A Flexible Forecasting Platform Enabling Zero Touch Networking and Digital Twinning

Luca Valcarenghi *, *Senior Member, IEEE, Piero Castoldi *, Senior Member, IEEE, Andrea Sgambelluri *, Emilio Paolini *†‡, Student Member, IEEE, Alessandro Pacini *, Student Member IEEE*

** TeCIP Institute, Scuola Superiore Sant'Anna, Via G. Moruzzi 1, Pisa, 56124, Italy*

† National Research Council of Italy, CNR-IEIIT, Pisa, 56122, Italy

‡ Sma-RTy Italia SRL, Carugate, 20061, Italy

e-mail: luca.valcarenghi@santannapisa.it

ABSTRACT

A proactive approach could bring several advantages in network operations to meet the increasingly strict requirements (e.g., in terms of latency, reliability) of emerging applications (e.g., XR-VR, xURLLC). This paper describes a flexible forecasting platform that provides estimates of the performance parameters involved in the Zero Touch Networking closed loop and Digital Twin. The advantages of using parameter estimation within the closed loop in reducing performance violations are also shown.

Keywords: Zero Touch Networking, Forecasting platform, Artificial Intelligence, Digital Twin.

1. INTRODUCTION

In general, current networks adopt a reactive approach to changes, where actions are performed after an event (e.g., soft failures, traffic pattern changes) detection or planning is based on overprovisioning. This paradigm does not allow to either meet the strict requirements (i.e., latency, reliability) of newly emerging applications, such as XR-VR and xURLLC [1], or effectively leverage available resources. Hence, to deal with these new requirements, a paradigm shift must be adopted.

In this scenario, a Zero Touch Networking [2] approach leveraging not only data monitored and directly collected from network nodes about past and current network states but also forecast data, might implement a proactive approach that can fulfil network autonomy. Indeed, thanks to observation and analytics made directly on networking data, scalability and fault recovery can be performed in a proactive and autonomous way. This approach will allow future networks to comply with newly emerging requirements, where a high level of automation in network operations is needed, with minimal (if any) human intervention.

In this paper we propose a forecasting platform, capable of providing estimates of key parameters involved in closed loop (e.g., Channel Quality Indicator (CQI)). Specifically, the developed forecasting block leverages both current and past measurement data to predict future values to implement a proactive approach.

We experimentally demonstrate the accuracy of the developed forecasting in several real-world scenarios, including a CQI forecasting and an orchestration and monitoring operations of 5G Network Services (NS) scenario.

2. RELATED WORKS

The increasing adoption of Zero Touch Networking paradigm has highlighted the needs for predictions of network parameters and the possibility to anticipate specific events. Several works discussing the use of Artificial Intelligence (AI)-based forecasting operations in the context of network management and operations have been proposed already.

For instance, in [3], the authors propose a solution to integrate time-series based predictive analytics in 5G Core Network and discuss a comparative study between two Time Series Forecasting Models: AutoRegressive Integrated Moving Average (ARIMA) and Facebook Prophet. One advantage of Prophet is related to the fact that it works considering trend, seasonality, and holiday effects on the data. However, in the comparison, it is highlighted that Prophet does not work well on highly irregular datasets. On the other hand, ARIMA models can work on both stationary and non-stationary data. In the context of Zero Touch Networking, in [4], an orchestration approach for network slicing is proposed. The discussed framework combines deep learning tools with classic optimization algorithms to provide a zero-touch anticipatory capacity forecasting for individual slices.

Another interesting work [5] proposes a deep learning-based forecasting to scale 5G core network resources by anticipating traffic load changes. The authors demonstrate that the forecast-based scalability outperforms the threshold-based solutions, in terms of latency to react to traffic changes.

Compared to the reported implementations, the forecasting platform proposed in this paper provides a higher level of flexibility, being capable of easily adapting to new use-cases.

3. PROPOSED FORECASTING SOLUTION

In this section the architecture and the details of the proposed forecasting platform (FP) are highlighted. The FP, shown in Figure 1, is composed by two main modules: (i) the forecasting engine and (ii) the data collector. The forecasting engine exposes a REST API for the submission of forecasting request. Each request includes all the data required to activate a forecasting job. The input parameters include the metric to be forecast, the service type, the device type, and the Kafka topic to be used to retrieve monitoring data.

The data collector is the module responsible for the allocation of Kafka Collector instances, to retrieve data from the monitoring systems. The module has been designed to allocate multiple instances, reading from different Kafka topic in parallel, according to the forecasting jobs allocated.

