless Computing
[2] sendmsg(), recvmsg(), SCM_RIGHTS
[3] KVM
[4] Connection pooling
[5] Above the Clouds: A Berkeley View of Cloud Computing
[6] Deverlay: Container Snapshots For Virtual Machines
[7] Less Boot is better than cold: Scaling out by scaling up

Σχετικά Μαθήματα: Λειτουργικά Συστήματα, Εργαστήριο Λειτουργικών Συστημάτων

Επικοινωνία: Ορέστης Λάγκας Νικολός, olagkas@cslab.ece.ntua.gr
Στράτος Ψωμαδάκης, psomas@cslab.ece.ntua.gr

8 Serverless/FaaS Infrastructure Evaluation & Optimization

Ενδεικτικές εργασίες:

- Integration of FaaSRail[1] with Knative[2]
- 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[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

III. Παράλληλα Συστήματα

9 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

10 Ανάθεση πόρων σε υπερυπολογιστικά περιβάλλοντα

Τα υπερυπολογιστικά συστήματα περιλαμβάνουν εκατοντάδες ή και χιλιάδες υπολογιστικούς κόμβους που με τη σειρά του περιλαμβάνουν πολυπύρηνους επεξεργαστές γενικούς σκοπού (CPUs) και επιταχυντές (τυπικά GPUs). Η ανάθεση πόρων σε εργασίες διαφορετικών χρηστών σε αυτά τα περιβάλλοντα είναι μια ιδιαίτερα σημαντική και απαιτητική διαδικασία, την οποία αναλαμβάνει κατάλληλο λογισμικό διαχείρισης και δρομολόγησης εργασιών (Resource and Job Management System - RJMS). Οι διπλωματικές εργασίες στη συγκεκριμένη θεματική περιοχή θα καταπιαστούν σε ζητήματα όπως η μελέτη συμπεριφοράς παράλληλων (MPI) εφαρμογών σε καθεστώς συνεκτέλεσης (co-scheduling), η δημιουργία μοντέλων πρόβλεψης της συμπεριφοράς τους, η επέκταση κατάλληλων εργαλείων προσομοίωσης και ο σχεδιασμός νέων, εξελιγμένων αλγορίθμων ανάθεσης πόρων.

Σχετικά Μαθήματα: Συστήματα Παράλληλης Επεξεργασίας

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

11 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

12 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

13 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

14 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

IV. Κατανεμημένα Συστήματα - Προχωρημένα θέματα βάσεων δεδομένων

15 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

16 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

V. Εργασίες σε συνεπίβλεψη με εξωτερικούς συνεργάτες

17 Αυτόματη επιτάχυνση προγραμμάτων κελύφους

Τα προγράμματα κελύφους (shell scripts) παραμένουν εξαιρετικά δημοφιλή (8η πιο δημοφιλής γλώσσα στο Github) εξαιτίας χαρακτηριστικών που βρίσκονται μόνο στο συγκεκριμένο περιβάλλον. Πρόσφατα, μια οικογένεια συστημάτων μέρος του του Linux Foundation, όπως όπως το PaSh και το DiSh, επιταχύνουν τέτοια προγράμματα αυτόματα — με τεχνικές παράλληλης, κατανεμημένης, και εκτός-σειράς επεξεργασίας.
Δυστυχώς, τα συστήματα αυτά υποθέτουν μια συγκεκριμένη μορφή και ιδιότητες από τα προγράμματα κελύφους. Ο στόχος διπλωματικών εργασιών σε αυτή την περιοχή είναι (1) η άρση αυτών των υποθέσεων, ώστε να αξιοποιηθούν ευρύτερα αυτές οι τεχνικές, (2) η επέκταση των προγραμμάτων και εφαρμογών, αντλώντας από πολλαπλές επιστημονικές περιοχές, (3) η βέλτιστη παραμετροποίηση αυτών των συστημάτων με χρήση ευριστικών, στατιστικών, και αναλυτικών μεθόδων, (4) η αξιοποίηση νέων μορφών υπολογισμού όπως αρχιτεκτονική microservice και υπολογισμός serverless, και (5) η αξιολόγηση τεχνικών και συστημάτων αιχμής σε πραγματικά περιβάλλοντα υπερυπολογιστών ή/και το cloud.

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

Επικοινωνία: Γιώργος Γκούμας, goumas@cslab.ece.ntua.gr
Νίκος Βασιλάκης, nikos@vasilak.is, Brown University
Κωνσταντίνος Καλλάς, kkallas@cs.ucla.edu, UCLA

18 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 βιβλιοθηκών 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