CASE STUDY

AI for 5G slicing

From a reactive to proactive approach

ERICSSON

Published by

MOBILE WORLD LIVE

Network slicing is one solution that has emerged as an opportunity with 5G. It provides the capability to enable new business models across a wide range of industries and allows operators to segment the network to support particular services and deploy multiple logical networks for different service types over one common infrastructure.

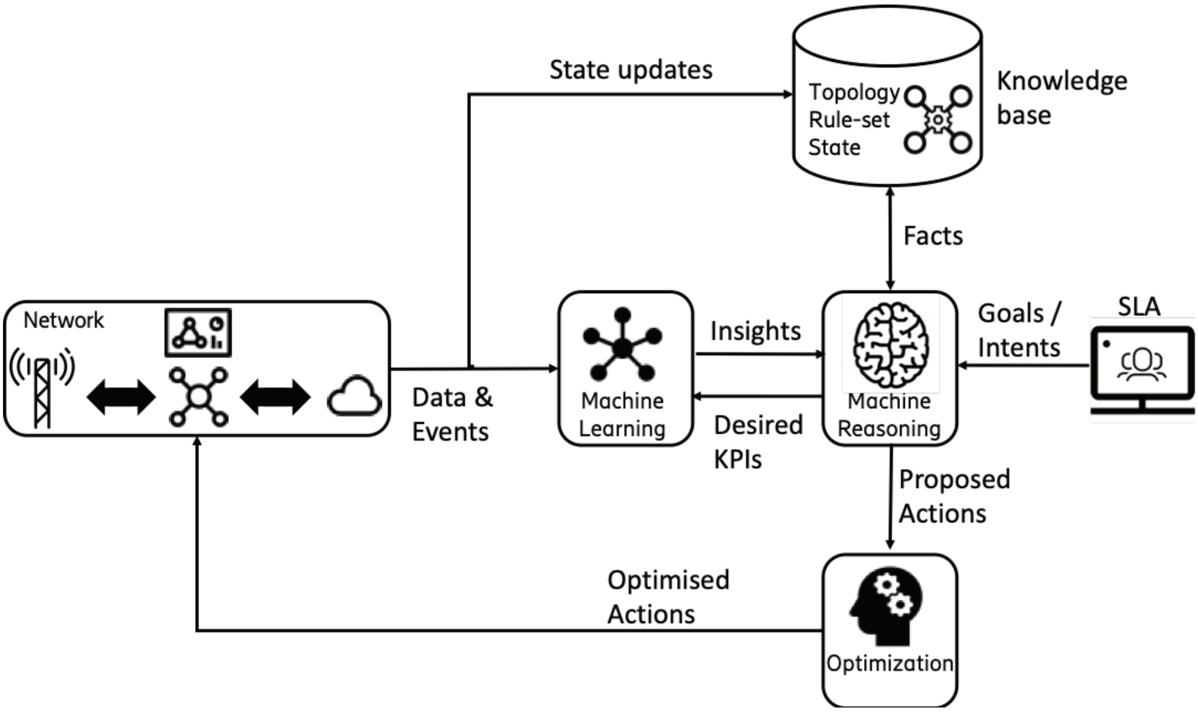
Dedicated network slices give a multitude of opportunities for service providers to grow their business especially towards industrial use cases. These new use cases bring new and more diverse network requirements, which increases the complexity beyond human control. In order to ensure the networks are agile and flexible, as well as easy to operate, artificial intelligence (AI) and automation becomes a cornerstone. AI makes it possible to create self-healing networks and Ericsson intends to embed AI-infused solutions that can identify problems, and suggest and apply solutions in a proactive manner.

Intent-based solution for 5G slicing

The most important thing in 5G slicing is guaranteeing the KPIs (as agreed with the customers in SLAs) throughout the lifecycle of the slice from its admission till its deactivation. This complex and time-consuming activity drives the demand for automation of problem identification/prediction and even proposing solutions. AI and ML (Machine Learning) technologies continue to be integral to the development of these technologies, such as enabling the automation of slice admission and its assurance across the mobile network platform. Ericsson´s continued research in this space will ensure we remain the market´s most advanced partner when it comes to AI-infused automation of the network platform.

