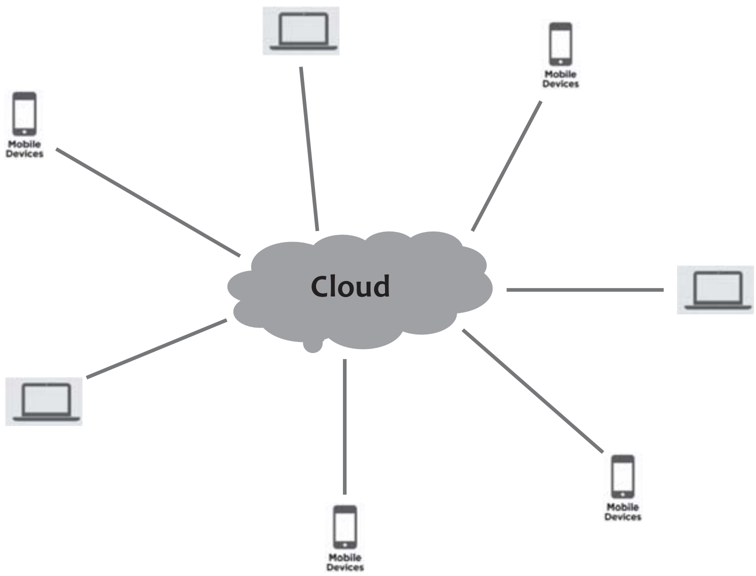
EDGE COMPUTING CONCEPTS, MOTIVATIONS, AND CHALLENGES

Erik Elmroth

Umeå University & Elastisys
erik.elmroth@umu.se & erik.elmroth@elastisys.com
www.cloudresearch.org & www.elastisys.com

UMEÅ UNIVERSITY

Centralized cloud model



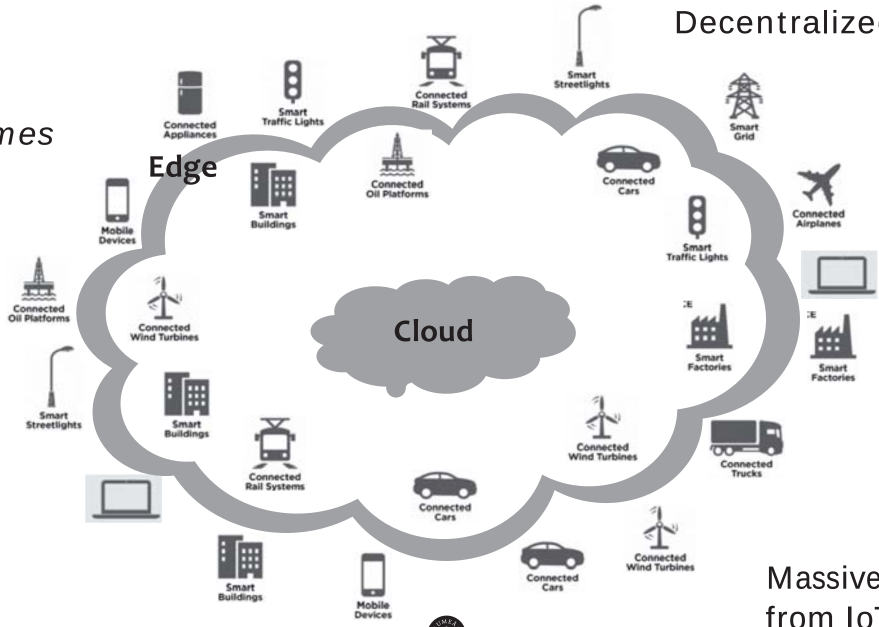
UMEÅ UNIVERSITY

Near future expectation:

