Autonomous Capacity Adjustment with Dynamic Margin Allocation for Optical Enterprise Links

Mihail Balanici, Behnam Shariati, Pooyan Safari, Geronimo Bergk*, Johannes Karl Fischer
Fraunhofer Institute for Telecommunications, Heinrich-Hertz-Institut, Einsteinufer 37, 10587 Berlin, Germany
Author e-mail address: mihail.balanici@hhi.fraunhofer.de

Abstract: This work presents a novel machine learning-based dynamic capacity allocation scheme for efficient bandwidth provisioning of optical links. It offers an average hourly capacity saving of over 75% compared to traditional static capacity allocation mechanisms. © 2024 The Author(s)

1. Introduction

The large number and variety of emerging online applications, along with the resulting increase of user- and machine-generated traffic, lead to a high necessity for autonomous optical networks to support greater dynamics and flexibility in network resource allocation [1]. In traditional bandwidth allocation scenarios, the provisioned port capacity remains constant over the entire link operation lifecycle, meaning that a huge portion of the available throughput remains unused at all times (Fig. 1). As opposed to this, the dynamic link-capacity adjustment offers a higher flexibility in the network resource allocation, allowing for several benefits, such as: an enhanced network operation efficiency by maximizing the utilization of available network resources; higher bandwidth provisioning on demand; and a reduction of energy consumption costs of the optical network, among others. A promising way to accomplish the task of autonomous and dynamic link-capacity allocation is by means of efficient forecasting of traffic behavior [2,3], related to its intensity, variability, and burstiness over time [4]. In this regard, a few machine learning (ML) algorithms for network traffic prediction have been previously proposed [5-7]. One common aspect peculiar to all these works is that the investigated traffic is often synthetic in nature [6,8], in some cases with very low or missing burstiness, lack of traffic peaks and outliers specific to real traffic flows. Moreover, even with collected real traffic samples, these are typically aggregated at 1h intervals or longer, with the task being to predict the mean traffic load in the next hour [5,7]. Besides the resulting smoothening of the aggregated traffic envelope, the prediction of mean traffic does not suit the task of capacity allocation, since the provisioned bandwidth has to fit the traffic peaks as well.

To address these limitations, we present a novel link-capacity adjustment scheme using the so-called dynamic capacity margin allocation (DCMA) that relies on traffic forecasting using LSTM-NNs. The technique adds dynamic safety margins based on the predicted variability of the traffic flows. Secondly, we show how this approach minimizes over-provisioning of real, fine-granular and bursty traffic flows, and quantify the resulting capacity gains.

2. Traffic Forecasting-based Adaptive Link-Capacity Adjustment

While network traffic is to a larger extent predictable when it comes to its periodicity and seasonality [9], forecasting its precise behavior at fine time granularities, is a challenge. Moreover, taking into account the uncertainty of its prediction (which primarily depends on the underlying traffic burstiness), it becomes utterly important to consider the traffic's short-term variability in order to allocate enough capacity to minimize the undesired over-provisioning, while simultaneously preventing under-provisioning. The traffic forecasting-based capacity adjustment schemes presented in the literature, sometimes consider a *static* safety margin to mitigate the risk of under-provisioning [8]. However, this approach eventually leads to suboptimal performance, since the variability and burstiness of traffic flows change quite significantly over time [4], and a traffic outlier, as we also show in the following, will certainly remain under-provisioned. Hence, a flexible, i.e., *dynamic* capacity margin allocation has more potential to solve this limitation. The performance of such an adaptive scheme depends on a couple of factors and parameters, presented in Fig. 2. In this work, we take these evaluation factors and metrics into account to present a concrete use-case of adaptive link-capacity adjustment. An important detail worth noting is that since the allocated capacity is set up step-wise and is fixed within

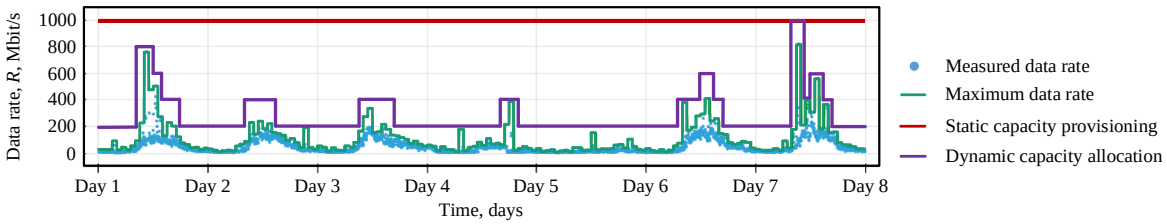

Fig. 1. Traditional static vs. dynamic capacity allocation schemes. Notice the total amount of unutilized bandwidth resources over a period of one week on one of Fraunhofer HHI's optical enterprise links.

