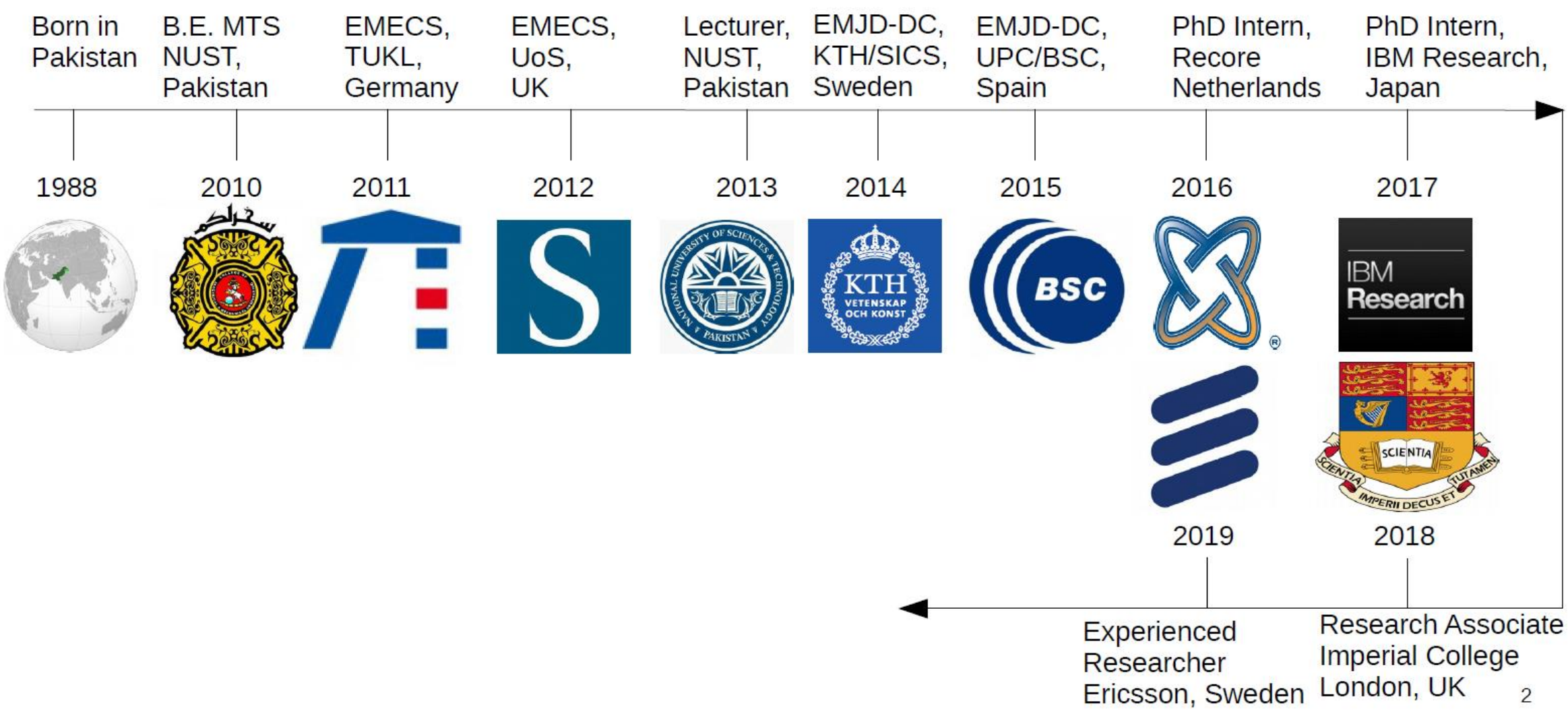
# Towards High Performance Cloud Infrastructure for 5G and Beyond