- 1 GB/person per day

Decentralized edge model

Data volumes



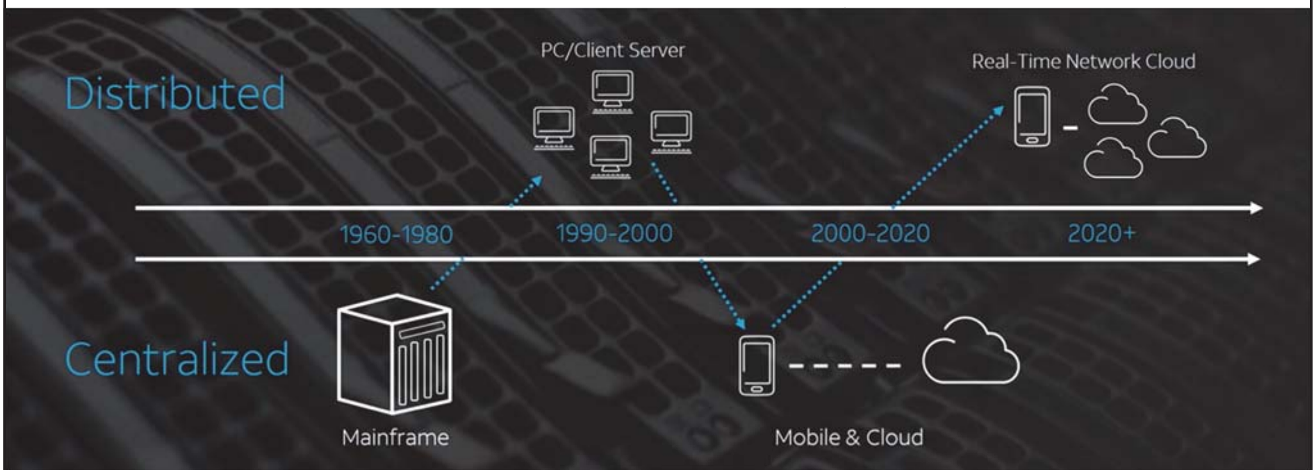
Latency

Privacy

Massive data volumes from IoT, IIoT, self-driving cars, drones, smart-*, etc

UMEÅ UNIVERSITY

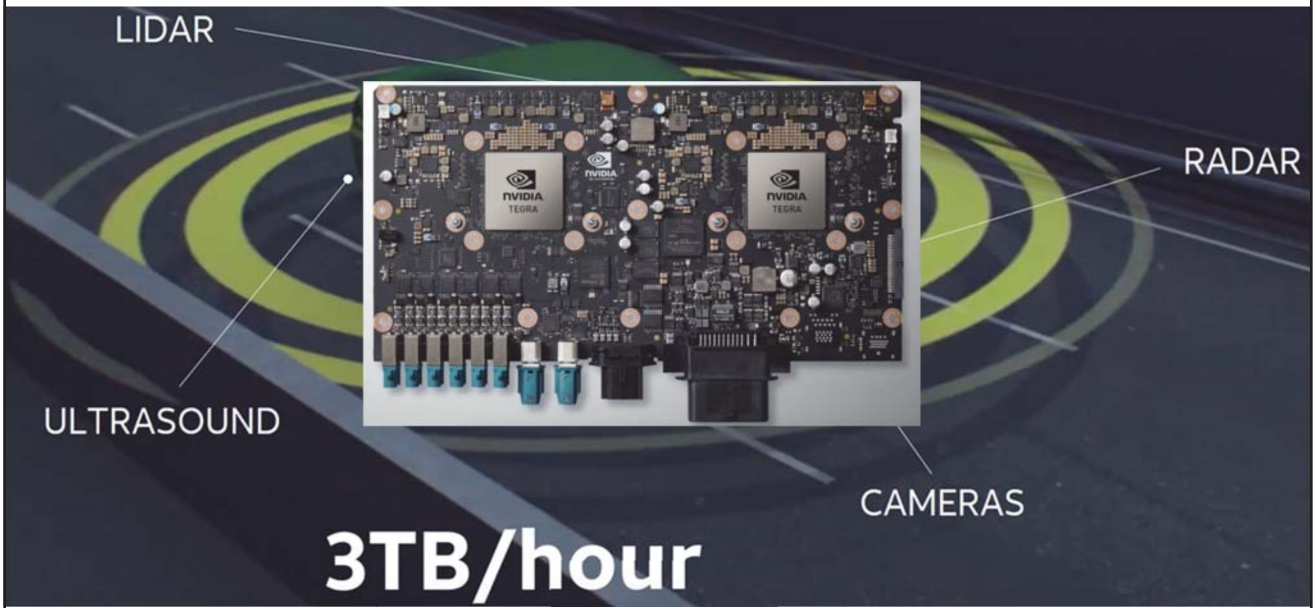
CENTRALIZED OR DISTRIBUTED? WE HAVE SEEN IT ALL BEFORE!