* Geronimo Bergk is now with Horváth AG, Rotebühlstraße 100, 70178 Stuttgart, Germany.

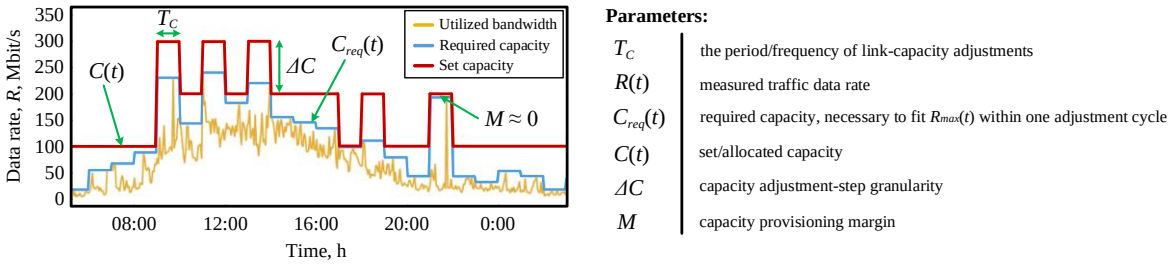

Fig. 2. Adaptive link-capacity adjustment and its corresponding parameters.

the allocation cycle/period, T_C (Fig. 2), and is not following the traffic envelope, a certain amount of over-provisioning will always be present.

3. Feature Engineering and Traffic Prediction

In order to schedule and effectively allocate the link-capacity capable to support the upcoming traffic flows, the future traffic behavior has to be predicted based on its past and current values. As such, the traffic forecasting represents the first phase (P1) of the capacity adjustment cycle followed by the dynamic margin allocation (P2) to compensate for the prediction under-estimates. Next, a new capacity level is selected and scheduled for the next cycle (P3), and eventually set at the end of the current cycle (P4). To choose the most feasible capacity adjustment period, T_C, which directly depends on the prediction frequency, a few factors have to be considered, such as the amount of total available measured traffic data, and the fact that longer T_C results in higher over-provisioning (Fig. 2). Given these considerations, a capacity adjustment period of T_C = 1h was one of the most feasible in terms of the underlying over-provisioning ($\sum_{t\in T_c} C(t) - R(t)$) and mean capacity utilization ($\sum_{t\in T_c} R(t)/C(t)$).

The traffic prediction is carried out on a real traffic data set (Fig. 3) collected from a local Fraunhofer HHI's optical P2P enterprise interface consisting of data rate variation over time, $R(t)$. The collected traffic traces represent business customer traffic sampled at intervals of about 3 min, measured in Mbit/s (Fig. 3) and labelled with a timestamp, resulting in 6-weeks' worth of traffic, i.e., 20160 samples overall. The dataset was split into training (4 weeks), validation (1 week), and test (1 week) subsets, respectively. In order to schedule the capacity for 1h ahead, the max traffic rate within each hour has to be accurately predicted. However, forecasting the future maxima based solely on the previous ones, when the traffic variance is large, is a highly challenging task. Therefore, to account for the high traffic burstiness, it is also essential to predict the future traffic variability. To accomplish this task, the traffic is decomposed into two parallel feature series of R_{max} and δ_{max} representing the hourly maximum traffic rate and hourly maximum difference between the consecutive traffic samples, calculated for each time window of 1h using the original 3 min-granular (20/hour) traffic samples. Two similar LSTM-NNs are then trained in parallel on these two feature subsets, followed by the actual $\tilde{R}_{max}(t_{i+1})$ and $\tilde{\delta}_{max}(t_{i+1})$ predictions (Fig. 3). Based on these forecasted values, the next hour capacity, $C(t_{i+1})$ is computed applying the equation presented in Fig. 3. Worth noting is that the two LSTMs have one hidden layer and one output, and are trained using 48 consecutive 1h-maxima fed into its input, while applying a sliding window over the training set. Also, a capacity adjustment-step granularity of ΔC = 50 Mbit/s was considered for demonstration purposes, but is a parameter primarily dependent on the hardware capabilities of the optical link modules, such as transceivers or transponders. Last but not least, an anticipation time of τ = 15 min has also been considered, meaning that 15 min before the next hour starts, the LSTMs predict the two expected maxima