Dr. Ahsan Javed Awan
Ericsson Research
ahsan.javed.awan@ericsson.com

# About me

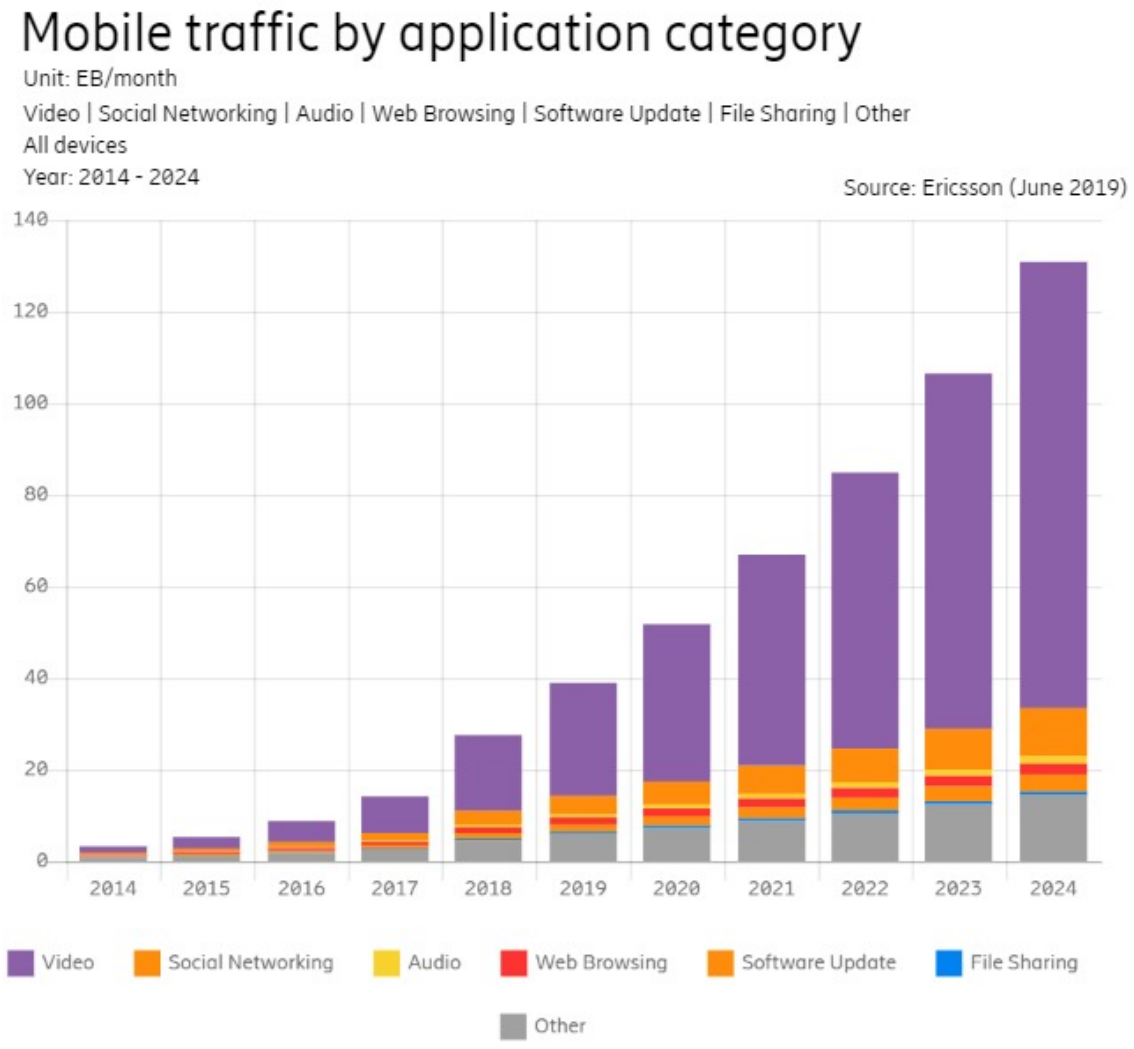

Born in Pakistan
1988
B.E. MTS NUST, Pakistan
2010
EMECS, TUKL, Germany
2011
EMECS, UoS, UK
2012
Lecturer, NUST, Pakistan
2013
EMJD-DC, KTH/SICS, Sweden
2014
EMJD-DC, UPC/BSC, Spain
2015
PhD Intern, Recore Netherlands
2016
PhD Intern, IBM Research, Japan
2017
Research Associate Imperial College London, UK
2018
Experienced Researcher Ericsson, Sweden
2019

# Cloud Systems Research @ Ericsson


- Cloud services and technologies for service delivery:
  - Focuses on cloud middle-wares and application components, technologies for cloud service delivery and adaptation of components to a cloud service.
- Core Cloud Platform:
  - Focuses on core technologies for current and future cloud operating systems, cloud native application execution environments, and distributed cloud system software.
- Future computing platform:
  - Focuses on maximizing the impact of disruptive HW capabilities on cloud systems and platforms
- Intelligent Cloud Operations:
  - Researches efficient and smart automation solutions for distributed cloud.

https://www.ericsson.com/en/blog/2019/5/cloud-evolution-the-era-of-intent-aware-clouds

