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 2025-2026

Περιεχόμενα

1. Αρχιτεκτονική Υπολογιστών

1.1 IoT Security Integrity

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 και η εξερεύνηση του CC, κυρίως

- Μηχανισμοί Security για low-end devices σε IoT - Remote Attestation (RA): Εξερεύνηση, μελέτη και σύγκριση υπάρχουσων λύσεων, αναγνώριση κενών. Εξερεύνηση πραγματικών use Εναλλακτικοί μηχανισμοί ασφαλείας, ελάχιστες αρχιτεκτονικές ασφαλείας [1,2]. Use Case Analysis των υπάρχουσων υλοποιήσεων σε πραγματικά περιβάλλοντα, σύγκριση και μελέτη λύσεων, αναγνώριση κενών. Έρευνα υλοποιήσεων σε embedded IoT για πολλαπλά domain (π.χ. automotive, healthcare, intermittent, space κλπ) και σύγκριση των χαρακτηριστικών τόσο των συσκευών όσο και των μηχανισμών ασφαλείας τους.

- Μηχανισμοί Security για low-end devices σε IoT - Control Flow Integrity και Control Flow Attestation (CFI/CFA): για embedded low-end devices. Μελέτη και επέκταση μηχανισμών CFI/CFA, μελέτη και επέκταση υλοποιήσεων τους σε ενσωματωμένα συστήματα [3].

Παραπομπές:

[1] Sprogø Banks, M. Kisiel, and P. Korsholm. Remote attestation: A literature review.
[2] T. Yagawa et al. Delegating Verification for Remote Attestation Using TEE."
[3]C-FLAT: Control-Flow Attestation for Embedded Systems Software

Σχετικά Μαθήματα: Αρχιτεκτονική Υπολογιστών, Προηγμένα Θέματα Αρχιτεκτονικής Υπολογιστών

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

1.2 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

1.3 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

1.4 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

1.5 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

2. Λειτουργικά Συστήματα

2.1 Τεχνικές βελτιστοποίησης εικονικής μνήμης (virtual memory)

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.

Παραπομπές:

1) Elastic Translations: Fast Virtual Memory with Multiple Translation Sizes [1], [2], [3], [4]
2) Design, Implementation and Evaluation of the SVNAPOT Extension on a RISC-V Processor [1]
3) Enhancing and Exploiting Contiguity for Fast Memory Virtualization [1]
4) eBPF-mm: Userspace-guided memory management in Linux with eBPF [1], [2]
5) SnapBPF: Exploiting eBPF for Serverless Snapshot Prefetching [1], [2]
6) WASP: Workload-Aware Self-Replicating Page-Tables for NUMA Servers

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

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

2.2 AI for Systems / Systems for AI

Στο πλαίσιο της προτεινόμενης διπλωματικής (AI for Systems) θα γίνει αξιολόγηση της χρησιμότητας τεχνικών μηχανικής μάθησης (ML) στο πλαίσιο του Λειτουργικού Συστήματος. Ενδεικτικές κατευθύνσεις αφορούν την ενσωμάτωση τεχνικών μηχανικής μάθησης: i) για την τοποθέτηση σελίδων (page placement) σε πολυεπίπεδες (multi-tiered) αρχιτεκτονικές μνημών (DRAM, PMEM, CXL), ii) για την χρονοδρομολόγηση (scheduling) διεργασιών, και iii) για την γενικότερη βελτιστοποίηση πολιτικών και ρυθμίσεων του Λειτουργικού Συστήματος.

Μια δεύτερη κατεύθυνση (Systems for AI) αφορά στην μελέτη και αξιολόγηση των συστημάτων τα οποία τρέχουν μεγάλα γλωσσικά μοντέλα (LLMs), κυρίως στο κομμάτι του inference. Συγκεκριμένα, θα δοθεί έμφαση στο κομμάτι της KV cache, που παίζει κρίσιμο ρόλο για την επίδοση του inference, και πώς τεχνικές διαχείρισης μνήμης από τα ΛΣ μπορούν να επιταχύνουν την λειτουργία της KV cache.

Παραπομπές:

1) Υλοποίηση και Αξιολόγηση Δρομολογητή Σελίδων με Χρήση Τεχνικών Μηχανικής Μάθησης για Υβριδικά Συστήματα Μνήμης [1]

2) Integrating Artificial Intelligence into Operating Systems: A Comprehensive Survey on Techniques, Applications, and Future Directions [1]

3) Towards Agentic OS: An LLM Agent Framework for Linux Schedulers [1]

4) How I learned to stop worrying and love learned OS policies [3] vLLM, TensorRT-LLM, llm-d

5) Efficient Memory Management for Large Language Model Serving with PagedAttention [1]

6) vAttention: Dynamic Memory Management for Serving LLMs without PagedAttention [1, 2]

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

Επικοινωνία: Στράτος Ψωμαδάκης, psomas@cslab.ece.ntua.gr
Κωνσταντίνος Μορές, kmores@cslab.ece.ntua.gr
Χλόη Αλβέρτη, xalverti@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, διαδιεργασιακή επικοινωνία, διαχείριση μνήμης, διαχείριση αρχειων κ.α., ωστε να αποκτήσουν επίγνωση της κατανεμημένης φύσης του νέου συστήματος.

