UCL INSTITUTE OF COMMUNICATIONS AND CONNECTED SYSTEMS

HUAWEI

Inter- and Intra-Datacenters for AI Training and Inference

4th June 2025 | UCL Workshop
sponsored by Huawei

Friends House, 173 – 177 Euston
Road London NW1 2BJ

Inter- and Intra-Datacenters for AI Training and Inference

UCL 4th June 2025 | Workshop sponsored by Huawei
Friends House, 173 – 177 Euston Road London NW1 2BJ

Agenda

UK Time	Topic	Speakers	Institution
9:00 – 9:10	Welcome Speech	Prof. Miguel Rio	UCL
9:10 – 9:20		Dr Liushaowei	Huawei ERI president
Morning Session Chair： Prof. Miguel Rio			
9:20 – 10:00	Key Challenges in Large-Scale AI Training and Inference Clusters	Dr. Guan Zixuan	Huawei
10:00 – 10:40	AI for generate and optimize networks	Dr. Laura Toni	UCL/DeepMind
10:40 – 11:00	Tea Break		
11:00 – 11:40	"Photon: Establishing a new SOTA in decentralized foundation model training"	Prof. Nic Lane	Cambridge
11:40 – 12:20	Wafer-scale LLM	Prof. Luo Mai	Edinburgh
12.20 – 13.00	Rethinking Datacenter Network Bandwidth Allocation from an Application's Perspective	Prof. Boris Grot	Edinburgh
13:00 – 14:00	Lunch Break		
	Afternoon Session Chair： Javier Picorel		
14:00 – 14:40	Routing Telemetry in Large AI/ML Clusters	Prof. Gianni Antichi	Politecnico di Milano
14:40 – 15:20	Lossy Network Transport for Large-Scale AI: Insights and Future Directions	Dr. Ran Bassat	UCL
15:20 – 16:00	Designing HPC Architectures to Accelerate AI at BSC	Prof. Miquel Moreto	Barcelona Super Computing Center
16:00 – 16:20	Tea Break & Group Photo		
16:20 – 17:00	Trustworthy AI... for Systems Security	Prof. Lorenzo Cavallaro	UCL
17:00 – 17:45	Panel discussion : The Battle of Future Computer Networks: Scale-up vs. Scale-out. Lead by Dr. Javier Picorel		

Welcome drinks and dinner from 18:00 at Hilton, Euston,
17-18 Upper Woburn Place, London WC1H 0HT

Inter- and Intra-Datacenters for AI Training and Inference

UCL 4th June 2025 | London UK

Key Challenges in Large-Scale AI Training and Inference Clusters

Dr. Zixuan Guan
Senior Engineer at Huawei

Zixuan is the DCN traffic modelling and algorithm expert at Huawei's DCN Technology Lab. He completed his Ph.D. at Stanford University in 2018. His research has focused on traffic modelling, load balancing, and cache management in DCN. To date, he has filed more than 10 invention patents and authored 3 academic papers in the DCN field. Recently, his work has concentrated on optimizing AI training and inference system and developing simulation platforms for over ten thousand GPUs.

Abstract

In the era of AI large-language-models (LLMs), data center infrastructure faces new opportunities and challenges. This talk, from the perspective of AI training clusters and in terms of cluster scale, efficiency and reliability, explores the challenges. Then, the talk combines the current hot inference services and related LLMs to elaborate on the trends and challenges of inference clusters. Monthly even weekly model iteration, the Mixture-of-Experts (MoE) sparse architecture, multimodal inference, and AI Agent applications are set to reshape AI data center architectural design.

AI for generate and optimize networks

Dr. Laura Toni
Associate Professor at UCL/Visiting Senior Research Scientist at Google DeepMind at UCL/DeepMind

Dr. Laura Toni is an associate professor in the Department of Electronic and Electrical Engineering at University College London (UCL) and a visiting senior research scientist at Google DeepMind. She received her PhD degree in electrical engineering in 2009 from the University of Bologna, Italy. She was a Post-Doc at the University of California at San Diego (UCSD) from 2011-2012 and at the Swiss Federal Institute of Technology (EPFL), Switzerland from 2012-2016. Her major contributions are in the area of large-scale signal processing for machine learning, graph signal processing, decision-making strategies under uncertainty, and multimedia processing. She has (co)-authored over 60 high-impact publications, and she is co-inventor of 2 patents on low-delay video processing and streaming. She is significantly involved in scientific committees of world-leading conferences/journals (e.g., Program Chair of ACM MM 2021 and MMSys 2018, general chair of ACM MMSys 2019 and MM 2026). She recently received the UCL Future Leadership Award, Royal Society Research Fellow and Cisco Academic grant and online optimization on irregular domains with application to smart cities.
She also received the Adobe System academic donation on graph-based processing for point clouds. Since at UCL (2016), she has been PI /Co-PI in over 5 projects sponsored by EPSRC, Royal Society, and industrial partners with cumulative funding for my research exceeding £500k as PI, all centred around media processing and online learning on irregular domains.