# Big Data @ Ericsson

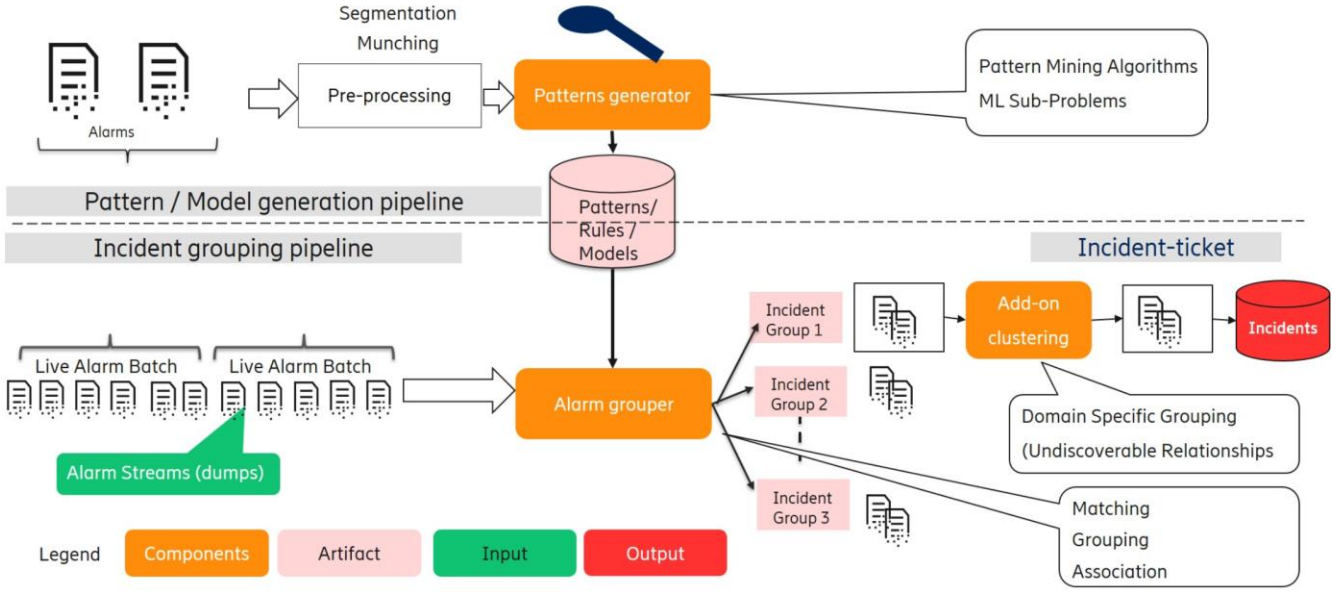



https://www.ericsson.com/en/mobility-report/reports/june-2019

# Big Data Analytics @ Ericsson

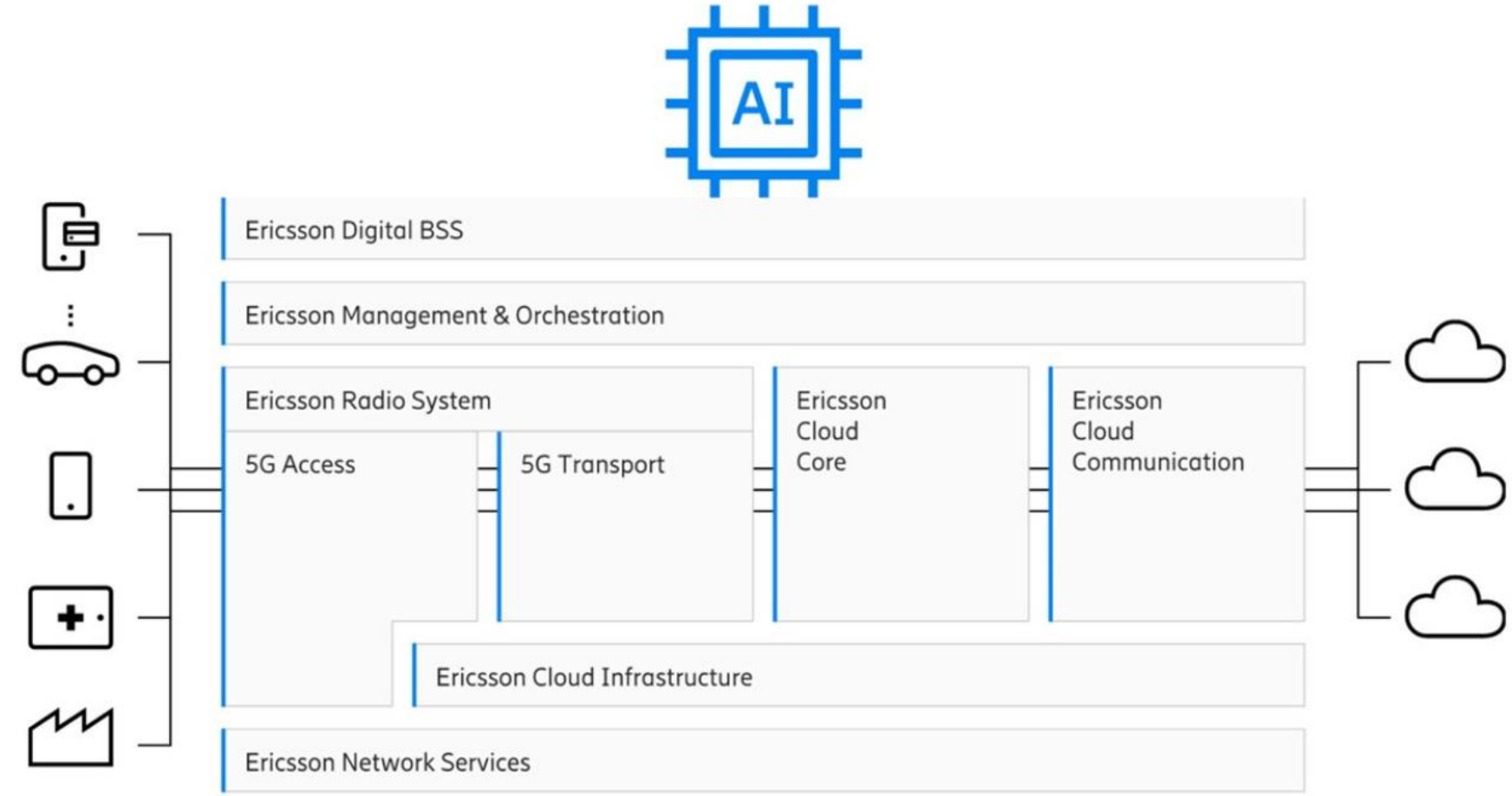
- Network Design and Optimization
  - Capacity Planning for the operator: Traffic forecast and optimal decision of where and when to add CAPEX
  - Performance Diagnostics: Advanced root cause analysis involving 100+ KPIs over time
  - Mobility optimization: Intelligent detection of high mobility cells for tailored parameterization.
  - Centrailized/Elastic RAN Design: Grouping of cells into baseband units and inter-baseband connectivity for optimal performace.

