Program

14th Cloud Control Workshop

Vindeln, Sweden
April 2nd – 4th, 2019

http://cloudresearch.org/workshops

- The schedule is deliberately done with a single track to leave room for additional planned or spontaneous sessions to be organized in parallel, i.e., for work in on-going projects or to initiate new collaborations.
- Some discussions topics will be announced depending on the the group topics selected by the participants
- Check http://cloudresearch.org/workshops/14th for updates

	Tuesday April 2nd
10.30	Arrival at Vindeln
10.30	**Coffee**
11.00	Workshop Opening *Erik Elmroth, Umeå University*
11.15	Keynote – Cloud-enabled News Distillation for the Age of Overload *Tarek Abdelzaher, University of Illinois at Urbana-Champaign*
12.05	**Lunch**
	Analytics and Orchestration Session *Chair: Karl-Erik Årzén*
13.00	Controlling Data Centers with Reinforcement Learning Albin Heimerson, *Lund University*
13.20	Root Cause Analysis in Container-based Micro-Service Environment *Li Wu, Technical University of Berlin & Elastisys AB*
13.40	Massively federated container orchestration *Lars Larsson, Umeå University & Elastisys AB*
14.00	**Discussion 1:** Problems in Cloud well-suited for control theory *Leaders: Victor Millnert and Tommi Nylander, Lund University*
15.00	**Coffee**
	Applications and Optimization Session *Chair: Maria Kihl*
15.30	Autonomous camera systems *Alexandre Martins, Lund University & Axis Communications AB*
15.50	Power-Aware Cloud Brownout: response time and power consumption control *Jakub Krzywda, Umeå University*
16.10	One Way Delay Measurements in Software Defined Radio Access Network *Haouri Peng, Lund University*
16.50	**Group Work:** Introduction *Victor Millnert, Lund University*
17.00	Team building activities
19.00	**Dinner**

Wednesday April 3rd

Edgy Session

Chair: Cristian Klein

8.30	Management Beyond the Edge: Gap and Research Direction *Danlami Gabi, Umeå University*
9.00	**Discussion 2:** What are the Killer Apps for the Mobile Edge Clouds? *Leader: Cristian Klein, Elastisys AB & Umeå University*
10.00	**Coffee**

Senior Session

Chair: Danlami Gabi

10.30	What can we learn from Network Calculus? *Victor Millnert, Lund University*
10.50	JitterTime – A tool for analyzing transient behavior in networked estimation and control applications *Anton Cervin, Lund University*
11.30	Modelling and simulation of virtualized Content Delivery Networks *P-O Östberg, Umeå University*
12.00	**Lunch**

DC Management Session

Chair: Tarek Abdelzaher

13.00	On long-time resource prediction for task scheduling *Johan Ruuskanen, Lund University*
13.20	Modeling of Request Cloning in Cloud Data Centers *Tommi Nylander, Lund University*
13.40	Hybrid Resource Management for HPC and Data Intensive Workloads *Abel Souza, Umeå University*
14.00	**Discussion 3:** API-driven Infrastructure: Kubernetes, Serverless, Disaggregation, Performance, and everything else *Leader: Abel Souza, Umeå University*
15.00	**Coffee**
15.30	**Group Work – Continuation** *Tommi Nylander, Lund University*
16.30	Team building activities
19.00	**Dinner**

Thursday April 4th	
Autoscale, Model, and Control Session *Session Chair: P-O Östberg*	
8.30	Cloud-Assisted Model-Predictive Control *Per Skarin, Lund University & Ericsson Research*
8.50	Autoscaling in Container Clusters Mulugeta Ayalew Tamiru*, University of Rennes 1 & Elastisys AB*
9.10	Are SRE's Afraid of Adversarial Attack Against Anomaly Detection in Cloud Datacenters? *Monowar Bhuyan, Umeå University*
9.40	Toward Elasticity Control for Edge Data Centers *Chanh Nguyen, Umeå University*
10.00	**Coffee / Checkout**
Session Chair: Monowar Bhuyan	
10.30	**Group Work:** Summary & Presentations *Victor Millnert, Lund University*
12.00	**Lunch**
13.00	**Discussion 4**: Topic TBD *Tommi Nylander, Lund University*
13.45	Closing Remarks *Erik Elmroth, Umeå University*
14.00	**Departure**