According to the input parameters (e.g., the metric, the device type, and the service type), the FP performs the model selection, searching among the known model types, with the related parameters (e.g., epochs, history window, steps ahead, input features to be collected). Then, if an already trained model is available, it deploys a forecasting job for the request, otherwise it activates a training phase for the job.

Each forecasting job is equipped with an instance of Kafka Consumer [6], to retrieve all the required data from the monitoring system, reading form the Kafka topic included in the request. The data includes the main feature (f) and the additional input features required by the considered model.

Considering the execution of a forecasting job at time *t*, Figure 2 shows the data structures in input and output manipulated by a forecasting job. In particular, each forecasting job receives in input a matrix of data. Each raw of the matrix represents one observation of the samples retrieved from the monitoring system, including one instance of all the values to be fed to the forecasting model. Each raw has a size of N and includes the main feature (the one to be forecast), labelled with *f* in the figure, and the additional input features required by the model (i.e., *f1*, *f2*, *fN-1*). The forecasting job temporarily stores last *K* observations (i.e., from *t* to *t-K*) and generates one forecasting value (f_f) referred to *X* steps ahead *t+X*.

The forecasting values are provided to external system, acting as a regular probe, by acting as Kafka producer, pushing the data to a specific forecasting topic (topic fx). In this way, starting from the retrieved data, the forecasting platform exposes forecast data to the decisional elements to take a proactive approach in reacting to change.

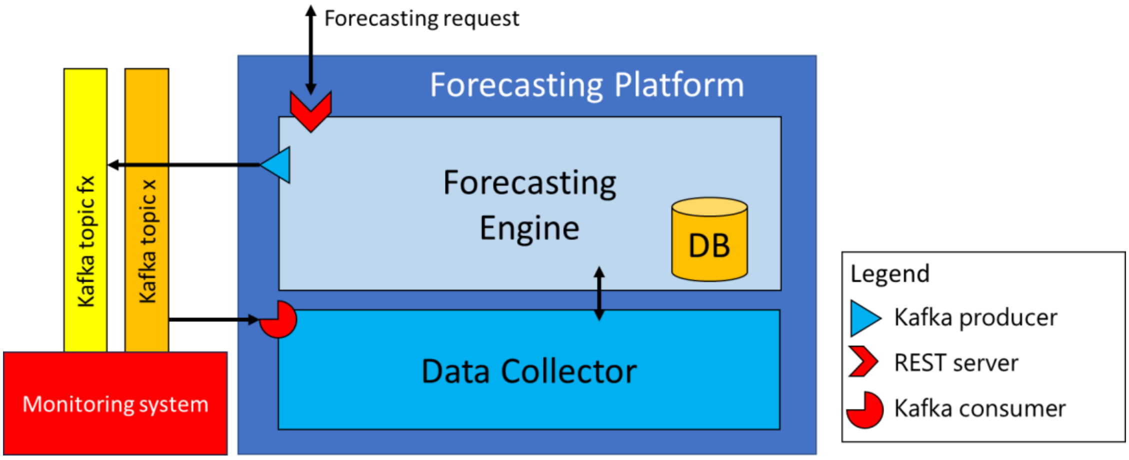


Figure 1. Forecasting platform architecture and interfaces.

Concerning forecasting operations, different model architectures have been considered, to optimize the forecasting accuracy, at the varying of the model. Available models include Long-Short Term-Memory (LSTM) [7], Gated-Recurrent Unit (GRU) [8], and Convolutional Neural Networks [9]. These models are optimized according to the input features and their manipulation (e.g., data preprocessing, normalization, filtering). The models can be trained using different optimizers, such as Adam [10], and at the varying of the number of epochs. Considering the case shown in Figure 2, given *N* features and targeting the forecasting of the main feature *f*, based on the past *K* observations, then the forecasting model will have as input *K x N* matrix, where each row represents a past observation, up to the current one. The output of the forecasting job will produce the forecasting value of feature *f (i.e.,* f_f*) X* steps ahead in the future. During the training phase, the loss of the model is evaluated by feeding the system the input matrix and comparing the real value of *f(t+X)* with the forecasting value at the same time $f_f(t+X)$. Then, the model parameters are adjusted consequently, by considering the loss produced.