Παραπομπές:

1) Memory disaggregation: why now and what are the challenges [SIGOPS 2023]

2) CXLfork: Fast Remote Fork over CXL Fabrics [ASPLOS 2025]

3) Partial Failure Resilient Memory Management System for Distributed Shared Memory [SOSP 2023]

4) HydraRPC: RPC in the CXL Era [ATC 2024]

5) An Introduction to the Compute Express Link (CXL) Interconnect [Computing Surveys 2024]

Σχετικά Μαθήματα: Λειτουργικά Συστήματα, Εργαστήριο Λειτουργικών Συστημάτων

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

2.4 Operating System principles for Serverless Cloud / AI Infrastructures

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, GPU, 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 και η ραγδαία ανάπτυξη του AI. Στην περιοχή ενδιαφέροντος που αναφερόμαστε (Λειτουργικά Συστήματα για πλατφόρμες Serverless / Cloud / AI) περιλαμβάνεται η μελέτη, αξιοποίηση και προσαρμογή βασικών αρχών και μηχανισμών των Λειτουργικών Συστημάτων όπως η εικονικοποίηση μέσω hardware (KVM [3]) και μηχανισμών πυρήνα (cgroups, namespaces), η διαχείριση μνήμης (memory allocation) και ο διαμοιρασμός των συνδέσεων προς απομακρυσμένες μονάδες αποθήκευσης (π.χ. databases) [4].
Παράλληλα σε μια πιο πρόσφατη προσέγγιση, μελετάμε την αποδοτική χρήση των GPUs από μοντέλα LLMs που εκτελούνται σε Serverless περιβάλλοντα [6].
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), memory snapshots [7], Linux / kernel (virtio, memory management, sockets, cgroups, namespaces)
- την εξοικείωση με την εκτέλεση Serverless μοντέλων LLMs σε GPUs [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] ServerlessLLM: Low-Latency Serverless Inference for Large Language Models
[3] KVM
[4] Connection pooling
[5] Above the Clouds: A Berkeley View of Cloud Computing
[6] ServerlessLLM
[7] Benchmarking, Analysis, and Optimization of Serverless Function Snapshots

Σχετικά Μαθήματα: Λειτουργικά Συστήματα, Εργαστήριο Λειτουργικών Συστημάτων

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

2.5 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

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 αλγορίθμων co-scheduling στον εξομοιωτή ELiSE (`https://github.com/cslab-ntua/elise`)
2. Πειραματική μελέτη quarter-socket co-scheduling στον υπερυπολογιστή ARIS
3. Υλοποίηση αλγορίθμων co-scheduling με το Flurm framework (`https://github.com/cslab-ntua/flurm`)
4. Συνεκτέλεση MPI εφαρμογών χρησιμοποιώντας κατάτμηση της cache και του εύρους ζώνης μνήμης (memory bandwidth) στην υπολογιστική υποδομή Grid5000

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. Βελτιστοποίηση της χρονοδρομολόγησης εργασιών με ευέλικτο τρόπο ανάθεσης πόρων (moldable jobs)

2. Δημιουργία γεννήτορα συνθετικών μετροπρογραμμάτων (synthetic benchmark generator) για τη μελέτη και πρόβλεψη της κλιμακωσιμότητας (scalability) εργασιών

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

3.2 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

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

5. Εργασίες σε συνεπίβλεψη με εξωτερικούς συνεργάτες

5.1 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] 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

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

[3] 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

5.2 Beyond Offloading: CPU-GPU Collaboration for High-Performance LLM Serving

Large Language Models (LLMs) have demonstrated remarkable generative capabilities, revolutionizing natural language processing and enabling sophisticated applications across diverse domains. However, these models require substantial memory capacity due to their large parameter counts, often exceeding hundreds of billions of parameters. These high memory footprints pose significant challenges for GPU-based deployment, as GPUs typically have limited memory capabilities (memory capacity). Consequently, fitting these large LLMs necessitates multiple expensive GPUs, while the GPU compute resources are not always fully utilized due to communication and memory bottlenecks.
To address the memory capacity problem of GPUs, recent works have explored CPU offloading capabilities, where model parameters or intermediate data are stored in CPU memory connected to GPUs via high-speed interconnects. Although this approach potentially resolves capacity limitations, prior works leave CPU cores idle or fail to fully leverage their computational capabilities. Moreover, data transfers between CPU and GPU memory introduce additional communication overheads that can degrade overall system performance. The goal of this thesis project is to explore well-crafted CPU-GPU LLM inference optimizations that fully utilize both CPU and GPU devices. This research project aims to design an intelligent LLM serving system optimized for memory-constrained environments that provides dynamic and adaptive policies for CPU-GPU execution, maximizing overall system performance and resource efficiency. The proposed system will also support Mixture-of-Experts (MoE LLM models and it can also be evaluated on GPU Hopper NVIDIA architectures (if such a platform is available) that feature unified memory capabilities between CPU and GPU, enabling more effective data movement optimizations.

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] Seonjin Na, Geonhwa Jeong, Byung Hoon Ahn, Aaron Jezghani, Jeffrey Young, Christopher J. Hughes, Tushar Krishna, Hyesoon Kim, “FlexInfer: Flexible LLM Inference with CPU Computations”, MLSys 2025

[2] Yinmin Zhong, Shengyu Liu, Junda Chen, Jianbo Hu, Yibo Zhu, Xuanzhe Liu, Xin Jin, Hao Zhang, ”DistServe: Disaggregating Prefill and Decoding for Goodput-optimized Large Language Model Serving”, OSDI 2024

[3] 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

[4] Yichao Yuan, Lin Ma, Nishil Talati, “MoE-Lens: Towards the Hardware Limit of High-Throughput MoE LLM Serving Under Resource Constraints”, arXiv 2025

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

5.3 System-Level Performance Optimizations for Physical AI

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

[2] 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

[3] 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

[4] 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