Program

18th Cloud Control Workshop

Skåvsjöholm, Åkersberga, Sweden
June 17 – 19, 2025

Practical information

Hotel contact
Skåvsjöholm Hotell och Konferens
Address: Skåvsjöholmsv. 80, 184 94 Åkersberga
e-mail: info@skavsjoholm.se
Telephone: +46-8-540 267 00

Hotel reception opening hours
07:00-22:00

Check-in and -out
Hotel room check-in from 15:00. Check-out until 10:00. (For check-out before 07.00, talk to the reception).

Payment
Rooms, meals, and some drinks are covered by the workshop registration fee. Please ensure to pay at check-out any additional costs put on your room, additional drinks, etc.

Cafeteria
Cafeteria on entrance floor with coffee, tea, fruit, and biscuits available free of charge 24 hours a day.

Luggage room
There is a dedicated luggage room on entrance floor on arrival.

Relax facilities and other leisure activities
The relax facilities includes a sauna and outdoor whirlpool and is by default open 17.00-22.00.
For use at other hours, tell the reception an hour in advance.
Equipment for various social games (e.g., boule, crocket, kubb, tennis) are available by the relax facilities (on shelfs near the bar)
Bicycles: ask at the reception
Canoes and stand-up paddle boards: ask at the reception – do not forget the life jackets!

Bar
The main bar is located at the ground floor (near the relax facilities) and is open 17.00-24.00 (last order at 23.30).
There is also a lobby bar by the hotel entrance.

Breakfast
Breakfast is served in the main dining room at 07.00-09.00

Coffee breaks
Coffee during coffee breaks is served in the hotel's main dining room

Meals and drinks
Lunches are served in the main dining room
Tuesday 2-course dinner is served in the main dining room. Dessert is from a buffet and can be brought to outside or to the lounge area. Wednesday dinner will be served on a tiny Island, a short boat ride away. The boat, carrying everyone, will leave at 18.00. Drinks for dinners and The Welcome Poster Reception are included. Further drinks are on your own cost. (If putting costs on your room, please ensure to cover the bill at check-out.)

Internet
WiFi throughout the premises
In main building: "SKAVAN" – No password
In Villa Skåvsjöholm: ”VILLAN” - No password
In Husarö: "HUSARO" - Password provided on site

Weather
Forecasts for all three days are a mix of clouds and sunshine with temperature highs being 20-23 degrees Celcius and lows being 12-14 degrees. There is a slight chance of light rain on Wednesday afternoon.

In case everything else fails
Erik's phone number is +46 70 315 3928. Anticipate limited ability to answer around bus departure on June 17.

Tuesday June 17th

8:40 Gathering immediately outside exit of Terminal 4, Arlanda airport
9:00 Bus departure

9:40 Coffee, registration, and luggage storage

Session Chair:Jose Fortes, University of Florida, USA

10:10 *Workshop Introduction*
Erik Elmroth, Umeå University and Elastisys, Sweden

10:20 *10 years on from the first Borg paper*
John Wilkes, Google, USA

10:50 *Decarbonizing Cloud Computing: Opportunities and Tradeoffs*
Prashant Shenoy, University of Massachusetts Amherst, USA

11:10 *Smart Agentic Memory: Challenges and Opportunities*
Jayaram Radhakrishnan, IBM Research, USA

11:30 *Synchrony in Datacenters?*
Patrick Eugster, Università della Svizzera italiana (USI), Switzerland

11:50 *Plenary Discussion: What I Always Wanted to Know*
Thijs Metsch, Intel Labs, Germany

12:30 Lunch

Session Chair:Karl-Erik Årzén, Lund University, Sweden

13:45 *Applications of Multivariable Feedback Control in Online Advertising*
Niklas Karlsson, Amazon, USA

14:15 *Enabling Secure Rolling Updates for Confidential Kubernetes*
Christof Fetzer, TU Dresden, Germany

14:35

Discussion 1 (Resarö)	Discussion 2 (Rindö)	Discussion 3 (Lillskär)	Discussion 4 (Storskär)	Discussion 5 (Huvudskär)
Edge systems: overestimated opportunity or underestimated necessity? Aaron Ding, TU Delft, The Netherlands & Marteen Van Steen, University of Twente, The Netherlands	*Attacks and defences in 5G and beyond (edge) infrastructures (with or without ML)* Simin Nadjm-Tehrani, Linköping University, Sweden	*Cloud Intelligence: AI/ML for cloud & cloud for AI/ML* Sara Bouchenak, INSA Lyon, France, Jian Zhang, Microsoft, USA & Cristian Klein, Umeå University and Elastisys, Sweden	*Predictable energy-efficient cloud/edge infrastructures* Tommaso Cucinotta, Scuola Superiore Sant'Anna, Italy	*Distributed computing challenges in blockchain and cloud computing* Yvonne-Anne Pignolet, DFINITY Foundation, Switzerland & Rodrigo Rodrigues, INESC-ID, ULisboa, Portugal

15.55 Coffee, Hotel Room Check-in, Poster preparation

Session Chair: Jie Xu, University of Leeds, UK

16.25 *SLOs (Slow down, & Service Level Objectives) before PROs (Processors, Performance, etc.)*
Indranil Gupta, University of Illinois Urbana-Champaign, USA

16.45 *Programmable Data Plane 101: a Network Switch with More Computing Power Than Your Cloud Server*
Tim Wood, George Washington University, USA

17.05	***Poster Reception with Refreshments*** (Svavelsö)

1. *Control Offloading for UAVs,* Emil Sundström, Lund University, Sweden
2. *A Stable Matching Approach to Energy Efficient and Sustainable Serverless Scheduling for the Green Cloud Continuum,* Yashwant Singh Patel, Umeå University, Sweden
3. *Offloading of Time-Optimal Motion Planning with Jerk Constraints,* Ahmed Al Bayati, Lund University, Sweden
4. *One Shot Federated Learning with Classifier-Free Diffusion Models,* Obaidullah Zaland, Umeå University, Sweden
5. *Automated Updates Over LoRaWAN for Energy-Efficient Edge AI,* Clément Courageux-Sudan, Umeå University, Sweden
6. *Decentralised carbon-aware scheduling for large-scale microservice-based infrastructures,* Yangyang Wen, Umeå University, Sweden
7. *Keeping Apps Afloat: Predicting and Mitigating Shared Resource Bottlenecks in the Cloud,* Oliver Larsson, Umeå University, Sweden
8. *A Fault-tolerance Strategy for Mission-Critical Application at the Edge,* Tanaz (Nayereh) Rasouli, Umeå University, Sweden
9. *Over-sampling based control & Aperiodic sampling,* Max Nyberg Carlsson, Lund University, Sweden
10. *Cloud continuum resource monitoring,* Angela Checa, Umeå University, Sweden
11. *Energy Efficient and QoS-Aware Model Selection for DNN Inference in Edge Intelligence,* Hajar (Maryam) Siar, Umeå University, Sweden
12. E*fficient scheduling to reduce energy use in a heterogeneous cluster,* Ileet Malla, Umeå University, Sweden
13. *Adaptive Deviation Learning for Visual Anomaly Detection with Data Contamination,* Anindya S Das, Umeå University, Sweden
14. *Spilling the Load: A Hydraulic Model Predictive Controller for Distributed Serverless Platforms,* Chanh Nguyen, Umeå University, Sweden
15. *Energy Forecasting with SNN-Transformer Integration for Cloud and Edge Environments,* Zulfiqar Ahmad Khan, Umeå University, Sweden
16. *Uncover Compromised AI Functions in a Serverless Device-Edge-Cloud Continuum,* Adil Bin Bhutto, Umeå University, Sweden
17. *Energy-efficient Right-sizing of Kafka-like Message Brokers,* Govind KP, INRIA, France
18. *Measure and Protect: Enhancing the Web PKI,* Carl Magnus Bruhner, Linköping University, Sweden
19. *Integrated Modelling of Energy & Cloud-Edge Systems for Carbon Analysis*, Rohail Gulbaz, Umeå University, Sweden
20. *Tiny Minds, Big Shields: Detecting DDoS across the Cloud-Edge Continuum Systems,* Yinuo Zhang, Umeå University, Sweden
21. *Efficient Cluster Scheduling for Fine-tuning LLMs Using Historical Configuration Data*, Matthijs Jansen, Vrije Universiteit Amsterdam, The Netherlands
22. *Accurate and Low-Overhead Workload Prediction for Cloud Management,* Lidia Kidane, Umeå University, Sweden

19:30 **Dinner**

Wednesday June 18th

7:00	**Breakfast**				
	Session Chair: Dilma Da Silva, Texas A&M University, USA				
8:30	*Re-imagining Cloud-Scale Telemetry from an Approximation-First Perspective* Vyas Sekar, Carnegie Mellon University, USA				
9:00	*The Modern PC: Personal Computer, or Personal Cloud?* Richard Mortier, Cambridge University, UK				
9:20	*Foundation Models: Emerging Bottlenecks and Research Implications* Tarek Abdelzaher, Indiana University of Illinois, USA				
9:40	**Coffee**				
	Session Chair: Evangelina Kalyvianaki, University of Cambridge, UK				
10:10	*Controlling AI Inference from the Edge to the Cloud* Henry Hoffman, University of Chicago, USA				
10:30	*Massivizing Computer Systems: VU on the Science, Design, and Engineering of Distributed Computer Ecosystems* Alexandru Iosup, Vrie Universitet Amsterdam, The Netherlands				
10:50	*Autonomics for the Digital Continuum* Manish Parashar, University of Utah, USA				
11:10	Discussion 6 (Resarö) *Federated learning and the next edge-to-cloud intelligence: Challenges, shifts, and opportunities* Jayaram Radhakrishnan, IBM Research, USA & Feras Awaysheh, Umeå University, Sweden	Discussion 7 (Rindö) *Is confidential computing living up to its potential?* Christof Fetzer, TU Dresden, Germany & Rüdiger Kapitza, FAU Erlangen-Nürnberg, Germany	Discussion 8 (Lillskär) *6G as part of a cyber-physical continuum* Johan Eker, Lund University and Ericsson Research, Sweden, Catrin Granbom, Ericsson Research, Sweden, & Wolfgang John, Ericsson Research, Sweden	Discussion 9 (Storskär) *Towards sustainable computing on the cloud/edge continuum* Abishek Chandra, University of Minnesota, USA	Discussion 10 (Huvudskär) *How can we democratize cloud systems research?* Jan Rellermeyer, Leibniz University Hannover, Germany
12:30	**Lunch**				
	Session Chair: Paul Townend, Umeå University, Sweden				
13:45	*Research Directions in Google Storage* Arif Merchant, Google, USA				
14.15	Discussion 11 (Resarö) *Agentic AI Systems and how they compares with the traditional Feedback control* Xiaoyun Zhu, NetApp, USA	Discussion 12 (Rindö) *Cloud native control - code changes as the source of disturbance?* Johan Tordsson, Elastisys, Sweden & Oliver Larsson, Umeå University, Sweden	Discussion 13 (Lillskär) *Reliability of cloud-scale systems and infrastructures* Tianyin Xu, University of Illinois Urbana-Champaign, USA& Indranil Gupta, University of Illinois Urbana-Champaign, USA	Discussion14(Storskär) *Can data compression reduce the TCO of next generation cloud systems and at what cost?* Angelos Arelakis, ZeroPoint Technologies AB, Sweden	Discussion 15 (Huvudskär) *Creating a roadmap for cyber-physical cloud* Johan Eker, Lund University and Ericsson Research, Sweden & Gerhard Fohler, RPTU Kaiserslautern, Germany
15:35	**Coffee**				