DATA DRIVES THE CHANGE! REAL-TIME, REAL-WORLD!

- Massive data from full-fledged IoT and IIoT
  - Cheap, powerful, and ubiquitous sensors. Video cams a good example
  - Data volumes to increase by several orders of magnitudes
- Autonomous cars, drones, robots are only early examples

DATA CENTER ON WHEELS!



Assume these are self-driving cars,
guided by on-line traffic control

DATA DRIVES THE CHANGE! REAL-TIME, REAL-WORLD!

- Massive data from full-fledged IoT and IIoT
  - Cheap, powerful, and ubiquitous sensors. Video cams a good example
  - Data volumes to increase by several orders of magnitudes
- Autonomous cars, drones, robots are only early examples

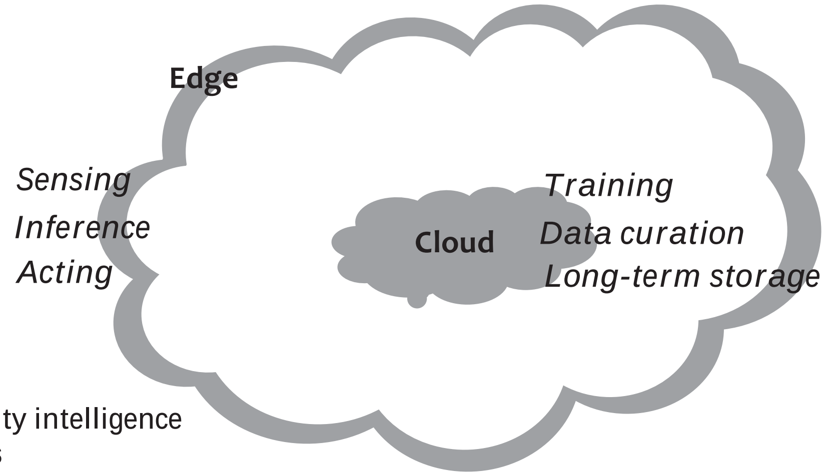
Response-time critical

- VR/AR applications
- Smart factories, control loops
- Smart hospitals, remote surgery
- Autonomous driving

Locality-sensitive data handling

- Data too sensitive for off-premise
- Data location legislations
- Data of pure local interest

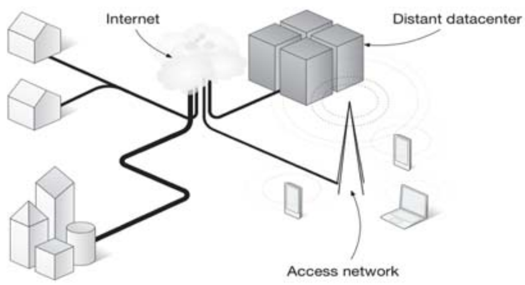
INTELLIGENCE AT THE EDGE



Low-capacity intelligence
Short loops

UMEÅ UNIVERSITY

FROM DATACENTERS TO EDGE COMPUTING



Today

SAMPLE CHALLENGES

- Performance & availability demands
- Cost
  - Investments
  - Maintenance in distributed hard-to-control environments
- Power consumption
- Programming models
- Security / Privacy / Compliance
- Scalability
- Complex multi-stakeholder scenarios
- Overall self-management

UMEÅ UNIVERSITY

RELATED & SUPPORTING TRENDS

- Containerization complementing traditional virtualization
- Orchestration – primarily Kubernetes and cloud-native design
- AI pushed out to the edge
- Edge development goes hand-in-hand with 5G development
- Functional safety - freedom from unacceptable risk of physical injury or of damage to health
- Time-sensitive networking - deterministic networking
- Privacy & compliance

