IISWC 2024 Program

Monday, September 16

8.45	Opening and Welcome
9.00	Keynote by Prof. Tor Aamodt (University of British Columbia)
10.00	Coffee Break
10.20	Session 1: Best Paper Nominees CRISP: Concurrent Rendering and Compute Simulation Platform for GPUs Junrui Pan, Timothy G. Rogers (Purdue University) LLMServingSim: A HW/SW Co-Simulation Infrastructure for LLM Inference Serving at Scale Jaehong Cho, Minsu Kim, Hyunmin Choi, Guseul Heo, Jongse Park (KAIST) Lotus: Characterization of Machine Learning Preprocessing Pipelines via Framework and Hardware Profiling Rajveer Bachkaniwala, Harshith Lanka, Kexin Rong, Ada Gavrilovska (Georgia Institute of Technology) Mediator: Characterizing and Optimizing Multi-DNN Inference for Energy Efficient Edge Intelligence Seung Hun Choi, Myung Jae Chung, Young Geun Kim, Sung Woo Chung (Korea University) Performance Modeling and Workload Analysis of Distributed Large Language Model Training and Inference Joyjit Kundu, Wenzhe Guo, Ali BanaGozar, Udari De Alwis, Sourav Sengupta (imec); Puneet Gupta (UCLA); Arindam Mallik (imec)
12.00	Lunch
13.20	Session 2: Performance Measurement Tools and Techniques CARM Tool: Cache-Aware Roofline Model Automatic Benchmarking and Application Analysis José Morgado, Leonel Sousa, Aleksandar Ilic (INESC-ID, Instituto Superior Técnico, Universidade de Lisboa) SHARP: A Distribution-Based Framework for Reproducible Performance Evaluation Viyom Mittal (University of California, Riverside); Pedro Bruel (Hewlett Packard Labs, USA); Michalis Faloutsos (University of California, Riverside); Dejan Milojicic, Eitan Frachtenberg (Hewlett Packard Labs, USA) Taming Performance Variability caused by Client-Side Hardware Configuration Georgia Antoniou, Haris Volos, Yiannakis Sazeides (University of Cyprus) HEX-SIM: Evaluating Multi-modal Large Language Models on Multi-chiplet NPUs Xinquan Lin (FuZhou University); Haobo Xu, Yinhe Han, Yiming Gan (ICT, Chinese Academy of Sciences)
14.40	Coffee Break
15.00	Session 3: Emerging Applications and Technologies Evergreen: Comprehensive Carbon Model for Performance-Emission Tradeoffs Tersiteab Adem, Andrew McCrabb, Vidushi Goyal, Valeria Bertacco (University of Michigan) Performance Analysis of Zero-Knowledge Proofs Saichand Samudrala, Jiawen Wu, Chen Chen (Texas A&M University); Jonathan Ku, Haoxuan Shan, Yiran Chen (Duke University); JV Rajendran (Texas A&M University) VelociTI: An Architecture-level Performance Modeling Framework for Trapped Ion Quantum Computers Alex Hankin (Harvard University); Abdulrahman Mahmoud (Harvard University/MBZUAI); Mark Hempstead (Tufts University); David Brooks, Gu-Yeon Wei (Harvard University) QRIO: Quantum Resource Infrastructure Orchestrator Shmeelok Chakraborty, Yuewen Hou, Ang Chen, Gokul Subramanian Ravi (University of Michigan)
16.20	Poster Lightning Talks
16.30	Poster Session
18.00	Conference Banquet @ Cecil Green Park House

Tuesday, September 17

9.00	Keynote by Prof. Vijay Janapa Reddi (Harvard University)
10.00	Coffee Break
10.20	Session 4: LLMs and Systems for Machine Learning Understanding Performance Implications of LLM Inference on CPUs Seonjin Na, Geonhwa Jeong (Georgia Institute of Technology); Byunghoon Ahn (University of California San Diego); Jeffrey Young, Tushar Krishna (Georgia Institute of Technology); Hyesoon Kim (Georgia Tech) Low-Bitwidth Floating Point Quantization for Efficient, High-Quality Diffusion Models Cheng Chen, Christina Giannoula, Andreas Moshovos (University of Toronto) Characterizing the Accuracy - Efficiency Trade-off of Low-rank Decomposition in Language Models Chakshu Moar, Faraz Tahmasebi (University of California, Irvine); Michael Pellauer (NVIDIA); Hyoukjun Kwon (University of California, Irvine) Understanding The Performance and Estimating The Cost Of LLM Fine-Tuning Yuchen Xia (University of Michigan); Jiho Kim (Georgia Institute of Technology); Yuhan Chen, Haojie Ye (University of Michigan); Souvik Kundu (Intel Labs); Cong "Callie" Hao (Georgia Institute of Technology); Nishil Talati (University of Michigan) Characterizing and Optimizing the End-to-End Performance of Multi-Agent Reinforcement Learning Systems Kailash Gogineni, Yongsheng Mei (George Washington University); Karthikeya Gogineni (Independent); Peng Wei, Tian Lan, Guru Venkataramani (George Washington University)
12.00	Lunch
13.20	Session 5: Caches and Memory Understanding Address Translation Scaling Behaviours Using Hardware Performance Counters Nick Lindsay, Abhishek Bhattacharjee (Yale University) Architectural Modeling and Benchmarking for Digital DRAM PIM Farzana Ahmed Siddique, Deyuan Guo, Zhenxing Fan, Mohammadhosein Gholamrezaei, Khyati Kiyawat, Morteza Baradaran, Alif Ahmed, Kyle Durrer, Abdullah T. Mughrabi, Hugo Abbot, Ethan Ermovick, Ashish Venkat, Kevin Skadron (University of Virginia) Enhanced System-Level Coherence for Heterogeneous Unified Memory Architectures Anoop Mysore Nataraja (University of Washington); Ricardo Fernández-Pascual, Alberto Ros (University of Murcia) Characterizing Emerging Page Replacement Policies for Memory-Intensive Applications Michael Wu (Yale University); Sibren Isaacman (Loyola University Maryland); Abhishek Bhattacharjee (Yale University) Kindle: A Comprehensive Framework for Exploring OS-Architecture Interplay in Hybrid Memory Systems Arun KP (Indian Institute of Technology Kanpur); Debadatta Mishra (IIT Kanpur, India)
15.00	Coffee Break
15.20	Session 6: GPUs and Heterogeneous Systems Characterizing CUDA and OpenMP Synchronization Primitives Brandon Burtchell, Martin Burtscher (Texas State University) Evaluating Performance and Energy Efficiency of Parallel Programming Models in Heterogeneous Computing Demirhan Sevim, Baturalp Bilgin, Ismail Akturk (Ozyegin University) Performance Impact of Removing Data Races from GPU Graph Analytics Programs Yiqian Liu, Avery VanAusdal, Martin Burtscher (Texas State University)
16.20	Closing and Best Paper Award