Supporting Beyond 5G Applications by Coordinating AI-based Intent Operation. An Example for Multilayer Metro Networks

Fatemehsadat Tabatabaeimehr, Marc Ruiz and Luis Velasco*

Optical Communications Groups, Universitat Politècnica de Catalunya, Spain, lvelasco@ac.upc.edu

Abstract *Intent-based Networking (IBN) promises facilitating autonomous decision-making for service assurance. In this paper, we extend IBN by coordinating AI-based intents targeting at supporting beyond 5G services, like immersive and Industry 4.0 applications, on metro infrastructures.*

Introduction

5G and beyond applications will transform current industries and create new ones. However, for this to happen, the network needs to be much more flexible and automated to allow anticipating future events and conditions. Artificial Intelligence (AI) -empowered Intent-based Networking (IBN) [1] simplifies network operation and provides the ideal framework for network automation[2]. Although in this paper we focus on the transport network, the conclusions are applicable to any segment; in fact, the network and the associated computing platform should behave as one single end-to-end entity, including radio and other access technologies, metro and core networks[3], as well as edge and cloud computing, to take full advantage of resources, wherever they are available, to provide the required Quality of Service (QoS) and Experience (QoE)[4].

In this paper, we propose coordinating intents by transferring knowledge[5] among them to be able to anticipate to ongoing events that however, cannot be predicted. The scenario that we focus on includes the autonomic operation of different beyond 5G applications, e.g., from industry 4.0 which requires time sensitive network (TSN) communications, sharing the same infrastructure with applications with different requirements. In that regard, authors in [6] showed that mixing best effort (BE) and TSN traffic might result in noticeable degradation of the performance of the former, which deserves special attention to assure BE QoS.

Coordinating Intent operation

Fig. 1 presents an example of a multilayer metro network with two customer and two infrastructure intents each operating independently. Intent *CI-A* is in charge of Customer A TSN connection between two factory networks in sites 1 and 2 (A1-A2) and *CI-B* operates on a service for a drone (mobile) application that captures video and requires high bitrate with low latency to a metro datacenter. *II-1* manages the virtual link (vlink) R1-R3 and *II-2* operates on vlink R2-R3; vlinks are supported by one or more lightpaths in the optical layer.

Let us imagine that customer intents CI-A and CI-B know about real service needs and they can demand for service reconfiguration, which would entails not only managing the connection capacity but also creating and releasing connections. In the example, CI-A can request additional capacity for connection A1-A2 during production peak hours and reduce such capacity when it is not more needed, and such changes can be as a consequence of factory management decisions. In addition, CI-B is aware of the geoposition of the drone and creates new connections to the metro data center as the drone moves.

Regarding infrastructure intents II-1 and II-2, they know about the services that they are supporting and their required performance. They can monitor the incoming traffic to model them and try to predict future variations, which is used to manage the capacity of the vlink (by adding or removing lightpaths). However, they cannot predict application decisions.

In the example in Fig. 1a, the factory management could make a decision to increase the production, which would require more capacity in the TSN connection A1-A2. Regarding the drone application, it moves from

