**Accepted Papers** (no specific order)

- Machine Learning-based Deep Packet Inspection at Line Rate for RDMA on FPGAs
  Authors: Maximilian Jakob Heer, Benjamin Ramhorst, Gustavo Alonso (ETH Zurich)

- Practical Federated Learning without a Server
  Authors: Akash Dhasade, Anne-Marie Kermarrec (EPFL); Erick Lavoie (University of Basel); Johan Pouwelse (Delft University of Technology); Rishi Sharma, Martijn de Vos (EPFL)

- Leveraging Approximate Caching for Faster Retrieval-Augmented Generation
  Authors: Shai Aviram Bergman, Zhang Ji (Huawei); Anne-Marie Kermarrec, Diana Petrescu, Rafael Pires, Mathis Randl, Martijn de Vos (EPFL)

- Efficient Federated Search for Retrieval-Augmented Generation
  Authors: Rachid Guerraoui, Anne-Marie Kermarrec, Diana Petrescu, Rafael Pires, Mathis Randl, Martijn de Vos (EPFL)

- Verifying Semantic Equivalence of Large Models with Equality Saturation
  Authors: Kahfi S. Zulkifli (University of Virginia); Wenbo Qian (Northeastern University); Shaowei Zhu, Yuan Zhou, Zhen Zhang (Amazon Web Services); Chang Lou (University of Virginia)

- NeuraLUT-Assemble: Hardware-aware Assembling of Sub-Neural Networks for Efficient LUT Inference
  Authors: Marta Andronic (Imperial College London); George A. Constantinides (Imperial College London, UK)

- Systems Opportunities for LLM Fine-Tuning using Reinforcement Learning
  Authors: Pedro F. Silvestre, Peter Pietzuch (Imperial College London)

- Decentralized Adaptive Ranking using Transformers
  Authors: Marcel Gregoriadis, Quinten Stokkink, Johan Pouwelse (Delft University of Technology)

- Performance Aware LLM Load Balancer for Mixed Workloads
  Authors: Kunal Jain (Microsoft); Anjaly Parayil (Microsoft Research); Ankur Mallick, Esha Choukse (Microsoft); Xiaoting Qin, Jue Zhang (Microsoft Research); Íñigo Goiri, Rujia Wang, Chetan Bansal (Microsoft); Victor Rühle (Microsoft Research); Anoop Kulkarni, Steve Kofsky (Microsoft); Saravan Rajmohan (Microsoft 365)

- Exploiting Unstructured Sparsity in Fully Homomorphic Encrypted DNNs
  Authors: Aidan Ferguson, Perry Gibson, Lara D'Agata (University of Glasgow); Parker McLeod, Ferhat Yaman, Amitabh Das, Ian Colbert (AMD); José Cano (University of Glasgow)

- Decoupling Structural and Quantitative Knowledge in ReLU-based Deep Neural Networks

Authors: José Duato (Qsimov Quantum Computing S.L.); Jose I. Mestre, Manuel F. Dolz (Universitat Jaume I); Enrique S. Quintana-Orti (Universitat Politècnica de València); José Cano (University of Glasgow)

- RMAI: Rethinking Memory for AI (Inference)
  Authors: Amir Noohi (University of Edinburgh); Mostafa Derispour (Isfahan University of Technology); Antonio Barbalace (The University of Edinburgh)

- Understanding Oversubscribed Memory Management for Deep Learning Training
  Authors: Mao Lin, Hyeran Jeon (University of California, Merced)

- Priority-Aware Preemptive Scheduling for Mixed-Priority Workloads in MoE Inference
  Authors: Mohammad Siavashi (KTH Royal Institute of Technology); Faezeh Keshmiri Dindarloo (Unaffiliated); Dejan Kostic, Marco Chiesa (KTH Royal Institute of Technology)

- Diagnosing and Resolving Cloud Platform Instability with Multi-modal RAG LLMs
  Authors: Yifan Wang, Kenneth P. Birman (Cornell University)

- FlexInfer: Breaking Memory Constraint via Flexible and Efficient Offloading for On-Device LLM Inference
  Authors: Hongchao Du, Shangyu Wu (City University of Hong Kong); Arina Kharlamova (Mohamed bin Zayed University of Artificial Intelligence); Nan Guan (City University of Hong Kong); Chun Jason Xue (Mohamed bin Zayed University of Artificial Intelligence)

- Deferred prefill for throughput maximization in LLM inference
  Authors: Moonmoon Mohanty, Gautham Bolar, Preetam Patil (Indian Institute of Science Bangalore); UmaMaheswari Devi, Felix George, Pratibha Moogi (IBM Research - India); Parimal Parag (Indian Institute of Science Bangalore)