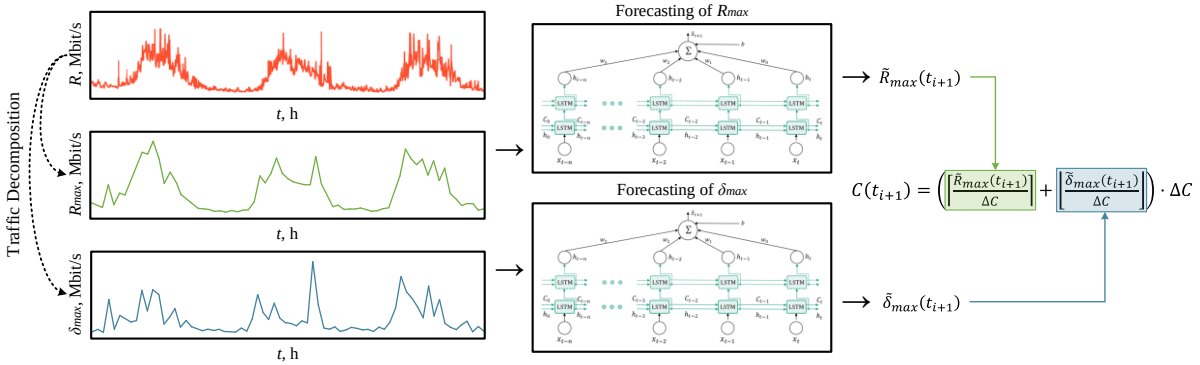

Fig. 3. Feature engineering by applying traffic decomposition into hourly R_{max} and δ_{max} with the subsequent features' prediction (using 2 similar LSTM-NNs), followed by the computation of the next-hour link-capacity.

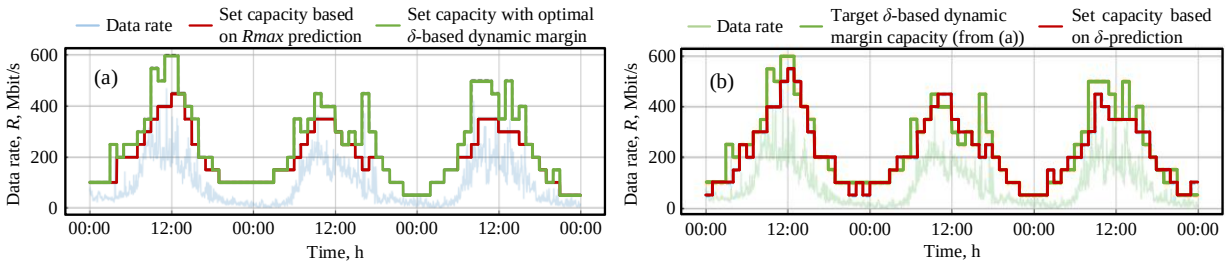

Fig. 4. (a) Optimal/theoretical capacity with δ-based margin. (b) Experimental DCMA based on forecasted δ_{max}.

(R_{max} and δ_{max}) of that hour, based on the maxima observed in the past 47 h and 45 min. The value of τ is chosen heuristically relative to T_C, and is accounting for the reconfiguration times of the aforementioned optical link modules.

4. DCMA Performance and Results' Analysis

The forecasting performance is evaluated on the testing set by predicting the maximum data rate and traffic variability within the next capacity adjustment cycle of 1h (Fig. 3) as:

$$\tilde{R}_{max}(t_{i+1}) = f(\{R_{max}(t) \mid t_0 \le t < t_{i+1} - \tau\}), \text{ and } \tilde{\delta}_{max}(t_{i+1}) = f(\{\delta_{max}(t) \mid t_0 \le t < t_{i+1} - \tau\}) \quad (1)$$

where $f_{\mathrm{LSTM}}(\cdot)$ are the LSTM prediction functions, and t_0 – the earliest data rate observation. Thus, the next-cycle's set capacity becomes:

