
# Program

## 17th Cloud Control Workshop

**Skåvsjöholm, Åkersberga, Sweden**
**June 18 – 20, 2024**

# 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) are available by the relax facilities (on shelfs near the bar)
Bicycles: ask at the reception
Canoes: ask at the reception – do not forget the life wests!

**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 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 barbecue buffet dinner is served outdoors, by the sea.
For drinks included in the registration, each participant receives 5 drink tickets with their hotel room key to be used at the Welcome Poster Reception and the two dinners. Further drinks are on own cost. (Participants without a hotel room, please contact the reception.)

**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**
Weather forecasts are mostly positive. Expect sunshine mixed with light clouds and possibly some minor risk for rain and temperatures up to 25 degrees and nights bright and somewhat chillier.

**In case everything else fails**
Erik’s phone number is +46 70 315 3928. Anticipate limited ability to answer around bus departure on June 18.

# Tuesday June 18th

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

| | | | | | |
|---|---|---|---|---|---|
| 9:40 | **Coffee, registration, and luggage storage** | | | | |
| | *Session Chair: Maarten van Steen, University of Twente, The Netherlands* | | | | |
| 10:10 | *Workshop Introduction*<br>Erik Elmroth, Umeå University and Elastisys, Sweden | | | | |
| 10:30 | *Multi-modal Agents for Autonomic Computing*<br>Jeffrey Kephart, IBM Research, Yorktown Heights, USA | | | | |
| 11:10 | *Can Clouds Reach into Space?*<br>Indranil Gupta, University of Illinois at Urbana Champaign, USA | | | | |
| 11:30 | *But is it Working?*<br>Narayan Desai, Google, USA | | | | |
| 11:50 | *Function as a Function*<br>Ana Klimovic, ETH Zürich, Switzerland | | | | |
| 12:10 | *Sharing is Caring (and also Efficient)*<br>Gonzalo P. Rodrigo Alvarez, Apple, USA | | | | |
| 12:30 | **Lunch** | | | | |
| | *Session Chair: Karl-Erik Årzén, Lund University, Sweden* | | | | |
| 13:45 | *On Optimization Opportunities for Future Cloud Computing*<br>Tarek Abdelzaher, University of Illinois at Urbana Champaign, USA | | | | |
| 14:15 | *How Can We Decarbonize the Power Grid and Meet AI's Exploding Power Demands?*<br>Andrew A. Chien, University of Chicago, USA | | | | |
| 14:35 | Discussion 1 (Resarö)<br>*Cloudy with a chance of offloading: The lighter side of edge computing*<br>Wolfgang John, Ericsson Research, Sweden & Johan Eker, Lund University and Ericsson Research, Sweden | Discussion 2 (Lillskär)<br>*Research challenges in multi-cloud networking*<br>David Breitgand, IBM Research, Israel | Discussion 3 (Storskär)<br>*Cloud Intelligence – AI/ML for cloud efficiency, quality, and experience*<br>Jian Zhang, Microsoft, USA; Pamela Delgado, University Applied Sciences Western Switzerland & Cristian Klein, Umeå University and Elastisys, Sweden | Discussion 4 (Huvudskär)<br>*Towards sustainable computing: Exploring green IT perspectives and solutions*<br>Monica Vitali, Politechnico di Milano, Italy & Abel Souza, University of Massachusetts Amherst, USA | Discussion 5 (Svavelsö)<br>*Collaborative connected computing: optimizing the computing continuum from the cloud to IoT sensors*<br>Geir Horn, University of Oslo, Norway |
| 15.55 | **Coffee, Hotel Room Check-in, Poster preparation** | | | | |
| | *Session Chair: Bhuvan Urgaonkar, Penn State University and Amazon, USA* | | | | |
| 16.25 | *Making Remote Attestation a Future Commodity for End Users*<br>Rüdiger Kapitza, FAU Erlangen-Nürnberg, Germany | | | | |
| 16.45 | *Reconciling High Accuracy, Cost-Efficiency, and Low Latency of Inference Serving Systems*<br>Pooyan Jamshidi, University of South Carolina, USA | | | | |

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

1. *Dynamic Offloading of Control Algorithms to the Edge using 5G and WebAssembly*, Ahmed Al Bayati, Lund University, Sweden
2. *Cheops: Externalizing Geo-Distribution for Cloud Applications at the Edge*, Geo Johns Antony, INRIA, France
3. *PerMFL: Personalized Multi-tier Federated Learning*, Sourasekhar Banerjee, Umeå University, Sweden
4. *Monogamous Relationships with Short-term Commitment are the Best,* Carl Magnus Bruhner, Linköping University, Sweden
5. *End-To-End Performance, Energy Consumption and Carbon Footprint of Fog Applications*, Clément Courageux-Sudan, Umeå University, Sweden
6. *Machine Learning for Anomaly Detection in Edge Clouds*, Javad Forough, Umeå University, Sweden
7. *Power Focused Green Energy Modeling in Cloud Edge Continuum Regions*, Rohail Gulbaz, Umeå University, Sweden
8. *Adaptive and Data Efficient Cloud Management: Enhancing Machine Learning for Efficient Resource Allocation*, Lidia Kidane, Umeå University, Sweden
9. *Avoiding the Heat Death of Kubernetes and the CNCF Landscape*, Lucas Käldström, Finland
10. *Application-Aware Hardware-Level QoS Enforcement Tuning Using Extremum Seeking Control*, Oliver Larsson, Umeå University, Sweden
11. *Stealthy Delay Attacks for Cyber-Physical Systems,* Talitha Nauta, Lund University, Sweden
12. *Modeling the Green Cloud Continuum: integrating energy considerations into Cloud–Edge models*, Yashwant Singh Patel, Umeå University, Sweden
13. *Edge Orchestration for Video Analytics Applications*, Ali Rahmanian, Umeå University, Sweden
14. *Mitigating Slow-Rate Data Plane Attacks using Model-Free Reinforcement Learning*, Kshira Sagar Sahoo, Umeå University, Sweden
15. *Constraint-Based Service Placement Optimization for IoT Applications in Fog environments*, Farah ait-Salaht, Léonard de Vinci Pôle Universitaire, France
16. *Resource Management of IoT Applications in Edge-based Infrastructures*, Hajar Siar, Umeå University, Sweden
17. *An Edge-Cloud Infrastructure for Fault-Tolerance Mission-Critical Applications,* Nayereh Rasouli, Umeå University, Sweden
18. *Analysis of control systems subject to timing discrepancies*, Yde Sinnema, Lund University, Sweden
19. *Eco-Efficiency in Cloud Computing: A Carbon-Conscious Reinforcement Learning Approach to Federated Learning*, Antonio (Eunil) Seo, Umeå University, Sweden
20. *Dynamic Computation Offloading for a Cluster of Drone,* Emil Sundström, Lund University, Sweden
21. *Energy-aware Network Model in Cloud-Edge Continuum*, Yangyang Wen, Umeå University, Sweden
22. *Federated Learning for Cloud Robotic Manipulation*, Obaidullah Zaland, Umeå University, Sweden
23. *Securing Serverless Edge AI for the Cloud-Edge Continuum*, Zhou Zhou, Umeå University, Sweden

19:00 **Dinner**

# Wednesday June 19th

| | | | | | |
|---|---|---|---|---|---|
| 7:00 | **Breakfast** | | | | |
| | *Session Chair: Azer Bestavros, Boston University, USA* | | | | |
| 8:20 | *Metrics for Building Warehouse Scale Computers*<br>John Wilkes, Google, USA | | | | |
| 9:00 | *Dynamically Controlling Infrastructure as Code*<br>Claus Pahl, Free University of Bozen-Bozano, Italy | | | | |
| 9:20 | *GPU-Cluster Management for a Stream of Distributed Deep Learning Jobs*<br>George Kesidis, Pennsylvania State University, USA | | | | |
| 9:40 | **Coffee** | | | | |
| | *Session Chair: Catrin Granbom, Ericsson Research, Sweden* | | | | |
| 10:10 | *Multi-scale Feedback Architectures for Decentralised Control*<br>Ada Diaconescu, Télécom Paris, France | | | | |
| 10:30 | *Low-Latency Privacy-Preserving Authentication in the Edge*<br>Luis Rodrigues, INESC-ID University of Lisboa. Portugal | | | | |
| 10:50 | *A Journey above the Clouds: Cloud Control in new Contexts*<br>Hermann De Meer, University of Passau, Germany | | | | |
| 11:10 | Discussion 6 (Resarö)<br>*Overcommitment and workload autoscaling in cloud datacenters: a provider's perspective*<br>Krzysztof Rzadca, Google, Poland | Discussion 7 (Rindö)<br>*Geo-distributed storage: addressing low latency, weak throughput, and disconnections*<br>Thomas Fahringer, University of Innsbruck, Austria; Mina Sedaghat, Ericsson, Sweden & Adrien Lebré, Centre Inria de l'Université de Rennes, France | Discussion 8 (Storskär)<br>*The future of orche-stration and resource management: steering innovation in the era of hardware abstraction*<br>Thijs Metsch, Intel, Germany & Oliver Larsson, Umeå University, Sweden | Discussion 9 (Huvudskär)<br>*Observability and diagnostics in the 5G networks and beyond when embracing the intelligent and autonomous networks paradigm*<br>Monowar Bhuyan, Umeå University, Sweden and Johan Forsman, Tietoevry, Sweden | Discussion 10 (Svavelsö)<br>*Foundation models for edge intelligence*<br>Feras M. Awaysheh, Tartu University, Estonia |
| 12:30 | **Lunch** | | | | |
| | *Session Chair: Xiaoyun Zhu, NetApp, USA* | | | | |
| 13:45 | *A CarbonFirst Approach for Decarbonizing Cloud Computing*<br>Prashant Shenoy, University of Massachusetts Amherst, USA | | | | |
| 14:15 | *The US National Artificial Intelligence Research Resource (NAIRR) Pilot*<br>Dilma Da Silva, Texas A&M University, USA | | | | |
| 14.35 | Discussion 11 (Resarö)<br>*Above the clouds: Potentials of non-terrestrial edge and cloud computing*<br>Hermann de Meer, University of Passau, Germany | Discussion 12 (Rindö)<br>*iPaaS, the edge, the CDN and the multi-cloud*<br>Mazin Yousif, University of Michigan, USA | Discussion 13 (Storskär)<br>*How to scale academic cloud research*<br>Karl-Erik Årzén, Lund University, Sweden & Khuzaima Daudjee, University of Waterloo, Canada | Discussion 14 (Huvudskär)<br>*Navigating design Space of modular-composed ML systems for multi-objective performance optimization under uncertainty*<br>Pooyan Jamshidi, University of South Carolina, USA | Discussion 15 (Svavelsö)<br>*Simulation of edge-cloud continuum systems*<br>P-O Östberg, Umeå University, Sweden |
| 15:55 | **Coffee** | | | | |

16:15 Social outdoor activities

Equipment and facilities available at no cost:

- Equipment for various social games (e.g., boule, crocket, kubb) are available by the relax facilities (on shelfs near the bar)
- Outdoor whirlpool and indoor sauna are available by the relax facilities
- Bicycles: ask at the reception
- Canoes: ask at the reception – do not forget the life wests!