- Incident Detection in Network Operations Center




https://www.ericsson.com/en/blog/2018/6/machine-intelligence-at-the-noc

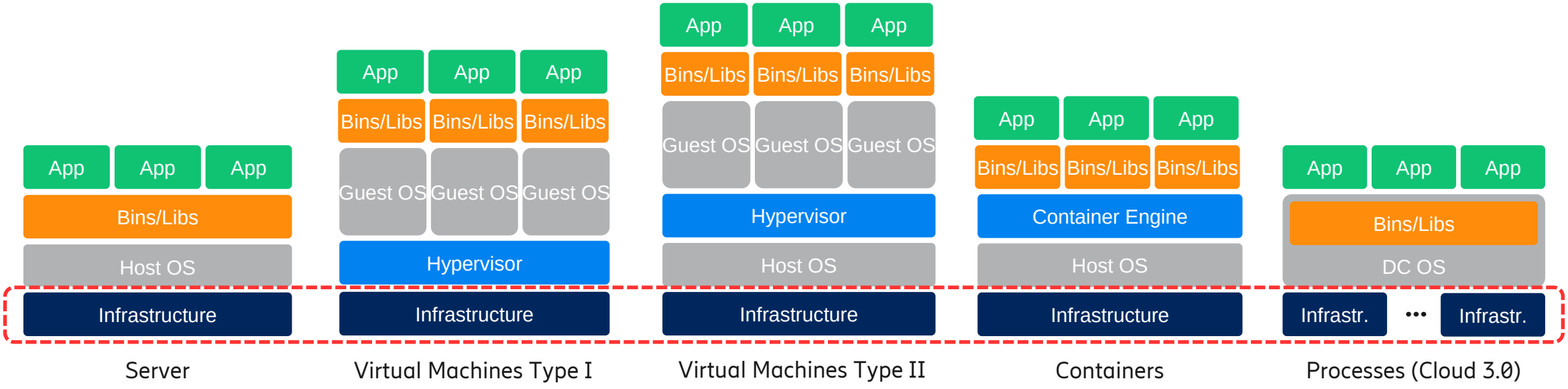
# Towards Zero Touch Networks



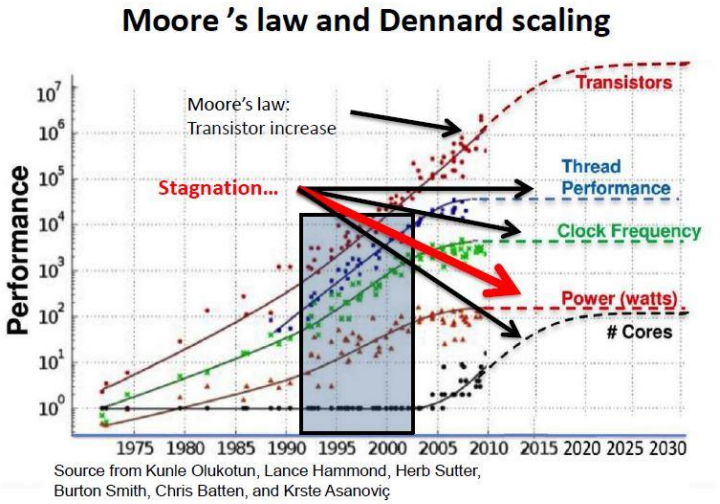

Ericsson 5G platform

https://www.ericsson.com/en/ai-and-automation

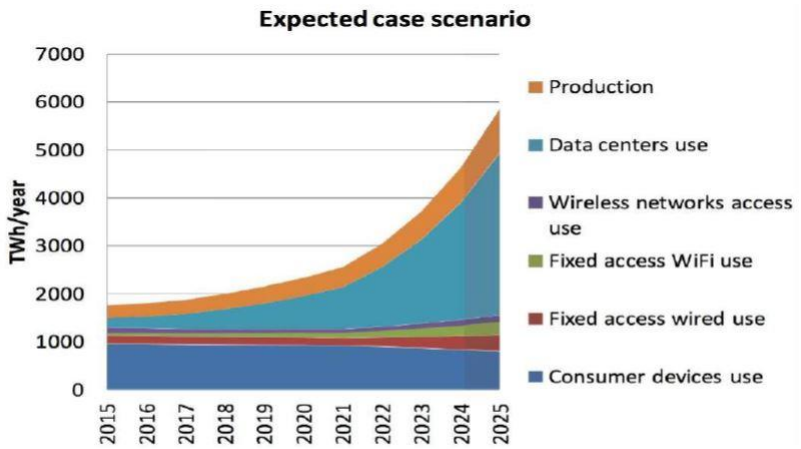
# Cloud deployments are largely based on COTS HW