16:00 Social outdoor activities
See 2 of this program for information about opportunities and available equipment

18:00 **Ferry Departure for Barbeque Dinner**

22:00 **Ferry returns to hotel**

Thursday June 19th

7:00 **Breakfast**

Session Chair: Frédéric Desprez, NRIA Grenoble - Rhône-Alpes, France

8:30 *Distilling The World Through Private Ubiquitous Computing*
John Kubiatowicz, UC Berkeley, USA

9:00 *A Fractionalized Supply Chain of Varying Yield - One Hyperscaler's View of Capacity*
Narayan Desai, Google, USA

9:20 *Predicting Failures of Autoscaling Distributed Applications*
Mauro Pezze, Università della Svizzera Italiana, Switzerland

9:40 **Coffee & Hotel Room Check-out (before 10.00)**

Session Chair: Catrin Granbom, Ericsson, Sweden

10.10 *The Value of Information in Multi-scale Management: The Cloud Case*
Ada Diaconescu, Télécom Paris, France

10:30 *Embracing non-linearity for Cloud Control*
Thijs Metsch, Intel Labs, Germany

10:50 *Optimizing the Full Lifecycle of Large Language Models: Balancing System Trade-Offs*
Jie Xu, University of Leeds, UK

11:10

Discussion 16 (Resarö)	Discussion 17 (Rindö)	Discussion 18 (Lillskär)	Discussion19(Storskär)	Discussion 20 (Huvudskär)
An exabyte a day: throughput-oriented, large scale, managed data transfers with Effingo Krzysztof Rzadca, Google, Poland	*Unpredictability in the Cloud: Sources and Solutions?* Ada Diaconescu, Télécom Paris, France	*Civil and military resilience through Cloud technology* *Exploring the potential of cloud robotics: Resource allocation, challenges, and platforms* Per Skarin, Saab Kockums, Sweden & Mikael Lindberg, Saab Kockums, Sweden	*Edge orchestration: Challenges, principles & patterns* Stefan Vallin, Avassa, Sweden	*Open discussion opportunity*

12.30 **Lunch**

Session Chair: Maria Kihl, Lund University, Sweden

13.45 *Toward Energy-Optimal AI Systems*
Mosharaf Chowdhury, University of Michigan, Ann Arbor, USA

14:05 *Towards Fine-grained Agile Elasticity in Inference Serving*
Alexey Tumanov, Georgia Institute of Technology, USA

14:25 *Software Reliability in Emerging Cloud Computing Paradigms*
Tianyin Xu, University of Illinois Urbana-Champaign, USA

14:45 *Closing*
Erik Elmroth, Umeå University and Elastisys, Sweden

14:50 **Coffee**

15:15 **Bus departure**

16:30 **Bus arrival at Arlanda airport, Terminal 4**

Abstracts for
Presentations, Posters, and Discussion Sessions
in order of Appearance
18th Cloud Control Workshop
June 17– 19, 2025

Tuesday, June 17th

10.10 *Workshop Introduction*
Erik Elmroth, Umeå University and Elastisys, Sweden

10:20 *10 years on from the first Borg paper*
John Wilkes, Google, USA

The first Borg paper from 2015 just won the EuroSys 10-year test of time award. The second one ("Borg: The Next Generation") came out in 2020. I thought it would be fun to complement a few slides from the first two, to set the scene, with a much less-widely seen short talk on "Lessons learned from Omega" (Omega = 'brave new Borg'), which was given at HotOS'2021.

10:50 *Decarbonizing Cloud Computing: Opportunities and Tradeoffs*
Prashant Shenoy, University of Massachusetts Amherst, USA

The exponential growth of cloud computing has been a defining trend of our time, fueled by rapidly growing demands from data-intensive and machine learning workloads. Despite the end of Denard scaling, the cloud's energy demand grew more slowly than expected over the past decade due to the aggressive implementation of energy-efficiency optimizations. Unfortunately, there are few significant remaining optimization opportunities using traditional methods, and moving forward, the cloud's and AI's continued exponential growth will translate into rising energy demand, which, if left unchecked, will translate to an increasing carbon footprint.

In this talk, I will discuss recent developments in decarbonizing cloud platforms and workloads. I will discuss how AI workloads have contributed to the rising demand for cloud computing and the promise that AI holds for enhancing the sustainability of cloud platforms. I will discuss how recently proposed approaches for designing carbon-efficient systems introduce new tradeoffs in performance, cost, and energy-efficiency when reducing the carbon footprint of modern cloud applications. I will present approaches for navigating these tradeoffs and end with open research challenges in the emerging field of computational decarbonization.

11:10 *Smart Agentic Memory: Challenges and Opportunities*
Jayaram Radhakrishnan, IBM Research, USA

AI agents are rapidly becoming integral to the control and management of cloud infrastructures, supporting tasks ranging from dynamic resource allocation to periodic maintenance and root-cause analysis. As the deployment of AI agents across datacenters, edge clouds, and fog infrastructures grows, the need for robust memory systems to support these agents is becoming critical.

In this talk, I will explore the emerging concept of smart agentic memory — memory systems designed to persist, organize, and reason over long-lived agent knowledge. Such memory systems must not only store domain knowledge and user interactions across time, but also enable agents with limited context windows to operate effectively over large datasets, personalize their actions based on user preferences, manage information about past failures, and enforce behavioral guardrails. These use cases pose significant new challenges for cloud-hosted agent platforms, including managing heterogeneous backing stores, dynamically optimizing placement and retrieval, minimizing computational costs, and ensuring system reliability and privacy at scale.

This talk will frame smart agentic memory as a new frontier for research at the intersection of AI, systems, and cloud control, outline key open problems and invite discussion on how to architect the memory foundations for the next generation of cloud-based AI agents.

11:30 *Synchrony in Datacenters?*
Patrick Eugster, Università della Svizzera italiana (USI), Switzerland

For decades distributed systems have been conservatively assumed to be asynchronous, leading to complex and brittle solutions for distributed coordination. With the advent of advanced hardware for "programmable networks" we believe that it is time to revisit this assumption in the context of datacenters.

11:50 **Plenary Discussion**
What I Always Wanted to Know
Thijs Metsch, Intel Labs, Germany

One of the unique features of the cloud control workshop series is the diversity of its participants—bringing together researchers, engineers, and practitioners from academia and industry alike. This discussion session leverages that diversity to spark a dynamic, curiosity-driven conversation under the theme "What I Always Wanted to Know."

Participants are encouraged to bring forward the questions they've always wanted to ask—but rarely had the forum to do so. From "How do you solve this in practice?" to "Is this research direction useful to industry?", all types of questions are welcome. The session aims to uncover knowledge gaps, clarify assumptions, and identify unmet needs on both sides.

Together, we'll collect and prioritize a set of actionable research directions and collaboration opportunities rooted in real-world challenges and cross-disciplinary insight.

13:45 *Applications of Multivariable Feedback Control in Online Advertising*
Niklas Karlsson, Amazon, USA

Online advertising is the core business model for many well-known large companies and involves the management of so-called ad campaigns. Ad campaigns in programmatic advertising are typically defined as extremely high-dimensional optimization problems. The solution to these problems relies heavily on cloud control with extensive use of distributed real-time control and bidding. In particular, the optimization problems are modularized in a manner that makes feedback control a critical component of the solution. However, the control problem is challenging due to plant uncertainties, nonlinearities, time-variance, and noise. Multi-constraint optimization problems are especially difficult to solve via feedback control because of the dynamic interaction across feedback loops. In this brief presentation I will demonstrate how the multivariable control problem appears from the optimization problem and I will highlight some of the many intriguing challenges to deal with in order to solve the problem. I will also showcase a solution to one special case of the optimization problem. My wish is that the presentation will inspire the audience to work on some of the more general problem statements.

14:15 *Enabling Secure Rolling Updates for Confidential Kubernetes*
Christof Fetzer, TU Dresden, Germany

Enabling confidential computing across entire Kubernetes clusters ensures data remains encrypted at all times for cloud customers. This approach, adopted by major cloud providers such as Microsoft Azure and Google Cloud, significantly enhances security and privacy. However, state-of-the-art systems face challenges in providing secure rolling updates, particularly in managing complex coordination across multiple nodes, ensuring compatibility between different hardware and software components and maintaining service integrity and availability during the update process. This talk will present our solution based on the confidential computing framework SCONE to address these challenges and enable secure rolling updates for confidential Kubernetes clusters. This solution is deployed in context of confidential applications that serve millions of end users.

14:35 **Discussion 1** (Resarö)
Edge systems: Overestimated opportunity or underestimated necessity?
Aaron Ding, TU Delft, The Netherlands & Marteen Van Steen, University of Twente, The Netherlands

Active communities around the 'Edge' have been growing in recent years, running from conferences such as ACM SEC, IEEE EDGE, EdgeSys to Dagstuhl and Lorentz workshops. The subject of 'edge' emerged as a paradigm to bridge cloud and IoT/mobile systems, whose roots can be traced back to the 2000s. Despite all the efforts to explore this new paradigm, numerous concerns have raised in the systems community, to name a few here:

- If latency and bandwidth are not a problem (which is often the case), why do we need edge systems?
- Should edge systems really be anything more than just intelligent caches? Why would we want more if that means more complexity?
- What do we really want from the edge? Are they for decentralization (and if so, why do we want to decentralize through edges) or are they here for just management and control?
- Is this edge-cloud continuum an artificial problem, or an unavoidable consequence of systems design?
- Will exploitation of AI on edge systems lead to a sustainable decentralization, or will it drive us into anarchy (given the security concerns for edge devices)?

