ANIARA Project - Automation of Network Edge Infrastructure and Applications with Artificial Intelligence

Wolfgang John, Ali Balador, Jalil Taghia, Andreas Johnsson, Johan Sjöberg
Ericsson, Stockholm, Sweden; email: {wolfgang.john, ali.balador, jalil.taghia, andreas.a.johnsson, johan.sjoberg}@ericsson.com
Ian Marsh, Jonas Gustafsson
RISE Research Institute of Sweden, Stockholm, Sweden; email: {ian.marsh, jonas.gustafsson}@ri.se
Federico Tonini, Paolo Monti
Chalmers University, Sweden; email: {tonini, mpaolo}@chalmers.se
Pontus Sköldström
Qamcom AB, Sweden; email: {pontus.skoldstrom}@qamcom.com
Jim Dowling
Hopsworks AB, Sweden; email: jim@hopsworks.ai

Abstract

Emerging use-cases like smart manufacturing and smart cities pose challenges in terms of latency, which cannot be satisfied by traditional centralized infrastructure. Edge networks, which bring computational capacity closer to the users/clients, are a promising solution for supporting these critical low latency services. Different from traditional centralized networks, the edge is distributed by nature and is usually equipped with limited compute capacity. This creates a complex network to handle, subject to failures of different natures, that requires novel solutions to work in practice. To reduce complexity, edge application technology enablers, advanced infrastructure and application orchestration techniques need to be in place where AI and ML are key players.

1 Introduction

ANIARA project is based on two use case families addressing smart manufacturing and smart cities. Figure 1 illustrates the functional technologies for evolving 5G edge systems. The application scenarios are drawn upon the synergies between 5G, cloud computing and edge computing. Our studies include elaborating detailed use case descriptions for verticals using 5G and the edge cloud. Concerning manufacturing, ANIARA focuses on two main aspects [1]. These are: environmental monitoring and control of the factory floor concerning properties such as air quality; temperature and power management and operations monitoring and control such as robot cell control, logistics and safety. These use cases drive the high-level system architecture requirements and serve as the basis for technology-specific use cases such as private 5G, edge micro datacenters and AI at the edge.

The remainder of this paper is organized to provide additional details of the components in Figure 1. Section 2 provides details about edge platform infrastructure and services that have been developing in ANIARA, including lightweight and portable execution environments and fast, dependable feature stores. Apart from edge infrastructure, developing edge AI enablers is also another important objective of the ANIARA. Section 3 describes AI/ML methods designed and developed within the project so far, including ML models for life-cycle management, intelligent feature selection and distributed learning in the edge considering privacy. ANIARA also addresses management and orchestration including both edge and cloud scenarios, to satisfy the specific needs of the use cases, detailed in Section 4. Section 5 presents research on smart power for building a large-scale 5G edge system, shown in the lower portion of Figure 1. Section 6 provides a comprehensive related work. Finally, Section 7 gives an overview of works planned to be done in the future.

2 Edge platform components

2.1 Programmable, light containerisation

Lightweight and portable execution environments have been identified as a crucial enabler for higher flexibility and dynamism of application deployment in a distributed network [2, 3]. In ANIARA, we experiment with WebAssembly technologies to fill this gap. WebAssembly [4] is an *open* binary instruction format for a stack-based virtual machine, designed to support existing programming languages in a web browser environment. As an example, Edgedancer, offers infrastructure support for portable, provider-independent, and secure migration of edge services, it is a lightweight and generic execution

Joint publication with Ada User Journal, Vol. 43(4), December 2022, with permission.