19:00 **Barbeque Dinner**

# Thursday June 20th

| | | | | | |
|---|---|---|---|---|---|
| 7:00 | **Breakfast** | | | | |
| | *Session Chair: Thomas Fahringer, University of Innsbruck, Austria* | | | | |
| 8:20 | *Towards Active Inference for Distributed Intelligence in the Computing Continuum*<br>Schahram Dustdar, TU Wien, Austria | | | | |
| 9:00 | *Edge Challenges and Opportunities: Data, Latency, Resilience*<br>Richard Mortier, Cambridge University, UK | | | | |
| 9:20 | *Tell Me What You Want, Not How to Do It!*<br>Thijs Metsch, Intel, Germany | | | | |
| 9:40 | **Coffee & Hotel Room Check-out (before 10.00)** | | | | |
| | *Session Chair: Eyal de Lara, University of Toronto, Canada* | | | | |
| 10.10 | *Exploring Cloud Control for Urgent Computing*<br>Manish Parashar, University of Utah, USA | | | | |
| 10:30 | *Fog Computing Challenges in Natural Environment Observatories*<br>Guillaume Pierre, Université de Rennes, France | | | | |
| 10:50 | *Towards Trustworthy AI in untrusted (Cloud) Environments*<br>Sonia Ben Mokhtar, CMRS, France | | | | |
| 11:10 | Discussion 16 (Resarö)<br>*Experimental driven research*<br>Christian Perez, INRIA, France; Manish Parashar, University of Utah, USA & Laurent Lefevre, INRIA, France | Discussion 17 (Rindö)<br>*Serverless cloud programming models*<br>Ana Klimovic, ETH Zürich, Switzerland | Discussion 18 (Storskär)<br>*Exploring the potential of cloud robotics: Resource allocation, challenges, and platforms*<br>Chanh Nguyen and Antonio Seo, Umeå University, Sweden | Discussion 19 (Huvudskär)<br>*Securing the foundation of the cloud through recent advances in hardware isolation mechanisms, formal verification, safe programming languages, static analys, and fuzzing*<br>Anton Burtsev, University of Utah, USA | Discussion 20 (Svavelsö)<br>*Open discussion opportunity* |
| 12.30 | **Lunch** | | | | |
| | *Session Chair: Maria Kihl, Lund University, Sweden* | | | | |
| 13.45 | *Lifting the Fog of Uncertainties: Dynamic Resource Orchestration for the Containerized Cloud*<br>Hans-Arno Jacobsen, University of Toronto, Canada | | | | |
| 14:05 | *Revolutionizing Datacenter Networks via Reconfigurable Topologies*<br>Stefan Schmid, TU Berlin, Germany | | | | |
| 14:25 | *Reliable Fast Process Failure Detection in Data Centers*<br>Patrick Eugster, Lugano, Switzerland | | | | |
| 14:45 | *Closing*<br>Erik Elmroth, Umeå University and Elastisys, Sweden | | | | |
| 14:50 | **Coffee** | | | | |

15:20 **Bus departure**

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

# Abstract for
# Presentations, Posters, and Discussion Sessions
# in order of Appearance
## 17th Cloud Control Workshop
## June 18 – 20, 2024

## Tuesday, June 18th

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

We will start by a short welcome and to give a brief historic flashback on how the project meetings in a project named Cloud Control (2nd largest project grant ever awarded by The Swedish Research Council) grow into an international workshop with no connections to that finished project other than the name (and the fact that Umeå and Lund Universities are still over-represented among participants). The main driver behind that transformation – *To run meetings whose forms dynamically adjust to make them as useful as possible for leading researchers in our field* – is still what drives the development of the series.

10:30 *Multi-modal Agents for Autonomic Computing*
Jeffrey Kephart, IBM Research, Yorktown Heights, USA

How should self-adaptive systems manage themselves in dynamic environments? In years past, I had advocated an approach in which humans express high-level goals as utility functions, and the system uses optimization and/or feedback control techniques in conjunction with models to adjust resources and tuning parameters to maximize the utility.

In hindsight, this body of work was flawed because it assumed the existence of utility functions without considering where they came from in the first place.

In this talk, I will explain why I believe my current research area, multi-modal agents, provides a framework and technology that addresses this shortcoming. Multi-modal agents assist people with data analysis, diagnosis, and decision making by interacting with them through a combination of speech and non-verbal modalities such as pointing. I will demonstrate a few of these agents, including a prototype assistant that helps a systems administrator with a resource allocation task while dynamically learning the administrator's preferences.

11:10 *Can Clouds Reach into Space?*
Indranil Gupta, University of Illinois, USA

Space is now the final frontier. Low earth orbit (LEO) satellite constellations are blanketing the earth for observation and communication. This talk will focus on networked systems protocols that live not (just) in the cloud but in the mix among LEO satellites, ground stations, backhaul, and datacenters. LEO constellations are an exciting new class of edge systems with predictable and fast mobility, and many real-time requirements that are challenging under bandwidth and resource limitations.

11:30 *But is it Working*
Narayan Desai, Google, USA

Over the last 5 years, we've run a pathfinding team focused on producing useful insights for infrastructure services and their customers. Along the way, we've uncovered a variety of important aspects of workloads and system performance that have previously been ignored by reliability analytics (SLOs, alerting via thresholds, and the like). In this talk, I'll describe these issues, how they confound our analytics and how they can be addressed.

11:50 *Function as a Function*
Ana Klimovic, ETH Zürich, Switzerland

Serverless computing raises the level abstraction to the cloud, making the cloud easier to use and enabling the cloud platform to optimize for performance and energy efficiency under the hood. Although the serverless paradigm holds great promise, the system software that serverless platforms are built on today is still rooted in the very different, more traditional execution model of long-running containers or virtual machines. Today's Functions as a Service (FaaS) platforms provide secure isolation by running functions inside sandboxes like 'lightweight' VMs, which still incur significant startup times, context switch overheads, and memory duplication. FaaS platforms also orchestrate function sanboxes with conventional cluster managers like Kubnernetes, which add significant overhead due to the high churn of

short-lived sandboxes in FaaS environments. Hence, rather than retrofitting existing cloud software infrastructure, we propose a clean slate system software design for serverless computing.

I will present Dandelion, a new serverless platform that rethinks the FaaS programming model to unlock performance and resource efficiency benefits. Dandelion treats user functions as pure functions, and hence separates untrusted user computations from I/O. Users build Dandelion applications by composing and expressing the dataflow between pure compute functions and trusted I/O functions implemented by the Dandelion platform, which enable interaction with external cloud services. This new programming model enables Dandelion to: 1) securely isolate functions with minimal overhead (without relying on VMs), 2) leverage dataflow information to optimize function scheduling, 3) offload user computations and I/O to specialized hardware.

12:10 *Sharing is Caring (and also Efficient)*
Gonzalo P. Rodrigo Alvarez, Apple, USA

Batch compute powers research and development in companies, universities, and research institutions. These organization are composed of numerous teams that harness large, powerful, and expensive batch resources in different ways. Organizations strive to be efficient and require processes and mechanisms to plan capacity and control it allocations to teams to ensure that research goals are met while resources are used efficiently.

In this talk, we'll review different models to allocate resources to teams within organizations. We'll focus on the effect of sharing resources and we'll review scheduling mechanisms that make sharing possible. We'll do a deep dive on models that represent resource allocations, allow sharing, and also mimc organizations. Finally, we'll discuss effects of these models on key metrics (such as meeting business goals and efficiency), but also other not so obvious metrics like management overhead and capacity predictability.

13:45 *On Optimization Opportunities for Future Cloud Computing*
Tarek Abdelzaher, Indiana University of Illinois, USA

This talk offers an overview of optimization and control needs and opportunities in distributed and cloud computing, driven by the emergence of novel applications that are hungry for computing and communication capacity. While artificial intelligence tops the list of rapidly expanding application domains with the proliferation of large language models, foundation models, and generative AI, other important application areas contribute unique needs as well. For example, the democratization of content sharing media allows individuals to offer real-time content for potentially global consumption, creating the potential for large bottlenecks as five billion Internet users generate “globally accessible” content. There arises a need to bridge the gap between the growing data generation volume on one end and consumers’ limited capacity to process it on the other. This need is further magnified by the increasing demands on more resource-consuming content modalities (e.g., video, as opposed to text), the increasing need for personalized information processing and summarization, and the rise of increasingly more immersive content. Additional consumption is envisioned from the growth of physical sensing devices driven by the rise of autonomous machines and Internet of Things (IoT) applications. Pervasive automation will increasingly delegate physical functions to drones, robotic assistants at home or business, and autonomous vehicles. At the same time, a growing number of IoT applications continue to contribute to major growth in connected devices. Eventually, data generation and consumption in the world may become dominated by untethered Edge AI. How do these application domains shape future computing demand? What challenges arise in the resulting landscape of computing resource optimization and control? The talk offers a vision of potential challenges and directions in this field.

14:15 *How Can We Decarbonize the Power Grid and Meet AI’s Exploding Power Demands*
Andrew A. Chien, University of Chicago, USA

AI and Cloud computing is growing rapidly and projected to increase US electric power consumption by as much as 2% by 2026. This is an explosive increase against a backdrop of zero growth in US power consumption from 2007-2020 (EIA). Such rapid power consumption growth presents a significant challenge to power grid stability and decarbonization. In several power grids, datacenter load exceeds 20% of total with many more grids to follow.

We will discuss how computing flexibility could help the grid decarbonize (Zero-carbon cloud), by aligning compute load to curtailed and stranded power (a rapidly growing challenge). Today, datacenter loads view power as an “on-demand” service, a difficult model for renewable-based grids to support. We will show a new framework that creates cooperation between datacenter loads and power grids with continuous matching. Critically, these new approaches address the fundamental conflict between loads and renewable power grids, supporting both corporate goals (efficient computing) and societal goals (power grid decarbonization). We encourage computing community to solve this problem by shaping power growth to support grid decarbonization, not retard it.