Join us to discuss these concerns. We encourage you to bring your insights and questions to what promises to be a lively and informative session. The goal of this discussion session is to get a better picture of the issues we should really focus on when it comes to edge systems research. In slightly provocative terms: moving away from things we can do, to things we should do.

14:35 **Discussion 2** (Rindö)
Attacks and defences in 5G and beyond (edge) infrastructures (with or without ML)
Simin Nadjm-Tehrani, Linköping University, Sweden

The future of the modern society is tightly intertwined with accessibility of the communication networks and transmission of data to and from cloud-based services. Current trends in development of communication infrastructures point in two directions: more software with open interfaces and (potentially) more deployment of AI.
The future of AI deployments in critical sectors is at a crossroad due to a number of open issues on which researchers and policy makers by and large agree. Among these is the lack of robustness due to adversarial use of benign AI models, lack of reliable methods to quantify the risk of unexpected model failures, and lack of monitoring methods to ensure secure operation of AI-based systems. It is not clear how to use AI to successfully monitor AI-based systems, but it is certain that attackers will use AI methods for more effective attacks and to do it faster.
Using more software in future generations of communication networks (5G and beyond) is, however, certain. The disaggregation of access and core services and the subsequent disaggregated fronthaul from multiple vendors, as promoted by the O-RAN Alliance, connected through switched Ethernet topologies brings new access and attack surfaces. There are opportunities for new spoofing, denial of service (DoS), time-desynchronisation, and replay attacks.
In this session we would like to discuss the potential and existing research to address some of these problems. With the added complexity of open interfaces and use of complex orchestrators to manage load and optimise performance, especially with deployed AI components, we can consider answers to any of the following questions:

- How to make sense of adverse behaviour at run-time?
- What are the most urgent attacker strategies we need to find effective means to block?
- How to distinguish benign but unusual load changes from DoS attacks with reasonable certainty and what mitigation options to pursue?
- How to collaborate across academy-industry to create a safe space for studying attacks and defences?
- What specific AI-generated attacks are to prioritise?
- What specific attacks to AI models or their capabilities (e.g. by submitting unintended data) to study? How to mitigate them?

14:35 **Discussion 3** (Lillskär)
Cloud Intelligence: AI/ML for Cloud & Cloud for AI/ML
Sara Bouchenak, INSA Lyon, France, Jian Zhang, Microsoft, USA & Cristian Klein, Umeå University and Elastisys, Sweden

The purpose of this discussion is to raise the following questions:

AI/ML for Cloud:

- What are the challenges and opportunities in landing AI/ML in cloud services for better resource and operation efficiency, cloud security, service quality, and customer experience?
- What are the successful applications, challenges and opportunities in leveraging LLMs and generative AI for cloud control?

Cloud for AI/ML:

- What are the challenges and opportunities in devising new cloud services and technologies for more efficient, sustainable, and trustworthy AI/ML?
- What are the successful applications in leveraging novel cloud services for efficient, sustainable, and trustworthy AI/ML?

And overall:

- What can we do to better raise student awareness in the usage of AI/ML & cloud and its implications?
- What can we do to accelerate collaboration across academia and industry in cloud control?

14:35 **Discussion 4** (Storskär)
Predictable Energy-Efficient Cloud/Edge Infrastructures
Tommaso Cucinotta, Scuola Superiore Sant'Anna, Italy

Cloud Computing is attracting more and more use-cases, including several performance-critical ones with tight constraints on the end-to-end latency. These applications need a stable performance, which is threatened by several factors: the wide heterogeneity of hardware resources available within a large Cloud infrastructures, and especially within Cloud/Edge scenarios, including different CPU/memory/storage characteristics, the availability of hardware accelerators such as GPUs, TPUs, DPUs, FPGAs, etc..., and the need for leaving high flexibility degrees to the provider in resources allocation; the advantages arising from workload consolidation; and, the increasingly important need for exploiting energy-efficiency mechanisms available on the platforms, including DVFS, deep-idle states, and power-hungry/energy-efficient cores with compatible ISA.

These problems may be tackled by traditional mechanisms that can provide a stable performance in Cloud data centers, such as: allowing for a fine-grained selection of hardware resources types, locations and their acceleration characteristics; horizontal elasticity to adapt the number of instances implementing a service so to match desired response times characteristics; real-time scheduling of CPU supporting kernel/hypervisor guarantees in co-scheduling of VMs or containers on the same CPUs; providing access to isolated/dedicated instances, sacrificing multi-tenancy on individual hosts; disabling energy-saving features on the physical hosts, such as DVFS and/or deep-idle states, to provide stable computing abilities; using kernel-bypass techniques like DPDK, driving NICs directly from user-space drivers, keeping CPUs fully loaded at 100% in poll-based cycles, saving the overheads of complex in-kernel packet-processing and additional context-switch and wake-from-idle overheads, maximizing performance but effectively causing huge power consumption even in low-traffic conditions.

A variety of mechanisms are still lacking to be able to exploit effectively power-saving abilities available in nowadays complex Cloud/Edge infrastructures, still with the ability to deliver the needed performance whenever deploying time-sensitive/critical applications and services: richer provider/customer/consumer interfaces and APIs might be introduced to allow providers to save power whenever the desired end-to-end latency is within the specification, or to expose to the customers visibility of the power tunables of the underlying platform, or even more to allow for customizing the OS/hypervisor software stack tunables for the needed performance/power-efficiency trade-offs (e.g., using a PREEMPT_RT vs PREEMPT_VOLUNTARY kernel on the host); richer billing models might be used to encourage the use of power-saving techniques at the expense of performance and end-to-end latency, whenever this is convenient from an energy-efficient standpoint, e.g., billing not only based on the instance running times, but also based on the consumed energy over the billing period, so to push towards the use of reduced-power mechanisms.

14:35 **Discussion 5** (Huvudskär)
Distributed Computing Challenges in Blockchain and Cloud Computing
Yvonne-Anne Pignolet, DFINITY Foundation, Switzerland & Rodrigo Rodrigues, INESC-ID, ULisboa, Portugal

Blockchains offer a secure way to store the most important state of a distributed system without requiring trust in any centralized entity, by providing a fully distributed and immutable ledger of transactions.

Blockchains and cloud computing have the potential to work in a complementary and synergistic way, for instance by using blockchains to tackle the issues that arise from having to rely on a centralized cloud service provider (CSP). Furthermore, both classes of systems are instances of distributed computing systems running untrusted software with untrusted users, where the same underlying principles and techniques apply. Blockchains can bring security, transparency, and decentralization to cloud computing, while cloud computing approaches for scalability and efficiency can inspire blockchain systems.

In this discussion, we will explore these synergies, guided by the following questions:

- What messaging guarantees between applications running on different blockchains or CSPs are useful and efficiently implementable?
- How can resource allocation (blockspace, CPU, memory) be managed fairly while maintaining simple and stable fee structures for users?
- How can we ensure the integrity of workload migration in Byzantine environments
- How can incentive mechanisms commonly used in blockchain systems be reconciled with traditional BFT consensus protocols?
- In the communication layer of blockchains and cloud platforms, how can we effectively balance user and node anonymity with performance requirements?

16.25 *SLOs (Slow down, & Service Level Objectives) before PROs (Processors, Performance, etc.)*
Indranil Gupta, University of Illinois Urbana-Champaign, USA

Today's world — which is focused on maximizing accuracy and reducing costs of LLMs, I mean of cloud computing — is a full circle back to the 1990s and 2000s when processors and their performance were all the rage in industry. That rage died down, and we are left with unpleasant vestiges and (some pleasant) remnants — developers still wondering how much resources they need, bugs, systems that don't work well under failure or scale well, and systems that can still not be used by most humans without plenty of training. The current rage in LLMs, I mean cloud computing, is opening a similar world where these ML tools (optimized for performance) will pervade our lives, and most humans will continue to have no clue how to use them.

This is a call for work on systems-based work in cloud computing that is more closely connected with and motivated by human-facing aspects of cloud computing.

16:45 *Programmable Data Plane 101: a Network Switch with More Computing Power Than Your Cloud Server*
Tim Wood, George Washington University, USA

In recent years, networks have grown in power, providing not only connectivity between hosts, but the opportunity to monitor and manipulate data as it flows through cloud networks. This talk will provide a brief introduction to key programmable data plane technologies, focusing on programmable switches (e.g., PISA/Tofino switches) and network cards (SmartNICs and NetFPGAs). These technologies have the potential to change cloud management approaches including load balancing, telemetry, security, and performance monitoring. For example, recent programmable switches have the capability to perform computation on billions of packets per second, which is several orders of magnitude greater than a high performance server. While these devices can provide impressive performance, their restrictive programming model and tiny memory resources require new approaches to software and data structure design. The talk will survey the opportunities and challenges in using programmable data plane devices in cloud environments.

17:05 ***Poster Reception with Refreshments*** (Svavelsö)

Poster 1. *Control Offloading for UAVs*
Emil Sundström, Lund University, Sweden

Dynamically offloading computationally heavy control algorithms to an external data centre comes with numerous challenges, one of which is the resource allocation task. With limited external computation capacity, poor resource management will lead to a decreased system performance. Therefore, my research focuses on how game theoretical approaches can be used to solve the resource allocation problem involving multiple devices within a constrained area.

Poster 2. *A Stable Matching Approach to Energy Efficient and Sustainable Serverless Scheduling for the Green Cloud Continuum*
Yashwant Singh Patel, Umeå University, Sweden

Cloud infrastructures are evolving from centralised systems to geographically distributed federations of edge devices, fog nodes, and clouds - often known as the Cloud-Edge Continuum. Continuum systems are dynamic, often massive in scale, and feature disparate infrastructure providers and platforms; this greatly increase the complexity of developing and managing applications. The Serverless paradigm shows the potential to greatly simplify the process of building Continuum applications - however, current scheduling mechanisms for Serverless Continuum platforms pay little attention to reducing the energy consumption and improving the sustainability of function execution. This is a significant omission, made worse as computing nodes within a Continuum may be powered by renewable energy sources that are intermittent and unpredictable, making low-powered and bottleneck nodes unavailable. There is great opportunity to design a decentralized energy management scheme for scheduling Serverless functions that takes advantage of the different layers of the Continuum, such as IoT devices located at the Edge, on-premises clusters closer to the data sources, or directly on large Cloud infrastructures. To achieve this, we formally model a green energy-aware Serverless workload scheduling problem for the multi-provider Cloud-Edge Continuum. We then design stable matching-based technique for decentralized energy management (utilising a distributed controller) that considers the availability of green energy nodes and the QoS requirements of Serverless functions. We prove the complexity, stability and termination of the proposed heuristic algorithm, and also compare its performance with baseline scheduling techniques.