Titles and Abstracts

Tuesday, April 2nd

Time	Type	Information	
11.15	Keynote	**Title**	Cloud-enabled News Distillation for the Age of Overload
		Presenter	Tarek Abdelzaher - University of Illinois - zaher@illinois.edu
		Abstract	*In the age of information overload, a growing bottleneck becomes human attention. In order to allocate this limited resource more judiciously, we increasingly favor information that carries an element of surprise, as it is least "redundant" with our current state of knowledge. From an information-theoretic standpoint, surprise may be measured by the change in the probability distribution of belief that results from receiving the information. This observation leads to the question of whether one can automate news generation by developing a cloud service that automatically detects novel content according to language-agnostic information-theoretic metrics?* *This talk describes such a service, presents its information-theoretic foundations, comments on its implementation overhead on a Hadoop cluster, and gives examples of results of its application for the purpose of identifying and summarizing newsworthy Twitter content. Future challenges and directions are also discussed.*
13.00	Student Presentation	**Title**	Controlling Data Centers with Reinforcement Learning
		Presenter	Albin Heimerson, *Lund University*
		Abstract	*What problems would be interesting to apply RL to in the DC setting and what problems needs to be solved to be able to train efficiently and safely.*
13.20	Student Presentation	**Title**	Root Cause Analysis in Container-based Micro-Service Environment
		Presenter	Li Wu - TU Berlin & Elastisys AB - li.wu@elastisys.com
		Abstract	*Cloud services have recently shifted from monolithic architecture to microservice architecture rapidly as microservice could simplify and accelerate development and deployment as well as unlimit the development techniques. Meanwhile, the complexity of dependency and frequent updates pose critical challenges to fault analysis. This paper presents a novel root cause analysis system to not only identify and locate the abnormal microservice with a ranked list but only classify the real cause into four catalogs which will be mapped to different remediations for recovery.*

13.40	Student Presentation	Title	Massively federated container orchestration
		Presenter	Lars Larsson – Umeå University – larsson@cs.umu.se
		Abstract	*With thousands of geographically dispersed clusters on a global or continental scale, how do you manage them and the applications deployed onto them? Help me discover the interesting questions and identify prior work and known pitfalls and limitations!*
15.30	Student Presentation	Title	Autonomours Cameras Systems
		Presenter	Alexandre Martins, Lund University & Axis Communications AB - alexandre.martins@control.lth.se
		Abstract	Presentation of today's challenges with big camera systems and ideas to address it using local control and game theory.
15.50	Student Presentation	Title	Power-Aware Cloud Brownout: response time and power consumption control
		Presenter	Jakub Krzywda, Umeå University – jakub@cs.umu.se
		Abstract	*Cloud computing infrastructures are poweringmost of the web hosting services that we use at all times. Arecent failure in the Amazon cloud infrastructure (Feb 2017) made manyof the website that we use on a hourly basis unavailable. This illustrates the importance of cloud applications being able toabsorb peaks in workload, and at the same time to tune theirpower requirements to the power and energy capacity offeredby the data center infrastructure. In this paper we combinean established technique for response time control – brownout– with power capping. We use cascaded control to take intoaccount both the need for predictability in the response times(the inner loop), and the power cap (the outer loop). We executetests on real machines to determine power usage and responsetimes models and extend an existing simulator. We then evaluatethe cascaded controller approach with a variety of workloadsand both open- and closed-loop client models.*
16.10	Student Presentation	Title	One Way Delay Measurements in Software Defined Radio Access Network
		Presenter	Haouri Peng, Lund University - haorui.peng@eit.lth.se
		Abstract	*An approach to measure the One Way Delay (OWD) of a software Radio Access Network (RAN) over a prototyping Massive Multiple Input Multiple Output (MIMO) antenna system and FPGA-signal processing platform. The purpose is to evaluate the performance of such an antenna system and target the applicable low latency services over this network.*