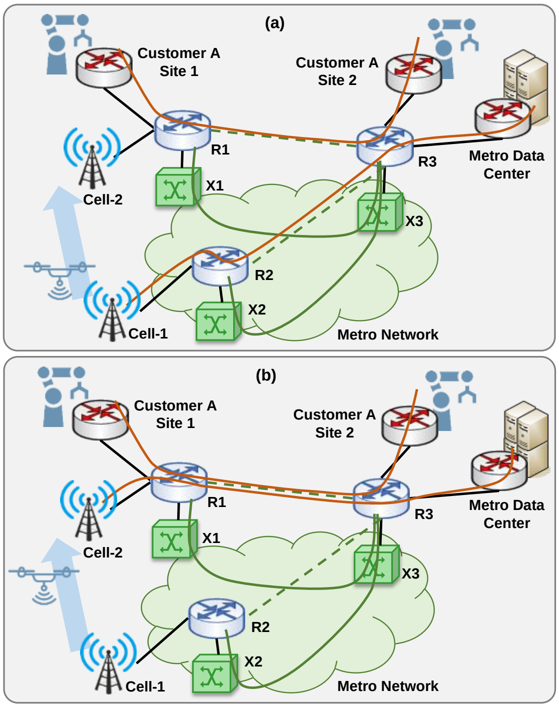

Fig. 1: Example of intent-triggered reconfiguration

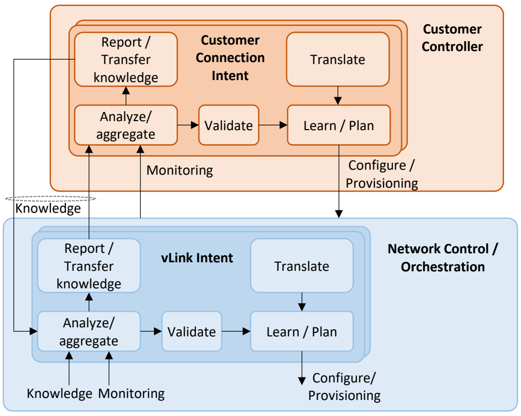

Fig. 2: Possible architecture supporting intent coordination

cell-1 in Fig. 1a to cell-2 in Fig. 1b, and so the traffic to the data center, as the intent CI-B manages the connectivity. Note that both customer connections in Fig. 1b use the same resources, which could result into poor service performance and even degrade the QoE of the drone application to an intolerable extent. Intent II-1 will realize of the poor performance by analyzing monitoring data and then it can create a parallel lightpath to increase the capacity of vlink R1-R3, which might take about 1 min.

Our solution for smooth operation is that customer and infrastructure intents coordinate among them. For instance, the intent in charge of connection A1-A2 can anticipate an increment of traffic in that connection, and the intent in charge of the connectivity for the drone can anticipate the need of a new connection from cell-2 to the metro data center, both with enough anticipation for the infrastructure intent in charge of vlink R1-R3 to react and increase the capacity or, on the contrary, reject the request if no resources are available.

Fig. 2 presents a possible architecture that supports a hierarchy of intents (the optical layer is omitted for the sake of simplicity), where the blocks in the intents are in line with [1]. Customer intents run in the customer / application controller, whereas infrastructure intents run in the network control and orchestration layer. Interfaces for provisioning and configuration are available at the network and orchestration layer for the customer controller to request new customer connections and to reconfigure them and monitoring data is collected by the network and orchestration layer and exported to the customer intents. In addition, knowledge is transferred bottom-up and top-down among intents to support coordination. Note that other possible option is to integrate customer intents in the network control and orchestration to facilitate data and knowledge exchange.

The next section focuses on the analyze and aggregate block of the vlink intent in Fig. 2 and presents a procedure that integrates the knowledge shared by customer connection intents for the prediction of vlink traffic.

Table 1: vlink traffic prediction with knowledge sharing

```
Input: $\hat{X}^C$, $x_t$, predRequest | Output: $\hat{x}_{t+\delta}$
1:  $Q$←getKnowledgeDB()
2:  $X$←getTrafficMonitoringDB()
3:  $f$←getVlinkTrafficModel()
4:  if $\hat{X}^C$ != ∅ then
5:    for each $<\hat{x}^i_{t+k}, p^i_{t+k}> \in \hat{X}^C$ do
6:      updateKnowledgeDB($Q(i)$, $<\hat{x}^i_{t+k}, p^i_{t+k}>$)
7:  if $x_t$ != ∅ then
8:    updateMonitoringDB($X$, $x_t$)
9:    validateTrafficModel($X$, $f$)
10: if !predRequest then return ∅
11: $<\hat{x}^C_{t+\delta}, p^C_{t+\delta}>$←predictFromKnowledge($Q$, $\delta$)
12: $<\hat{x}^V_{t+\delta}, p^V_{t+\delta}>$←predictFromModel($f$, $\delta$)
13: return ensemble($<\hat{x}^C_{t+\delta}, p^C_{t+\delta}>$,$<\hat{x}^V_{t+\delta}, p^V_{t+\delta}>$)
```