Picture 1 demonstrates a high-level architecture of AI-infused intent-based solution using ML, knowledge bases, and machine reasoning (MR) techniques. The customer intent is captured during the service level agreement (SLA) stage and is converted into desired KPIs that ML algorithms are configured to predict. At the start, a 5G slice is admitted based on the current resource status and admission policy. During the lifecycle of the slice, the data/events are captured from the network, ML analysis performed to predict potential KPI degradations and the insights are fed to MR that, based on its background knowledge about network state and operational procedures, proposes mitigation actions according to the given intent. The implementation of these actions can be further processed to optimize other non-functional SLA metrics such as cost or power consumption.

Picture 1: Overview of Intent-based way of performing 5G slicing



Machine learning agents for prediction and root-cause analysis (RCA)

Advances in ML and AI technologies will help to reach new levels in how 5G slices admission and service assurance can be maintained, not only without any manual intervention but also in a proactive manner. This will be possible by creating a new type of active element of the network called ML agents. These agents are expected to function on all places where the data flows; that can be antennas, network nodes, cloud infrastructure nodes or data lakes where the data is collected and aggregated. The ML agents may or may not fulfil the same purpose by using the same or different data. Some of the ML agents will provide input to others for RCA while other ML agents are expected to work in parallel, providing possibility to choose the prediction with the highest accuracy.

Different ML techniques could be used for building the ML agents, such as:

- Time series-based prediction and analysis using multivariate LSTM's and Vector Auto Regression for detecting violations of the SLA's.
- Proactive actions can be taken based on prediction and reasoning.
- Regression modelling with forest based like random forest or XGB and neural networks for finding causal relations between most important features for the predicted SLA violation. By comparing prediction with SLA.
- Some time series prediction algorithms are compared using a real-world dataset. The dataset contains mainly daily and weekly seasonality and lasts for one year. Compared algorithms are ARIMA, Fourier Extrapolation, Prophet, XGB, LSTM, GARCH. ARIMA is ok for very-short term prediction (for example less than 5 steps) with an acceptable variance. Fourier is good at capturing long-term seasonality, but only seasonality. Prophet is a combination of Fourier and scaling, which is good for seasonal prediction especially long-term prediction with a trend, but still not good for dynamic time-changing data. Prophet provides a higher variance and it performs bad for prediction with spikes/peaks. XGB is a general approach and comes with an acceptable variance which is suggested but may need to have updatable models. LSTM can also handle multivariate in a good manner but does not show obvious improvements compared to XGB. GARCH can be utilized to predict variance which is needed for confidence intervals and possibility of SLA breach. It also has the shortcomings of ARIMA but a little better so a mid-term variance prediction can use GARCH.
- Bayesian networks could be used for RCA.

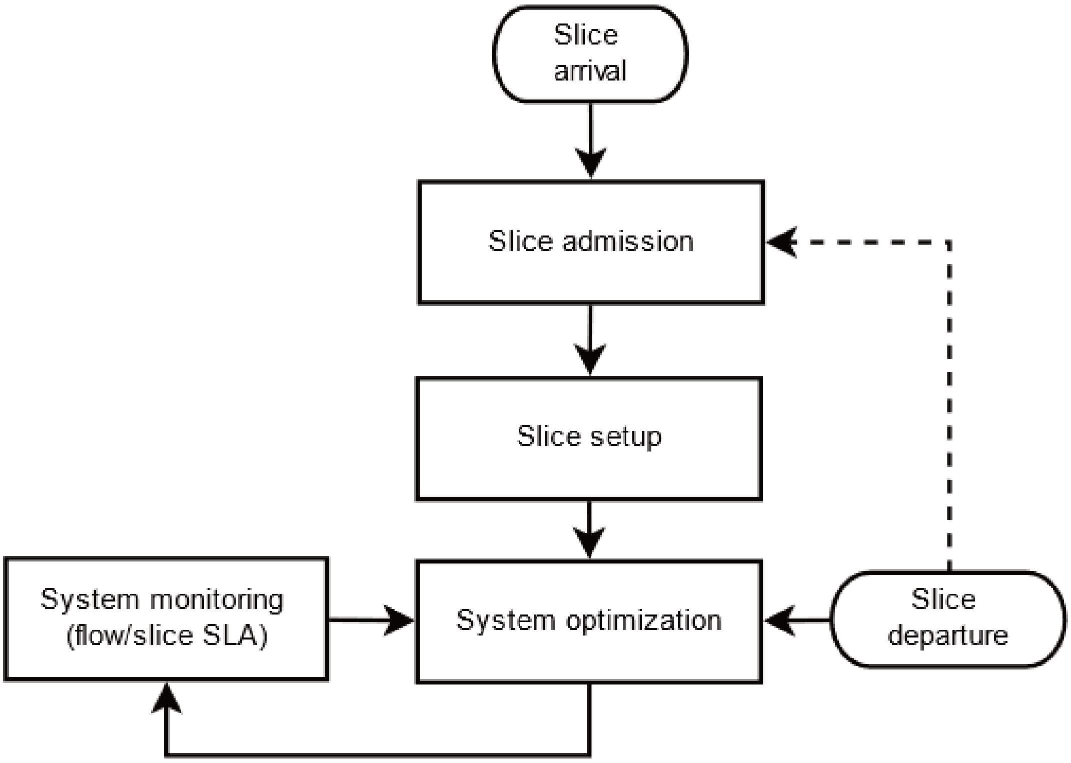
Machine reasoning for best solution according to intent and the ML output