14:35 **Discussion 1** (Resarö)

*Cloudy with a chance of offloading: The lighter side of edge computing*

Wolfgang John, Ericsson Research, Sweden & Johan Eker, Lund University and Ericsson Research, Sweden

Computational device offloading expands an application's functionality from devices like XR headsets, smartphones, or IoT devices to remote computing environments. This is contrasting with typical edge computing which moves cloud functions closer to the user. An offloading service would allow flexible and dynamic task deployment, enabling users to run their code remotely, anytime, anywhere. This approach is versatile in applicability, reduces device strain, and can improve application QoE, targeting a broader range of enterprises and developers without the complexities of traditional edge computing.

Different variants of device offloading solutions have popped up recently in both academic and industrial research. In this session we want to discuss the technical implications and challenges, promising use-case application scenarios, and potential business opportunities related to device offloading.

14:35 **Discussion 2** (Lillskär)

*Research challenges in multi-cloud networking*

David Breitgand, IBM Research, Israel

Enterprises, especially those from regulated industries, exercise complex communication policies that must be enforced consistently across their entire network. They require detailed visibility into their network traffic and trusted traceability/auditability. This is already a difficult problem in traditional data center networking when interconnecting branches across multiple domains and interacting with the external world. This problem is exacerbated when public Internet or cloud networks are used for intra-enterprise con-
nectivity, because the underlying network details are abstracted away and there is no standardized interface allowing to specify network policies to providers of the network services that would be consistently applied across multiple clouds. In essence, there is a growing need by enterprises to extend their private connectivity multiple clouds while maintaining an illusion of a dedicated network that provides broad control, customization, and optimization options. This would allow them to have continuity with respect to their processes and maintain the level of control that is required by regulations.

Creating, managing, and optimizing workload-aware connectivity for workloads that span multiple clouds, on-prem locations, and edge is difficult. Providers need to deal with the problem of virtual network embedding at scale. This problem is especially pronounced at the edge. Consumers need to minimize the cost of their virtual networks while obtaining consistent performance. This problem is more acute in the cross-cloud setting. Futhermore, connectivity spanning multiple clouds requires uniform approach to policing, private communnitaion, security, and observability. We posit that managing cross-cloud connectivity at the level of IP addresses is inadecvate, becase IP addresses do not reflect neither identity, nor properties of communicating endpoints. One reason for that IP addresses are transformed through multiple NATs and proxy appliances. Furthermore, we argue that imperative orchestration of connectivity is inferior to intent-driven orchestration that would automatically reconcile the observed and desired state of the cross-cloud connectivity to radically simplify management challenges. We will make a strawman proposition of a future cross-cloud connectivity architecture and outline some well-studied fundamental research challenges, e.g., network embedding, scheduling, load balancing, traffic engineering relevant to this new context.

14:35 **Discussion 3** (Storskär)

*Cloud Intelligence – AI/ML for cloud efficiency, quality, and experience*

Jian Zhang, Microsoft, USA; Pamela Delgado, University Applied Sciences Western Switzerland & Cristian Klein, Umeå University and Elastisys, Sweden

With uptick of digital transformation across industries, IT leaders are calling for technologies that can enable them to automate cloud service management to improve service quality, resource and engineering efficiency, and cut down COGS (Cost of Goods Sold). On the other hand, cloud service providers have been investing in developing Cloud Intelligence solutions, which leverage AI/ML to automate and optimize the design, build, and operation of cloud services for efficiency and quality at global scale.

In this session, we are going to discuss challenges and opportunities in leveraging AI/ML including Generative AI for cloud efficiency, quality and experience. Below are some sample questions for the discussion:

1. What are challenges and opportunities in landing AI/ML in cloud service for resource and operation efficiency, service quality, and customer experience?
2. What are the successful applications, challenges and opportunities in leveraging Generative AI for cloud control?
3. What can we do to accelerate collaboration across academia and industry in cloud control space?

14:35 **Discussion 4** (Huvudskär)
*Towards sustainable computing: Exploring green IT perspectives and solutions*
Monica Vitali, Politechnico di Milano, Italy & Abel Souza, University of Massachusetts Amherst, USA

The relevance of IT's environmental impact is increasingly relevant, where cloud data centers alone are responsible for 3% of the world's electricity consumption. On the one hand, the energy consumption of data centers and the demand for computational resources are significantly increasing. On the other hand, improvements in hardware efficiency have not resulted in decreases in data center Greenhouse Gas (GHG) emissions. Additionally, international regulations and growing environmental awareness across several countries are propelling the quest for sustainable computing solutions.

This discussion session aims to engage participants in exploring the current challenges of Green IT and sustainable computing. We'll detail various perspectives to understand the potential and limitations of the state-of-the-art in scheduling, scaling and software tools and techniques currently available for building greener, holistically solutions. S

Specifically, we'll examine three key perspectives:

Infrastructural Perspective: We'll assess the maturity level of the energy-related performance and management of cloud data centers versus edge devices.

Orchestration Perspective: We'll explore how the management of data and applications impacts resource usage efficiency.

Software Perspective: This will involve analyzing the current state of the art in designing and coding environmentally friendly applications.

Through a collaborative effort, we'll identify current trends and solutions for each perspective, considering several aspects such as monitoring tools and mitigation techniques and their associated environmental and economical costs. The outcome will be a comprehensive map delineating available solutions for each perspective and aspect analyzed. This exercise will highlight existing gaps and challenges towards more sustainable IT practices.

14:35 **Discussion 5** (Svavelsö)
*Collaborative connected computing: optimizing the computing continu-um from the cloud to IoT sensors*
Geir Horn, University of Oslo, Norway

The vision of collaborative connected computing is to allow applications to process data closer to where the data is generated, combine data from federated sources, and to use virtualised on-demand computing resources and data from the establish Cloud model through to the resource restricted devices close to the sensors.

16.25 *Making Remote Attestation a Future Commodity for End Users*
Rüdiger Kapitza, FAU Erlangen-Nürnberg, Germany

Confidential computing eases the concerns of distrustful cloud customers by removing the provider and large parts of the cloud infrastructure software from their trusted computing base. This is facilitated by new hardware extensions, like AMD's SEV Secure Nested Paging, which can run a whole virtual machine with confidentiality and integrity protection against a potentially malicious hypervisor owned by an untrusted cloud provider. However, the assurance of such protection to either the service providers deploying sensitive workloads or the end-users passing sensitive data to services requires sending proof to the interested parties. Service providers can retrieve such proof by performing remote attestation, while end users typically have no way to obtain or validate this proof and must therefore rely on the trustworthiness of the service providers.

This talk will discuss how remote attestation can be exposed to end users in a meaningful way. This includes integrating remote attestation into standard browsers for widespread adoption and ease of use. It will also evaluate how remote attestation results can be processed to provide useful assurances to users, thereby offering a basis for increased trust when accessing a remote service protected by confidential computing. Finally, Revelio is presented as a first building block towards more secure and confidential access to remote services. Revelio provides two main contributions: i) it allows confidential virtual machine-based workloads to be designed and deployed in a way that prevents any tampering, even by service providers, and ii) it allows users to easily validate their integrity via browser-based remote attestation.

16.45 *Reconciling High Accuracy, Cost-Efficiency, and Low Latency of Inference Serving Systems*
Pooyan Jamshidi, University of South Carolina, USA

ML inference services serve user requests directly, requiring fast and accurate responses. Moreover, these services face dynamic workloads of requests, imposing changes in their computing resources, and failing to right-size computing resources results in either latency service level objectives (SLOs) violations or wasted computing resources. Adapting to dynamic workloads for ML inference pipelines is a difficult problem because of the exponentially large design space (multiple interacting and interdependent components, each with different knobs that influence performance) and multiple competing performance objectives, and conflicting user preferences (accuracy, latency, and cost). In addition,

the specific reconfiguration must be decided in real-time and with incomplete and imperfect knowledge due to dynamic workload, variable network latency, and variable resource availability. In this talk, I will present our recent solutions to the abovementioned challenges: InfAdapter [1] combines model-switching and auto-scaling to enable a more granular design space to trade accuracy, cost, and latency in inference serving systems; IPA [2] dynamically re-configures ML inference pipelines to achieve the tradeoff; and Sponge [3] deals with dynamic SLOs and achieves its goal by applying in-place vertical scaling, dynamic batching, and request reordering.

[1] Mehran Salmani, Saeid Ghafouri, Alireza Sanaee, Kamran Razavi, Max Mühlhäuser, Joseph Doyle, Pooyan Jamshidi, and Mohsen Sharifi. 'Reconciling High Accuracy, Cost-Efficiency, and Low Latency of Inference Serving Systems.' In: Proceedings of the 3rd Workshop on Machine Learning and Systems. EuroMLSys. 2023.

[2] Saeid Ghafouri, Kamran Razavi, Mehran Salmani, Alireza Sanaee, Tania Lorido-Botran, Lin Wang, Joseph Doyle, and Pooyan Jamshidi. 'IPA: Inference Pipeline Adaptation to Achieve High Accuracy and Cost-Efficiency.' In: Journal of Systems Research (JSys) (2024)

[3] Kamran Razavi, Saeid Ghafouri, Max Mühlhäuser, Pooyan Jamshidi, and Lin Wang. 'Sponge: Inference Serving with Dynamic SLOs Using In-Place Vertical Scaling.' In: Proceedings of the 4th Workshop on Machine Learning and Systems. EuroMLSys. 2024.

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

***Poster 1.*** *Dynamic Offloading of Control Algorithms to the Edge using 5G and WebAssembly*,
Ahmed Al Bayati, Lund University, Sweden

The aim of this work was to test if WebAssembly universal byte code and 5G communication technology is suitable in the context of offloading control missions of real-time systems. To test these tools a new dynamic offloading framework was implemented and tested on a Furuta pendulum, an inherently unstable and time-critical process using, among other computing units, an edge node as the offloading target. The implementation is considered dynamic because: 1) The local device which interacts with the I/O of the process dynamically send the control application to be used by the edge node. 2) The local device dynamically decides on which controller should control the process, either the local fallback LQR controller or the CVXGEN Model Predictive Control (MPC) solver written in an Ahead-of-Time (AOT) WebAssembly (Wasm) format, which is used by the edge node.