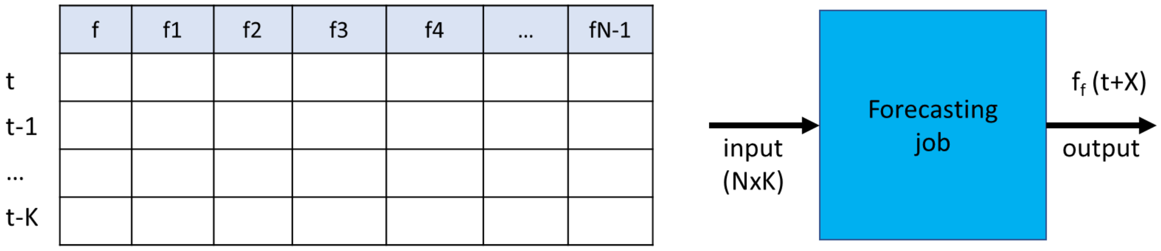


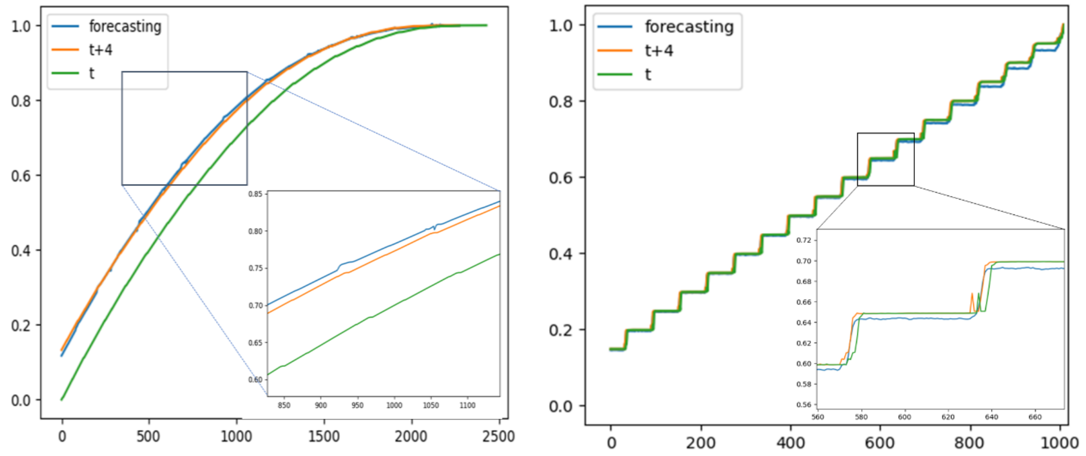
Figure 2. Forecasting data structures.

4. USE-CASES AND EXPERIMENTAL VALIDATION

The proposed solution has been evaluated considering two main scenarios: (i) a robotic Digital Twin use-case and (ii) a CQI forecasting in a Software Defined (SD)-Radio Access Network (RAN) scenario. In the following of this section, these two scenarios are discussed, highlighting the results obtained by the forecasting platform.

4.1 Robotic Digital-Twin

This use-case concerns the forecasting of metrics to efficiently enable the automatic scaling operations to meet a certain Service Level Agreement (SLA). Specifically, the scenario in which the forecasting platform has been tested involves the scaling in/out of an NFV-Network Services (NFV-NS). The forecasting operations concerned two metrics, one related to the CPU usage and the other related to memory occupancy. For both metrics, a 4-step ahead forecasting has been performed. As depicted in Figure 3-a, the forecast CPU (blue line) follows almost exactly the actual t+4 CPU value (orange line), with slight detours. Furthermore, in Figure 3-b, the forecast memory is depicted against the actual t+4 memory. As it can be observed, the memory occupancy is estimated with a very high degree of accuracy with respect to the actual value, with very few distortions in the forecast value.



(a) CPU metric.

(b) Memory metric.

Figure 3. Forecasting results on the Robotic Digital-Twin scenario.

4.2 Channel Quality Indicator

In this second use-case, the forecasting platform is validated to perform prediction on the CQI of a User Equipment (UE) in a 5G SD-RAN. First, a dataset has been collected to train the forecasting platform, considering a sampling inter-arrival time of 1s. Then, a training phase was performed considering as input the t-100 samples in the past and as output the t+4 CQI value. Once this phase has been accomplished, forecasting operations have been performed working on real-world data.

Figure 4 compares values for t (in green), forecast t+4 (in blue) and the actual t+4 (in orange) CQI value. Comparing the actual and the forecast value, we can observe a Mean Squared Logarithmic Error (MLSE) of 0.074. If we observe the forecast and the actual values, we can see that the two lines follow the same trend, with the forecast metrics more prone to fluctuations. However, this does not hamper the exploitation of the forecasting platform in this scenario, since for this use-case, we are mainly interested in trend forecasting and not in the actual value. Indeed, the knowledge on the trend can be used to perform specific actions, such as proactive handover to prevent CQI degradation.