- AMPLE: Event-Driven Accelerator for Mixed-Precision Inference of Graph Neural Networks
  Authors: Pedro Gimenes (Imperial College London, UK); Aaron Zhao (Imperial College London); George A. Constantinides (Imperial College London, UK)

- Client Availability in Federated Learning: It Matters!
  Authors: Dhruv Garg, Debopam Sanyal (Georgia Institute of Technology); Myungjin Lee (Cisco Systems); Alexey Tumanov, Ada Gavrilovska (Georgia Institute of Technology)

**Accepted Poster Papers** (no specific order)

- Global-QSGD: Allreduce-Compatible Quantization for Distributed Learning with Theoretical Guarantees
  Authors: Jihao Xin, Marco Canini, Peter Richtárik (KAUST); Samuel Horváth (MBZUAI)

- Hybrid Task Scheduling for Optimized Neural Network Inference on Skin Lesions in Resource-Constrained Systems
  Authors: Diogen Babuc, Teodor-Florin Fortiş (West University of Timişoara)

- Cross-Domain DRL Agents for Efficient Job Placement in the Cloud-Edge Continuum
  Authors: Theodoros Aslanidis (University College Dublin); Sokol Kosta (Department of Electronic Systems, Aalborg University Copenhagen); Spyros Lalis (University of Thessaly); Dimitris Chatzopoulos (University College Dublin)

- Towards a Unified Framework for Split Learning
  Authors: Boris Radovič (KAUST & University of Ljubljana); Marco Canini (KAUST); Samuel Horváth (MBZUAI); Veljko Pejović (University of Ljubljana); Praneeth Vepakomma (MBZUAI & MIT)

- Manage the Workloads not the Cluster: Designing a Control Plane for Large-Scale AI Clusters
  Authors: Ruiqi Lai, Siyu Cao, Leqi Li (NTU Singapore); Luo Mai (University of Edinburgh); Dmitrii Ustiugov (NTU Singapore)

- Harnessing Increased Client Participation with Cohort-Parallel Federated Learning
  Authors: Akash Dhasade, Anne-Marie Kermarrec (EPFL); Tuan-Ahn Nguyen (Independent Researcher); Rafael Pires, Martijn de Vos (EPFL)

- Accelerating MoE Model Inference with Expert Sharding
  Authors: Oana Balmau (McGill); Anne-Marie Kermarrec, Rafael Pires, André Loureiro Espírito Santo, Martijn de Vos, Milos Vujasinovic (EPFL)

- TAGC: Optimizing Gradient Communication in Distributed Transformer Training
  Authors: Igor Polyakov (VK, ITMO University); Alexey Dukhanov (ITMO University); Egor Spirin (VK Lab)

- β-GNN: A Robust Ensemble Approach Against Graph Structure Perturbation
  Authors: Haci Ismail Aslan (Technical University of Berlin); Philipp Wiesner, Ping Xiong, Odej Kao (Technische Universität Berlin)

- May the Memory Be With You: Efficient and Infinitely Updatable State for Large Language Models
  Authors: Excel Chukwu, Laurent Bindschaedler (Max Planck Institute for Software Systems)

- Towards Asynchronous Peer-to-Peer Federated Learning for Heterogeneous Systems
  Authors: Christos Sad (Aristotle University of Thessaloniki); George Retsinas, Dimitrios Soudris (National Technical University of Athens); Kostas Siozios (Aristotle University of Thessaloniki); Dimosthenis Masouros (National Technical University of Athens)

- Beyond Test-Time Compute Strategies: Advocating Energy-per-Token in LLM Inference
  Authors: Patrick Wilhelm, Thorsten Wittkopp, Odej Kao (Technische Universität Berlin)

- Utilizing Large Language Models for Ablation Studies in Machine Learning and Deep Learning
  Authors: Sina Sheikholeslami, Hamid Ghasemirahni, Amir H. Payberah, Tianze Wang (KTH Royal Institute of Technology); Jim Dowling (Hopsworks AB); Vladimir Vlassov (KTH Royal Institute of Techonology, Sweden)

- Rethinking Observability for AI workloads on Multi-tenant public clouds
  Authors: Theophilus A. Benson (Carnegie Mellon University)

- OptimusNIC: Offloading Optimizer State to SmartNICs for Efficient Large-Scale AI Training
  Authors: Achref Rebai, Marco Canini (KAUST)

- Analysis of Information Propagation in Ethereum Network Using Combined Graph Attention Network and Reinforcement Learning to Optimize Network Efficiency and Scalability
  Authors: Stefan Behfar, Richard Mortier, Jon Crowcroft (University of Cambridge)