# Free Lunch is over



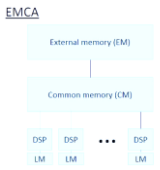
Source from Kunle Olukotun, Lance Hammond, Herb Sutter, Burton Smith, Chris Batten, and Krste Asanoviç



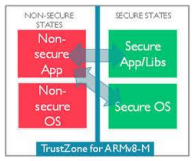
From “Total Consumer Power Consumption Forecast”, Anders S.G. Andrae, October 2017

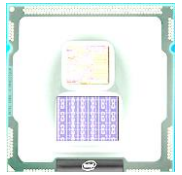
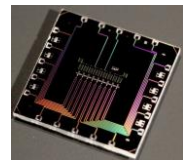
- Domain specific accelerators are rapidly joining the class of commodity hardware at the end of Moore's law.
- Energy efficiency is the primary concern in green/sustainable cloud deployments.
- Future data centers will be highly heterogenous.

ASIC GPU FPGA P4/SmartNIC TEE Memory Tech. Neuromorphic HW Quantum Comp. ...





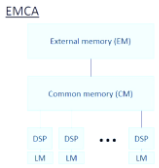




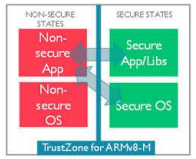
A Performant Cloud-stack has to leverage HW-based Optimization

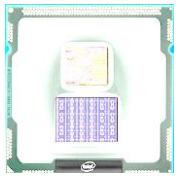
# The Scope of Future Computing Platforms

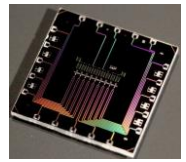
- Maximizing the impact of disruptive HW capabilities on cloud systems and platforms by
  - Designing models and abstractions of HW innovations to make them accessible by Cloud stacks
  - Providing low level management services for these HW resources, including integration with distributed Cloud infrastructures
  - Developing interface specifications towards higher level services to expose them via distributed Cloud infrastructures

ASIC GPU FPGA P4/SmartNIC TEE Memory Tech. Neuromorphic HW Quantum Comp. ...





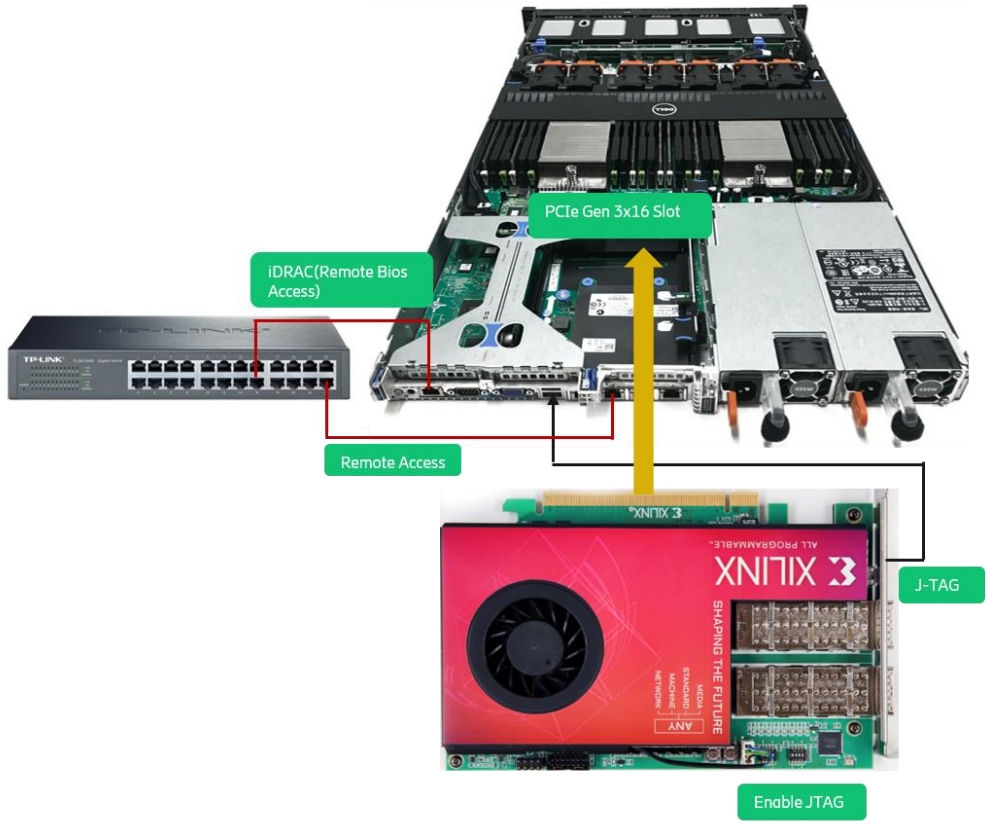
Mature HW:
lacking proper abstractions for virtualization