Abstract

This talk explores the intersection of Artificial Intelligence with network generation and optimization. We will delve into the primary challenges inherent in developing generative models for graph structures, highlighting their practical application in accelerating drug discovery. The discussion will then transition to how AI can be effectively employed to optimize existing networks. We will examine various methodologies and showcase diverse potential applications, ranging from enhancing sustainability initiatives to further advancements in pharmaceutical research and development. The session aims to provide insights into both creating novel network architectures and refining current ones through intelligent algorithms.

Inter- and Intra-Datacenters for AI Training and Inference

UCL 4th June 2025 | London UK

Photon: Establishing a new SOTA in decentralized foundation model training

Prof. Nicholas D. Lane
University of Cambridge | Flower Labs

Nic Lane (http://niclane.org) is a full Professor in the department of Computer Science and Technology at the University of Cambridge and holds a Royal Academy of Engineering Chair in De-centralized AI. He is also a Fellow of St. John's College. At Cambridge, Nic leads the Cambridge Machine Learning Systems lab (CaMLSys; https://mlsys.cst.cam.ac.uk/). The mission of CaMLSys is to invent the next-generation of breakthrough ML-centric systems. Alongside his academic roles, Nic is the co-founder and Chief Scientific Officer of Flower Labs (https://flower.ai), a venture-backed AI company (YCW23) behind the Flower open-source federated learning framework. Flower Labs seeks to enable an AI future that is collaborative, open and decentralized. Nic has received multiple best paper awards, including ACM/IEEE IPSN 2017 and two from ACM UbiComp (2012 and 2015). In 2018 and 2019, he (and his co-authors) received the ACM SenSys Test-of-Time award and ACM SIGMOBILE Test-of-Time award for pioneering research, performed during his PhD thesis, that devised machine learning algorithms used today on devices like smartphones. Nic was the 2020 ACM SIGMOBILE Rockstar award winner for his contributions to "the understanding of how resource-constrained mobile devices can robustly understand, reason and react to complex user behaviors and environments through new paradigms in learning algorithms and system design."

Abstract

As established scaling laws indicate, the future performance improvements of AI depend on the amount of computing and data sources we can leverage. Where will we get the necessary compute and data to drive the continued advances in AI that the world now has grown to expect? I believe all roads lead to federated learning, and approaches of this kind. In the relatively near future, decentralized and federated techniques in machine learning will be how the strongest LLMs (and foundation models more generally) are trained; and in time, aspirational capabilities like AGI will finally be achieved, in part, due to the adoption of federated methodologies. In this talk, I will describe why the future of AI will be federated, and describe early solutions developed by Flower Labs and CaMLSys that address the underlying technical challenges that the world will face as we shift from a centralized data-center mindset to de-centralized alternatives.

Inter- and Intra-Datacenters for AI Training and Inference

UCL 4th June 2025 | London UK

Rethinking Datacenter Network Bandwidth Allocation from an Application's Perspective

Prof. Boris Grot
Professor in the School of Informatics at University of Edinburgh

Professor Boris Grot is a Professor in the School of Informatics at the University of Edinburgh, where he leads the EASE Lab. His research is focused on server hardware and software stacks, and datacenter-scale computing. Boris is a member of the MICRO and HPCA Halls of Fame and a recipient of multiple awards for his research. Boris was the Program Chair for MICRO 2022, General Chair for HPCA 2024, and an Area Chair for ISCA 2025.

Abstract

Today's datacenter workloads increasingly comprise distributed data-intensive applications, including data analytics, graph processing, and machine-learning training. These applications are bandwidth-hungry and often congest the datacenter network, degrading network performance for co-running workloads. We observe that various workloads exhibit different sensitivity to network bandwidth; for some, even a small reduction in the available bandwidth significantly increases completion time; for others, the completion time is largely insensitive to the available bandwidth. Building on this observation, I will make the case for an application-aware approach to allocating network bandwidth in datacenters. Our approach considerably improves application completion time while requiring only light-weight software support and no modifications to existing network hardware or protocols.