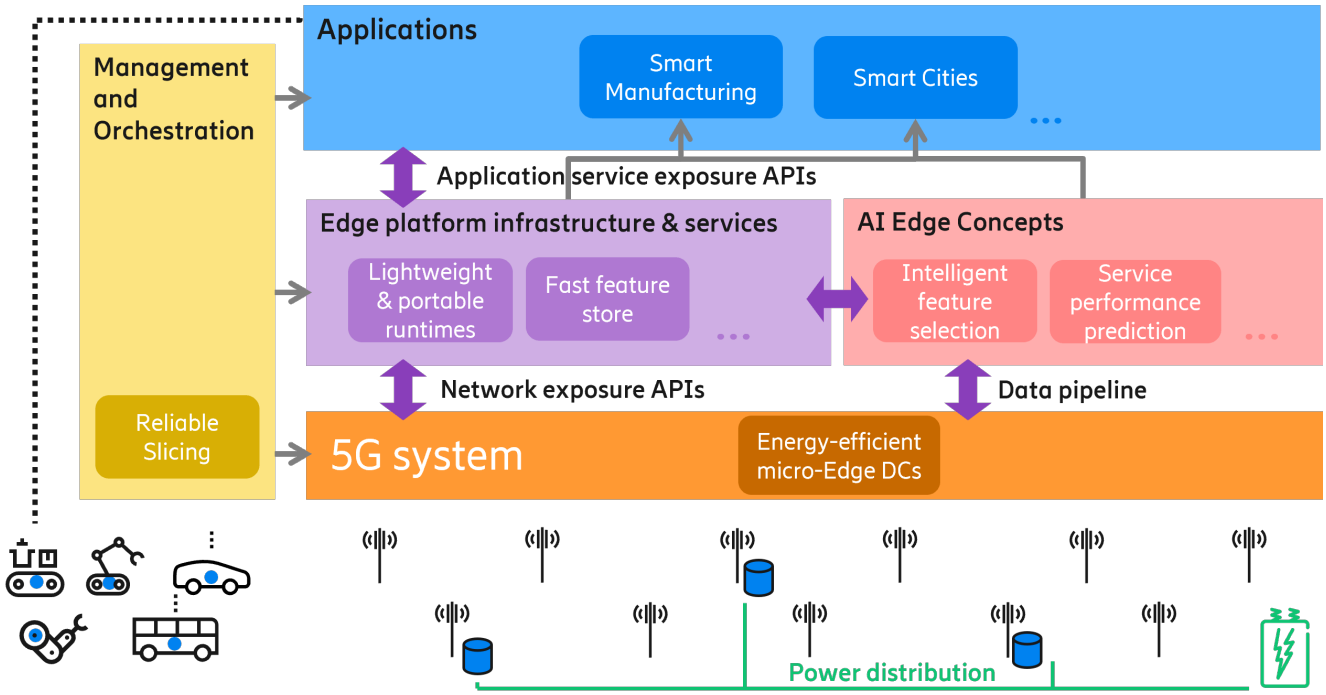

Figure 1: Functional structure of the ANIARA project.

environment by utilising WebAssembly [5]. In the browser, WebAssembly is generally faster than JavaScript due to its more compact format and manual memory management. The virtual machine is, however, not bound to browsers, but can be used standalone on various platforms. A deployment flow is where a high-level language such as C, C++, Rust, Python and so on can be compiled to WebAssembly (using clang) and executed on various hardware or software platforms. Webassembly allows edge applications to execute within a wide range of devices and operating systems. In ANIARA, we wrote an application in Rust and executed it on a WebAssembly runtime using a $30 microcontroller, to prove the versatility and low footprint. The same application was also written in Python that runs (on WebAssembly again) in an ASCII terminal.

2.2 Fast, dependable, feature store

A production-grade AI solution for edge computing, known as Feature stores, is a part of ANIARA. They act as a halfway house between data scientists and data engineers, enabling the same feature computation code to be used for model training and inference. Feature stores also act as a centralized repository of AI models and have been adapted to perform in distributed scenarios for real-time data synchronization between geographically distributed Edge locations.

The underlying database is called RonDB, evaluated using **LATS**, meaning low **L**atency, high **A**vailability, high **T**hroughput, scalable **S**torage principles to list RonDB's performance:

- RonDB x3 times lower latency
- RonDB 1.1M reads/sec, REDIS 800k reads/sec
- RonDBs processed 200M reads in a 30 node cluster

3 AI for Edge Management

We divide AI for edge management into three main research tracks: (i) ML model life-cycle management, (ii) Intelligent feature selection, and (iii) Distributed learning in the edge. The three tracks have been studied within a use case for service performance prediction from edge statistics available to an operator with the objective to automate parts of the edge management process.

ML model life-cycle management: Ensuring high-performing ML models that are continuously updated and correctly deployed is critical for successful integration of ML in the edge. Transfer learning [6, 7], which is one approach, allows for structured incorporation of previously acquired knowledge enabling timely and robust model adaptation, especially when data are scarce for reliable training of ML models. In ANIARA, we studied concepts and methods for adapting and selecting ML models for mitigating ML-model performance degradation due to expected vertical and horizontal scaling in the edge infrastructure and 5G system, with specific focus on performance models for networked services [8, 9, 10].

Intelligent feature selection: A challenge for AI at the edge is related to network overhead with respect to measurements and monitoring, and feature selection for improved model performance [11]. A key enabler for ML models is timely access to reliable data, in terms of features, which require pervasive measurement points throughout the network. However, excessive

monitoring is associated with network overhead. Using domain knowledge provides hints to find a balance between overhead reduction and identifying future ML requirements. A review on techniques for unsupervised feature selection is provided in [7] and authors in [12] show a comprehensive review of online feature selection techniques. In ANIARA, We implemented an unsupervised feature selection method that uses a structured approach in incorporation of the domain knowledge acquired from domain experts or previous learning experiences [13].