Analyze / Aggregate module for vlink intent

The procedure in Table 1 presents the main procedure that is executed either when: *i*) a new vlink traffic monitoring sample x_t or customer connection traffic prediction $\hat{X}^C$ (knowledge) becomes available; or *ii*) a vlink traffic prediction request is needed. It is worth highlighting that autonomic vlink management requires computing traffic prediction with enough anticipation to ensure that optical connections can be setup (if needed). Such prediction needs thus to be performed for a δ time window (e.g., if 1 min is required for setting up an optical connection, δ can be set to 2 min).

The procedure is in charge of updating the repositories with knowledge shared by intents (Q) and monitoring traffic (X). For the sake of simplicity, we can assume that both repositories centralize data for all the vlinks in the network. In addition, the vlink intent manages its own traffic prediction model (f). Lines 1-3 in the algorithm in Table 1 get access to the repositories and the traffic prediction model.

Since knowledge and monitoring might have in general different granularity, they are managed separately and specific models for each one needs to be obtained. When new knowledge $\hat{X}^C$ is provided in current time t, the block updates Q (lines 4-6). Every element in $\hat{X}^C$ contains traffic prediction of customer connection i for the next time interval [t, t+k], as well as a fitness score p that indicates how likely is the traffic prediction.

When a new monitoring sample x_t is provided, the monitoring repository X is updated, and the traffic model is evaluated to detect whether it needs to be retrained or some adjustment is needed (lines 7-9).

When a new prediction is requested (line 10), two different traffic estimations are computed. First, individual predictions for each customer connection are obtained from data in Q. To this aim, polynomial interpolation is used to obtain the estimation and score at exact time t+δ. The sum of all individual connection estimations and the most restrictive (smallest) score is the

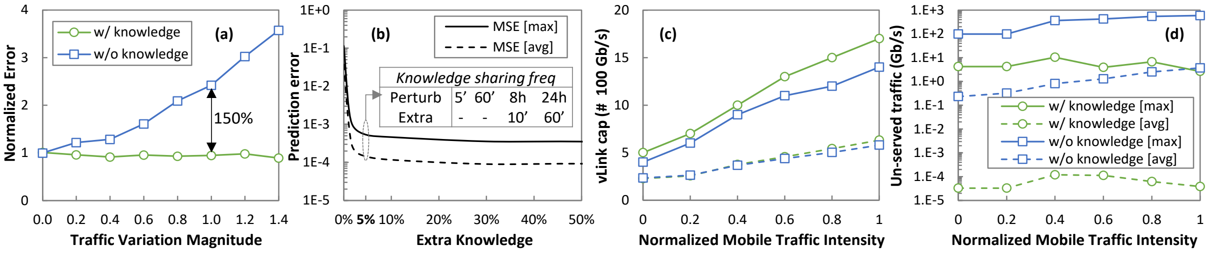

Fig. 3: Accuracy and vlink dimensioning performance results

result of knowledge-based prediction (line 11). Second, the prediction with the aggregated vlink model is computed (line 12). Both predictions are ensembled to return a single traffic estimation (line 13); in the ensemble, the estimation with highest score is selected.

Illustrative Results

Let us evaluate the benefits of implementing knowledge sharing between customer and infrastructure intents, as compared to a baseline IBN approach without such knowledge sharing, where traffic prediction is based on a vlink traffic model trained with the observed traffic.

For the numerical evaluation, scenarios reproducing that depicted in Fig. 1 have been generated by means of a customer connection traffic generator based on the models and methodology in[6] to emulate industry 4.0 and immersive mobile applications. The vlink traffic prediction model was based on artificial neural network (ANN) with 10 hidden neurons and a sigmoid activation function that uses short-term traffic characterization features, in line with [5]. The ANN achieves accurate prediction for δ=2 minutes (max error <5%) for aggregated traffic. Sudden traffic changes (*perturbations*) due to customer operation were introduced using a Markov-based model and a multiplicative factor to increase or decrease the generated traffic.