***Poster 2.*** *Cheops: Externalizing Geo-Distribution for Cloud Applications at the Edge*
Geo Johns Antony, INRIA, France

The advent of edge computing introduces a new challenge for cloud applications: how to leverage geo-distribution while managing the constraints of wide-area network links. Traditionally, this requires modifying them to integrate geo-distribution into the business logic, complicating the code and contradicting the software engineering principle of externalizing concerns. We propose a different approach using the modularity of microservices: (i) deploying an instance at each edge location ensures system robustness against network partitions by satisfying local requests, and (ii) enabling collaborations between instances outside the application. Collaboration brings challenges related to synchronization and dependencies between multiple instances. Existing approaches for it are intrusive, but our proposal includes to externalize these challenges. This non-invasive method is enabled by (i) a DSL extending the application API, (ii) an orchestrator to manage geo-distribution.

***Poster 3.*** *PerMFL: Personalized Multi-tier Federated Learning*
Sourasekhar Banerjee, Umeå University, Sweden

The key challenge of personalized federated learning (PerFL) is to capture the statistical heterogeneity properties of data with inexpensive communications and gain customized performance for participating devices. To address these, we introduced personalized federated learning in multi-tier architecture (PerMFL) to obtain optimized and personalized local models when there are known team structures across devices. Moreau envelopes are used as the devices' and teams' regularized loss function. We provide theoretical guarantees of PerMFL for smooth, strongly convex, and non-convex settings. PerMFL achieves linear convergence rates for strongly convex problems and sub-linear rates for non-convex problems. We have conducted empirical experiments and confirmed that PerMFL achieves faster convergence and outperforms the current state-of-the-art methods.

***Poster 4.*** *Monogamous Relationships with Short-term Commitment are the Best*
Carl Magnus Bruhner, Linköping University, Sweden

Certificates are the foundation of secure communication over the internet. However, not all certificates are created and managed in a consistent manner and the certificate authorities (CAs) issuing certificates achieve different levels of trust. Furthermore, user trust in public keys, certificates, and CAs can quickly change. Combined with the expectation of 24/7 encrypted access to websites, this quickly evolving landscape has made careful certificate management both an

important and challenging problem. In this paper, we first present a novel server-side characterization of the certificate replacement (CR) relationships in the wild, including the reuse of public keys. Our data-driven CR analysis captures management biases, highlights a lack of industry standards for replacement policies, and features successful example cases and trends. Based on the characterization results we then propose an efficient solution to an important revocation problem that currently leaves web users vulnerable long after a certificate has been revoked.

This poster presents the work of the following paper:
Carl Magnus Bruhner, Oscar Linnarsson, Matus Nemec, Martin Arlitt, Niklas Carlsson, Changing of the Guards: Certificate and Public Key Management on the Internet, Proc. Passive and Active Measurement Conference (PAM), Mar. 2022. https://doi.org/10.1007/978-3-030-98785-5_3 Available: https://www.ida.liu.se/~nikca89/papers/pam22.pdf

***Poster 5.*** *End-To-End Performance, Energy Consumption and Carbon Footprint of Fog Applications*
Clément Courageux-Sudan, Umeå University, Sweden

Deploying applications close to end-users through Fog computing can reduce network latency and contention. However, distributed applications in the Fog have intricate interactions between heterogeneous network and processing devices. To understand the impact of application and infrastructure parameters on performance, the literature offers end-to-end models lacking granularity and validation, or fine-grained models missing part of the infrastructure. We combine a collection of validated models to obtain comprehensive metrics regarding microservice applications operating in the fog. Our approach can investigate fog environments from application latencies to greenhouse gas emissions.

***Poster 6.*** *Machine Learning for Anomaly Detection in Edge Clouds*
Javad Forough, Umeå University, Sweden

The IoT and 5G advancements create opportunities for intelligent applications across various domains like public services, transportation, augmented reality, automation, and healthcare. Traditional centralized cloud computing struggles with bandwidth and latency for edge applications. Edge clouds, comprising edge nodes, fog nodes, and distant clouds, aim to address these challenges. However, there are numerous security and performance challenges with edge clouds. This project focuses on designing decentralized anomaly detection and security measures using machine learning to enhance performance and security in edge clouds.

***Poster 7.*** *Power Focused Green Energy Modeling in Cloud Edge Continuum Regions*
Rohail Gulbaz, Umeå University, Sweden

The continuum environment is evolving towards powerful infrastructures and the complex nature of continuum has kept researchers to integrate all energy aspects in model. This raises the need to investigate, what are the model components and their interactions within the continuum. The vast continuum environment is granulated into separable dependent layers including the Power Infrastructure Layer. Different components identified are diversity of power sources, intermittent production, peak hours, cost, thermal aspects, impact of weather, and usage behavior etc.

Following categories of regions define a continuum:

1) Cloud Service Provider Regions
2) Power Source Provider Regions
3) Geographical Regions

Our hypothesis is that formal modeling of Power Infrastructure, holistically across the continuum can assist to achieve better energy efficiency. Hence, to find the impact of modeling on energy, we formulated a formally modeled task scheduler, and contrasted with another existing approach.

***Poster 8.*** *Adaptive and Data Efficient Cloud Management: Enhancing Machine Learning for Efficient Resource Allocation*
Lidia Kidane, Umeå University, Sweden

Autonomous Cloud and Edge Management systems utilize machine learning to optimize resource allocation and application deployment, ensuring quality of service while minimizing costs and energy usage. The systems need to be frequently retrained and to calibrate their models to capture the latest information available. This eventually enables them to adapt to changing contexts, like software or hardware upgrades, and accommodate increasing workloads. Moreover, the amount of monitoring data that can be available from massive-scale cloud data centers is virtually unlimited. Thus, selecting relevant monitoring data amidst massive data volumes is challenging, requiring methods to minimize storage needs and training time. Enhancing machine learning techniques for this scenario involves addressing both data selection and training efficiency, to meet the demands of large-scale cloud environments effectively.

***Poster 9.*** *Avoiding the Heat Death of Kubernetes and the CNCF Landscape*
Lucas Käldström, Finland

The concept of entropy is a measure of disorderliness or chaos. The second law of Thermodynamics states that the Universe evolves spontaneously towards more chaotic states, eventually to "the Heat Death". Kubernetes started off as a container orchestrator for stateless web apps. But now, what once was an orderly list of use-cases, has become a turbulent sea of possibility and complexity. This is also the case for the CNCF Landscape as a whole. With novel use cases in e.g. AI, cloud native will also need to evolve, increasing entropy. However, as we navigate these possibilities with Kubernetes at the base, it is critical that we talk about some of the philosophies and early decisions of the project, as well as how they have fared with an evolving industry. In doing so, we understand what we can rely on it for and what we can't. Continuing from Tim Hockin's KubeCon keynote, join us as we talk about the physics of cloud native and how our community can deal with unseen use cases and scale.

***Poster 10.*** *Application-Aware Hardware-Level QoS Enforcement Tuning Using Extremum Seeking Control*
Oliver Larsson, Umeå University, Sweden

Recent advances in chip technology have enabled the dynamic tuning of shared memory resources such as last level cache and memory bus bandwidth. However, the indirect and unmodelable nature of these QoS enforcement features have limited their adoption in real-world cloud computing environments. We propose a novel approach to platform QoS enforcement tuning based on extremum seeking control that places applications and ease-of-use in focus. We show that our controller can efficiently tune hardware-level QoS to maintain a setpoint in application-level metrics while improving best-effort workload throughput without profiling or prior knowledge of a system.

***Poster 11.*** *Stealthy Delay Attacks for Cyber-Physical Systems*
Talitha Nauta, Lund University, Sweden

Researchers have identified that control systems are vulnerable to security attacks, which discreetly alter control signals or output measurements to avoid detection. Successful attacks need unauthorised access to the system's components and a deep understanding of its internal workings, including its intrusion detection mechanisms. While theoretical models explain and determine how to conduct these stealthy attacks, their practical implementation is challenging because it requires overcoming security defences and maintaining updated knowledge of the system. A more feasible method of attack involves manipulating the timing of controller operations, subtly impacting system performance without altering data or control signals directly. This timing attack, hard to detect, poses a significant threat as it circumvents the standard security measures that focus on data integrity. We explore the attack theoretically and present preliminary results.

***Poster 12.*** *Modeling the Green Cloud Continuum: integrating energy considerations into Cloud–Edge models*
Yashwant Singh Patel, Umeå University, Sweden

The energy consumption of Cloud–Edge systems is becoming a critical concern economically, environmentally, and societally; some studies suggest data centers and networks will collectively consume 18% of global electrical power by 2030. New methods are needed to mitigate this consumption, e.g. energy-aware workload scheduling, improved usage of renewable energy sources, etc. These schemes need to understand the interaction between energy considerations and Cloud–Edge components. Model-based approaches are an effective way to do this; however, current theoretical Cloud–Edge models are limited, and few consider energy factors. This poster analyses all relevant models proposed between 2016 and 2023, discovers key omissions, and identifies the major energy considerations that need to be addressed for Green Cloud–Edge systems (including interaction with energy providers). We investigate how these can be integrated into existing and aggregated models, and conclude with the high-level architecture of our proposed solution to integrate energy and Cloud–Edge models together.

***Poster 13****. Edge Orchestration for Video Analytics Applications*
Ali Rahmanian, Umeå University, Sweden

Ever-increasing and compute-intensive demands for real-time video analytics applications with stringent latency constraints necessitate cutting-edge solutions. Edge computing emerges as a promising infrastructure offering constrained compute resources in close proximity to end users and data sources. However, as demand for video analytics continues to surge, addressing latency constraints on constrained edge resources becomes increasingly imperative. We aim to investigate innovative application and system solutions for resource orchestration, facilitating the efficient deployment of real-time video analytics on the edge.

***Poster 14.*** *Mitigating Slow-Rate Data Plane Attacks using Model-Free Reinforcement Learning*
Kshira Sagar Sahoo, Umeå University, Sweden

SDN have revolutionized networking with programmability, but limited TCAM space in switches exposes them to Slow Rate Data Plane (SRDP) attacks. These attacks flood flow tables with recurrent flow rules, leading to packet drops. To counter this, a model is proposed, using Q-learning to evict repeated flow rules based on attack strength. P4 SDN switch

experiments show effective SRDP attack mitigation with minimal overhead, achieving a practical 2% malicious rule restriction.