Poster 3. *Offloading of Time-Optimal Motion Planning with Jerk Constraints*
Souras Ahmed Al Bayati, Lund University, Sweden

In this Poster offloading an advanced motion planning pipeline consisting of a path planner and a trajectory planner is investigated. The path planner generates an obstacle-free geometric path, while the trajectory planner computes time-optimal trajectories which respect velocity, acceleration, and jerk constraints for path tracking. The complete motion planner is integrated into an offloading framework, deployed on an edge cluster, and communicates with a robot via 5G to evaluate its performance on a physical system. Results show that when offloaded, the advanced motion planner achieves execution speeds comparable to the heuristic approach running directly on the device.

Poster 4. *One Shot Federated Learning with Classifier-Free Diffusion Models*
Obaidullah Zaland, Umeå University, Sweden

Federated learning (FL) enables collaborative learning without data centralization but introduces significant communication costs due to multiple communication rounds between clients and the server. One-shot federated learning (OSFL) addresses this by forming a global model with a single communication round, often relying on the server's model distillation or auxiliary dataset generation - mostly through pre-trained diffusion models (DMs). Existing DM-assisted OSFL methods, however, typically employ classifier-guided DMs, which require training auxiliary classifier models at each client, introducing additional computation overhead. This work introduces OSCAR (One-Shot Federated Learning with Classifier-Free Diffusion Models), a novel OSFL approach that eliminates the need for auxiliary models.

OSCAR uses foundation models to devise category-specific data representations at each client which are integrated into a classifier-free diffusion model pipeline for server-side data generation. In our experiments, OSCAR outperforms the state-of-the-art on four benchmark datasets while reducing the communication load by at least 99%.

Poster 5. *Automated Updates Over LoRaWAN for Energy-Efficient Edge AI*
Clément Courageux-Sudan, Umeå University, Sweden

This paper proposes an approach to update edge applications over LoRaWAN, addressing distinct and significant challenges compared to typical scenarios over Wi-Fi or cellular networks. First, we fully automate the FUOTA process to propagate model updates to selected machines with heterogeneous communication capabilities. Then, we propose three parameter selection policies to balance the energy consumed by end devices and the time taken to update entire networks. An evaluation on a hardware testbed demonstrates the effectiveness of our approach in propagating updates automatically. Large-scale simulations of up to 300 end devices show that our energy-oriented update policy reduces the energy consumption of end devices between 2.2x and 2.7x compared to default baselines.

Poster 6. *Decentralised Carbon-aware Scheduling for Large-scale Microservice-based Infrastructures*
Yangyang Wen, Umeå University, Sweden

In cloud-edge systems, especially in microservice architectures, as the number of services continues to grow, traditional centralized scheduling systems face scalability bottlenecks. Since all requests rely on the central scheduler's decision, the system's response efficiency is limited. If each service can make independent decisions based on local information, the system as a whole will have higher scalability and elastic processing capabilities.
To this end, we proposed a decentralized carbon-aware scheduling algorithm, using a mainstream centralized scheduling algorithm as a control benchmark, and designed comparative experiments to evaluate the differences between the two in multiple dimensions such as carbon emissions, energy consumption, and latency under the same topology and request load conditions. The results reveal the potential advantages of decentralized scheduling in terms of performance and carbon emission balance, and further verify its feasibility and scalability in dealing with large-scale microservice scenarios.

Poster 7. *Keeping Apps Afloat: Predicting and Mitigating Shared Resource Bottlenecks in the Cloud*
Oliver Larsson, Umeå University, Sweden

Modern clouds see large numbers of applications co-located together on a single host. Advantages are of course many, including improved utilization and cost-effectiveness. However, congestion is likely to happen in shared system resources such as last level cache (LLC) and the memory bus, affecting applications KPIs and ultimately triggering cluster-level scaling events at lower-than-optimal utilization levels. By analyzing workloads at runtime, we attempt to predict resource bottlenecks and take mitigating action before cluster-level interception is required. This allows higher levels of utilization to be reached while ensuring workload performance targets are met.

Poster 8. *A Fault-tolerance Strategy for Mission-Critical Application at the Edge*
Tanaz (Nayereh) Rasouli, Umeå University, Sweden

Mission-critical applications, such as real-time healthcare, and transport systems, depend heavily on low latency and reliability provided by Mobile Edge Computing (MEC). The failure of these applications can lead to severe consequences, including loss of life, financial catastrophe, or operational disruption. However, the dependability of edge clusters, particularly in terms of placement and fault recovery, crucial to these applications, is often overlooked. In this work, we focus on loosely coupled IoT applications and propose a fault-tolerance scheduler for edge that balances the trade-off between the availability of edge clusters and the latency of containerized mission-critical applications. We considered estimation latency between IoT devices/users to nodes. Our results demonstrate that this approach maintains a high availability of edge clusters during failure while minimizing latency for containerized mission-critical applications. To validate the proposed approach, we compared it to standard Kubernetes.

Poster 9. *Over-sampling based Control & Aperiodic sampling*
Max Nyberg Carlsson, Lund University, Sweden

First half of poster: Some advanced control algorithms take long time to calculate compared to plant dynamics. Their latency inhibits the usage in practical applications, as control performance deteriorates. We propose interlacing the output of multiple controllers, and sample faster than the execution time. This way, distributed computing resources can be utilized for improved control.

Second half of poster: Most controllers sample and actuate periodically. While easy to analyze, their period must be lower bounded by the worst case latency and therefore may be conservative. But why is it not the norm to sample immediately after actuation is done? Preliminary experiments indicate potential for better control performance compared to classic periodic control.

Poster 10. *Cloud Continuum Resource Monitoring*
Angela Checa, Umeå University, Sweden

Cloud continuum can be considered as the natural evolution of cloud computing. A consequence of the cloud adoption expansion, massive end devices and hyper connectivity that opens the gate for a new generation of services and applications deployed transparently across enormous resource ecosystems. An aggregation of heterogeneous clouds as such generates operational data at large scale that requires autonomous capabilities for proactive monitoring and smart management of all its components. Manual monitoring, legacy tracking utilization and offline prediction methods are no longer feasible to effectively harness the potential of the cloud continuum. Therefore, a lightweight monitoring mechanism of streaming operational data in large scale systems is a critical aspect to address the limitations of the current approaches. The challenge is to develop an inter-cloud monitoring overlay system that reacts to the dynamic changes and perform accurate predictions for making intelligent decisions regarding resource allocation, workload placement and energy consumption at the same time minimizing the overload of its own monitoring data to avoid saturation on the overall continuum.

Poster 11. *Energy and QoS Aware Deep Learning Model Selection in Edge Computing*
Hajar (Maryam) Siar, Umeå University, Sweden

Edge intelligence aims to enable the execution of deep learning applications on edge computing platforms. These applications often have specific Quality of Service (QoS) requirements, such as deadline and average accuracy, while edge servers are typically heterogeneous and constrained in terms of resources and energy. Energy efficiency can be managed by considering the number of floating-point operations (FLOPs) required by deep learning models. Additionally, the availability of models with varying sizes provides flexibility in selecting appropriate models to serve inference requests. This paper investigates efficient resource allocation and scheduling for deep learning model selection in a multi-edge server environment. The objective is to minimize the total energy consumption of the system while maximizing the number of completed inference requests. To address this, we introduce a polynomial-time many-to-many stable matching algorithm and explore its behavior under different preference functions defined for both inference requests and edge replicas. Experimental results across various scenarios demonstrate the effectiveness of the proposed approach in improving request completion rates and reducing total energy consumption in edge computing system.

Poster 12. E*fficient Scheduling to Reduce Energy use in a Heterogeneous Cluster*
Ileet Malla, Umeå University, Sweden

The rapid growth of data centers has made them major consumers of electrical power, raising concerns about operational costs and environmental impact. One promising way to address these challenges is to explore the performance heterogeneity of servers within a cluster. A workload can exhibit different energy footprints depending on which nodes they are executed. This work focuses on scheduling workloads in a cluster of nodes with heterogeneous performance profiles, based on high-level objectives referred to as intents. Intents are declarative specifications of workload such as latency, that guide how resources should be allocated. The target is to build a scheduler that schedules workloads to suitable nodes by considering their intent and node performance profiles, aiming to meet performance objectives while minimizing energy consumption. This approach aims to enable more adaptive, efficient and sustainable workload management in a cluster.

Poster 13. *Adaptive Deviation Learning for Visual Anomaly Detection with Data Contamination*
Anindya S Das, Umeå University, Sweden

Visual anomaly detection targets to detect images that notably differ from normal pattern, and it has found extensive application in identifying defective parts within the manufacturing industry. These anomaly detection paradigms predominantly focus on training detection models using only clean, unlabeled normal samples, assuming an absence of contamination; a condition often unmet in real-world scenarios. The performance of these methods significantly depends on the quality of the data and usually decreases when exposed to noise. We introduce a systematic adaptive method that employs deviation learning to compute anomaly scores end-to-end while addressing data contamination by assigning relative importance to the weights of individual instances. In this approach, the anomaly scores for normal instances are designed to approximate scalar scores obtained from the known prior distribution. Meanwhile, anomaly scores for anomaly examples are adjusted to exhibit statistically significant deviations from these reference scores. Our approach incorporates a constrained optimization problem within the deviation learning framework to update instance weights, resolving this problem for each mini-batch. Comprehensive experiments on the MVTec and VisA benchmark datasets indicate that our proposed method surpasses competing techniques and exhibits both stability and robustness in the presence of data contamination.

Poster 14. *Spilling the Load: A Hydraulic Model Predictive Controller for Distributed Serverless Platforms*
Chanh Nguyen, Umeå University, Sweden

Serverless computing offers fine-grained, event-driven execution and has gained substantial traction due to its simplicity and elasticity. However, it faces two persistent challenges: (i) cold start delays, where a function experiences significant initialization latency before handling requests, particularly severe for customized or blackbox functions not based on provider-managed runtimes; and (ii) load imbalance across distributed servers, especially in heterogeneous cloud–edge deployments, where some nodes may become overloaded while others remain underutilized.

To address these issues, we introduce a novel hydraulic control model for serverless systems. Drawing inspiration from hydroelectric spillway control, we model each data center as a dam and workloads as inflowing water. Using this analogy, we develop a Model Predictive Control (MPC)-based offloading and prewarming strategy that dynamically decides when to "open the gates" (offload or prewarm functions) and where to redirect requests (water flow) across the cloud–edge continuum. Our approach reduces cold start impact, improves load distribution, and maintains SLA targets, as demonstrated through a real OpenWhisk-based serverless testbed deployed across distributed edge and cloud nodes.