UMEÅ UNIVERSITY

INTELLIGENT EDGE CLOUD AUTOMATION

UMEÅ UNIVERSITY

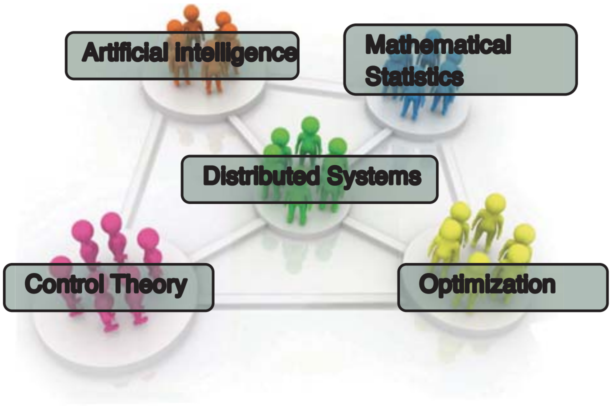
Complexity

Operations

Faults at scale

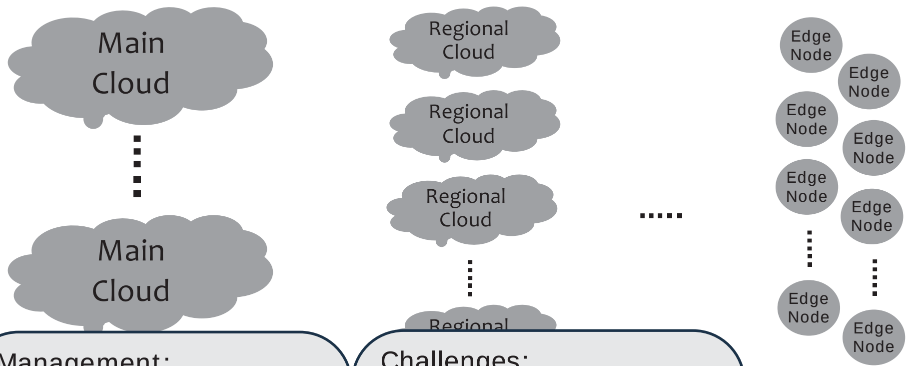
UMEÅ UNIVERSITY

OUR APPROACH

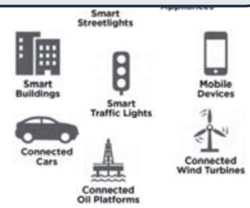


Google, intel, ERICSSON, IBM, SAP, redhat

MULTI-STAKEHOLDER MANAGEMENT SCENARIO



Stakeholders:
- Cloud-providers
- Telecom operators
- Virtual operators
- Application-owners
- IoT-owners

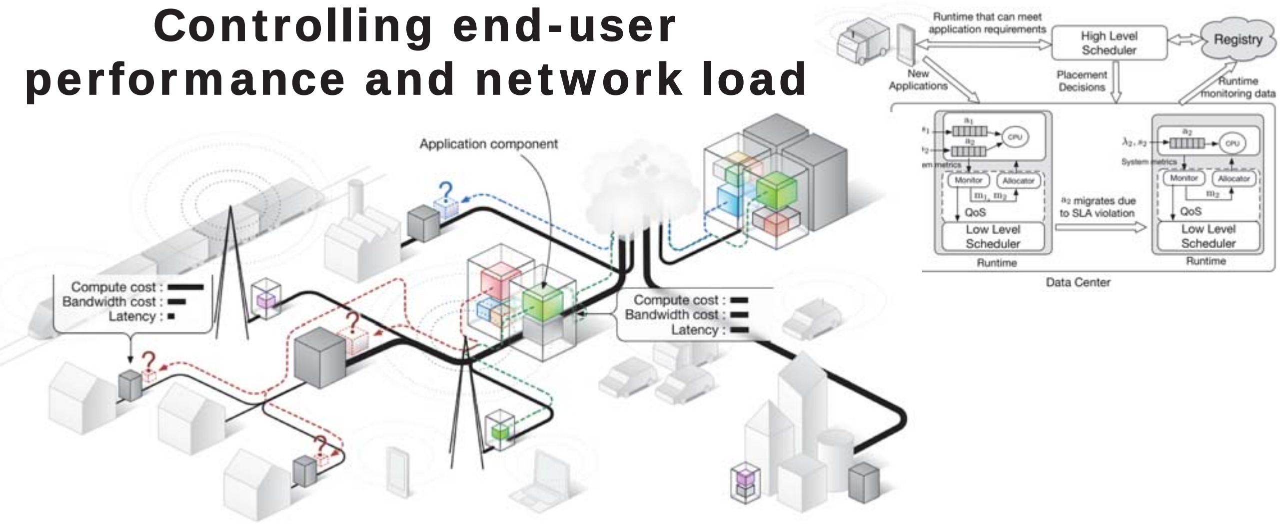


Management:
- App-owner has contracts with multiple cloud & telco providers
- Management by virtual operator or telco/cloud provider
- Management decisions left-to-right