Fig. 3a shows the maximum prediction error of both approaches normalized w.r.t. that when no perturbations are added vs. perturbation magnitude (computed as the relative traffic variation introduced). Let us assume that both traffic monitoring data and knowledge sharing are received with a frequency of 1 per minute. We observe that sharing knowledge provides virtually zero added error regardless of perturbation magnitude, in contrast to no sharing knowledge, which presents large inaccuracy as soon as perturbation magnitude increases (150% of added prediction error for perturbations introducing 100% traffic variation).

Assuming such high frequency for knowledge sharing can be not realistic. In fact, knowledge could be shared only to anticipate large traffic changes. In such case, the accuracy of interpolation is critical to retrieve valuable knowledge for the desired prediction time. Fig. 3b shows the difference in terms of mean square error (MSE) between the prediction using all available knowledge and that when shared knowledge includes only a percentage of that traffic between perturbations (extra knowledge). The curves aggregate several cases, where consecutive perturbations were spaced from 5 minutes to 24 hours. We observe that negligible MSE (<0.001) is achieved when ~5% of extra data is shared. The embedded table shows that for frequent perturbations (from 5 to 60 min) no extra data is actually required, whereas frequencies of minutes are enough when perturbations are less frequent. In conclusion, knowledge sharing enormously improves traffic prediction and requires small amount of data to be shared between intents.

Let us now evaluate the impact of traffic predictions in terms of optical capacity resources utilization. We emulated the vlink intent II-1 operation in the scenario in Fig. 1b, where the vlink supports two types of flows: *i*) a fixed large industry 4.0 customer connection (400Gb/s average traffic) subject to large perturbations every 8 hours; and *ii*) a variable number of 100 Gb/s mobile connections that are dynamically routed through the vlink.

Fig. 3c shows the average and max capacity in terms of 100 Gb/s optical connections required to support the traffic vs. the intensity of mobile traffic (the larger the intensity is, the larger the number of dynamic mobile connections to be supported). In addition, Fig. 3d shows the amount of customer traffic that cannot be served as a result of unavailable capacity in the vlink. The results in Fig. 3c show that both approaches allocate similar amount of optical connections on average with higher peaks of capacity when knowledge is shared. Such extra capacity is needed to serve the requested traffic, as shown in Fig. 3d; in fact, unserved customer traffic drops several orders of magnitude (~2 in maximum and ~5 in average) when knowledge is shared.

Conclusions

In conclusion, IBN exploiting knowledge sharing between customer and infrastructure intents is a promising solution for network operators to support B5G applications.

Acknowledgements

The research leading to these results has received funding from the Spanish MINECO TWINS project (TEC2017-90097-R) and from ICREA.

References

[1] A. Clemm *et al*. (Eds.), “Intent-Based Networking - Concepts and Definitions,” IRTF draft work-in-progress, Mar. 2020.

[2] D. Rafique and L. Velasco, “Machine Learning for Optical Network Automation: Overview, Architecture and Applications,” (Invited Tutorial) IEEE/OSA Journal of Optical Communications and Networking (JOCN), vol. 10, pp. D126-D143, 2018.

[3] A. Bernal, M. Richart, M. Ruiz, A. Castro, and L. Velasco, “Near Real-Time Estimation of End-to-End Performance in Converged Fixed-Mobile Networks,” Elsevier Computer Communications, vol. 150, pp. 393-404, 2020.

[4] X. Jiang *et al.*, “Low-Latency Networking: Where Latency Lurks and How to Tame It,” Proceedings of the IEEE, vol. 107, pp 280-306, 2019.

[5] M. Ruiz, F. Tabatabaeimehr, and L. Velasco, “Knowledge Management in Optical Networks: Architecture, Methods and Use Cases [Invited],” IEEE/OSA Journal of Optical Communications and Networking, vol. 12, pp. A70-A81, 2020.

[6] L. Velasco and M. Ruiz, “Supporting Time-Sensitive and Best-Effort Traffic on a Common Metro Infrastructure,” in IEEE Communications Letters, 2020.