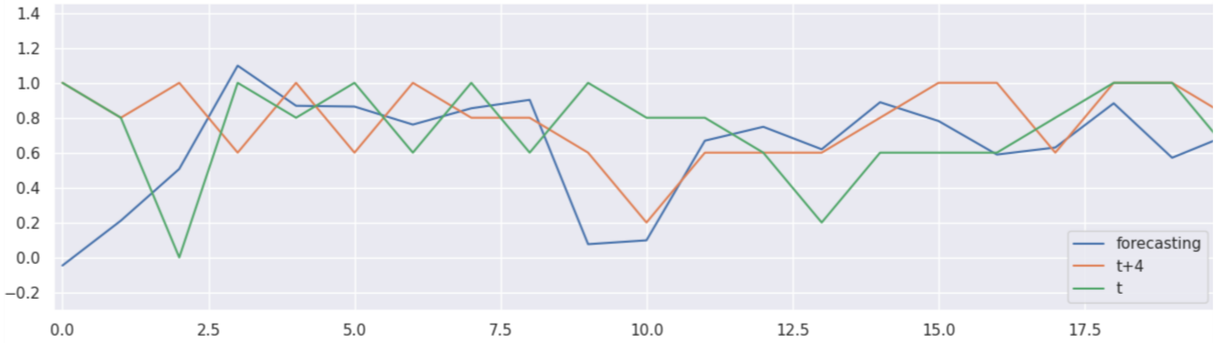


Figure 4. Forecasting results on CQI data.

5. CONCLUSIONS

Forecasting operations are becoming of primary importance in the context of Zero Touch Networking, since they can enable a proactive network management. Hence, in this paper, we have proposed a novel forecasting platform that can be easily implemented in many use-cases to perform predictions on many networking metrics.

We have experimentally validated the proposed forecasting platform in two scenarios, one concerning a robotic Digital Twin and the other related to CQI prediction. In both use-cases the forecasting platform has proved to be effective in providing accurate forecasts, paving the way for being implemented in future networks to enable proactive approaches.

Finally, this works is just a first step towards a real proactive network operation and management. Indeed, the forecasting platform will be enhanced to consume data from more than one data stream, therefore able to forecast more than one input parameter at the same time. Furthermore, an online model retraining through Neural Architecture Search (NAS) will be considered, capable of automatically optimizing forecasting architectures.

ACKNOWLEDGEMENTS

This work was partly supported by the project CLEVER (Project ID 101097560), that is supported by the Key Digital Technologies Joint Undertaking and its members (including top-up funding by Italian Ministry of University and Research --- MUR), and partly by the European Union under the Italian National Recovery and Resilience Plan of NextGenerationEU, partnership on "Telecommunications of the Future" (PE00000001 - program "RESTART")."

REFERENCES

[1] "ETSI Zero touch network & Service Management (ZSM)." [Online]. Available: https://www.etsi.org/technologies/zero-touch-network-service-management

[2] Du, Hongyang, et al. Attention-aware resource allocation and QoE analysis for metaverse xURLLC services. *IEEE Journal on Selected Areas in Communications*, 2023.

[3] Chakraborty, Pousali; Corici, Marius; Magedanz, Thomas. A comparative study for Time Series Forecasting within software 5G networks. In: *2020 14th International Conference on Signal Processing and Communication Systems (ICSPCS)*. IEEE, 2020. p. 1-7.

[4] Bega, Dario, et al. AZTEC: Anticipatory capacity allocation for zero-touch network slicing. In: *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*. IEEE, 2020. p. 794-803.

[5] Alawe, Imad, et al. Improving traffic forecasting for 5G core network scalability: A machine learning approach. *IEEE Network*, 2018, 32.6: 42-49.

[6] Garg, Nishant. *Apache kafka*. Birmingham, UK: Packt Publishing, 2013.

[7] Hochreiter, Sepp; Schmidhuber, Jürgen. Long short-term memory. *Neural computation*, 1997, 9.8: 1735-1780.

[8] Dey, Rahul; Salem, Fathi M. Gate-variants of gated recurrent unit (GRU) neural networks. In: *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*. IEEE, 2017. p. 1597-1600.

[9] Gu, Jiuxiang, et al. Recent advances in convolutional neural networks. *Pattern recognition*, 2018, 77: 354-377.

[10] Zhang, Zijun. Improved adam optimizer for deep neural networks. In: *2018 IEEE/ACM 26th international symposium on quality of service (IWQoS)*. Ieee, 2018. p. 1-2.