Challenges:
- Scarce edge resources (capacity & power)
- Co-management – both resources & applications
- Different application characteristics
- Different performance demands

Many telcos / private
Specialized, heterogeneous
In neighborhood/on-premise
Power-constrained
Limited multi-tenancy

Controlling end-user performance and network load



A. Mehta, E.B. Lakew, J. Tordsson, E. Elmroth. Utility-based Allocation of Industrial IoT Applications in Mobile Edge Clouds, The 15th IEEE Conference on Autonomic Computing (ICAC 2018), pp. 121-130.

W. Tärneberg, A. Mehta, E. Wadbro, J. Tordsson, J. Eker, M. Kihl, and E. Elmroth. Dynamic Application Placement in the Telco-cloud, *Future Generation Computer Systems*, Elsevier, Vol. 70, pp. 163-177, 2017.

A. Mehta, R. Baddour, H. Gustafsson, F. Svensson, and E. Elmroth. Calvin Constrained - A Framework for IoT Applications in Heterogeneous Environments, *The 37th IEEE International Conference on Distributed Computing (ICDCS 2017)*, pp. 1063-1073, 2017.

WORKLOAD PREDICTION, MOBILITY PREDICTION

- Edge clouds modelling using the real geographical distribution of the network base stations
- Multivariate Long Short Term Memory networks (recurrent neural network) on historical workload information of edge resources

Emulated edge clouds with resource distribution in San Francisco & Rome

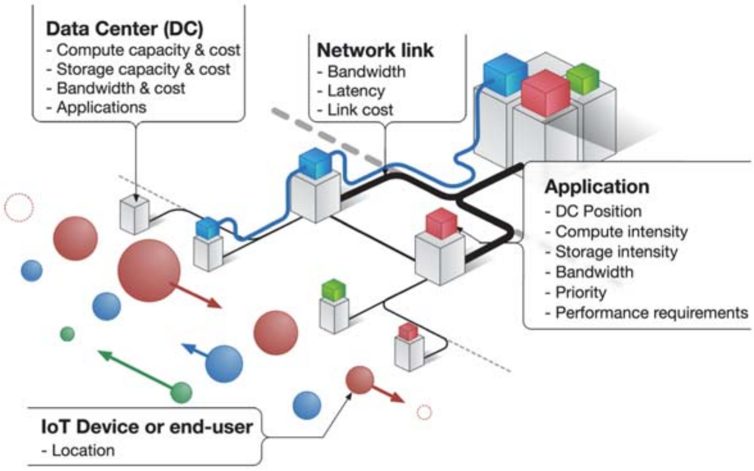
C. Nguyen, C. Klein, and E. Elmroth. Location-aware Load Prediction in Edge Data Centers. In *Proceedings of the Second International Conference on Fog and Mobile Edge Computing (FMEC)*, pp. 25-31, IEEE Computer Society, 2017.

C. Nguyen, C. Klein, and E. Elmroth. Multivariate LSTM-based Location-aware Load Prediction in Edge Data Centers . In *Proceedings of the 19th Annual IEEE/ACM International Symposium in Cluster, Cloud, and Grid Computing (CCGrid 2019)*, pp.341-350, IEEE, 2019.

EDGE CLOUD DESIGN

Example:
Quantify the benefit of additional Data Centers (DCs) closer to the network edge for the optimal application placement

- How close?
- How many layers?
- What relative size?