***Poster 15.*** *Constraint-Based Service Placement Optimization for IoT Applications in Fog environments*
Farah ait-Salaht, Léonard de Vinci Pôle Universitaire, France

Fog Computing enhances Cloud Computing by extending it to the network edge, fulfilling the real-time needs of IoT applications. This poster presents the CP-SPP model, which uses Constraint Programming to tackle the Service Placement Problem in Fog Computing infrastructures. The CP-SPP model employs the flexible language of CP to minimize effort in placement description and adapt to varied Fog and applications needs. The poster also highlights the model's comprehensiveness, flexibility, and upgradeability, addressing joint service replication and placement through the CP-SRP model. Building on CP-SPP, the CP-SRP model addresses workload variability through one-stage service replication and placement optimization. This approach enables elastic and parallel service processing, adjusting to the demand fluctuations of IoT data stream applications with low response times. Evaluations with Choco-solver demonstrate our model's scalability and performance. Future work will explore the adaptability of our model for service sharing and dynamic deployment adjustments, including migrations.

***Poster 16.*** *Resource Management of IoT Applications in Edge-based Infrastructures*
Hajar Siar, Umeå University, Sweden
This poster presents two papers addressing resource management of IoT applications in edge-based infrastructures, considering the QoS requirements. The first one proposes a game-theoretic solution for efficient allocation and scheduling of independent IoT applications with indivisible loads in edge computing. By considering job demands, and deadlines, the algorithm minimizes response time and maximizes the number of jobs completed within their deadlines helping a weighted time sharing scheduling. The second paper investigates offloading and allocation in fog computing for workflow ensembles.
A coalition game-based mechanism is introduced, aiming to minimize execution time and cost while meeting Quality of Service (QoS). Considering the diverse applications of machine learning in edge computing, and their complexity, especially deep learning, we will focus on resource management of these applications in the future works to facilitate their deployment in edge-based infrastructures.

***Poster 17.*** *An Edge-Cloud Infrastructure for Fault-Tolerance Mission-Critical Applications*
Nayereh Rasouli, Umeå University, Sweden

Fault tolerance is vital in IT systems, particularly during natural disasters such as earthquakes and hurricanes, which can damage components or disrupt networks, affecting essential applications and complicating disaster relief. Although prior research has addressed disaster management in Mobile Edge Computing (MEC) systems, integrating robust fault tolerance remains understudied. Our study utilizes contemporary technologies like Kubernetes to manage node failures effectively. We introduce an infrastructure with RabbitMQ as a resilient message broker, ensuring reliable communication even during severe disruptions. Our tailored fault tolerance solution for MEC systems includes a holistic disaster recovery strategy, validated by a case study with weather stations in urban-forest adjacent areas, demonstrating the system’s capability to sustain dual node failures and maintain 99.966% availability for critical operations

***Poster 18.*** *Analysis of control systems subject to timing discrepancies*
Yde Sinnema, Lund University, Sweden
Timing discrepancies affect the performance of real-time systems. For example, a timing misalignment between sensor channels can deteriorate the outcome of a sensor fusion algorithm. Even in a less complex setting, the presence of sensing delays can compromise the stability and performance of a control system.

This ongoing work aims to analyse the impact of sensing delays on control systems that receive measurements from more than one sensor channel. We address temporal misalignment and time-varying delays in particular. We present an analysis framework using switching systems and illustrate the effect of delays in a case study.

***Poster 19.*** *Eco-Efficiency in Cloud Computing: A Carbon-Conscious Reinforcement Learning Approach to Federated Learning*
Antonio (Eunil) Seo, Umeå University, Sweden

This research addresses the urgent need to mitigate climate change by pioneering a dual-objective optimization strategy within a federated learning framework. Our approach integrates a novel carbon-conscious methodology to reduce carbon emissions while optimizing energy consumption and maintaining high-quality intelligent applications. We employ reinforcement learning (RL) to dynamically adjust client contributions based on environmental impact and alignment with global model parameters. Extensive experiments demonstrate that our strategy significantly lowers carbon emissions and enhances energy efficiency, outperforming conventional methods with a 61.78% improvement in energy conservation and a 64.23% reduction in carbon emissions. This research validates the effectiveness of incorporating

environmental metrics into computing processes, highlighting substantial benefits for ecological sustainability in cloud computing.

***Poster 20.*** *Dynamic Computation Offloading for a Cluster of Drone*
Emil Sundström, Lund University, Sweden

With an increasing demand for computationally intensive tasks on drones, such as video analysis and advanced control algorithms, dynamic offloading becomes an important concept to achieve that behavior. Hence, this WASP-funded research project aims to explore dynamic offloading for drones in harsh environments from a control systems perspective. The current focus is to investigate the decision-making process of when to offload computations from the individual devices in a cluster of drones.

***Poster 21.*** *Energy-aware Network Model in Cloud-Edge Continuum*
Yangyang Wen, Umeå University, Sweden

This research addresses the urgent need for sustainable and energy-efficient computing solutions within Cloud-Edge environments. It introduces a formal model that integrates energy sustainability and consumption considerations into Cloud-Edge systems and workloads, with a primary focus on optimizing energy efficiency and resource utilization.

The study's initial phase concentrates on the network layer within Cloud-Edge systems. By categorizing various energy, network, hardware, and software components into integrating layers, the research aims to comprehensively understand their dynamics and interactions, Leveraging metrics from the Ericsson Research Data Center and prior WCIB project efforts, the methodology includes creating a formal model, identifying workload types, and estimating energy consumption within data centers.

Moving forward, the research will emphasize decentralized and scalable monitoring and metric collection, exploring lightweight monitoring services integration across Cloud-Edge infrastructures.

***Poster 22.*** *Federated Learning for Cloud Robotic Manipulation*
Obaidullah Zaland, Umeå University, Sweden

Traditional machine learning algorithms enable mobile robots to understand and interact with their environments. Considering the limited resources for on-device training, the learning process can be carried out either on-device or in-cloud. Federated learning enables cloud robotics systems to share their learned knowledge with their peers without explicitly sharing their data. As concerns for data privacy have grown, federated learning has become one of the go-to methods for learning distributed models. While federated learning has been the focus of extensive research in the last few years, problems such as handling heterogeneous data, efficient communication for low-computation devices, and scalability still require improvement, especially when many clients have been deployed. In this project, we will mainly focus on 1) Representation learning, 2) Scalability, and 3) Robustness in federated learning. As the main application of this project is cloud robotics, in representation learning, we want to explore creating uniform representation for heterogeneous data and representation alignments. In 2, we will investigate deploying federated learning models at scale from thousands of devices to millions. Furthermore, in 3, we will create robust federated learning algorithms that can specifically work better in a cloud robotics environment and can also generalize to other applications.

***Poster 23.*** *Securing Serverless Edge AI for the Cloud-Edge Continuum*
Zhou Zhou, Umeå University, Sweden

Serverless is an emerging technology that significantly empowers the cloud-edge continuum paradigm since its intrinsic characteristics include user-friendly, pay-as-you-go, strong elasticity, and on-demand resource availability. Edge AI (artificial intelligence) aims to bring intelligence close to the device, where data are generated that includes diverse machine learning applications. The combination of serverless technology and edge AI becomes a modern technosphere termed serverless edge AI. However, ignoring the security of serverless edge AI hinders its widespread applications and the systems as well. Based on our investigation and analysis, the security issue in serverless edge AI is two-fold: secure edge AI and serverless system. For the former, the AI models, in principle, can be employed for diverse tasks, either for optimising resource usage or for orchestration of resources on the serverless platform when deployed in the cloud-edge continuum. The vulnerabilities of AI models could be exploited by attackers and, for example, manipulated data, model parameters or decisions. For the latter, the current serverless system also has some distinctive vulnerabilities, like shared resources and intermittent communication between services, which could also expose the serverless system to attackers. Further, when it combines serverless edge AI and cloud-edge continuum, these problems boost difficulty. Therefore, when deployed in the cloud-edge continuum, devising offensive strategies for assessing and exploring security disputes in the serverless edge AI are challenging and plan to address by providing secure solutions. In this presentation, the primary focus will be on assessing the impact of poisoning attacks in AI models and serverless systems.

# Wednesday June 19th

8:20 *Metrics for building warehouse scale computers*
John Wilkes, Google, USA

If you cannot measure something, the only safe thing to assume is that it is out of control. Putting up (and filling) a warehouse-scale computer is a monumental task, rife with complications and opportunities for mistakes - some of which can cost $x-xxM. I'll talk about some of the issues we face in doing this at scale, and a few of the approaches we use to keep the process more or less on track.

9:00 *Dynamically Controlling Infrastructure as Code*
Claus Pahl, Free University of Bozen-Bozano, Italy

Software management and its quality in the operation of infrastructures is becoming increasingly important. Sample contexts are general automation, cloud and edge, and various software-defined networking applications. DevOps practices that are already applied in the Infrastructure as Code (IaC) context need to be extended towards dynamic, automated control. The ultimate objective would lead towards full self-adaptation of IaC code. The presentation aims to review state-of-the-art and challenges.

9:20 *GPU-Cluster Management for a Stream of Distributed Deep Learning Jobs*
George Kesidis, Pennsylvania State University, USA

For many years and to the present day, managing cloud-spend has remained an important and open problem for cost-conscious customers of the public cloud running complex workloads. One challenge of managing a cluster of procured cloud services executing a stream of jobs is the gap between user-specified service-level objectives (SLOs, performance and cost criteria) and various resource-management decisions for the cluster that must be made, e.g., selecting the service suite and assigning resources to the currently active jobs. The case of a stream of diverse Distributed Deep-Learning (DDL) jobs is increasingly important, e.g., to perform nightly updates of recommender models, or from-scratch DL by a DLaaS cluster. Typical DDL jobs require a plurality of GPUs (or similar hardware accelerators), have a periodic resource-usage profile, and can involve execution times on the order of hours to weeks. Also, deep learning is a highly heuristic process that is rather robust to, e.g., intermittent "irregularities" in the participation of the training dataset in parameter updates (so long as they are not persistent and cause or exacerbate model bias). Prior work has shown how different DDL jobs are distinctly sensitive to network delays in the exchange of the GPU-computed gradient information to update model parameters. There is a need to create practical and comprehensive cluster management platforms for a stream of heterogeneous DDL jobs to satisfy user-specified SLOs including spending constraints. The cluster management will be customized to take advantage of special properties of DDL jobs. We will discuss the role that online optimization methods can play including those that employ near-future workload forecasting. Also, we will describe complications due to "TinyML" operations to reduce the model size (e.g., requiring workload re-profiling) and due to methods which detect and mitigate biases and backdoors. In addition to accommodating smaller hardware form factors, smaller models are more energy efficient. Finally, we will discuss approaches to managing congestion in the cluster's RDMA-over-Ethernet infrastructure.