$$C(t_{i+1}) = \left(\left\lceil \frac{\tilde{R}_{max}(t_{i+1})}{\Delta C} \right\rceil \cdot \Delta C + M\right) \quad (2)$$

with $\lceil\cdot\rceil$ being the ceiling function. As mentioned earlier, in order to compensate for underestimates of the maximum data rate due to highly variable traffic bursts as much as possible, the DCMA adds a flexible/dynamic capacity margin M acc. to eq. (2) and Fig. 3, where recommended by the predictor, i.e., based on the predicted traffic variability:

$$M(t_{i+1}) = \left\lfloor \frac{\tilde{\delta}_{max}(t_{i+1})}{\Delta C} \right\rfloor \cdot \Delta C, \quad (3)$$

with $\lfloor\cdot\rfloor$ being the floor function. The results of DCMA application show that the δ_{max} prediction-based margin, M, matches the target, i.e., the optimal/theoretical margin (Fig. 4 (a), green curve) in over 97% of cases (Fig. 4 (b), red curve), yielding an average hourly capacity saving of over 77%. Nevertheless, two traffic outliers manage to "escape" both predictions ($\tilde{R}_{max}(t_{i+1})$ and $\tilde{\delta}_{max}(t_{i+1})$) though, leading to a low risk of under-provisioning of roughly 0.45%.

5. Conclusions

Applying DCMA in parallel with the main traffic forecasting for adaptive link-capacity allocation is a promising way to compensate for a large proportion of prediction under-estimates. Nevertheless, a small percentage of traffic outliers still represents a challenge to be accurately predicted, however, omitting them through synthetic data sets or averaging them out makes the provisioning solution less applicable for a real-world scenario.

6. Acknowledgements

This work was partly funded by the German Ministry of Education and Research (BMBF) in the framework of the project 6G-RIC (FKZ 16KISK020K).

7. References

[1] N. E. Frigui, T. Lemlouma, S. Gosselin, B. Radier, R. Le Meur and J.-M. Bonnin, "Dynamic Reallocation of SLA Parameters in Passive Optical Network Based on Clustering Analysis," *21st Conference on Innovations in Clouds, Internet and Networks*, pp. 1-8, 2018.

[2] M. Balanici, G. Bergk, P. Safari, B. Shariati, J. K. Fischer and R. Freund, "Demonstration of a Real-Time ML Pipeline for Traffic Forecasting in AI-Assisted F5G Optical Access Networks," *European Conf. on Optical Commun.*, Tu2.5, pp. 1–4, 2022.

[3] F. Musumeci, C. Rottondi, A. Nag, I. Macaluso, D. Zibar, M. Ruffini and M. Tornatore, "An overview on application of machine learning techniques in optical networks," *IEEE Commun. Surv. Tutorials 21*, pp. 1383-1408, 2019.

[4] M. Balanici and S. Pachnicke, "Classification and forecasting of real-time server traffic flows employing long short-term memory for hybrid E/O data center networks," *J. Opt. Commun. Netw. 13*, pp. 85-93, 2021.

[5] W. Jiang, "Internet traffic prediction with deep neural networks," *Internet Technology Letters 5*, pp. 1-6, 2021.

[6] S. K. Singh and A. Jukan, "Machine-Learning-Based Prediction for Resource (Re)allocation in Optical Data Center Networks," *J. Opt. Commun. Netw. 10*, pp. D12–D28, 2018.

[7] D. Andreoletti, S. Troia, F. Musumeci, S. Giordano, G. Maier and M. Tornatore, "Network Traffic Prediction based on Diffusion Convolutional Recurrent Neural Networks," *IEEE Conf. on Commun. Workshops INFOCOM*, pp. 246-251, 2019.

[8] L. Velasco, S. Barzegar, F. Tabatabaeimehr and M. Ruiz, "Intent-based networking and its application to optical networks [Invited Tutorial]," *J. Opt. Commun. Netw. 14*, pp. A11–A22, 2022.

[9] S. Troia, G. Sheng, R. Alvizu, G. Maier and A. Pattavina, "Identification of tidal-traffic patterns in metro-area mobile networks via Matrix Factorization based model," *IEEE Intern. Conf. on Pervasive Comput. and Commun. Workshops*, pp. 297-301, 2017.