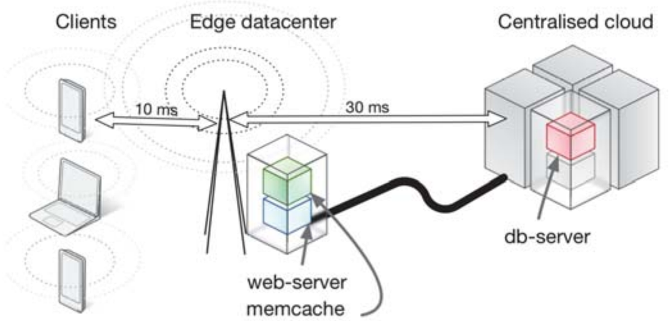


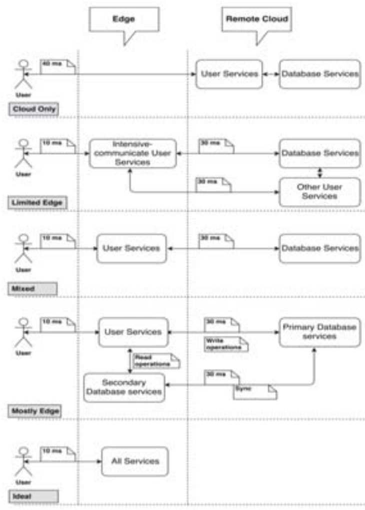
A. Mehta, W. Tärneberg, C Klein, J. Tordsson, M. Kihl, E. Elmroth. How beneficial are intermediate layer Data Centers in Mobile Edge Networks? In *Foundations and Applications of Self* Systems (FAS* 2016)*, 2016.

J. Krzywda, W. Tärneberg, P-O. Östberg, M. Kihl, and E. Elmroth. Telco Clouds: Modelling and Simulation, *Proceedings of the 5th International Conference on Cloud Computing and Services Science (CLOSER 2015)*, SCITEPRESS, pp. 597-609, 2015.

APPLICATION DESIGN/PROGRAMMING MODELS

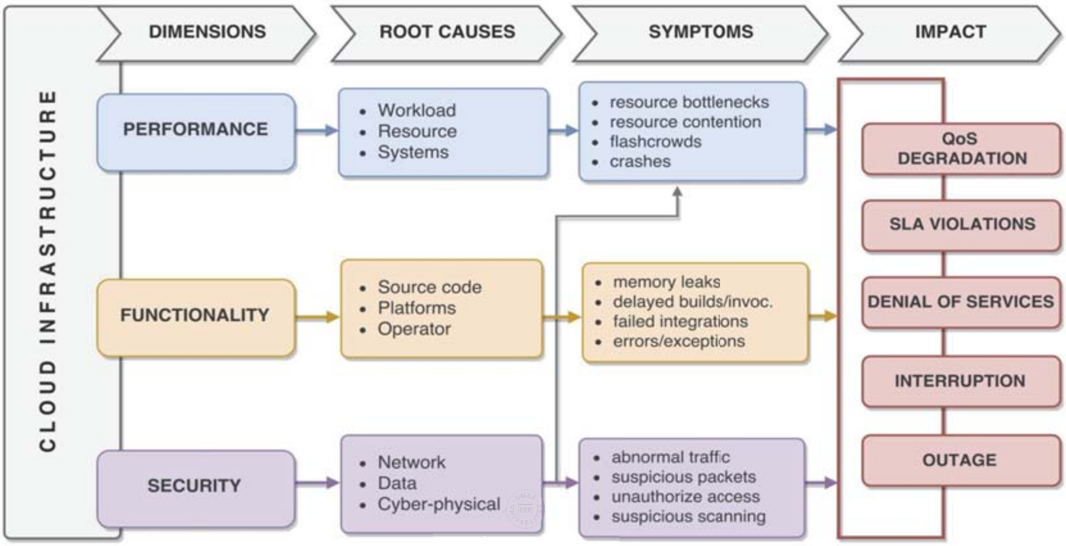
Software design for highly distributed edge clouds