10:10 *Multi-scale Feedback Architectures for Decentralised Control*
Ada Diaconescu, Télécom Paris, France

Control decentralisation may enhance resource distribution, resilience and adaptability. Yet, in large systems, it introduces the difficult issue of coordination among numerous, quasi-independent controllers. This problem has been addressed in both natural and artificial systems via various forms of hierarchical structures – also referred to as holonic, multi-level, micro-macro or fractal systems. We refer to these generically as “multi-scale” feedback systems. Notable examples include complex adaptive systems such as organisms, societies, eco-systems, or socio-technical organisations. Within the Cloud Control domain, this concerns the coordination of controllers in Cloud Data Centres (thousands), Fog Nodes (millions) and Edge Devices (billions). While multi-scale solutions have been extensively studied and implemented within domain-specific contexts, a general theory of multi-scale feedback systems applicable cross-domain is still missing.

This talk will present our research towards establishing such theory. It proposes a generic design pattern – Multi-scale Feedbacks System. It consists of multiple information flows that merge and split through the system, abstracting and reifying information at different scales, and forming inter-scale cycles that ultimately link data collection to control action through multi-scale feedback loops.

We show how varying such information flows (e.g. abstraction functions, topology, inter-scale communication delays) impact the overall system properties (e.g. stability, adaptability, resilience, scalability). These generic tendencies may provide guidance to system engineers when selecting appropriate design variants for specific application domains. Such

theory allows transfer of expertise across application domains concerned with complex adaptive control – notably including cloud systems embedded within highly-competitive and continuously-evolving socio-technical contexts.

10:30 *Low-Latency Privacy-Preserving Authentication in the Edge*
Luis Rodrigues, INESC-ID University of Lisboa. Portugal
Authentication based on pseudonyms offers both accountability and privacy, protecting clients from curious application providers. A challenging task in this context is to support revocation with- out violating privacy. In this talk we discuss strategies to support revocation while ensuring backward unlinkability. We describe a novel abstraction that we have named Range-Revocable Pseudonyms (RRPs). RRPs are a new class of pseudonyms whose validity can be revoked for any time-range within its original validity period. The key feature of RRPs is that the information provided to revoke a pseudonym for a given time-range cannot be linked with the information provided when using the pseudonym outside the revoked range. We provide an algorithm to implement RRPs using efficient cryptographic primitives where the space complexity of the pseudonym is constant, regardless of the granularity of the revocation range, and the space complexity of the revocation information only grows logarithmically with the granularity; this makes the use of RRPs far more efficient than the use of many short-lived pseudonyms.

10:50 *A Journey above the Clouds: Cloud Control in new Contexts*
Hermann De Meer, University of Passau, Germany

Efficient resource management is vital for data centers and communication networks to meet quality of service standards. This importance extends to emerging non-terrestrial networks that use satellites and base stations as resources. Additionally, non-routed networks such as energy grids hold immense potential for optimizing resource utilization.

In the realm of satellite technology, driven by initiatives like Starlink and Amazon's Kuiper project, the focus expands beyond Internet provision to encompass 6G mobile communications and direct satellite-device connections. Satellites equipped with onboard processing capabilities streamline data flow, reminiscent of the autonomy demonstrated by Ingenuity's Mars AI. Laser communications promise swift, efficient connectivity across vast distances, which requires novel network architectures.

The integration of cloud systems into decentralized critical infrastructures necessitates innovative control mechanisms. Applying techniques of network virtualization can facilitate efficient resource sharing but presents distinct challenges in cyber-physical systems. Adapting cloud control for these contexts requires a fusion of traditional methods with innovative approaches to navigate both opportunities and obstacles effectively.

11:10 **Discussion 6** (Resarö)
*Overcommitment and workload autoscaling in cloud datacenters: a provider's perspective*
Krzysztof Rzadca, Google, Poland

Overcommitment and workload autoscaling are closely related mechanisms that significantly enhance the efficiency of large-scale datacenters. I'll begin the discussion session by outlining the production systems used in Google's internal cloud. These systems use fairly straightforward yet robust resource estimation methods. Next, I'd like to engage the audience in exploring open questions and design decisions, such as the interpretability of ML-based predictions and the trade-off between efficiency and robustness.

In many public and private cloud systems, users must specify resource limits (CPU cores and RAM) for their workloads. Jobs exceeding these limits risk throttling or termination, potentially delaying or dropping end-user requests. Consequently, operators often request higher limits, leading to significant resource wastage. Google addresses this with Autopilot, which automatically configures resources, adjusting both the number of concurrent tasks (horizontal scaling) and individual task limits (vertical scaling). Autopilot's primary goal is to reduce slack—the difference between the limit and actual usage—while minimizing the risk of out-of-memory (OOM) errors.

To further increase utilization, data center schedulers often overcommit resources where the sum of resources allocated to the tasks on a machine exceeds its physical capacity. Determining the right overcommitment level is challenging: low levels waste resources, while high levels degrade task performance. We approach overcommit policy design and evaluation from first principles, asking: Assuming perfect knowledge of future task resource usage, what's the safest policy yielding the highest utilization? We term this the "peak oracle" policy. We then develop practical policies that emulate this oracle by predicting future machine resource usage. Deployed within Google's data centers, these policies increase usable CPU capacity by 10-16% compared to no overcommitment.
Joint work with Noman Bashir, Nan Deng, David Irwin, Sree Kodak, and Rohit Jnagal (overcommitment); and with Pawel Findeisen, Jacek Swiderski, Przemyslaw Zych, Przemyslaw Broniek, Jarek Kusmierek, Pawel Nowak, Beata Strack, Piotr Witusowski, Steven Hand, John Wilkes (autoscaling).

11:10 **Discussion 7** (Rindö)

*Geo-distributed storage: addressing low latency, weak throughput, and disconnections*

Thomas Fahringer, University of Innsbruck, Austria; Mina Sedaghat, Ericsson, Sweden & Adrien Lebré, Centre Inria de l'Université de Rennes, France

Continuum computing enables applications and services to utilize the advantages offered by both the Edge and the Cloud: low latency, data locality, and low network traffic at the Edge, and capacity and capability computing, high availability, and reliability provided by the Cloud.

In this discussion, we want to explore solutions to support scalable data processing and storage across the continuum, considering the limitations of the infrastructure and application requirements, such as HW limits for data storage, costly data transmission between sites, and applications requirement on data access, retrieval times, data consistency and data resiliency.

Numerous types of storage systems are used for the computing continuum which Includes key-value stores, object stores, graph databases, distributed files systems, relational databases, time series databases, and many more. This discussion topic puts the focus on usability, high flexibility, consistency, low latency, disconnections, and scalability.

We propose to discuss the following questions in the context of the computing continuum:

- How can distributed storage systems be optimized for real-time processing?
- How can distributed storage systems improve energy efficiency?
- What is the role of storage systems to improve resilience of distributed applications?
- What other optimization goals are relevant/interesting?
- How to quantify and optimize costs of data synchronization, in particular in case of network partitions?
- What are the implications of using different types of storage systems (e.g., key-value, time series, object stores) on the performance and functionality of IoT applications?
- What are the challenges and solutions for achieving data consistency in distributed storage systems?
- What data consistency models should be offered by modern distributed data repositories?
- How can scalability be achieved without compromising on performance of distributed storage systems?
- How can distributed storage systems facilitate collaboration among dispersed IoT devices and applications?
- Should there be some control provided at the API-level to control on which resource, layer, location data should be stored?
- What are good data placement approaches to decrease latency and reasonable costs?

11:10 **Discussion 8** (Storskär)

*The future of orchestration and resource management: steering innovation in the era of hardware abstraction*

Thijs Metsch, Intel, Germany & Oliver Larsson, Umeå University, Sweden

Low-code platforms, managed services, and serverless deployments share a common concept: they all minimize the user's need to manage or even understand the underlying hardware. This shift raises a pivotal question about efficiency: "Are we still using our resources as effectively as possible?" As the technology landscape evolves, traditional control systems like Kubernetes, which is nearing its decade mark, are being reassessed. What will the next generation of control planes look like?

In this interactive discussion, we aim to delve into the challenges and opportunities that future, or even current, control plane implementations must address. With advancements in AI, the rise of intent-driven systems, and other innovations, what are the next big steps for managing distributed systems? How can these technologies contribute to more intuitive and efficient infrastructures?

Join us as we explore these questions and more. We encourage you to bring your insights and questions to what promises to be a lively and informative session. Let's collaborate to uncover what the future holds for control plane technology. Your ideas can help shape the next big breakthrough. We're excited to learn together: What's next?

11:10 **Discussion 9** (Huvudskär)

*Observability and diagnostics in the 5G networks and beyond when embracing the intelligent and autonomous networks paradigm*

Monowar Bhuyan, Umeå University, Sweden and Johan Forsman, Tietoevry, Sweden

The advent of 5G networks and the anticipated transition beyond 5G (B5G) and 6G networks mark a paradigm shift towards intelligent and autonomous network operations. This shift necessitates enhanced observability and sophisticated diagnostic mechanisms to ensure network reliability, performance, and security. This discussion wants to explore the critical role of observability and diagnostics within the context of intelligent and autonomous networks. We investigate

the unique challenges posed by the increased complexity and dynamic nature of these networks, including the integration of artificial intelligence (AI) for real-time monitoring and anomaly detection. By embracing advanced observability and diagnostic strategies, the next generation of networks can achieve unprecedented levels of automation and intelligence, paving the way for robust, self-healing, and resilient communication infrastructures.

11:10 **Discussion 10** (Svavelsö)
*Foundation models for edge intelligence*
Feras M. Awaysheh, Tartu University, Estonia

Edge-AI facilitates distributed training and fine-tuning of models, enabling the creation of personalized and adaptive models tailored to specific edge environments or user preferences without transferring substantial amounts of data to centralized servers. Additionally, edge computing platforms streamline foundation models' deployment, management, and scaling across distributed edge infrastructure. The session entails advancing and practically applying foundation models in Edge-AI services. The symposium places significant emphasis on training and inference optimization techniques aimed at substantially enhancing the speed and efficiency of these models for edge deployment while also facilitating sophisticated on-device learning.