Non-existent/very experimental HW: understanding the potential impact on future compute / communication

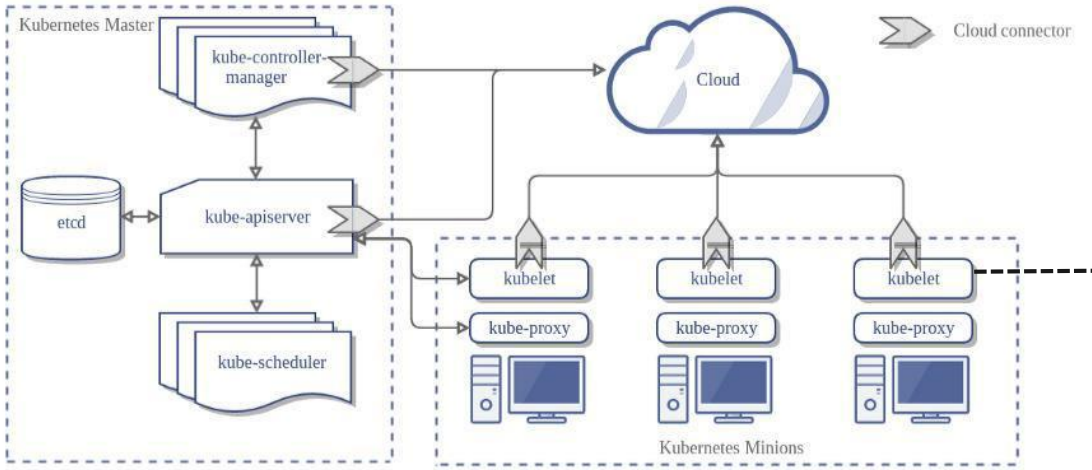
# Acceleration Enablers: FPGAs

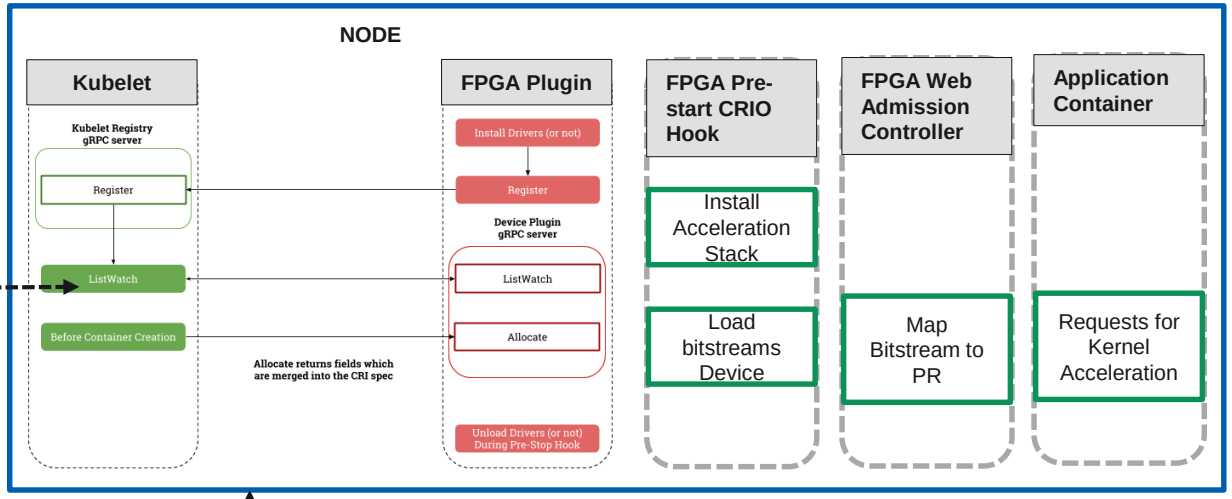
- **Why**: Cloud Deployment of RAN on COTS hardware does not meet the performance and energy efficiency requirements.
  - Edge data-centers (Radio base stations) will run third-party accelerators together with RAN accelerators.