Machine reasoning can be used to make rational decisions based on all the available knowledge about the network, its topology and current state as well as the domain knowledge captured in the form of rules that describe what procedures are applicable in various situations. Such rules may be both manually provisioned into the system by knowledge engineers with the help of SMEs, and automatically learned and generalized over time from the results of heuristic search performed by MR algorithms. The way MR proposes an action is by generating and estimating potential situations using all available knowledge as constraints that define the exploration space and heuristics that efficiently direct such exploration towards the goal. The techniques employed by MR may include forward and backward chaining, inductive, abductive and probabilistic reasoning. For example, forward chaining is traditionally used in expert system to derive immediate conclusions and actions based on recently asserted facts. Backward chaining allows to infer implicit facts (not asserted in the knowledge base) but logically follow from the existing facts and rules. Using probabilistic reasoning it is possible to predict likelihood of outcomes if a certain action is taken, thus, enabling action planning by considering historical information. One valuable property that MR algorithms have is that, unlike wide-spread statistical ML approaches, the MR algorithms are symbolic in nature, i.e. they operate explicit concepts and relations and can be easily augmented with explainability and traceability features that contribute to system transparency and trust.

Close-loop operation for slice management

Network management operations such as slice admission or orchestration can be modeled as a closed-loop operation, as depicted in Picture 2. In general, a slice arrives in the network and then goes through an admission policy. The policy decides whether it is in the interest of the infrastructure provider to accept or reject the slice. Slices that are admitted in the network require setup and continuous monitoring and optimization. The closed-loop monitoring and control mechanism ensure that if a deployed slice instance requires more resources than it did at the deployment time, those additional resource requirements can be accommodated; or, if the slice instance requires less resources, that they can be allocated to another slice in the network. Additionally, in the case of slice completion and departure from the network, the closed-loop operation ensures that not only the resources are taken away from departing service but also the state of remaining services is optimized (e.g., the network management discussed in Picture 1). Besides, the slice experienced performance is reported to the admission module, so it can evaluate the consequences of the admission as well as the slice behavior. Consequently, enabling not only the admission control to learn how to make better admission decisions in the future but also enabling proactive slice management during its lifecycle.

Picture 2. An overview of the general network management loop. The dashed line represents the exchange of information, for example, reporting the overall satisfaction experienced by the service during its lifetime



AI for slice admission

A slice deployment request made from the tenant contains an SLA requirement, and the infrastructure provider (InP) must provide enough resources to fulfill the SLA. Examples of SLA include network coverage over a certain geographical area, minimum network bandwidth, latency, etc. The InP has limited resources, therefore, in some situations, it cannot fulfill the SLA for all the tenants requesting slices.