This session aims at assembling leading researchers actively engaged in discussion of the following areas:

- The exploration of quantization strategies at the edge, aiming to effectively reduce the size of AI models to better fit the unique constraints of edge devices. Additionally, discussing the critical need for parameter-efficient fine-tuning, enabling large-scale distributed models to adapt to specific tasks while minimizing the requirement for extensive computational resources.
- The study and application of Model Distillation seek to streamline complex models to be compatible with the limitations of edge computing without compromising their effectiveness. The session will also discuss the growing importance of hardware acceleration, examining how the integration of specialized hardware can dramatically boost AI performance at the edge.
- A focus on Energy-Efficient Inferencing Architectures, highlighting the necessity for sustainable, power-aware solutions in edge computing environments.
- An in-depth exploration of Federated Learning (FL) and its pivotal role in harmonizing the expansive capabilities of large-scale models with the imperative of safeguarding data privacy, achieved through the distributed training of models across multiple nodes.

13:45 *A CarbonFirst Approach for Decarbonizing Cloud Computing*
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 Dennard 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 continued exponential growth will translate into rising energy demand, which, if left unchecked, will translate to increasing carbon emissions.

In this talk, I will argue for a CarbonFirst approach to designing cloud computing systems by making carbon efficiency a first-class design metric, similar to traditional metrics of performance and reliability. I will explain how today's systems can be made first carbon-aware by exposing energy and carbon usage information to software platforms and then made carbon-efficient by providing control over the system's carbon usage. I will present an initial design of a system to enable such carbon awareness and management and present several application case studies on how modern cloud applications can employ these mechanisms to reduce their carbon footprint. I will end with open research challenges in the emerging field of computational decarbonization.

14:15 *The US National Artificial Intelligence Research Resource (NAIRR) Pilot*
Dilma Da Silva, Texas A&M University, USA

This talk describes the recently launched US National AI Research Resource (NAIRR) Pilot effort led by US National Science Foundation. The NAIRR was envisioned as a widely accessible AI research cyberinfrastructure that brings together computational resources, data, testbeds, algorithms, software, services, networks, and expertise. Such a resource would help to democratize the AI research and development (R&D) landscape in the United States.

14.35 **Discussion 11** (Resarö)
*Above the clouds: Potentials of non-terrestrial edge and cloud computing*
Hermann de Meer, University of Passau, Germany

Low-earth orbit satellite technology is expanding at an exponential pace, driven by ambitious ventures from companies like Starlink and Amazon with their Kuiper project. Beyond merely providing internet connectivity through large antennas, the next frontier lies in 6G mobile communication, which envisions direct satellite-to-handheld device communication. In fact, the feasibility of high-bandwidth communication has already been demonstrated using 5G technology in pioneering experiments, such as those conducted in Hawaii [1].

One intriguing development in this domain is the increasing processing power embedded within satellites themselves. This trend opens up a plethora of possibilities for remote computing services. For instance, the pre-processing of satellite image data could be efficiently conducted directly onboard satellites or within a network of interconnected satellites, thereby minimizing the volume of data that needs to be transmitted back to Earth. A remarkable precedent for this lies in the autonomous capabilities of the Ingenuity Mars helicopter, which autonomously performs image processing and movement control calculations using its onboard AI processor. Looking ahead, this processing power could be shared among satellites or even extend to undertaking terrestrial tasks, such as hosting webpages directly from space.

Satellites offer several inherent advantages for cloud computing resources. Firstly, they benefit from efficient and predictable renewable energy sources, courtesy of photovoltaic systems, without the interference of atmospheric conditions or cloud cover. Moreover, the advent of laser-based communication systems enables direct, high-speed communication between satellites in line of sight and end-users on Earth, the Moon, or even other celestial bodies. This fosters the emergence of new dynamic network structures, transcending traditional boundaries and ushering in an era of interconnectedness that spans the cosmos.

In essence, non-terrestrial cloud computing represents a paradigm shift in how we harness the potential of space-based assets, revolutionizing communication, data processing, and enabling breakthroughs in Earth observation and interplanetary exploration. As humanity continues to push the boundaries of technological innovation, the sky is no longer the limit — it's just the beginning.

[1] https://www.vodafone.com/news/technology/vodafone-ast-spacemobile-world-first-space-based-5g-call-conventional-smartphone

14.35 **Discussion 12** (Rindö)
*iPaaS, the edge, the CDN and the multi-cloud*
Mazin Yousif, University of Michigan, USA

The landscape of iPaaS (Integrated Platform as a Service) is undergoing continual evolution, marked by escalating intricacies stemming from the interplay of are steadily increasing due to the increased complexities around Edge-CDN-Cloud architectures, the increased dependency on multi-cloud, and the complexities around enterprises business and operational models. More importantly, the need to integrate anything, anywhere at any time. Additionally, an iPaaS is expected to provide support for Application integration, data integration, API management, B2B integration, and Event integration. Furthermore, the need to integrate technology trends such as AI (with its flavors) and the need to support data sources with varying specifications and geographical locations exacerbates complexities. Our discussion will venture into all of the above.

14.35 **Discussion 13** (Storskär)
*How to scale academic cloud research*
Karl-Erik Årzén, Lund University, Sweden & Khuzaima Daudjee, University of Waterloo, Canada

A common comment when presenting cloud research performed in academia to industry is that the proposed solution does not scale. This may very well be the case but it is difficult for the researcher to handle. How should one do academic research so that it scales? Can simulator help or do one need deep access to a real data center?

14.35 **Discussion 14** (Huvudskär)
*Navigating design Space of modular-composed ML systems for multi-objective performance optimization under uncertainty*
Pooyan Jamshidi, University of South Carolina, USA

We are now seeing the tip of a paradigm shift for building AI systems. Traditionally, AI systems are built by training a large model that interacts with users to perform requested tasks, such as translation or coding. Modern machine learning systems, however, tackle AI/ML tasks using multiple interacting and interdependent components, including multiple calls to models, search & retrieval algorithms, and external tools. This shift to modular-composed systems opens many interesting systems questions:

- Exploring large design space of composed ML systems with multiple interacting and interdependent components, each with different knobs that influence performance. Deciding what knobs (e.g., compute resource, model variant, batch size) to use for adaptations and reconfigurations determine the size of the search space for performance optimization. However, such search spaces are overconstrained by multiple competing performance objectives and conflicting user preferences (accuracy, latency, and cost).
- ML systems operate in an uncertain environment and have an imperfect and incomplete understanding of their environment, which is a fundamental limitation of any optimization. For example, the specific reconfiguration must be decided with incomplete and imperfect knowledge due to dynamic workload, variable network latency, and variable resource availability.
- The reconfiguration decisions are made in real-time. Therefore, the multi-objective optimization problem should be solved fast. However, the search space for such optimization problems is exponentially large due to the multi-component aspect of ML systems.

14.35 **Discussion 15** (Svavelsö)

*Simulation of edge-cloud continuum systems*

P-O Östberg, Umeå University, Sweden

Simulation of edge-cloud continuum systems, i.e. distributed systems operating across both cloud and edge resources, is inherently challenging due to the scale and heterogeneity of the resource landscape. Continuum systems are often used to deploy or integrate with a broad range of computation- (e.g., training of machine learning models) and communication-intensive (e.g., IoT and NFV) distributed applications, and are often managed using cloud technologies such as virtualization and deployment management tools such as docker and kubernetes. Simulations can be used to help design, evaluate, and facilitate experimentation with distributed resource management systems, e.g. autonomic control systems. In this session we will discuss goals, challenges, technologies, and perspectives with edge-cloud continuum simulation.

# Thursday June 20th

8.20 *Towards Active Inference for Distributed Intelligence in the Computing Continuum*
Schahram Dustdar, TU Wien, Austria

Modern distributed systems also deal with uncertain scenarios, where environments, infrastructures, and applications are widely diverse. In the scope of IoT-Edge-Fog-Cloud computing, leveraging these neuroscience-inspired principles and mechanisms could aid in building more flexible solutions able to generalize over different environments. A captivating set of hypotheses from the field of neuroscience suggests that human and animal brain mechanisms result from a few powerful principles. If proved to be accurate, these assumptions could open a deep understanding of the way humans and animals manage to cope with the unpredictability of events and imagination.

9:00 *Edge Challenges and Opportunities: Data, Latency, Resilience*
Richard Mortier, Cambridge University, UK

Rather like the End in End-to-End, the Edge in Edge Computing is not very well-defined. It is typically characterised in relation to the datacenters that provide cloud computing, with the Edge having some or all of lower bandwidth, higher latency, and more unreliable networks as well as lower compute, memory and storage resources. These factors create an environment which incentivises greater distribution of applications to match these resource constraints. They also create opportunities: greater distribution gives opportunities to better match application deployments to data characteristics, to reduce latency increasing responsiveness, and to improve system resilience by distributing components across multiple independent and geographically distributed resources. In this talk I will try to articulate some of these challenges and opportunities, illustrating with recent work done in my group.

9.20 *Tell Me What You Want, Not How to Do It!*
Thijs Metsch, Intel, Germany

In this presentation, we will explore the concept of 'intents' and how to apply them for application and resource owners. As application owners tend to aim for maximize performance per dollar, resource owners might seek optimal efficiency. In this presentation we'll look into how to enable intent-driven systems and balance the intents of the stakeholders at play.

We will look into how intents can be applied in edge and cloud deployments, and showcase - maybe with a quick demo - on how they can help with optimizing for performance among other variables (such as sustainability goals).

By the end of this presentation, we hope to underscore the need for advancements in resource-assured orchestration, paving the way for more efficiency in the data center, cloud, or edge.

10.10 *Exploring Cloud Control for Urgent Computing*
Manish Parashar, University of Utah, USA

Urgent science describes time-critical, data-driven scientific workflows that can leverage distributed data sources in a timely way to facilitate important decision making. Despite the exponential growth of available digital data sources and the ubiquity of non-trivial computational power for processing this data, realizing such urgent science workflows remains challenging. This talk will explore application usecases, their requirements and challenges of urgent computing, and how the cloud control approaches such as autonomic computing can harness resources across the computing continuum (i.e., at the edges, in the core and in-between) to support urgent science.

10:30 *Fog Computing Challenges in Natural Environment Observatories*
Guillaume Pierre, Université de Rennes, France

Natural environment observatories allow a wide range of scientists such as biologists, botanists and hydrologists, to observe a zone of particular interest from the points of view of their different scientific disciplines. They follow a data-driven approach based on a variety of sensors and/or actuators deployed in the natural environment, coupled with different techniques to report the produced data to a public or private cloud for further analysis. The constraints which stem from the specificities of such observatories however deviate from traditional IoT use-cases deployed in urban environments. Observatories are often created in remote locations, where energy supply, cellular networking coverage and human maintenance are more challenging than usual. Based on a survey of current practice in 34 observatories in France and abroad, I will make a case for the introduction of fog technologies in these systems, and outline a research agenda to adapt existing fog computing platforms to their specific requirements.