Distributed learning in the edge: Multi-domain service metric prediction is a key component in the ANIARA framework. It should enable privacy-preserving sharing of knowledge between operators, and low-overhead training of models within an operator in terms of data sharing. The approach requires in-network processing capabilities to enable federated learning. The works in [14, 15] review federated learning methods in mobile edge networks and edge computing, respectively. Our work in ANIARA extends the scope of these works specifically towards edge clouds for telecommunication industry. We are working on a multi-domain service metric prediction framework using federated learning, corresponding to a scenario where several services are managed by a number of operators in geographically distributed locations.

4 Edge-cloud orchestration

Slicing allows provisioning of multiple services over the same infrastructure, where virtual or physical resources are interconnected to form end-to-end logical networks, also known as slices [16]. Orchestrators run resource allocation algorithms to select the most suitable set of resources to satisfy the specific needs of the clients. In the edge cloud, compute resources are located close to the users, allowing provisioning of low latency services and enabling 5G Ultra-Reliable Low-Latency Communication slices. Allocating backup resources requires protecting the slices against link or node failures. Backup resources can be provided by means of a dedicated protection scheme, where resources are dedicated for each slice. Since backup resources are accessed only in case of failures, shared protection schemes can be developed, where backup resources are shared among different slices to decrease the overall amount of required resources.

Resource allocation strategies for 5G networks and reliable services have been investigated recently. In particular, different techniques for backup protection of optical network resources, relying on both DP and SP schemes, have been presented in [17, 18]. Works propose efficient DP and SP algorithms for cloud and baseband resources in 5G access and metro networks, see [19, 20]. Considering connectivity and compute resources separately may lead to impractical solutions, especially when resources are scarce. Our work focuses on the dynamic slice provisioning where both type of resources are jointly allocated. A whitepaper presents an overview of the market and implementation trends, see [21].

In ANIARA, we developed a heuristic-based shared protection to encourage sharing of backup connectivity and cloud resources. We also evaluated this against a dedicated protection scheme using a Python simulator, published in [22]. Results show that the shared approach reduces the blocking probability by order of magnitude, and is especially beneficial when in-node processing resources are scarce.

5 Edge-cloud power research

Installing thousands of edge data centers, primarily in cities will require significant amounts of power, however many power grids are already utilized close to 100%. To maintain high availability, the edge-data centers need to be complemented with alternative power sources and pro-active power management systems. This requires tailored hardware solutions integrated with the power grid and on-site power generation. Supporting active load balancing by going off-grid for shorter periods of time. We are working on the design and implementation of a series of micro-edge-data center demonstrators for deployment at industrial sites. The first generation consists of a double rack configuration including, cooling, UPS-system with batteries, multiple power source inputs and IT-hardware.

To build out a large-scale 5G edge system, smart power utilization is required. One approach is to utilize the on-site UPS installation to go off-grid during peak power periods. The battery storage needs to be dimensioned to address this active usage. Incentivizing the active participation from the edge data centers in the load-balance activity is necessary. Moving away from fixed to dynamic prices will permits battery charging during off-peaks and discharging during higher-priced periods. Discharging implies less grid power. The value for a power grid operator will be larger than that reflected by the customer energy price, if the data center is placed in a particularly energy-hungry section of the grid. Therefore, we have initiated a dialog with a major power grid operator.

6 Future work

Going forward in the AI for edge field, we will study distributed learning under various sources of data and system heterogeneity. The objective is to tackle concerns with data privacy, resource heterogeneity among AI actors, challenges in re-usability of previously learned ML models, and difficulties in effective incorporation of domain knowledge. Experiments with WebAssembly based runtimes to implement *code-once, execute anywhere* approaches across the device-edge-cloud continuum are ongoing. The idea is to support offloading applications from the user equipment device to the edge node. Future work for the Feature store is a Kubernetes operator for RonDB and using it to store and serve our WASM containers. Improving RonDB and implementing an *evaluation store*, feature drift detection is planned. Before a widespread edge data center can be used, we will need to work on power integration aspects that means deployment/installation of the physical hardware. Installation at hard-to-reach places, requiring easy assembly on-site and an autonomous operation with minimal on-site maintenance

is ongoing (a demo was shown at the mid-term review, April 2022). We will also investigate the potentials and limitations of resource sharing in bare metal deployments of containers, and enhanced scaling strategies to improve utilization.

Acknowledgements

This work was supported by EU Celtic Plus (ID C2019/3-2), Vinnova (under the project ID 2020-00763), Bundesministerium für Bildung und Forschung (under the name "AI-NET ANIARA 16KIF1274K") and InnovateUK (under the project ID 106197: ukANIARA) via the ANIARA project.