Poster 15. *Energy Forecasting with SNN-Transformer Integration for Cloud and Edge Environments*
Zulfiqar Ahmad Khan, Umeå University, Sweden

Power forecasting is critical for efficient resource management in cloud and edge environments, where energy consumption patterns are highly dynamic and distributed. This work introduces an adaptive energy forecasting model that integrates the Spiking Neural Network (SNN) with transformer architectures to address forecasting challenges in such complex systems. The proposed model integrates the event-driven nature of SNNs, which effectively capture sudden spikes and irregularities in energy consumption, with the transformer self-attention mechanism, which models long-term temporal dependencies. The SNN-Transformer integration enables the model to adapt to real-time fluctuations in energy consumption, providing efficient and effective forecasting, while its ability to learn from short-term variations and long-term trends makes it well-suited for cloud and edge environments, where energy management needs are continuously evolving. However, cloud-specific energy consumption data is not yet publicly available, therefore, alternative power consumption datasets are used for the model evaluation. Experimental results demonstrate that the proposed model outperforms traditional forecasting models, leading to more effective energy management, reduced operational costs, and improved resource allocation in dynamic computing environments.

Poster 16. *Uncover Compromised AI Functions in a Serverless Device-Edge-Cloud Continuum*
Adil Bin Bhutto, Umeå University, Sweden

AI (Artificial Intelligence) functions deployed as serverless functions as a service (FaaS) in the device-edge-cloud continuum are increasingly vulnerable because functions are short-lived, hard to isolate continuum layers, multi-layered data, and so on. This vulnerability is compounded by the reliance on data obtained from cross-layers and the dynamic, distributed infrastructure, making conventional detection methods insufficient. This work uncovers the compromised AI functions through system logs analysis using transformer based models. This approach enables rapid, automated alerts for compromised functions to strengthen security for stateful AI workloads in serverless device-edge-cloud environments.

Poster 17. *Energy-efficient Right-sizing of Kafka-like Message Brokers*
Govind KP, INRIA, France

Modern data-intensive applications depend on reliable, high-performance streaming data brokers like Apache Kafka and Redpanda. However, operators often over-provision resources to ensure reliability, leading to inefficiency.

We present a practical methodology for right-sizing Kafka and Redpanda deployments, balancing performance and energy efficiency. Our analysis shows that storage technology is the main factor in broker performance, with both systems scaling throughput nearly linearly and remaining I/O-bound.

Notably, Kafka's energy use plateaus at moderate loads, while Redpanda's scales more proportionally. Our approach predicts performance at scale and provides actionable guidelines for efficient, cost-effective broker deployments

Poster 18. *Measure and Protect: Enhancing the Web PKI*
Carl Magnus Bruhner, Linköping University, Sweden
Certificates are the foundation of secure communication over the Internet, but not all certificates are created and managed in a consistent manner. Certificate authorities (CAs) issuing certificates achieve different levels of trust, and user trust in public keys, certificates, and CAs can quickly change. This has made careful certificate management both an important and challenging problem.

Our work presents novel characterization of certificate replacement (CR) relationships in the wild including reuse of public keys, head-to-head comparison of CRs where the replaced certificate was revoked vs. not revoked, and longitudinal analysis of certificate chains for popular domains, examining their evolution over time and across categories.

Based on the results, we highlight shortcomings in existing revocation protocols and practices, emphasizing the need for improvements, proposing efficient solutions, and discussing other ongoing efforts aimed at addressing these problems.

Poster 19. *Integrated Modelling of Energy & Cloud-Edge Systems for Carbon Analysis,*
Rohail Gulbaz, Umeå University, Sweden

The Edge-Cloud Continuum has evolved into a vast, loosely coupled system involving multiple providers, dynamic infrastructures, and conflicting objectives. As reliance on these complex federated Continuum systems grows, the energy consumption of data centres has risen significantly. To address this increasing distributed computing load more effectively across distributed power infrastructures, integrated models that link the Cloud Continuum and Smart Grids are needed. Different scheduling strategies are engineered to demonstrate how knowledge from both the Smart Grid and Edge-Cloud Continuum can be leveraged for carbon-aware and green energy-sensitive management. Our simulation-based evaluation highlights the impact of focusing on independent and energy-integrated QoS on carbon footprints.

Poster 20. *Tiny Minds, Big Shields: Detecting DDoS across the Cloud-Edge Continuum Systems*
Yinuo Zhang, Umeå University, Sweden

The cloud–edge continuum is rapidly becoming the backbone of critical infrastructures—spanning education, transportation, and healthcare—where distributed components must process data and make decisions in real time. Emerging threats like Distributed Denial-of-Service (DDoS) attacks exploit both cross-layer services and resource bottlenecks across these systems, and benign and malicious loads often appear indistinguishable until service quality degrades. To address this, we propose a lightweight LLM—tinyBERT—to detect DDoS attacks by analyzing the semantic behavior of system traffic. Our experiments show that tinyBERT achieves high detection accuracy while significantly reducing inference time compared to state-of-the-art methods.

Poster 21. *Efficient Cluster Scheduling for Fine-tuning LLMs Using Historical Configuration Data*
Matthijs Jansen, Vrije Universiteit Amsterdam, The Netherlands
Job scheduling for machine learning (ML) has received significant attention, targeting objectives such as job completion time, utilization, and fairness through techniques such as heterogeneity, elasticity, and task merging. However, the quality of the configuration space in which these scheduling policies make decisions and choose between GPU placements and hyperparameters for jobs is under-explored. We find that performance between configurations for the same ML fine-tuning job can differ over 40x, and, as a result, schedulers with under-explored views of the configuration space cannot find those configurations that optimize their objective, regardless of the quality of the scheduling policy. Hence, we make a case that configuration knowledge should be treated as a first-class citizen in ML schedulers. We propose CoTune, a scheduling framework for ML fine-tuning that exposes various configuration knowledge classes, qualities, and quantities for scheduling policies to integrate and evaluate. As a foundation, we propose a methodology for defining, gathering, and predicting configuration knowledge, which we apply to build a comprehensive configuration knowledge database. Using production traces on a simulated GPU cluster, we demonstrate how full configuration knowledge from our database reduces average and tail job completion time for fine-tuning jobs by 73.9% and 79.0% across policies while decreasing average GPU utilization by 27.9%.

Poster 22. *Accurate and Low-Overhead Workload Prediction for Cloud Management Systems*
Lidia Kidane, Umeå University, Sweden

This study proposes a comprehensive data-driven framework for automating workload prediction in cloud computing, primarily focusing on minimizing the overhead associated with such predictions. This is critical because the cost of maintaining and executing workload prediction models is solely on the resource providers, making efficiency a key priority. The framework addresses key challenges, including adaptability to workload variability, predictive accuracy, computational efficiency, storage optimization, and knowledge reusability. A knowledge base is employed to store and reuse prediction models to meet these challenges, enabling rapid adaptation to workload changes while avoiding redundant computations. Additionally, historical workload data is compressed using a pre-trained autoencoder, significantly reducing storage requirements without compromising the quality of information needed for accurate predictions. These methods are integrated into an extensive framework, systematically evaluating the impact of various configurations on model accuracy, computational performance, and resource utilization. Results demonstrate the ability of each component to minimize operational costs while maintaining high prediction accuracy, offering substantial benefits to resource providers. By reducing the overhead of workload prediction, the proposed framework enhances operational efficiency and improves service quality and cost-effectiveness in dynamic cloud ecosystems

Wednesday June 18th

8:30 *Re-imagining Cloud-Scale Telemetry from an Approximation-First Perspective*
Vyas Sekar, Carnegie Mellon University, USA

Telemetry systems are widely used to collect data from distributed endpoints, analyze data in conjunction to gain valuable insights, and store data for historical analytics. With increasing volumes of data to be collected and analyzed, and the increasing need for real-time analytics, telemetry costs are rising across the stack. Thus, simply collecting all data, transmitting it for analysis, and storing it exactly has become prohibitively expensive. Instead of existing solutions that leverage exact telemetry or leverage myopic solutions in isolation, we take a holistic bird's eye view of the telemetry stack and considers approximation primitives like sketches as first-class primitives. We will demonstrate early results on how this paradigm unlock orders of magnitude reduction in cost without significant deployment effort.

9:00 *The Modern PC: Personal Computer, or Personal Cloud?*
Richard Mortier, Cambridge University, UK

Computation is now ubiquitous, people commonly now carrying multiple networked computation devices almost constantly. But in most cases we still operate and manage these devices as independent from each other, at best using cloud services to support limited integration within a closed ecosystem. I think it's time we thought more fundamentally about what a modern PC should be, by revisiting a range of core operating system design decisions that were made in the early days of digital computing. I will try to motivate this position, and note some of the challenges and opportunities that I think could be interesting to explore in this space.

9:20 *Foundation Models: Emerging Bottlenecks and Research Implications*
Tarek Abdelzaher, Indiana University of Illinois, USA

Recent advances in AI, such as Foundation Models and LLMs, culminate a gradual shift in science and engineering away from strong reliance on the exploitation of algorithmic and symbolic knowledge towards new data-driven approaches. While early LLMs aimed to cater to the common user, new foundation models will be increasingly application-specific and customized, possibly trained with proprietary data. How does the emerging customization of data-centric machine intelligence impact research challenges in computing? Data-centricity affects the types of bottlenecks that arise in AI training and inference. We posit that, at training time, customized intelligence will be generally bottlenecked by the data needed for training. New data sources will be used that differ from those commonly available on the Internet, raising new challenges and opportunities in efficient learning from limited data. At inference time, the bottleneck is computational, calling for improved resource economy and novel scheduling policies. The reliance on data as the primary citizen in intelligent computational systems changes the conventional task model and thus one's view of resource allocation, leading to new challenges and opportunities in resource management. The talk discusses research directions that arise in the data-centric world of customized AI, covering training, inference, and additional LLM-specific challenges.

10:10 *Controlling AI Inference from the Edge to the Cloud*
Henry Hoffman, University of Chicago, USA

This talk introduces a control-theoretic approach to managing AI inference in dynamic environments from the could to the edge---and back. By embedding adaptive controllers into neural network pipelines and structuring models for real-time flexibility, we enable inference systems that intelligently reconfigure themselves on the fly---responding to shifting workloads, resource limits, performance constraints, and even security. We discuss methods for building and training controllable AI, techniques for controlling it within larger applications, and then how such techniques introduce new side channels, which we then work to close. Given the increasing importance of AI as a workload, we advocate for controlling it for accuracy, energy, and security.

