Emmanouil Ioannis Farsarakis

farsarakis@gmail.com | +44-746-281-1213 | linkedIn/farsarakis | Google Scholar

PROFILE

Systems Architect & AI Researcher | Systems architect and AI researcher in Intel's Systems Architecture and Engineering team, focused on memory-centric AI inference: disaggregated and in-network memory, KV-cache placement across heterogeneous memory, and dataflow versus heterogeneous accelerators for VLA and LLM serving. Physics and HPC background, with peer-reviewed systems research (IEEE ISPASS) and a US patent. US, UK and EU work entitlement.

PROFESSIONAL EXPERIENCE

INTEL | AI Systems & Solutions Engineer (Remote) Edinburgh, UK | Nov 2023 – Present
Technical lead for SPEC Cloud benchmark simulation and system-level performance projection, leading a team of five. Driving next-generation, inference-focused architecture research for large-memory-capacity serving: evaluating silicon photonics and rack-scale optical interconnects, and dataflow accelerators (e.g. Groq) versus heterogeneous GPUs across latency, throughput and cost; disaggregated/in-network memory; and VLA-model serving for robotics.

INTEL | Artificial Intelligence Researcher (Remote) Edinburgh, UK | Feb 2022 – Nov 2023
Coordinated research with leading scientists on next-generation AI algorithms — GNNs, NLP and recommendation systems — for graph analytics and security.

INTEL | Platform Architect (Hybrid) Edinburgh, UK | Feb 2019 – Feb 2022
Drove novel AI optimizations (SIMT, sparse-compute acceleration, quantization) and led AI workload characterization for the Intel PiUMA sparse accelerator. Based at the University of Edinburgh, collaborating with K. Heafield and P. Boyle.

THE UNIVERSITY OF EDINBURGH | Applications Consultant Edinburgh, UK | May 2016 – Feb 2019
HPC and data-science leadership across finance, cyber security and exascale, bridging research and industrial challenges.

THE UNIVERSITY OF EDINBURGH | Applications Developer Edinburgh, UK | Sep 2014 – May 2016
EPCC Computational Science & Engineering team — porting, profiling and optimisation on academia–industry collaborations.

SELECTED PROJECTS

Inference Architectures for VLA Robotics Dataflow vs. heterogeneous accelerators, serving, KV-cache
Characterized VLA robotics models (e.g. π0.5, GR00T) as a two-frequency serving workload and mapped them onto candidate architectures, comparing dataflow accelerators (Groq) against heterogeneous GPUs; framed robotics and data-centre serving as a common FLOP/s-versus-TB/s balancing act.

Silicon Photonics & Disaggregated Memory Rack-scale optical interconnects, disaggregation
Assessed silicon-photonics and rack-scale optical interconnect technologies (e.g. Celestial AI) for memory-centric inference, comparing latency, throughput and cost across on- versus off-chip solutions; explored disaggregated compute/memory and in-network memory.

Sparse Computational Infrastructure for GNNs Research, Management, Engineering
Established and led a cross-org collaboration — around seven contributors across Intel, Intel Labs and Harvard SEAS — on graph-learning acceleration for Intel® PiUMA. Owned inception, stakeholder management, software and performance optimization; resulted in ISPASS 2023, a TASK Quarterly journal paper (2024) and a US patent.

Low-Precision CPU Inference --- Marian NMT in SPEC CPU int8 GEMM, quantization, CPU optimization
Part of the Intel–University of Edinburgh collaboration (K. Heafield) on CPU-efficient, low-precision (int8) Marian neural machine translation, winning the 2020 NGT efficiency task. The underlying int8 GEMM kernels were adopted into the SPEC CPU benchmark in 2026, where Marian machine translation is the sole NLP workload.

NextGenIO Exascale, NVM, system-level performance modelling
Built system-level workload characterization of ARCHER (UK national supercomputer) and researched data-aware, workflow-enabled HPC scheduling (SLURM-class) for Intel® 3D XPoint non-volatile memory.

Energy Efficiency --- ISC'14 Student Cluster Competition GPU acceleration, OpenACC, scientific computing

Designed and optimized the software stack of a custom GPU cluster — Highest Linpack at ISC'14 (3.38 Tflops/kW, est. 4th on the June 2014 Green500). Ported GADGET-3 kernels to GPU via OpenACC for 2× speedup.

PUBLICATIONS & PATENTS

Optimizing Graph Learning using Hierarchical Graph Adjacency Matrix (HGAM) — Journal | 2024
TASK Quarterly, vol. 28, no. 3

Neural Network training and inference with Hierarchical Adjacency Matrix — Patent | 2023
US Patent Office - app. US18/325,348

Characterizing the Scalability of Graph Convolutional Networks on Intel® PIUMA — Paper | 2023
2023 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)

Edinburgh's submissions to the 2020 machine translation efficiency task — Proceedings | 2020
Fourth Workshop on Neural Generation and Translation (pp. 218-224)

Optimized Data Decomposition for Reduced Communication Costs — Presentation | 2018
ISC'18, IXPUG

Resource requirement specification for novel data-aware and workflow-enabled HPC job schedulers — WiP | 2017
SC'17, PDSW-DISCS

Monitoring and evaluating I/O performance of HPC systems — Abstract | 2016
4th International Exascale Applications and Software Conference

Experiences Porting Production Codes to Xeon Phi Processors — Proceedings | 2015
Parallel Computing: On the Road to Exascale (pp. 575-583)

AWARDS AND DISTINCTIONS

Highest Linpack
Leipzig, Germany | Jun 2014
ISC'14 Student Cluster Competition
Leader of the record-breaking team from EPCC at ISC'14 Student Cluster Competition. The team took the first place mark for the "Highest Linpack" award by recording a score of 3.38 Tflops/kW with the system ranking at an estimated 4th on the June 2014 Green500 list. It is the first time a team has broken the 10 Tflops barrier in under 3kW.

Highly Skilled Workforce Scholarship
Edinburgh, UK | Sep 2013
Scottish Funding Council
Funded by the Scottish Funding Council, this prestigious scholarship is awarded to select students studying one of the University of Edinburgh's School of Physics and Astronomy's MSc programmes to undertake research projects which apply their skills to real-world problems drawn from the Higgs Centre for Theoretical Physics, EPCC, or industry.

EDUCATION

MSc in High Performance Computing
Edinburgh, UK | Aug 2014
University of Edinburgh
Academic Performance: Awarded with Distinction and elected Class Representative
Grade Point Average: 78%
Dissertation: "Energy Efficiency: Benefits and limitations of modern HPC architectures"
Coursework: Advanced parallel programming, Distributed computing, HPC architectures and ecosystem, Data Structures and Algorithms, Extreme computing, Performance programming, Software development

BSc Physics With Computational Science Specialization
Heraklion, Greece | Mar 2011
University of Crete
Academic Performance: Top 5% of class
Grade Point Average: 77%
Teacher Assistant (TA): Introduction to C++ and Java
Coursework: Classical mechanics, Electrodynamics, Optics, Thermodynamics, Atmospheric physics, Advanced Math, Modeling of complex networks, Computational science, Fortran, C, C++, Java, Object oriented analysis and design