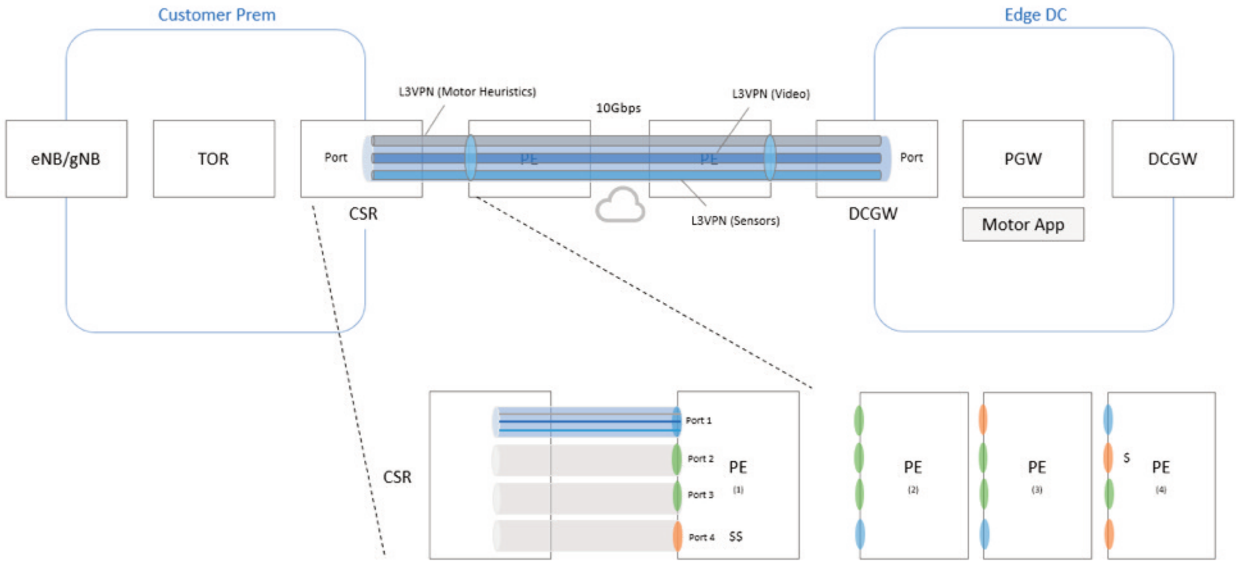
The limited resources motivate the existence of the slice admission module. Its objective is to admit as many slices as possible into the system, with the objective of maximizing resource utilization, consequently increasing the revenue for the infrastructure provider. The constraint is it should not admit slices that would have their SLA violated, or cause SLA violation for the other deployed services. Further insights of such strategy were discussed in a published study.

Important information to decide on the slice admission includes the state of the system, some features are: amount of resources allocated to network slices such as physical resources or data rate; timing such as starting time, duration and periodicity of requests and time window; the type of resources and Quality of Service (QoS) parameters such as radio/core bearer type, prioritization, delay, jitter and loss; maximum resource utilization and traffic classes. In this context, traffic class specifies some behavior of the traffic i.e., delay tolerance and if the bit rate should be guaranteed or not.

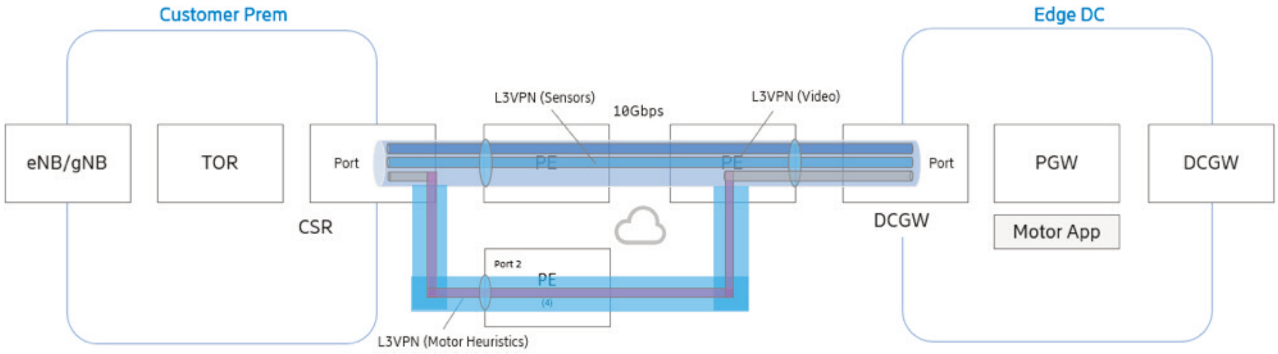
AI for slice assurance

Typically, when SLA breaches occur, engineers try to find a solution and allocate more resources. With this strategy, the system will use AI based approaches to predict possible problems and proactively provide recommendations of solutions for engineers to choose for high-impact issues, and/or automatically fix the problem for small-impact issues. This requires higher accuracy of prediction especially for delay and traffic volume, together with pre-defined SLA, so that SLA breaches can be predicted, and proactive actions can be taken.