10:30 *Massivizing Computer Systems: VU on the Science, Design, and Engineering of Distributed Computer Ecosystems*
Alexandru Iosup, Vrie Universitet Amsterdam, The Netherlands

Wherever we look, our society is turning digital. Science and engineering, business-critical and government operations, and online education, shopping, and gaming rely increasingly on digitalization---on core processes powered by data, AI, and computers. For digitalization to succeed, we must integrate computer systems into larger _ecosystems_, effectively and efficiently managed. However successful until now, we cannot take computer ecosystems for granted: the core does not yet rely on sound principles of science, design, and engineering, and we already see many warning signs about the scalability, dependability, and sustainability of engineered operations. Creating a core that can tame these signs for the upcoming unprecedented scale and demand is the fundamental challenge of massivizing computer systems.

In this talk, inspired by this challenge and our experience with distributed computer ecosystems for over 15 years, we focus on understanding, deploying, scaling, and evolving computer ecosystems. Together with our team and partners, we envision achieving this through an ambitious, comprehensive research, development, and innovation program, where the key is starting from combined computer ecosystems rather than individual, small-scale computer systems.

Using illustrative use cases, we define core concepts, formulate principles, and introduce a reference architecture for computer ecosystems supporting complex workloads, and diverse resources and backend services across the computing continuum. We synthesize a framework of resource management and scheduling (RM & S) techniques, which we, alongside the community, should explore systematically in the next decade. We show early results obtained experimentally through controlled real-world experiments, long-term observation, and what-if analysis of short- and long-term scenarios, and focus on reproducible results. We introduce a new operational instrument for RM & S, digital twins that can mimic operations, and conduct and explain meaningful counterfactual experiments. On the data side, we introduce the computing continuum Memex to record long-term operational signals, and make them available efficiently and meaningfully.

When further pursued, these conceptual and practical advances enable holistic improvements to applications, services, and entire computer ecosystems. They let us hope to cure the many illnesses abundant in today's ecosystems. So this is a call to the community spanning distributed systems, performance engineering, software engineering, data engineering, and more: there is much to discover and achieve!

This vision aligns with the IPN SIG on Future Computer Systems and Networking and their Manifesto [1] (but any mistakes remain the speaker's). Many of our examples come from real-world prototyping and experimentation, grand experiments in computer systems, and/or benchmarking and performance analysis work conducted with the Cloud group of SPEC RG [2]. Last but not least, the work presented here has been performed by the Massivizing Computer Systems team at the VU and our partners, whom the speaker gratefully acknowledges.

[1] Future Computer Systems and Networking Research in the Netherlands: A Manifesto, 2022. [Online] https://arxiv.org/pdf/2206.03259
[2] SPEC RG Cloud https://research.spec.org/working-groups/rg-cloud/

10:50 *Autonomics for the Digital Continuum*
Manish Parashar, University of Utah, USA

Autonomic, data-driven workflows, such as those underlying digital twins, integrate distributed data sources with pervasively available computing resources, spanning HPC to the edge, to understand end-to-end phenomena, drive experimentation, and facilitate critical decision making. This talk will explore a digital continuum spanning resources at the edges, in HPC centers and clouds, and in-between. It will present programming abstractions that can express what data should be processed and when and where it should be processed, as well as autonomic middleware services that automate the discovery of resources and the orchestration of computations across these resources.

11:10 **Discussion 6** (Resarö)
Federated Learning and the Next Edge-to-Cloud Intelligence: Challenges, Shifts, and Opportunities
Jayaram Radhakrishnan, IBM Research, USA & Feras Awaysheh, Umeå University, Sweden

This discussion session will explore critical open questions and emerging paradigms in Edge-to-Cloud intelligence, focusing on Federated Learning (FL) as a key enabler of privacy-preserving, distributed AI. Drawing from experience across heterogeneous edge environments, aggregation architectures, and real-world deployments, the session will examine what's working, what's missing, and what's next.

Topics will include the challenges of adapting FL to streaming and temporal environments, particularly under real-time edge constraints and asynchronous updates. The discussion will also address questions around model lifecycles, including retraining schedules and strategies for mitigating catastrophic forgetting in evolving deployments. Participants will consider extending FL beyond traditional neural networks to support more complex paradigms such as graph learning, transformers, LLMs, and diffusion models in resource-constrained settings. The session will also examine scalability challenges, including strategies for efficiently operating across diverse compute tiers, from microcontrollers to cloud infrastructures. Finally, participants will reflect on the potential of adapting foundation models to edge environments through federated fine-tuning and modular learning.

11:10 **Discussion 7** (Rindö)
Is Confidential Computing Living up to its Potential?
Christof Fetzer, TU Dresden, Germany & Rüdiger Kapitza, FAU Erlangen-Nürnberg, Germany

As cloud computing has evolved, concerns regarding loss of control over outsourced services and data persist. Users fear security vulnerabilities within the cloud infrastructure, compounded by potential threats from cloud providers themselves, such as rogue system administrators. Addressing these challenges remains a major area of research across academia and industry.

In recent years, Confidential Computing has emerged as a key solution, leveraging security hardware extensions to isolate customer workloads. When appropriately implemented, this approach can safeguard customer code and data against risks posed by the cloud provider's infrastructure and the teams operating the container orchestrators and the application.

Despite its promising potential and adoption by major cloud providers, several critical questions remain:

- Is confidential computing an effective way to protect applications?
- Does it undoubtedly protect against adversaries controlling the infrastructure and the system software?
- Can we indeed trust the attestation collaterals provided by cloud providers?

- What are the associated costs in terms of performance, resources, and energy consumption? Can they be tolerated?
- Does Confidential Computing introduce complexity by necessitating specific workload structuring that prevents widespread adoption?
- Will Confidential Computing evolve into a standard commodity within Cloud offerings?
- Beyond enhanced security, what additional benefits could Confidential Computing unlock for cloud-hosted workloads?

This discussion aims to explore these questions and engage participants in a critical examination of Confidential Computing's role in the future landscape of cloud services.

11:10 **Discussion 8** (Lillskär)
6G as part of a cyber-physical continuum
Johan Eker, Lund University and Ericsson Research, Sweden, Catrin Granbom, Ericsson Research, Sweden, & Wolfgang John, Ericsson Research, Sweden

Unsurprisingly, 6G networks are set to provide enhancements in classic performance metrics related to connectivity. However, 6G is envisioned to additionally offer new capabilities and services that go beyond pure connectivity. These services will position 6G as a platform supporting a cyber-physical world, fusing the physical world of actions and experiences with its programmable digital representation.

In this session, we want to introduce the current industry status on 6G, with a focus on the role of compute and cloud in 6G. This includes compute and AI related services that might be supported by future networks, like dynamic device offloading and Edge AI. We hope that this short primer will spark a discussion with the audience on the convergence of compute and communication, as well as related research directions

11:10 **Discussion 9** (Storskär)
Towards Sustainable Computing on the Cloud/Edge Continuum
Abishek Chandra, University of Minnesota, USA

It is well-documented that data centers are a major source of energy and water usage along with CO2 emission. At the same time, the advent of edge computing and data generation with the explosion in the number of user devices and sensors has created new challenges for sustainability. These devices have high embedded carbon and might contribute to other environmental harms through use of dirty operational energy and heavy reliance on batteries or other non-sustainable materials. The discussion will examine various research issues towards achieving sustainability in the cloud/edge continuum such as:

- How to mitigate the environmental impacts of edge devices and computing resources?
- How could the edge and cloud work together (e.g., for data transfer and storage, task placement and execution, etc.) to improve sustainability?
- How to design end-to-end sustainable computing solutions that cover the whole computational pipeline from edge data collection and computation to transmission and storage in the cloud?

11:10 **Discussion 10** (Huvudskär)
How can we Democratize Cloud Systems Research?
Jan Rellermeyer, Leibniz University Hannover, Germany

Cloud computing continues to be a driving force for various areas of science and business alike, eclipsing an assumed total market value of 330 Billion USD in 2024. Emerging technologies like LLMs demand more customized and efficient infrastructure and system software, which should provide us researchers with a myriad of interesting systems challenges. However, the disadvantages in terms of resources compared to hyperscalers (e.g., high-end network interconnects, optimized system stacks, GPUs, custom hardware) as well as the gap in operational knowledge is making cloud systems research at scale in academia perceivably more difficult. In this session, we would like to discuss the concrete challenges that practitioners experience in their work and propose ways to close the gap.

13:45 *Research Directions in Google Storage*
Arif Merchant, Google, USA

Google's exascale data centers are a very large consumer of storage, and these storage systems are geographically distributed, multi-layered, heterogeneous, and complex. We present a high-level overview of Google's storage systems, the common challenges faced by storage at Google, and explore several research directions for managing and optimizing the resources used. We conclude with a discussion of some open questions.

14.15 **Discussion 11** (Resarö)
Agentic AI Systems and how they Compares with the Traditional Feedback Control Systems
Xiaoyun Zhu, NetApp, USA

In his recent book The Coming Wave, DeepMind co-founder Mustafa Suleyman highlights artificial intelligence (AI) and synthetic biology as the two most transformative technological trends of the coming decades, bringing unprecedented opportunities alongside undeniable risks that will shape the future of humanity. Of the two, AI has already made a striking leap into mainstream life—thanks in large part to the emergence of ChatGPT backed by large language models (LLMs). These models have moved beyond the realm of technology enthusiasts and into everyday use.

The latest wave of innovation in LLM applications centers around the concept of AI agents. An AI agent is a software application powered by an LLM that can autonomously perform specific tasks. These tasks might include analyzing research questions, retrieving and synthesizing information, reasoning through complex problems, making decisions, or interacting with various services via APIs.

An agentic AI system takes this idea further. It is a system composed of multiple AI agents working collaboratively and adaptively to accomplish a broader goal. For instance, one agent might interpret a high-level objective and devise a detailed execution plan or workflow. Others might handle the execution of individual steps, while yet another agent monitors progress, evaluates outcomes, and adjusts the plan as needed. These systems are particularly well-suited for tasks that require deep analysis, operate in dynamic environments, or demand a high degree of automation and flexibility.

Sound familiar? These qualities—goal-driven behavior, autonomy, decision-making, complex reasoning, adaptivity, and memory/context awareness—are also core attributes of feedback control systems. Designed to automate complex processes while responding to changing inputs and environments, feedback control systems share many conceptual similarities with agentic AI. This raises a compelling question: Are agentic AI systems simply feedback control systems augmented by LLMs?

Despite the parallels, key differences set the two apart. Traditional feedback control systems prioritize predictability. Their control algorithms are built on deterministic mathematical models, ensuring that identical inputs yield identical outputs. In contrast, LLMs are inherently stochastic: the same prompt can produce different responses on different occasions, even under similar conditions.