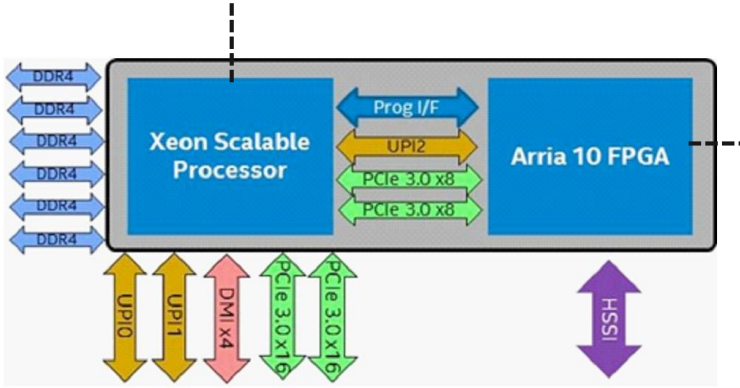
- **What**: Kubernetes based FPGA sharing system.
  - Time and Space sharing of FPGA resources
  - Performance isolation of shared resources, e.g. on-board DRAM and PCIe bandwidth.
  - Generation of bitstreams for partially reconfigurable part of the FPGA with-out exposing the static logic.
  - Extension of K8 Device Plugin with additional metrics for improved life cycle management and orchestration of accelerators.

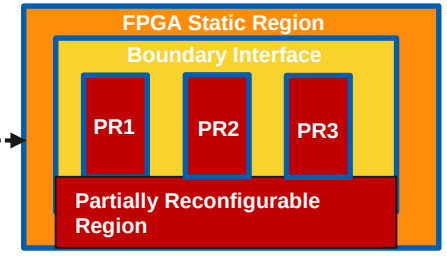


# FPGA Sharing in K8s



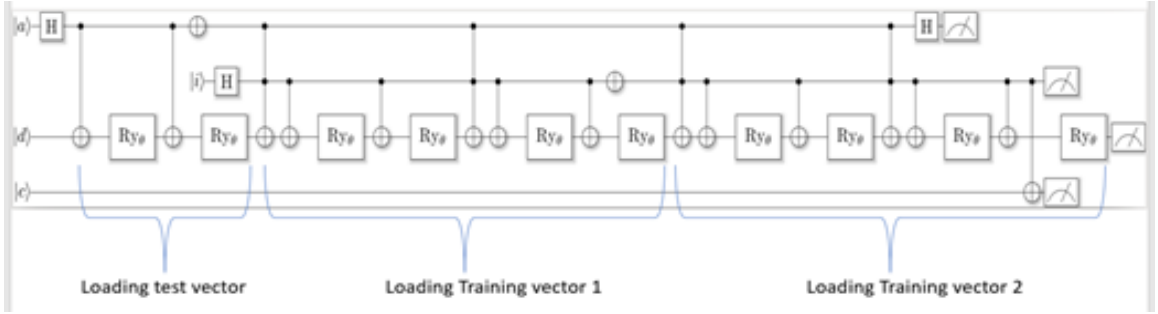



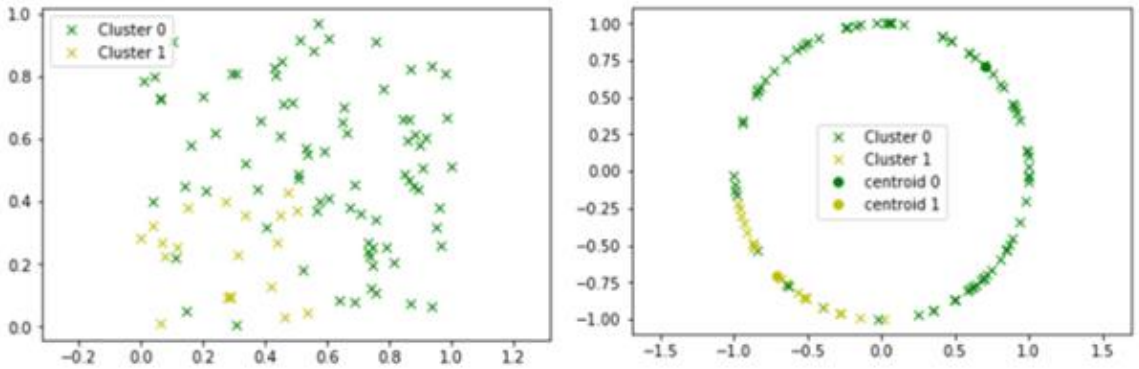




- FPGA is divided in to multiple partially reconfigurable regions.
- FPGA web admission controller maps user provided bit-stream to one of the partial regions
- FPGA device plugin is strarted in the region mode and plugin registers itself with the kubelet. It checks whether one of the regions are allocatable
- FPGA Pre-start CRIO hook installs the Accelerations Stack Runtime and it program the static and partial region
- Application container requesting kernel acceleration is launched and run

# Quantum Technologies: Q clustering algorithm

- E use case: clustering is used for automatic anomaly detection in network design optimization project @BMAS
- A 4-qubit implementation of QK-means algorithm has been performed on IBM simulator and *IBMQX2 machine*
  - Our own and novel 3-qubit implementation of the QK-means algorithm is able to perform clustering within the coherence time of the IBMQX2 machine
  - Clustering accuracy is similar to the classical K-means on a randomly generated data set.