Wednesday, April 3rd

Time	Type	Information	
8.30	Senior Presentation	**Title**	Management Beyond the Edge: Gap and Research Direction
		Presenter	Danlami Gabi, Umeå University - gabid@cs.umu.se
		Abstract	*Mobile Edge Clouds (MECs) is limited in its capacity to deal with resource contention. We reviewed existing research on MECs and discovered that the coexistence of multiple service providers can serve as a potential solution to addressing resource contention. However, a concern is resource allocation technique that will provide benefit to each service providers in terms of profits. This research is focus on providing an ideal solution.*
10.30	Senior Presentation	**Title**	What can we learn from Network Calculus?
		Presenter	Victor Millnert, Lund University
		Abstract	*I will give a brief introduction to the theory of network calculus, and how it could be used in new ways; for instance in how one should scale virtual machines, do admission control, or assign priorities to flows.*
10.50	Senior Presentation	**Title**	JitterTime – A tool for analyzing transient behavior in networked estimation and control applications
		Presenter	Anton Cervin, Lund University - anton@control.lth.se
		Abstract	*JitterTime is a Matlab toolbox for calculating the time-varying state covariance of a mixed continuous/discrete linear system driven by white noise. It also integrates a quadratic cost function for the system. The passing of time and the updating of the discrete-time systems are explicitly managed by the user in a simulation run. Since the timing is completely handled by the user, any complex timing scenario can be analyzed, including scheduling algorithms, timing jitter and drift, and asynchronous execution in distributed systems.*
11.30	Senior Presentation	Title	Modelling and simulation of virtualized Content Delivery Networks
		Presenter	P-O Östberg, Umeå University – p-o@cs.umu.se
		Abstract	*The presentation highlights some recent results in modelling for simulation and control of virtualized CDNs developed in the H2020 project RECAP, and will feature a demonstration of the modelling capabilities of a simulation framework, some simple autoscalers applied to a hierarchical control problem (placement of cache capacity in CDNs), and give an overview of some of the datasets we are working with (population data, geospatial data, infrastructure models, and real data from our partners CDNs).*
13.00	Student Presentation	Title	On long-time resource prediction for task scheduling

		Presenter	Johan Ruuskanen, Lund University - johan.ruuskanen@control.lth.se
		Abstract	*Re-scheduling tasks in a cloud environment due to their dynamic workloads is not without cost, and by taking future resource usage into account the number of expensive control actions can be reduced. In this presentation I will discuss some preliminary work on the potential of using Gaussian Processes in predicting workload over a long time horizon.*
13.20	Student Presentation	Title	Modeling of Request Cloning in Cloud Data Centers
		Presenter	Tommi Nylander, Lund University - tommi.nylander@control.lth.se
		Abstract	*We present a model that allows us to equivalently represent a system of servers with cloned requests, as a single server. The model is very general, and we show that no assumptions on either inter-arrival or service time distributions are required, allowing for, e.g., both heterogeneity and dependencies. Additionally, we show that the model holds for any queuing discipline as well. The key requirement that enables us to use the single server G/G/1 model is that the request clones have to receive synchronized service. We show examples of server systems that fulfill this requirement, and use our G/G/1 model to co-design traditional load-balancing algorithms together with cloning strategies, providing well performing, provably stable designs. Finally, we use our model to evaluate request cloning in heterogeneous server systems.*
13.40	Student Presentation	Title	Hybrid Resource Management for HPC and Data Intensive Workloads
		Presenter	Abel Souza – abel@cs.umu.se
		Abstract	*High Performance Computing (HPC) and Data Intensive (DI) workloads have been executed on separate clusters using different tools for resource and application management. With increasing convergence, where modern applications are composed of both types of jobs in complex workflows, this separation becomes a growing overhead and the need for a common platform increases. Executing both workload classes on the same clusters not only enables hybrid workflows, but can also increase system efficiency, as available hardware often is not fully utilized by applications.* *In here, we present the architecture of a hybrid system enabling dual-level scheduling for DI jobs in HPC infrastructures. Our system takes advantage of real-time resource profiling to efficiently co-schedule HPC and DI applications. The architecture is easily extensible to current and new types of distributed applications, allowing efficient combination of hybrid workloads on HPC resources with increased job throughput and higher overall resource utilization. The implementation is based on the Slurm and Mesos resource*