In this discussion, we'll explore both the commonalities and distinctions between agentic AI systems and feedback control systems. More importantly, we'll consider how decades of experience in control theory and engineering might inform the design of more effective, robust, and trustworthy agentic AI systems.

14.15 **Discussion 12** (Rindö)
Cloud Native Control - Code Changes as the Source of Disturbance?
Johan Tordsson, Elastisys, Sweden & Oliver Larsson, Umeå University, Sweden

Within the broad area of cloud control, dynamic allocation of virtualized hardware resources has been studied extensively, for example with objectives such as cost, performance, and energy efficiency. The dynamic parameter (source of disturbance) in such scenarios can be for example fluctuations in user demand, hardware performance, and resource availability. The recent advances in technology (Kubernetes etc.) and software development paradigm (DevOps, microservices, etc.), often summarized using the term cloud native, enables frequent and high-impact changes to large software systems. Such systems, combined with the rapidly rising popularity of generative AI tools to aid (or even replace) human software engineers, are creating a perfect storm of increasingly faster software changes with little or no human oversight or contextual understanding.

This discussion session wants to explore the idea of viewing the flow of rapid software updates as a source of disturbance in a large-scale software system. A few initial observations on the topic are that such a system has fine-grained actuators in the form of progressive software rollout strategies such as canary and blue/green deployments, detailed monitoring both of state of running system (performance, resource usage, etc.) as well as a complete history of software updates (version control systems such as git). Can control be used to make such a rapidly changing software system more stable?

A few alternative titles suggested by ChatGPT that may serve as a source of inspiration:
"Tracing Failures to Source: Intelligent Detection in Deployed Microservices" "Code-Aware Observability: Pinpointing Problems in Deployed Microservices" "Root Cause Intelligence: Connecting Microservice Failures to Code Changes" "When Systems Break: Observability and Root Cause Detection for Modern Microservices" "Smarter Observability: Linking Production Failures to the Developer's Fingerprint" "Commits in the Crosshairs: Identifying Problematic Code in Microservice Failures"

14.15 **Discussion 13** (Lillskär)
Reliability of Cloud-Scale Systems and Infrastructures
Tianyin Xu, University of Illinois Urbana-Champaign, USA & Indranil Gupta, Tianyin Xu, University of Illinois Urbana-Champaign, USA

Cloud reliability has been a grand challenge in the past decade, and is unfortunately continuing to be so, despite tremendous R&D efforts. In addition to many unsolved problems, new, emerging failure modes and patterns are surfaced, such as mercurial cores and accelerators, metastable failures, and cross-system interaction issues. This discussion session hopes to summarize what reliability problems have been solved and discuss what are the new challenges and research problems? Specifically, it looks for discussion of the following questions:

- Why is cloud reliability still a grand challenge?
- What is the most important cloud reliability problem today?
- What problems have been solved well and what are the approaches?
- What are the new fault modes, domains, and patterns?
- What research is mostly desired by industry?
- What are our views of formal verification of cloud system implementations? A pipedream or a truly practical direction?
- How can we understand the reliability problems of real-world cloud systems and make progress?

14.15 **Discussion 14** (Storskär)
Can data compression reduce the TCO of next generation cloud systems and at what cost?
Angelos Arelakis, ZeroPoint Technologies AB, Sweden

Data compression has been known for decades as a technique to reduce the storage size of data and improve their transmission time. In cloud systems, where data storage, bandwidth, and processing efficiency are paramount, compression can play a pivotal role in optimizing resource usage improving the overall total cost of ownership (TCO), as well as enhancing performance. In this session, we will explore the principles of data compression and its strategic application in cloud environments focusing on the benefits and the challenges. Some questions that will be discussed in the context of cloud systems, hoping to trigger more discussion topics, are:

- Compression is known for ages. What is really new here? And why is this an interesting topic at all?
- Is software compression enough to achieve the benefits, or do I need hardware accelerators and hardware-software co-designs?
- Are there application areas for lossy compression?
- What is the net benefit in TCO improvement?
- How much performance am I going to sacrifice to get this TCO benefit?
- What are the known challenges in deploying compression in cloud environments?

14.15 **Discussion 15** (Huvudskär)
Creating a Roadmap for Cyber-physical Cloud
Johan Eker, Lund University and Ericsson Research, Sweden & Gerhard Fohler, RPTU Kaiserslautern, Germany

Cloud services have become an integral of our daily routines, helping us navigate both the internet and the physical world. We are currently witnessing is a second wave of cloud services that goes beyond web storefronts and IT systems and includes digitalization of industrial systems. Automation and time-sensitive systems, that until recently have been confined to the factory floor and hidden from the external world, are now taking their first steps towards the cloud.

However, from a research perspective this evolution is happening in silos, both with respect to use case domains and research areas. To that end, we are trying to create a community around what we decided to call Cyber-Physical Cloud (CPC). The first step is to agree on a research agenda by a joint CPC roadmap. This session will present the initiative and use the opportunity to discuss both technical topics as well as overall direction.

.

Thursday June 19th

8.30 *Distilling The World Through Private Ubiquitous Computing*
John Kubiatowicz, UC Berkeley, USA

Our physical lives are mirrored by a constantly growing body of information -- generated by our actions, reflecting our lives, and charting our history. Ideally, we should have control over where this information is stored and who is allowed to access it. We should be confident that our information will survive disasters and repel ransomware. We should also have the ability to process our information wherever and however we see fit -- without risking exposure to unauthorized parties. As a particular example, with the rise of modern AI techniques, we should be able to train our own private AI models continuously but privately and securely.

To address our modern information needs, I discuss a mobile and embedded data format called a DataCapsule, along with a supporting data-centric infrastructure called the Global Data Plane (GDP). DataCapsules provide append-only semantics with a blockchain-like internal structure. Together, the GDP and DataCapsules provide for a secure substrate on which to host a wide variety of secure, mobile information processing. I will further describe a lightweight but stateful Function-as-a-Service (FaaS) framework called Paranoid Stateful Lambdas (PSLs). This framework exploits Trusted Execution Environments (TEEs) to permit secure and private processing of DataCapsules by unsecured (Edge) computers without risking exposure of this information. As a bonus, PSL provides rich programming language support for parallel execution using a familiar Key-Value Store (KVS) interface to DataCapsules. PSL provides a parallel execution model using eventually consistent state updates with release-consistent locking. As an example use-case, I will report how well PSLs support personalized machine learning.

I will also discuss a variant of the GDP designed for Fog Robotics environments, called FogROS2. FogROS2 combines a data-centric network with a proxy interface for the standard Robot Operating System v2 (ROS2). The result is a secure communication infrastructure that enables collaboration between geographically distributed robots, the cloud, and edge computing. I will present multiple use-cases of this distributed infrastructure.

9:00 *A Fractionalized Supply Chain of Varying Yield - One Hyperscaler's View of Capacity*
Narayan Desai, Google, USA

Google capacity evolved in several interesting ways, resulting in a somewhat unique view of capacity and resource management. Computing capacity is integral to its offerings, hence capacity is thought of more as a supply chain than a static pool. Capacity is fractionated into different classes of resources, and a variety of optimizations are employed that result in varying yields for classes of resources. In this talk, I'll walk through the implications of these approaches for analysis and modeling of capacity, and what this means for the process of capacity management.

9.20 *Predicting Failures of Autoscaling Distributed Applications*
Mauro Pezze, Università della Svizzera Italiana, Switzerland

Predicting failures at runtime allows service providers to activate countermeasures that prevent harming the users of the applications. The most successful approaches predict failures from time series of values of key performance indicators that they monitor at runtime. All current approaches work with time series of some fixed sets of values collected at each time stamp, thus, they work well with applications that run on a statically configured set of pods and nodes. However, they cannot handle time series of variable sets of values, as the ones monitored from cloud applications that exploit autoscaling on top of dynamically resizable sets of virtual nodes.

I present Preface, a novel approach to predict failures in autoscaling distributed applications. Preface combines descriptive statistics with a generative neural network. The descriptive statistics (count, mean, quartiles, maximum, and minimum across all the replicas) summarize the runtime behavior of the components, which can be replicated on multiple nodes via the autoscaling mechanism, and reduce the variable sets of metrics collected at each timestamp to a fixed set of key performance indicators. The generative

neural network uses the values of the summary statistics to spot symptoms of incoming failures and failing components in unsupervised fashion. We experimentally validated Preface against a commercial Learning Managing System that exploits the Kubernetes platform to autoscale eighteen microservices.

10.10 *The Value of Information in Multi-scale Management: The Cloud Case*
Ada Diaconescu, Télécom Paris, France

Information is essential to system management: from monitoring and analysis, to planning and action. In large complex systems, multiple controllers coordinate to meet global objectives, by sharing information at various scales of information and time. Information flows through the control system, merging, splitting and transforming through multiple interconnected cycles. “Limited rationality”, caused by limited resources, bounds the throughput of such information flows. To grasp what information to keep and what to ignore, one must assess the influence of different pieces of information on the control process. However, common measures from Information Theory (e.g. Shannon entropy or Kolmogorof complexity) focus on the quantity of information, while ignoring its value to the system that uses it. Taking inspiration from more recent developments in information theory – from complex systems, cognitive science, philosophy of information and biosemiotics domains – we suggest adopting wider notions of information and measures.

We propose “semantic information” to assess the information impact on models, or knowledge; and “pragmatic information” to assess the information impact on action and adaptation. We illustrate how these additional information types can be measured via several generic examples; and hope to provide inspiration for generating similar measures in the cloud computing domain.

10:30 *Embracing Non-linearity for Cloud Control*
Thijs Metsch, Intel Labs, Germany

Modern computer systems exhibit a wide range of non-linear relationships, spanning latency-throughput dynamics, algorithmic complexity, and e.g. power-performance trade-offs. These non-linearities arise from both hardware behavior and software interactions and are often overlooked in system-level design and orchestration.

This presentation is structured in two parts. First, we'll explore a variety of these non-linear behaviors for both educational value and a bit of fun, illustrating the mechanisms that produce them and their implications for performance and efficiency. In the second part, we discuss how these non-linearities can be intentionally leveraged—rather than mitigated—through the application of intent-aware programming models and control-theoretic approaches. We demonstrate how such methods enable more adaptive orchestration, so that system efficiency can be improved by aligning resource management strategies with workload intent and dynamic system behavior.

10:50 *Optimizing the Full Lifecycle of Large Language Models: Balancing System Trade-Offs*
Jie Xu, University of Leeds, UK

In this presentation, we will share our recent experience in designing and implementing a large-scale distributed system for supporting the full lifecycle of large language models (LLMs) with billions of parameters. Our focus will be on balancing various system requirements and design objectives, such as effectiveness and efficiency.