Picture 3 (a). AI for enabling mission critical 5G based use case



Picture 3 (b) Re-route NW traffic based in proactive manner



Picture 3 (a) presents an example of a network slice where an SLA violation is predicted using ML agent. The insights are fed to MR, which, based on the SLA requirements, cost, and using the current system states, proposes the solution in Picture 3 (b) to reroute that traffic from port1 to port 2 so that the SLA requirements and cost can be met in a proactive manner.

Conclusion

Efficient sharing of network resources while providing customers with superior SLA-based service quality is at the core of 5G slicing. The complexity imposed by it needs to be mitigated by employing a fundamentally different approach to management. Intent-based way of performing 5G slicing is revolutionary because it replaces manual configurations and maintenance steps with a declarative model of desired outcomes. We have presented the work in which instead of controlling the network imperatively, high-level business goals are provisioned, and the network automatically takes care of optimally implementing it. A significant shift in controlling the networks is possible by combining machine learning (ML) and machine reasoning (MR).

Ericsson enables communications service providers to capture the full value of connectivity. The company's portfolio spans Networks, Digital Services, Managed Services, and Emerging Business and is designed to help our customers go digital, increase efficiency and find new revenue streams. Ericsson's investments in innovation have delivered the benefits of telephony and mobile broadband to billions of people around the world. The Ericsson stock is listed on Nasdaq Stockholm and on Nasdaq New York.

Find out more www.ericsson.com

Produced by the mobile industry for the mobile industry, Mobile World Live is the leading multimedia resource that keeps mobile professionals on top of the news and issues shaping the market. It offers daily breaking news from around the globe. Exclusive video interviews with business leaders and event reports provide comprehensive insight into the latest developments and key issues. All enhanced by incisive analysis from our team of expert commentators. Our responsive website design ensures the best reading experience on any device so readers can keep up-to-date wherever they are.

We also publish five regular eNewsletters to keep the mobile industry up-to-speed: The Mobile World Live Daily, plus weekly newsletters on Mobile Apps, Asia, Mobile Devices and Mobile Money.

What's more, Mobile World Live produces webinars, the Show Daily publications for all GSMA events and Mobile World Live TV – the award-winning broadcast service of Mobile World Congress and exclusive home to all GSMA event keynote presentations.

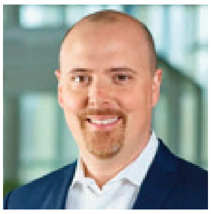
Find out more www.mobileworldlive.com

Dr. Elena Fersman, *Head of Research*

Dr. Elena Fersman is a Research Director in Artificial Intelligence at Ericsson. She is responsible of a team of researchers located in Sweden, US, India, Hungary and Brazil. She is a docent and an adjunct professor in Cyber-Physical Systems specialized in Automation at the Royal Institute of Technology in Stockholm. She holds a PhD in Computer Science from Uppsala University, a Master of Science in Economics and Management from St. Petersburg Polytechnic University and did a postdoc at the University Paris-Saclay. At Ericsson, she had various positions ranging from product management to research leadership. Since April 2019, Elena is a member of the Board of Directors of RISE Research Institutes of Sweden. Elena has co-authored over 50 patent families.

Dr. Rafia Inam, *Senior Project Manager*

Rafia Inam is a senior project manager at Ericsson Research in research area Artificial Intelligence, Sweden. She joined Ericsson research in 2015 and has worked as senior researcher and Single Point of Contact for Scania 2017-2018. Her research interests include 5G network slices and management, using AI for automation, use cases applied to 5G for industries, service modeling for Intelligent Transport Systems, automation and safety for CPS, reusability of real-time software and ITS. She received her Ph.D. from Mälardalen University, Sweden, in 2014. Rafia has co-authored 40+ scientific publications and 40+ patent families



Jonas Åkeson, *Head of Automation & AI*

Jonas Åkeson drives implementation of AI and automation in Ericsson's services business, realizing the company's vision of delivering superior user experience by transforming networks and IT operations and optimization through AI, automation and the power of data.

With 15 years' experience in the telecom industry. He strives to explain complex issues in simple terms, ensuring that his passion for automation and vision for artificial intelligence within Managed Services does not get lost in technical jargon.

Jonas holds a Master of Science degree in Engineering from Linköping Institute of Technology and a Business and Administration degree from Stockholm University. Before his current position, Jonas has held various executive positions in Ericsson.

Acknowledgement

A special thank you to the following co-authors for contributing to the article: Leonid Mokrushin, Pedro Batista, Valentin Kulyk, Bin Sun S, Sushanth David and Jia You J