Inter- and Intra-Datacenters for AI Training and Inference

UCL 4th June 2025 | London UK

WaferLLM: Large Language Models Inference at Wafer Scale

Prof. Luo Mai
Associate Professor at University of Edinburgh

Professor Luo Mai is an Associate Professor at the University of Edinburgh, where he leads the Large-Scale Machine Learning Systems Group and co-leads the UK EPSRC CDT in Machine Learning Systems and the ARIA project on Scaling AI Compute by 1000X. His research spans systems, machine learning, and data management. He has received awards from Google, Microsoft, Alibaba, and Tencent, and authored the open-source textbook Machine Learning Systems. He co-founded projects including TensorLayer, TorchOpt, and ServerlessLLM, with over 20,000 GitHub stars. He previously worked at Imperial College London and Microsoft Research, and earned his PhD with a Google PhD Fellowship at Imperial College.

Abstract

Emerging wafer-scale AI accelerators integrate hundreds of thousands of cores with massive on-chip memory and ultra-high bandwidth, yet existing LLM inference systems—designed for GPUs—fail to leverage their full potential. In this talk, I'll present WaferLLM, the first LLM inference system tailored for wafer-scale architectures. Guided by our new PLMR model, which captures key hardware characteristics, WaferLLM introduces wafer-scale parallelism and efficient LLM kernels—MeshGEMM and MeshGEMV—to optimize accelerator utilization. On real hardware (Cerebras WSE), WaferLLM achieves up to 200× better utilization, 606× faster and 22× more energy-efficient GEMV than advanced GPUs, enabling up to 2700 toks/sec/reqand 39× faster decoding for widely used LLMs. In the end, if time permits, I will discuss our ongoing work on scaling AI systems beyond a single wafer.

Inter- and Intra-Datacenters for AI Training and Inference

UCL 4th June 2025 | London UK

Routing Telemetry in Large AI/ML Clusters

Prof. Gianni Antichi
Associate Professor at Politecnico di Milano

Gianni Antichi is an Associate Professor at Dipartimento Elettronica, Informazione e Bioingegneria of Politecnico di Milano (Italy) and Senior Lecturer (Associate Professor) at the School of Electronic Engineering and Computer Science of Queen Mary University of London (United Kindgom). His research interests sit at the intersection of networks and systems and the goal is to develop hardware/software co-designs to improve performance and efficiency of end-host applications as well as packet-processing programs. He received a PhD in Information Engineering from the University of Pisa (Italy). His awards include the best paper at ACM SIGCOMM 2017, the ACM SOSR system 2019, the EPSRC New Investigator, the Facebook Networking Systems Research RFP in 2020 and the best student paper at ACM EuroSys 2015.

Abstract

Network applications from traffic engineering to path tracing often rely on the ability to transmit fine-grained telemetry data from network devices to a set of collectors. The problem is that this constant flow of telemetry shares the same underlying network of data packets. A clear trade-off is present: gaining more network visibility (i.e., acquiring more telemetry) at the cost of impacting application traffic or not? This becomes even more important in the presence of AI/ML workloads, presenting an on-off pattern that can be sensibly impacted. In this talk, I will start by introducing the design space for collecting data from network switches and then present InvisiFlow, a novel communication substrate to collect network telemetry data, silently. I will conclude the talk with my personal thoughts and next steps.

Inter- and Intra-Datacenters for AI Training and Inference

UCL 4th June 2025 | London UK

Lossy Network Transport for Large-Scale AI: Insights and Future Directions

Dr. Ran Ben Bassat
Associate Professor in Computer Science at UCL

Ran Ben Basat is an Associate Professor at the Computer Science Department of University College London. He completed his PhD at Technion, after which he has been a postdoctoral fellow at Harvard University. His research interests include algorithms for networking, data management, and machine learning. Ran has received the Meta "Network for AI" faculty award for his work in accelerating DNN training.

Abstract

Large-scale AI training frequently involves hundreds of thousands of GPUs spread across multiple data centers, where congestion and straggling nodes lead to performance bottlenecks. In this talk, I will present our work on loosening the traditional assumption of fully reliable, lossless data transfer. By allowing controlled loss—through mechanisms such as homomorphic and in-network compression, coded transmission, and selective skipping of straggler GPUs—we have observed significant improvements in training speed. I will conclude by outlining the next steps toward an integrated vision of “loss-aware” ML networking.