We introduce a theorem of "Impossible Trinity of System S" where S is a distributed system designed and implemented for enabling full lifecycle support for LLMs. This theorem provides guidance on balancing diverse design requirements and explores various architectural options to ensure the development and operation of large language models are both cost-effective and efficient. The theorem formalises three desirable objectives of LLM systems and asserts that it is impossible to achieve all three simultaneously within S. Therefore, system designers must prioritise two out of the three properties based on the specific requirements and constraints of their application. Drawing on our previous experience, we illustrate these principles through practical system examples that demonstrate how to design and implement a well-balanced LLM support system.

11:10 **Discussion 16** (Resarö)
An Exabyte a Day: Throughput-oriented, Large Scale, Managed Data Transfers with Effingo
Krzysztof Rzadca, Google, Poland

WAN bandwidth is never too broad - and the speed of light stubbornly constant. These two fundamental constraints force globally-distributed systems to carefully replicate data close to where they are processed or served. A large organization owning such systems adds dimensions of complexity with ever-changing network topologies, strict requirements on failure domains, multiple competing transfers, and layers of software and hardware with multiple kinds of quotas.

In the discussion session, I plan to discuss these requirements and the Google's solution, Effingo, a throughput-oriented, massively-parallel, logically-distributed data copy service. For its users, Effingo delivers high-throughput transfers with an scp-like interface. For Google, Effingo optimizes the network cost with a small footprint on datacenters. On a typical day, Effingo transfers over an exabyte of data between dozens of clusters spread across continents and serves more than 10,000 users.

Our experience shows that the key component of such a system is not its supposed global optimality, but rather its resiliency achieved through cluster-level failure domains and its reliable, efficient and convenient service.

The discussion is based on our 2024 SIGCOMM paper:

Ladislav Pápay, Jan Pustelnik, Krzysztof Rzadca, Beata Strack, Paweł Stradomski, Bartłomiej Wołowiec, and Michal Zasadzinski. 2024. An exabyte a day: throughput-oriented, large scale, managed data transfers with Effingo. In Proceedings of the ACM SIGCOMM 2024 Conference (ACM SIGCOMM '24). Association for Computing Machinery, New York, NY, USA, 970–982. https://doi.org/10.1145/3651890.3672262

11:10 **Discussion 17** (Rindö)
Unpredictability in the Cloud: Sources and Solutions?
Ada Diaconescu, Télécom Paris, France

Context: Cloud management aims to meet service-level objectives (SLO), at minimal cost, while dealing with fluctuating demands, erratic resource availability and exponential growth. It involves reaction to changes on the short-term; and change prediction, planning and pre-emptive action over the long-term.

Both reactive and predictive adaptations habitually involve complex system modelling and forecasting to deal with uncertainty, within an estimated range (e.g. trade-offs between over-provisioning and costs; and between updating hardware too soon or too late).

Problem: Forecasting generally assumes that the future will be an extrapolation of the past. When the conditions that generated past observations shift, such assumptions about the future may no longer hold. Unpredictable events, or changes, may emerge under the new conditions, evading the forecasted uncertainty range, and raising acute challenges.

E.g., cloud warehouses must scale-up with AI-driven demand; while necessitating reliable supplies of cool water, energy, hardware, and workforce – all potentially influenced by unpredictable environmental, social and geo-economic variations; all out of the direct control of cloud managers.

While major unexpected events may appear too improbable or indeterministic to be worth considering in advance, their vital impacts may push stakeholders to reconsider such risky shortcuts.

Discussion points:

- Imagine potential cloud-related incidents, or bifurcations, with serious repercussions, at various magnitudes and time-scales;
- Argue whether we should ignore such potentialities because they are unlikely or unquantifiable; or investigate them because we are unwilling to take the risk;
- Consider alternative, metric-free prediction and assessment approaches, e.g., based on prospective, simulation, or strategic foresight;
- Reconsider the historic priority that efficiency has had over robustness, sustainability and survival.

Discussion objectives and importance:
In a world of accelerating inter-dependence and change, unpredictability might become the only predictable thing. As modern societies increasingly rely on cloud-provided digital services, considering the above discussion seems not entirely inconsequential. The aim is to estimate the awareness of the aforementioned issues in the Cloud Computing community and envisage appropriate research directions.

11:10 **Discussion 18** (Lillskär)
Civil and Military Resilience through Cloud Technology
Per Skarin, Saab Kockums, Sverige & Mikael Lindberg, Saab Kockums, Sverige

This discussion group will focus on digital resiliency, distributed systems and Clouds in civil and military defence. Participants are interested in the technology but also in how we enable collaboration, build competence, and the concepts of civil-military synergies and dual use in this domain (see for example the Vinnova call for civil-military synergies). The organisers will bring the context of research programs and collaborations with academia, cyber-security and availability during a crisis, the naval cloud perspective, and technology for multi-domain services (see EDF projects EDOCC and EDINAF).

Today, infrastructure is as much about the digital domain as it is about transport logistics, access to clean water and electricity. To that end, Cloud technology know-how and the availability of necessary ICT has become increasingly important and sparked several initiatives nationally, regionally, and within organizations. In the new landscape we need to revisit best practices for Cloud and edge, federated systems, standards and services, security and software development. What are the unique challenges put on suppliers and developers? Are there new research avenues that need to be explored or some that should be accelerated? Who can supply the necessary capability? Efforts to strengthen the domain of cyber security is resulting in an explosion of standards and regulations that are both cumbersome to work with and slow to adapt to changing situations. How can we find better methods for increasing quality and security, while minimizing the impediment to technology adoption? These discussions will apply to commodity systems for everyday use and bridging that with availability during an eventual crisis. How far can we stretch civil-military synergies and dual use in this domain?

We will open up for a discussion on massive global compute and communication infrastructure to specifics about services and agents in edge clouds. The organisers look forward to a broad and lively discussion on how to position cloud technology and its practitioners in this landscape. We hope to hear about some past experiences within industry and academia, understand more about how we can and should cooperate and perhaps touch upon some technologies and actors we should be focusing on for cyber-security, confidentiality, robustness, and agility.

11:10 **Discussion 19** (Storskär)
Edge Orchestration: Challenges, Principles & Patterns
Stefan Vallin, Avassa, Sweden

As cloud orchestration matures, edge computing presents a new set of challenges: highly distributed deployments, constrained compute environments, and the absence of traditional perimeter security. This session examines how cloud orchestration models must evolve to support workloads at the edge, where operational environments differ fundamentally from those in centralized cloud environments.

Examples of real-world edge deployments include:

- Retail chains running software across thousands of stores
- Industrial machines with globally distributed update points
- Software-defined vehicles receiving over-the-air updates
- AI inference deployed close to sensors and real-world signals
- Distributed 5G/6G nodes

We'll examine the emerging patterns for edge-native orchestration and ask:

- What are the defining constraints of edge environments?
- Should orchestration extend from the cloud or evolve from the device layer?
- Which solution patterns are proving most effective at the edge?
- What kind of tools and platforms are fit for purpose?
- Can we converge on a shared taxonomy for "the edge"?

Bring your experience and perspective—join the discussion to help shape the principles guiding orchestration at the edge.

11:10 **Discussion 20** (Huvudskär)
Open opportunity – please send proposals to erik.elmroth@cloudresearch.org

13.45 *Toward Energy-Optimal AI Systems*
Mosharaf Chowdhury, University of Michigan, Ann Arbor, USA

Generative AI adoption and its energy consumption are skyrocketing. Power availability has become the gating factor in building new AI datacenters, limiting the deployment of AI services in regions without high-capacity electricity grids. This creates asymmetries in service accessibility and raises both national and international security concerns. Unfortunately, we don't even have a clear understanding of where all the energy is going!

In this talk, I will briefly introduce the ML Energy Initiative, which focuses on understanding and optimizing AI's energy consumption. I will then highlight the key challenges that remain—particularly around measuring and interpreting energy use in closed-source models, as well as understanding the broader implications of AI-energy interactions across the economy.

14:05 *Towards Fine-grained Agile Elasticity in Inference Serving*
Alexey Tumanov, Georgia Institute of Technology, USA

The increasing deployment of ML models on the critical path of production applications requires ML inference serving systems to serve these models under unpredictable and bursty request arrival rates. Serving many models under such conditions requires a careful balance between each application's latency and accuracy requirements and the overall efficiency of utilization of scarce resources. Faced with this tension, state-of-the-art systems either choose a single model representing a static point in the latency-accuracy tradeoff space to serve all requests or incur latency target violations by loading specific models on the critical path of request serving. Our work instead resolves this tension through a resource-efficient serving of the entire range of models spanning the latency-accuracy tradeoff space. Our novel mechanism, SubNetAct, achieves this by carefully inserting specialized control-flow operators in pre-trained, weight-shared super-networks. These operators enable SubNetAct to dynamically route a request through the network to actuate a specific model that meets the request's latency and accuracy target. Thus, SubNetAct can serve a vastly higher number of models than prior systems while requiring upto 2.6× lower memory. More crucially, SubNetAct's near-instantaneous actuation of a wide-range of models unlocks the design space of fine-grained, reactive scheduling policies. We design one such extremely effective policy, SlackFit, and instantiate both SubNetAct and SlackFit in a real system, SuperServe.

14:25 *Software Reliability in Emerging Cloud Computing Paradigms*
Tianyin Xu, University of Illinois Urbana-Champaign, USA

Cloud system reliability has been a grand challenge in the past decade due to prevalent, inevitable hardware faults, software bugs, and misconfigurations. Emerging computing paradigms such as microservices, serverless, and sky computing further expand reliability challenges by significantly increasing the complexity of system interactions and dependencies, with new failure domains. In this talk, I will share some of our recent efforts on improving the reliability of cloud infrastructure systems such as Kubernetes and cloud-native applications running atop them via software testing, model checking, and formal verification. Specifically, I will discuss the promises of building new components with proven correctness towards a vision of truly reliable cloud infrastructures.

14:45 *Closing*
Erik Elmroth, Umeå University and Elastisys, Sweden

Hotel Floor Plan

Floor plan showing location of conference rooms being used. Notably, entrance floor and reception are located on the third floor ("plan 3" in Swedish).

Resarö (3rd floor, next to Hotel Reception) is used for all plenary and breakout sessions.

For breakout sessions, we are also using:
- **Lillskär** and **Rindö** on 3rd floor.
- **Huvudskär** and **Storskär** on 2nd floor.

Welcome Poster Reception, weather depending, is being held inside or outside **Svavelsö** on 1st floor