			managers for HPC and DI jobs. Experimental evaluations in a real cluster based on a set of representative HPC and DI applications demonstrate that our hybrid architecture improves resource utilization by 20%, with 12% decrease on queue makespan while still meeting all deadlines for HPC jobs.

Thursday, April 4th

Time	Type	Information	
8.30	Student Presentation	**Title**	Cloud-Assisted Model-Predictive Control
		Presenter	Per Skarin, Lund University & Ericsson Research
		Abstract	*In the presentation we develop a computational offloading strategy with graceful degradation for executing Model PredictiveControl using the cloud. Using available works as an empirical basis we simulate the control of a cyber-physical-system at high frequency and illustrate how the system can be improved using the edge while keeping the computational burden low.*
8.50	Student Presentation	**Title**	Autoscaling in Container Clusters
		Presenter	Mulugeta Ayalew Tamiru, University of Rennes 1 & Elastisys AB
		Abstract	*In the last few years, Kubernetes has become the de-facto container orchestration platform for large scale deployment of microservices. It has enjoyed wide acceptance in industry because of the ease of setting it up not only in the public cloud but also on private data centers, hybrid clouds and fog/edge environments. Kubernetes enables true elasticity of applications by offering autoscaling capabilities on two layers, namely the container and virtual machine layers. Horizontal - and vertical pod autoscaling have been proposed for autoscaling of applications at the container level, whereas the Kubernetes cluster autoscaler leverages public cloud infrastructures to offer autoscaling of nodes at the virtual machine layer. However, how these autoscaling mechanisms behave under different workloads is not well studied. In this (work in progress) we evaluate the Kubernetes autoscaling mechanisms to uncover interesting findings by stressing the cluster with workloads modeled based on the Google cluster traces.*
9.10	Senior Presentation	**Title**	Are SRE's Afraid of Adversarial Attack Against Anomaly Detection in Cloud Datacenters?
		Presenter	Monowar Bhuyan, Umeå University – monowar@cs.umu.se
		Abstract	*Due to increasing volume and rapid changing behaviour of metric streams (e.g., latency, CPU, memory) in the cloud datacenters, cloud service providers encounter difficulties to*

			ensure high availability, capacity, latency and performance. In order to establish this, SRE measures everything and often attempt to detect system anomalies in cloud datacenters by employing machine learning algorithms. While injecting a fraction of well-crafted malicious samples in training data, attackers can subvert the learning process and results in unacceptable false positives. These security issues cause threats to all categories of anomaly detection. Hence, it is imperative to assess these techniques against adversaries to improve robustness and scalability. We assess the vulnerability of gray-box anomaly detection techniques against adversaries explicitly designed for linear regression and statistical learning, where validate this framework using synthetic and real-time Yahoo! Webscope S5 datasets..
11.30	Senior Presentation	Title	Toward Elasticity Control for Edge Data Centers
		Presenter	Chanh Nguyen, Umeå University
		Abstract	*Elasticity is one of fundamental properties that MECs must inherently hold in order to become a mature computing platform hosting software applications. However, unlike the current cloud platform, MECs need to cope with more challenges from the densed distribution, the heterogeneity and limitation of resource capacity in Edge Data Centers (EDCs), and the end-user mobility. In this presentation, we will discuss our proposed auto scaling technique to help MECs overcome the aforementioned challenges to automatically scale its resource in a proactive manner. The technique utilizes the location information of EDCs to first estimate request arrival rate at EDCs, then the two level controllers (a local controller, and a neighbor-cluster controller) decide the right amount of EDC's resource to be scaled up/down which ensure persistently the Quality of Service (QoS), while maximizing resource utilization.*