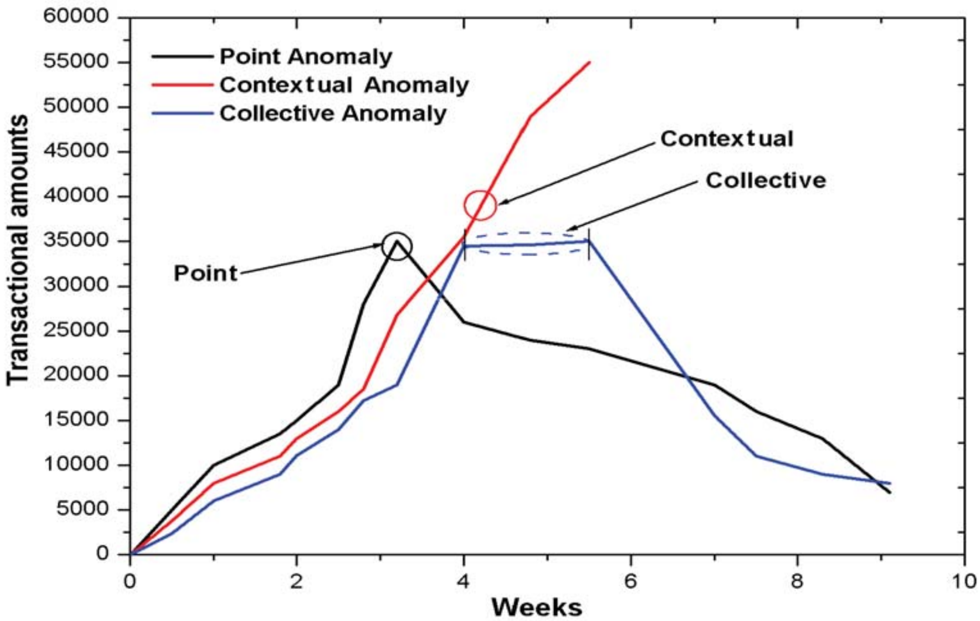


C. Nguyen, A. Mehta, C. Klein, and E. Elmroth. Why Cloud Applications Are not Ready for the Edge (yet). *The Fourth ACM/IEEE Symposium on Edge Computing (SEC 2019)*.

AUTOMATED ANOMALY DETECTION



ANOMALIES: ILLUSTRATION

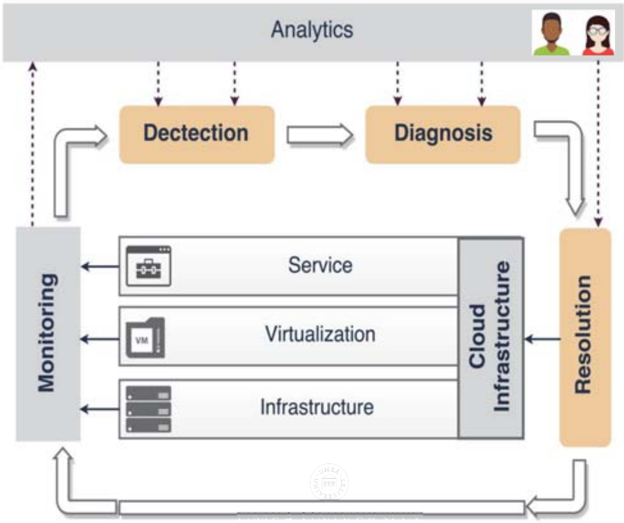


22

ANOMALY DETECTION PIPELINE
(THE MAPE LOOP OF ANOMALY DETECTION)

Anomalies
- *increased latency, SLA violation, excessive stolen CPU cycles, unreachable endpoints*

KPIs
- *response time, throughput, resource usage, errors and workload rates*



Root-causes
- *crashed servers, resource contention, code bugs, DDoS attacks*

Corrective actions
- *passive: rebooting, replacement, reconfiguration*
- *active: micro-booting, patching, throttling, scaling, migration,*

The Autonomous Distributed Systems Lab

Principal Investigators

Erik Elmroth, Professor · Johan Tordsson, Assoc. Prof · Paul Townend, Assoc. Prof · P-O Östberg, Assoc. Prof. · Monowar Bhuyan, Assist. Prof.