References

[1] A. Y. Ding *et al.*, "Roadmap for edge AI: A dagstuhl perspective," *CoRR*, vol. abs/2112.00616, 2021.

[2] G. Wikström *et al.*, "6g – connecting a cyber-physical world: A research outlook towards 2030," *Ericsson, White paper*, Feb. 2022.

[3] A. Sefidcon, W. John, M. Opsenica, and B. Skubic, "The network compute fabric – advancing digital transformation with ever-present service continuity," *Ericsson Technology Review*, June 2021.

[4] A. Haas *et al.*, "Bringing the web up to speed with webassembly," in *Proceedings of the 38th ACM SIGPLAN Conference on Programming Language Design and Implementation*, PLDI 2017, (New York, NY, USA), p. 185–200, Association for Computing Machinery, 2017.

[5] M. Nieke, L. Almstedt, and R. Kapitza, "Edgedancer: Secure mobile webassembly services on the edge," in *Proceedings of the 4th International Workshop on Edge Systems, Analytics and Networking*, EdgeSys '21, (New York, NY, USA), p. 13–18, Association for Computing Machinery, 2021.

[6] K. R. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big Data*, vol. 3, pp. 1–40, 2016.

[7] S. Solorio-Fernández, J. A. Carrasco-Ochoa, and J. F. Martínez-Trinidad, "A review of unsupervised feature selection methods," *Artificial Intelligence Review*, vol. 53, pp. 907–948, 2019.

[8] M. E. F. G. Sanz and A. Johnsson, "Exploring approaches for heterogeneous transfer learning in edge clouds," in *IEEE/IFIP Network Operations and Management Symposium (NOMS)*, 2022.

[9] H. Larsson, J. Taghia, F. Moradi, and A. Johnsson, "Source selection in transfer learning for improved service performance predictions," in *2021 IFIP Networking Conference and Workshops*, 2021.

[10] F. Moradi, R. Stadler, and A. Johnsson, "Performance prediction in dynamic clouds using transfer learning," *2019 IFIP/IEEE Symposium on Integrated Network and Service Management (IM)*, 2019.

[11] X. Wang, F. S. Samani, A. Johnsson, and R. Stadler, "Online feature selection for low-overhead learning in networked systems," in *2021 17th International Conference on Network and Service Management (CNSM)*, pp. 527–529, IEEE, 2021.

[12] X. Hu, P. Zhou, P. Li, J. Wang, and X. Wu, "A survey on online feature selection with streaming features," *Frontiers of Computer Science*, vol. 12, pp. 479–493, 2016.

[13] J. Taghia, F. Moradi, H. Larsson, X. Lan, M. Ebrahimi, and A. Johnsson, "Policy-induced unsupervised feature selection: A networking case study," in *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*, 2022.

[14] W. Y. B. L. et. al, "Federated learning in mobile edge networks: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 22, pp. 2031–2063, 2020.

[15] Q. Xia, W. Ye, Z. Tao, J. Wu, and Q. Li, "A survey of federated learning for edge computing: Research problems and solutions," *High-Confidence Computing*, 2021.

[16] A. A. Barakabitze, A. Ahmad, R. Mijumbi, and A. Hines, "5G network slicing using SDN and NFV: A survey of taxonomy, architectures and future challenges," *Computer Networks*, vol. 167, p. 106984, 2020.

[17] N. Shahriar, S. Taeb, S. R. Chowdhury, M. Zulfiqar, M. Tornatore, R. Boutaba, J. Mitra, and M. Hemmati, "Reliable slicing of 5G transport networks with bandwidth squeezing and multi-path provisioning," *IEEE Transactions on Network and Service Management*, vol. 17, no. 3, 2020.

[18] A. Marotta, D. Cassioli, M. Tornatore, Y. Hirota, Y. Awaji, and B. Mukherjee, "Reliable slicing with isolation in optical metro-aggregation networks," in *2020 Optical Fiber Communications Conference and Exhibition (OFC)*, pp. 1–3, 2020.

[19] B. M. Khorsandi, F. Tonini, and C. Raffaelli, "Centralized vs. distributed algorithms for resilient 5G access networks," *Photonic Network Communications*, vol. 37, pp. 376–387, Jun 2019.

[20] H. D. Chantre and N. L. Saldanha da Fonseca, "The location problem for the provisioning of protected slices in NFV-based MEC infrastructure," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 7, pp. 1505–1514, 2020.

[21] T. L. F. Project, "State of the edge 2021: A market and ecosystem report for edge computing," whitepaper, 2021.

[22] E. Amato, F. Tonini, C. Raffaelli, and P. Monti, "A resource sharing method for reliable slice as a service provisioning in 5G metro networks," in *2021 International Conference on Optical Network Design and Modeling (ONDM)*, pp. 1–3, 2021.