	EuroMLSys (March 31, 2025)		
Time(EST)	Session	Presenter	Affiliation
08:50 - 09:00	Opening		
09:00 - 10:30	Session 1: GPUs, Hardware, LLM (15mins x 6) Chair (Amir Payberah - KTH)		
#31	Understanding Oversubscribed Memory Management for Deep Learning Training	Mao Lin	UC Merced
#38	FlexInfer: Breaking Memory Constraint via Flexible and Efficient Offloading for On-Device LLM Inference	Hongchao Du	City University of Hong Kong
#16	NeuraLUT-Assemble: Hardware-aware Assembling of Sub-Neural Networks for Efficient LUT Inference	Marta Andronic	Imperial
#39	Deferred prefill for throughput maximization in LLM inference	Moonmoon Mohanty	Indian Inst. Bangalore (video)
#41	AMPLE: Event-Driven Accelerator for Mixed-Precision Inference of Graph Neural Networks	Pedro Gimenes	Imperial
#6	Machine Learning-based Deep Packet Inspection at Line Rate for RDMA on FPGAs	Maximilian Jakob Heer	ETHZ
10:30 - 11:00	Coffee Break / Poster Session (Browsing)		
11:00 - 12:00	Session 2: LLM - 1 Optimisation, MoE (15mins x 4) Chair (Pedro Gimenes - Imperial College London)		
#20	Performance Aware LLM Load Balancer for Mixed Workloads	Esha Choukse	Microsoft
#15	Verifying Semantic Equivalence of Large Models with Equality Saturation	Kahfi S. Zulkifli	U. Virginia
#17	Systems Opportunities for LLM Fine-Tuning using Reinforcement Learning	Pedro F. Silvestre	Imperial
#32	Priority-Aware Preemptive Scheduling for Mixed-Priority Workloads in MoE Inference	Mohammad Siavashi	KTH
12:00 - 12:30	Poster Session: Elevator Pitch (2 minutes x 16) Chair (Eiko Yoneki - University of Cambridge)		
12:30 - 14:00	Lunch Break / Poster Session (Browsing - in Floyer)		
14:00 - 14:50	Keynotes Zhihao Jia (CMU)		
	Title:Superoptimizing Machine Learning Systems		
14:50 - 15:35	Session 3: LLM - 2 Optimisation, RAG (15mins x 3) Chair (Eiko Yoneki - University of Cambridge)		
#13	Leveraging Approximate Caching for Faster Retrieval-Augmented Generation	Mathis Randl	EPFL
#34	Diagnosing and Resolving Cloud Platform Instability with Multi-modal RAG LLMs	Yifan Wang	Cornell (video)
#42	Client availability in Federated Learning: It matters!	Dhruv Garg	Georgia Tech
15:35 - 16:00	Coffee Break / Poster Session		
16:00 - 16:45	Session 4: LLM - 2 Optiisation continued (15 mins x 3) Chair (Amir Payberah - KTH)		
#18	Decentralized Adaptive Ranking using Transformers	Marcel Gregoriadis	Deft
#24	Decoupling Structural and Quantitative Knowledge in ReLU-based Deep Neural Networks	Jose I. Mestre	Universitat Jaume I
#28	RMAI: Rethinking Memory for AI (Inference)	Amir Noohi	Edinburgh (video)
16:45 - 17:30	Session 5: Federated Learning and others (15 mins x 3) Chair (Pedro Silvestre - Imperial College London)		
#10	Practical Federated Learning without a Server	Rishi Sharma	EPFL
#21	Exploiting Unstructured Sparsity in Fully Homomorphic Encrypted DNNs	José Cano	Glasgow
#14	Efficient Federated Search for Retrieval-Augmented Generation	Diana Petrescu	EPFL
17:30 - 17:35	Closing		
12:00 - 12:30	Poster Session: Elevator Pitch (2 minutes x 16)		
#11	Harnessing Increased Client Participation with Cohort-Parallel Federated Learning	Akash Dhasade	EPFL
#22	β-GNN: A Robust Ensemble Approach Against Graph Structure Perturbation	Haci Ismail Aslan,	Technische Universität Berlin
#37	OptimusNIC: Offloading Optimizer State to SmartNICs for Efficient Large-Scale AI Training	Achref Rebai	KAUST
#7	Towards a Unified Framework for Split Learning	Boris Radovič	KAUST & University of Ljubljana
#12	Accelerating MoE Model Inference with Expert Sharding	Milos Vujasinovic	EPFL
#25	May the Memory Be With You: Efficient and Infinitely Updatable State for Large Language Models	Excel Chukwu	Max Planck Institute for Software Systems
#27	Beyond Test-Time Compute Strategies: Advocating Energy-per-Token in LLM Inference	Patrick Wilhelm	Technische Universität Berlin
#3	Global-QSGD: Allreduce-Compatible Quantization for Distributed Learning with Theoretical Guarantees	Jihao Xin	KAUST
#33	Utilizing Large Language Models for Ablation Studies in Machine Learning and Deep Learning	Sina Sheikholeslami	KTH
#35	Rethinking Observability for AI workloads on Multi-tenant public clouds	Theophilus A. Benson	CMU
#45	Analysis of Information Propagation in Ethereum Network with GAN and RL to Optimize Network Efficiency and Scalability	Stefan Behfar	University of Cambridge
#8	Manage the Workloads not the Cluster: Designing a Control Plane for Large-Scale AI Clusters	Ruiqi Lai	NTU Singapore
#19	TAGC: Optimizing Gradient Communication in Distributed Transformer Training	Akash Dhasade	EPFL
#26	Towards Asynchronous Peer-to-Peer Federated Learning for Heterogeneous Systems	Dimosthenis Masouros	National Technical University of Athens
#4	Hybrid Task Scheduling for Optimized Neural Network Inference on Skin Lesions in Resource-Constrained Systems	Diogen Babuc	West University of Timişoara
#5	Cross-Domain Adaptive DRL Agents for Efficient Resource Management in the Cloud-Edge Continuum	Theodoros Aslanidis	UCD
#45 full title*	Analysis of Information Propagation in Ethereum Network Using Combined Graph Attention Network and Reinforcement Learning to Optimize Network Efficiency and Scalability		