Inter- and Intra-Datacenters for AI Training and Inference

UCL 4th June 2025 | London UK

Designing HPC Architectures to Accelerate AI at BSC

Prof. Miquel Moreto
Associate Professor at Barcelona Supercomputing Center

Miquel Moreto is an Associate Professor (with tenure) at the Computer Architecture Departament (DAC) at the Universitat Politecnica de Catalunya-Barcelona Tech (UPC), where he teaches Computer Architecture. Since 2025, he is the Director of the High Performance Computer Architecture research area at the Barcelona Supercomputing Center (BSC), coordinating 10 research groups and over 150 researchers in computer architecture. He received the BSc, MSc, and PhD from the UPC. His PhD thesis advisors were Mateo Valero and Francisco J. Cazorla. During his PhD, he interned at IBM T. J. Watson Research Center for 4 months, and visited the Universities of Edinburgh and Cantabria for 3 months. After finishing the PhD, he spent 15 months at the International Computer Science Institute (ICSI), affiliated with UC Berkeley, as a Fulbright Postdoctoral Research Fellowship Holder during 2011 and 2012. In 2013, he returned to Barcelona to work on the RoMoL and Mont-Blanc projects. In 2017, he spent 2 months at Arm Research (Cambridge, UK) as a Visiting Professor. In 2018, he continued his career as a Ramón y Cajal fellow at UPC, leading the Lagarto initiative that developed the first open source processor in Spain based on the RISC-V ISA. Finally, he became Associate Professor at UPC in 2023.

Abstract

Since 2004, the Barcelona Supercomputing Center is leading the European efforts to develop High Performance Computing (HPC) designs based on domestic technology. In the context of the Mont-Blanc European projects (2011-2021) and in close collaboration with Arm and Atos, BSC deployed the first HPC cluster based on Arm technology. More recently, BSC led the design, verification and fabrication of RISC-V-based vector accelerators in the context of the European Processor Initiative (EPI) and RISC-V general purpose processors in the context of the DRAC project. Since March 2025, BSC is leading the Digital Autonomy with RISC-V in Europe (DARE) project that will develop prototype HPC and AI systems based on EU-designed and developed industry-standard chiplets. In this talk, we will provide an overview of the main achievements in these projects, focusing on the efforts to accelerate AI workloads with RISC-V official and custom ISA extensions. Finally, we will present current challenges to achieve European technological independence based on the RISC-V open instruction set architecture.

Trustworthy AI… for Systems Security

Prof. Lorenzo Cavallaro
Professor of Computer Science at UCL

Prof. Lorenzo Cavallaro grew up on pizza, spaghetti, and Phrack, and soon developed a passion for underground and academic research. He is a Full Professor of Computer Science at University College London (UCL), where he leads the Systems Security Research Lab. Lorenzo's research vision is to enhance the effectiveness of machine learning for systems security in adversarial settings. He works with his team to investigate the interplay among program analysis abstractions, representations, and ML models, and their crucial role in creating Trustworthy AI for Systems Security. Lorenzo publishes at and sits on the Program Committee of leading conferences in computer security and ML, received the Distinguished Paper Award at USENIX Security 2022, an ICML 2024 Spotlight Paper, and the Best Paper Award at the Deep Learning for Security and Privacy 2025 workshop. He is also Associate Editor of ACM TOPS and IEEE TDSC. In addition to his love for food, Lorenzo finds his Flow in science, music, and family.

Abstract

No day goes by without reading machine learning (ML) success stories across various application areas. Systems security is no exception, where ML's tantalizing performance leave one to wonder whether there are any unsolved problems left. However, machine learning has no clairvoyant abilities and once the magic wears off, we're left in uncharted territory. Is machine learning truly capable of ensuring systems security? In this talk, we will take malware research as a representative example of a long-studied, thriving, yet challenging subject and we will illustrate some of the open issues that are affecting learning-based malware models along with promising results to move forward. When relevant, and if time allows for it, we will also delve into behind-the-scenes aspects to encourage reflection on the reproducibility crisis. Our goal is to foster a deeper understanding of machine learning's role in systems security along with a discussion on promising directions the research community is and should be pursuing to address such challenges and open problems.