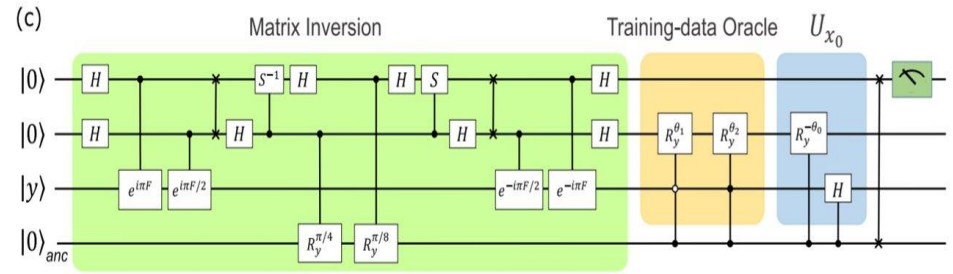



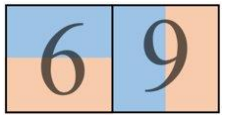

IBMQX2 machine results

https://www.ericsson.com/en/blog/2019/7/introduction-to-quantum-computer-technology

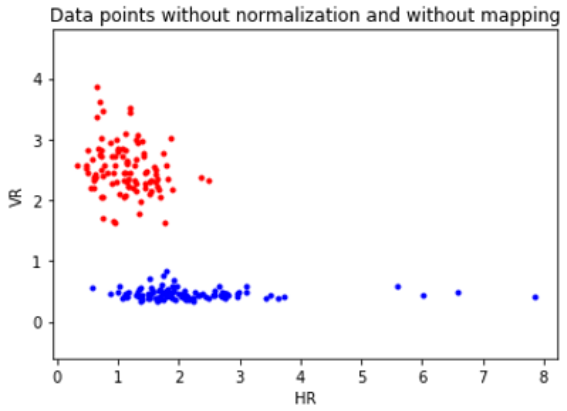
# Quantum Technologies: Q support vector machine

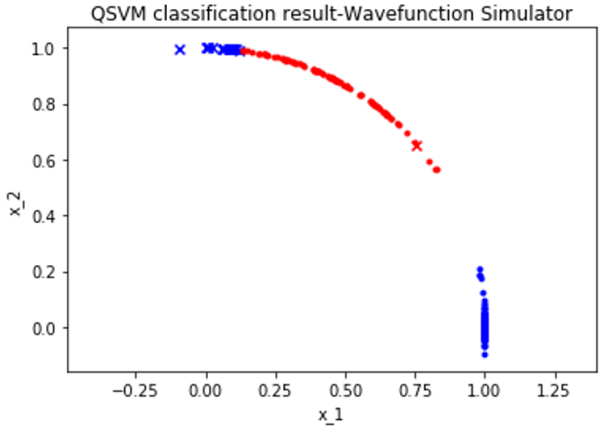
- E use case:  predict the quality of user experience for video streaming based on device and network level metrics @ER MIA
- The QSVM algorithm in its linear form it will separate a set of positive examples from a set of negative examples with a maximum margin.
  - Our 4-qubit implementation of linear Q support vector machine on IBM's simulator is able to separate two types of values with a ca 90% classification accuracy
  - use classical computation of the density matrix to reduce the quantum circuit depth



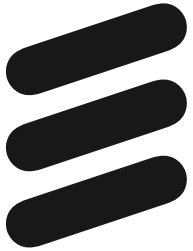
QSVM run on IBM's simulator



# Take Aways!

- Big data analytics in the cloud is one of key differentiators of Ericsson future offerings.
- Exploiting novel hardware in the cloud is necessary to support 5G usecases.
- We offer master thesis projects and summer internships.



Thank You!

# Take aways

- Big data analytics in the cloud is one of key differentiators of Ericsson future offerings.
- To support telco applications, we
- https://www.ericsson.com/en/blog/2019/6/applying-the-spark-streaming-framework-to-5g?utm_source=twitter&utm_medium=social_organic&utm_content=f21c92b8-8c95-4094-b67b-fc8d5fd9ad4d&utm_campaign=
- Artificial Intelligence in Next Generation Systems

https://www.ericsson.com/en/white-papers/machine-intelligence

https://www.ericsson.com/en/ai-and-automation

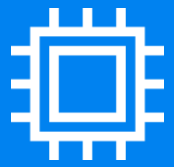
- Machine Intelligence at the Network Operations Center
- https://www.ericsson.com/en/blog/2018/6/machine-intelligence-at-the-noc
- Network Services and Automation
- https://www.ericsson.com/en/networks/offerings/network-services-and-automation
- Analytics
- https://www.ericsson.com/en/blog/?topics=380702
- AI in the telecom
- https://www.ericsson.com/en/blog/2019/6/ai-in-telecom

# Glimpse of our work

Advanced Memory and Storage Technologies

Quantum computing

Accelerator Enablers



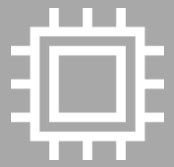
Resource Isolation

# Project Highlights

Advanced Memory and Storage Technologies

Quantum computing

Accelerator Enablers



Resource Isolation