Other Senior Researchers

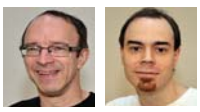
Cristian Klein, PhD · Abel Souza, PhD · Eunil Seo, PhD · Danlami Gabi, PhD · Mansour Khelghatdoust, PhD

Affiliated Researchers

Monica Vitali, PhD (Politecnico Di Milano) · Ahmed Ali-Eldin, PhD (Chalmers Univ of Tech.) · Jakub Krzywda, PhD (Elastisys)

PhD Students

Magdalena Viktorsson · Chanh Nguyen Le Tan · Lars Larsson · Tobias Sundqvist · Javad Forough · Charles Meyers · Mohammad Reza Saleh · Ali Rahmanian · Sourasekhar Banerjee · Lidia Kidane



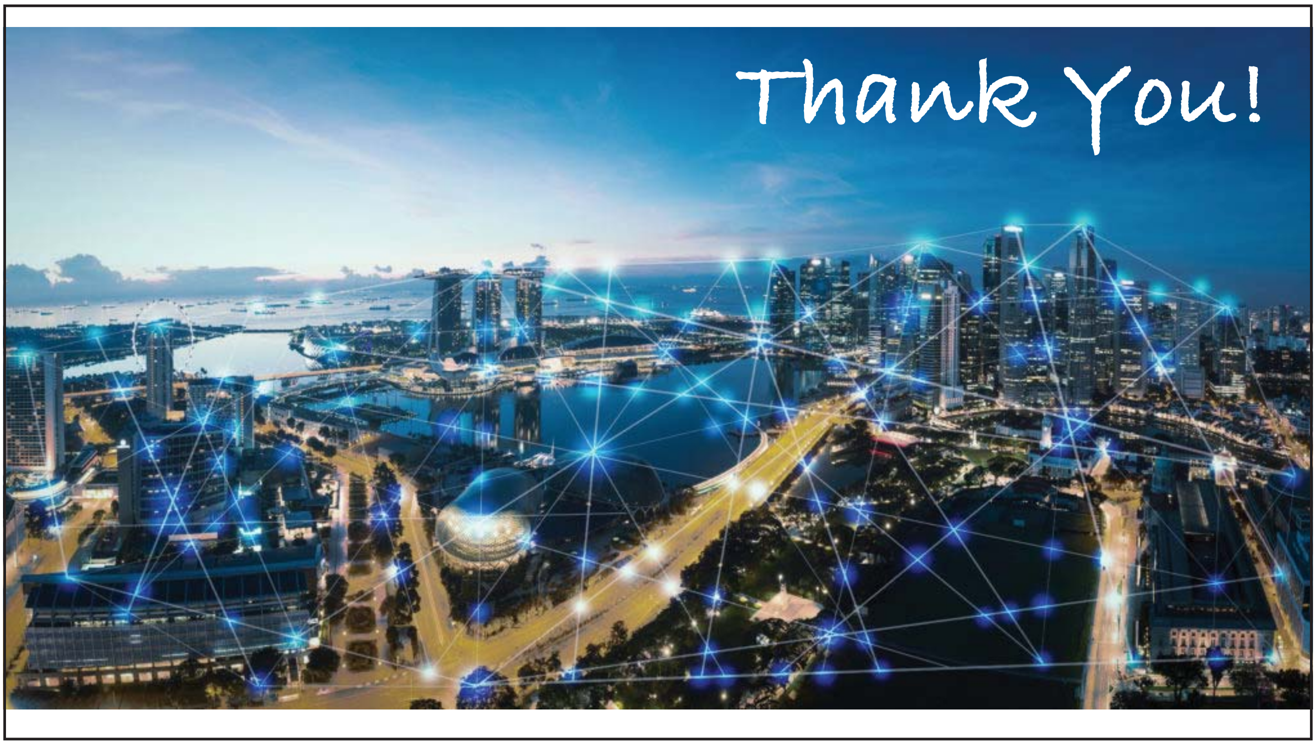
Others

Lennart Edblom, Senior Lecturer · Tomas Forsman, Systems Expert

Recent PhD/PostDoc Departures

- **Selome Kostentinos**, Ericsson Research (Stockholm)
- **Amardeep Mehta**, Ericsson Research (Stockholm)
- **Olumuyiwa Ibidunmoye**, IBM Advanced Analytics (Calgary)
- **Gonzalo Rodrigo**, Apple, USA (via Lawrence Berkeley Lab)
- **Mina Sedaghat**, Ericsson Research (Stockholm)
- **Luis Tomás**, Red Hat (Spain)
- **Wubin Li**, Ericsson Research (Montreal)
- **Lei Xu**, IBM Research (Dublin)
- **Daniel Bergström**, Ericsson Research (Luleå)
- **+ 9** involved in the local spin-off company

Visit us at:
www.cloudresearch.org


Thank You!