10:50 *Towards Trustworthy AI in untrusted (Cloud) Environments*
Sonia Ben Mokhtar, CMRS, France

There is a strong momentum towards data-driven services at all layers of society and industry. This started from large scale web-based applications such as Web search engines (e.g., Google, Bing), social networks (e.g., Facebook, TikTok, Twitter, Instagram) and recommender systems (e.g., Amazon, Netflix) and is becoming increasingly pervasive thanks to the adoption of handheld devices and the advent of the Internet of Things. New systems that aim at empowering users with the ability to gain back control over their personal data, and prevent a few economic actors from over concentrating decision power are becoming increasingly needed. In this presentation I will discuss novel opportunities for devising online services (and trustworthy AI algorithms that reside at the heart of them) in untrusted cloud environments through the use of distributed/decentralized AI and trusted execution environments (TEEs).

11:10 **Discussion 16** (Resarö)
*Experimental driven research*
Christian Perez, INRIA, France; Manish Parashar, University of Utah, USA & Laurent Lefevre, INRIA, France

Designing, developing and evaluating new software services required to have access to adequate hardware to be as close as possible to real live conditions with monitoring services. This is the goal of experimental platforms (or research infrastructures) such as ChameleonCloud in USA (https://www.chameleoncloud.org) or SLICES in Europe (https://www.slices-ri.eu). Among the numerous advantages, such research infrastructures enable to increase reproducibility as various researchers/engineers can redo experiments and/or vary parameters on the same infrastructure.

The topics of discussion of this session cover experimental driven research. A first topic is about the identification of the type of experiments that are still hard/challenging to achieve. There is a strong link with the evolution of emerging technologies/new usage such as IA, computing continuum, etc. A second topic is to understand the sources of difficulty in designing/conducting experiments such as a lack in hardware or software stack, including monitoring. A third topic could be to better understand the needed/desired reproducibility that should be achieved by such platforms.

Lastly, the session will explore methods used in the private sector as well to validate their new product at large-scale.

The expected outputs of the discussion are a better understanding of the challenges for experimental driven research, including but not limited to the platforms and the software to conduct experiments.

11:10 **Discussion 17** (Rindö)
*Serverless cloud programming models*
Ana Klimovic, ETH Zürich, Switzerland

The cloud has become the dominant platform for running all kinds of applications, from data analytics to web services. In the process, cloud platforms have evolved from renting virtual machines (VMs) on-demand to offering elastic compute and storage services. While the ability to support legacy applications was critical in the early days of cloud to ease migration from on-premise, today's users commonly develop cloud-native applications by composing cloud storage services (e.g., S3), compute services (e.g., AWS Lambda), data analytics services (e.g., BigQuery), machine learning services (Azure ML), and elastic databases (e.g., Snowflake). With this approach, users no longer need to explicitly provision CPU/memory/storage for their applications, as the elastic services automatically scale-out based on load and bill users for the resources consumed.

Opportunity and obstacle: By abstracting resource management from users, elastic cloud services have the potential to optimize resource allocation, task scheduling, and data movement under the hood to improve overall performance and energy-efficiency. However, a major optimization obstacle is that today's cloud programming model captures very little about the resource requirements and data access patterns of individual applications, leaving cloud services with little information to apply optimizations. Despite new cloud-native models like Functions as a Service (FaaS), today's cloud is still built around the principle of executing opaque user applications inside VMs. For example, FaaS platforms execute a user function as an opaque unit in a MicroVM. Each serverless function arbitrarily combines custom computation logic and calls to external cloud services for data passing. The platform is not aware of inter-function nor inter-service dependencies, making it difficult to optimize task scheduling and data prefetching.

Rethinking the programming model: A promising way to enable cloud platforms to improve performance and resource efficiency is to rethink the cloud-native programming model, such that users develop applications in ways that provide the cloud platform with key information to guide task scheduling and data prefetching optimizations. The goal of this discussion session is to explore different programming models that balance the goals of being easy to use yet expressive enough for users while also giving the platform enough information to optimize performance and resource efficiency under the hood.

11:10 **Discussion 18** (Storskär)
*Exploring the poten-tial of cloud robotics: Resource allocation, challenges, and platforms*
Chanh Nguyen and Antonio Seo, Umeå University, Sweden

In recent years, the utilization of mobile robots has experienced a significant upsurge, representing a critical milestone in their widespread adoption. However, many of these robots still rely on simplistic control mechanisms or are controlled remotely by humans. This limited intelligence stems largely from the high costs associated with onboard computation and storage, impacting both the affordability of the robots and their mobility and operational capacity due to increased size and weight requirements.

The evolving landscape of cloud computing and edge computing presents promising opportunities for the advancement of robotics. With advancements in wireless technology and the increasing availability of data centers, ranging from local edge nodes to remote data centers, integrating robots with edge-cloud computing resources holds significant potential. The concept of cloud robotics, wherein robots utilize remote computational resources, dates back to the 1990s. Since then, considerable efforts have been made to develop cloud computing frameworks tailored for robotics. Robotics applications typically involve multiple components, each with varying degrees of freedom, operating both independently and collaboratively towards shared objectives. Effective resource allocation, encompassing computing power, storage, and network resources, is crucial for optimizing performance across these diverse components.
In this discussion, we will explore several thought-provoking questions, beginning with:

1. Which components of robotics applications stand to derive the greatest benefits from employing edge nodes and cloud remote resources?
2. What are the principal challenges in resource management, considering factors such as response time, throughput, and reliability, when distributing resources for these diverse applications?
3. What advanced platforms can provide robust support for specific cloud robotics applications?

11:10 **Discussion 19** (Huvudskär)
*Securing the foundation of the cloud through recent advances in hardware isolation mechanisms, formal verification, safe programming languages, static analys, and fuzzing*
Anton Burtsev, University of Utah, USA

This discussion section aims to bring together researchers and developers from the field of operating systems, programming languages, security, computer architecture and verification with the goal to accelerate changes at the very core of the systems stack, i.e., operating system kernel and hypervisors, through a combination of novel hardware security mechanisms, programming language techniques like practical low-overhead safety, static analysis, fuzzing, and formal verification.

Specifically, our goal will be to:

- Discuss recent advances in the areas of formal verification, static analysis and fuzzing that enable development of more secure operating system kernels and hypervisors.
- Identify the challenges and opportunities in using hardware support for fine-grained isolation
- Discuss the opportunities of applying safe programming languages like Rust, and automated verification tools like Verus and Dafny for securing operating system kernels and hypervisors
- Understand advantages and limitations of modern hardware isolation mechanisms, such as ARM memory tagging extensions (MTE), ARM pointer authentication (PAC), Intel memory protection keys (MPK), Morello CHERI, etc.
- Discuss opportunities and limitations of combining fuzzing and static analysis for efficient exploration of kernel and hypervisor vulnerabilities
- Discuss the practical aspects of isolating kernel and hypervisor subsystems and developing parts of the kernel in a safe programming language

11:10 **Discussion 20** (Svavelsö)
*Open opportunity*

13.45 *Lifting the Fog of Uncertainties: Dynamic Resource Orchestration for the Containerized Cloud*
Hans-Arno Jacobsen, University of Toronto, Canada

Advances in virtualization technologies have sparked a growing transition from virtual machine (VM)-based to container-based infrastructure for cloud computing. From the resource orchestration perspective, the lightweight and highly configurable nature of containers not only enables opportunities for more optimized resource management strategies, but also poses greater challenges due to additional uncertainties and a larger configuration parameter search space.

In this talk, we introduce Drone, a container orchestration framework that adaptively configures resource parameters to improve application performance and reduce operational cost in the presence of cloud uncertainties. Built on Contextual Bandit techniques, Drone is able to achieve a balance between performance and resource cost on public clouds, and optimize performance on private clouds where a hard resource constraint is present. We show that our algorithms

achieve sub-linear growth in cumulative regret, a theoretically sound convergence guarantee, and our extensive experiments show that Drone achieves an up to 45% performance improvement and a 20% resource footprint reduction across batch processing jobs and microservice workloads.

This talk draws from a paper that is joint work with Yuqiy Zhang; SoCC'23 - https://dl.acm.org/doi/10.1145/3620678.3624646

14:05 *Revolutionizing Datacenter Networks via Reconfigurable Topologies*
Stefan Schmid, TU Berlin, Germany

With the growing popularity of cloud computing and data-intensive applications such as machine learning, datacenter networks have become a critical infrastructure for our digital society. Given the explosive growth of datacenter traffic and the slowdown of Moore's law, significant efforts have been made to improve datacenter network performance over the last decade. A particularly innovative solution is reconfigurable datacenter networks (RDCNs): datacenter networks whose topologies dynamically change over time, in either a demand-oblivious or a demand-aware manner. Such dynamic topologies are enabled by recent optical switching technologies and stand in stark contrast to state-of-the-art datacenter network topologies, which are fixed and oblivious to the actual traffic demand. In particular, reconfigurable demand-aware and "self-adjusting" datacenter networks are motivated empirically by the significant spatial and temporal structures observed in datacenter communication traffic. This talk presents an overview of reconfigurable datacenter networks. In particular, we discuss the motivation for such reconfigurable architectures, review the technological enablers, and present a taxonomy that classifies the design space into two dimensions: static vs. dynamic and demand-oblivious vs. demand-aware.

14:25 *Reliable Fast Process Failure Detection in Data Centers*
Patrick Eugster, Lugano, Switzerland

In distributed systems, failure detection is one of the most fundamental primitives on which fault tolerant services and applications rely to achieve liveness. Typical failure detectors resort to using timeouts that have to take into account the unpredictability in interaction times among remote processes, caused by resource contention in the network and in endhost processors. As a consequence, failure detectors use prohibitively large timeouts which fail to meet the microsecond scale required by modern data center services and/or are designed as unreliable/eventually reliable where a smaller chance of a false failure suspicion comes at the cost of even more conservative timeouts, and hence decreased performance. We propose a novel, fast, fully reliable failure detector (FiDe) that can report a failure of a remote process in a data center within less than 5 μs on average and less than 50 μs at maximum, respectively 3.7× and 7.2× faster than the current state of the art. FiDe is built ground-up using kernel bypass, process isolation, traffic engineering, and network redundancy to achieve timely stable end-to-end process interactions providing strong reliability guarantees. We further showcase the practical relevance of FiDe by using it as a foundation for two novel highly efficient consensus algorithms which we integrate into a key-value store for safe replication. Our algorithms achieve up to 1.7× throughput speedup against Raft-based replication at scale.

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