Discussion 1: Problems in Cloud well-suited for control theory
Leader: Victor Millnert and Tommi Nylander, Lund University
Abstract: TBD

Discussion 2: What are the Killer Apps for the Mobile Edge Clouds? Leader: Cristian Klein, Elastisys AB & Umeå University
Abstract: Mobile Edge Clouds (MECs) are clouds in which distant data-centers are complemented with computing and storage capacity located at the edge of the network. Much literature focuses on ``killer apps'' -- we call them MEC-native applications -- that could drive investment into MECs. However, considering that adoption of traditional clouds was fostered by legacy, non-cloud-native applications, we argue that MECs also need to focus on bringing benefits to non-MEC-native applications. Failing to do so risks leading to a dead-lock: Infrastructure investment is slow due to insufficient MEC-native applications, and development of MEC-native applications is postponed until more MECs are available. While working on placement algorithms for MECs, we discover that non-MEC-native applications seem to be slowed down by MECs. In this session, we will discuss

suitable benchmarks for evaluating placement algorithms on MEC. We will discuss both non-MEC-native applications, as well as MEC-native ones.

Discussion 3: API-driven Infrastructure: Kubernetes, Serverless, Disaggregation, Performance, and everything else
Leader: Abel Souza, Umeå University
Abstract: TBD

Discussion 4: Topic TBD
Leaader: Tommi Nylander, Lund University
Abstract: TBD

Grouping

Topic	Description
Blockchain and Cloud?	How could smart-contracts be used within the cloud to simplify some aspects of the cloud business? Free Papers? Montization, Security, Smart contract...
Serverless — open problems?	How does what we have derived for scaling of VMs or Containers change when we move towards a serverless architecture? Example: when renting VMs the price is the same during the deployment. With serverless, you only pay when the code runs—how does this change the strategies of the user? How does hosting Serverless functions change the behavior from the providers perspective?
Abstractions to connect different clouds	How can we design abstractions so that we can reason about what happens when we connect different cloud systems together? Example: provide end-to-end latency guarantees when the route goes through multiple different cloud systems and vendors.
Save the environment	Cloud computing is using significant energy resources, and is becoming a major impact for the environment. How should we tackle this? The illusion of infinite processing capacity in the cloud can lead us down a path of writing less efficient code (because it is easier to spin up another VM than optimize your code). But are there ways to use the cloud to actually save energy when doing the computations?
What else looks like a cloud?	Are there any other fields/areas where the problems resembles the problems we face? Can we learn anything from there field or their techniques and apply them to cloud control?
New Topics for a PhD	What could be a good topic for new PhD students to focus on? Or, in other words: where do you think the “Cloud” will be in 5-7 years from now, and what do you foresee as the biggest challenges to get there (or once we get there)?
Stateful Applications in the Edge Cloud	How can we allow stateful applications in the edge cloud. How should we handle the case when an edge node (with stateful application running within it) goes down?
Artificial Intelligence and Cloud Control	We are seeing a lot of efforts in combining AI and cloud control. One prominent area is the use of reinforcement learning for scaling the allocated resources. What other areas within Cloud can AI be useful for? Is it possible to use AI and still have any performance guarantees?
Open Problem	What do you think is still an open, and unsolved, problem within the field of cloud control? Why do you think this is the case—is it not an important problem, or is it perhaps too difficult to solve?
Abstract the core design principles	What are the core design principles of the work we have done within this community? Is it possible to look at all the methods, techniques, and theory and abstract some core design principles that are useful when designing a cloud system (or a control-method for a cloud system)?