High Quantile Traffic Forecasting in Backbone Networks for Adaptive Proactive Protection

Attila Dobai-Pataky, Lehel Csató, László Németh, Balázs Vass

***Abstract*—In this paper, we revisit adaptive proactive protection schemes for backbone networks to guarantee high service availability. We aim to predict available – spare – link capacities and to use them as protection bandwidth to meet the service availability requirements of each connection. With a reasonable discretization of time, the advent of highly scalable risk-aware routing algorithms could enable minimizing the lookahead at 2. In other words, we claim the spare bandwidth has to be predicted for only 2 time slots ahead. While a shorter horizon improves accuracy, traditional forecasting objectives centered on the conditional mean provide insufficient safety margins for high-reliability targets. To address this, we propose a hybrid forecasting framework that combines Temporal Convolutional Networks (TCNs) with Extreme Value Theory (EVT). The TCN captures complex temporal patterns in the traffic, while the Peaks Over Threshold (POT) method is used to characterize the extreme, stochastic bursts in the residuals. Evaluations on real-world traffic data from the Energy Sciences Network (ESnet) show that for target quantiles of $\tau \geq 0.99$, the hybrid TCN-POT model provides more stable and accurate high-quantile estimates than direct quantile-regression networks and empirical heuristics.**

***Index Terms*—time series prediction, extreme value theory, adaptive proactive protection**

I. Introduction

In today's cloud era, communication networks are topmost critical infrastructures [2]. The new mission-critical applications, such as telesurgery or stock market prediction, clearly demand high Quality of Service (QoS) of the underlying network infrastructure [3]. Today's networks are operated with approximately four-nines availability, that is, the services are available at least 99.99% of the time, which translates to a maximum of around 53 minutes downtime a year. Four-nines is often considered a reasonable cost-benefit ratio, however, the lack of highly reliable Internet has become a limiting factor for development. Current technology trends point toward the advent of the era of the Internet of Everything (IoE) with an enormous amount of data [4], where, e.g., augmented and virtual reality applications not only require high dependability, but are also extremely latency-sensitive. Thus, QoS enhancement to reach and surpass the requirements of ultra-reliable and low-latency communication (URLLC) is a must, where six-nines is a typical required availability [5].

There are two fundamentally distinct strategies to enhance connectivity in backbone networks. *Proactive* methods aim to prepare the system for potential failures so that, when a failure occurs, no internal network reconfiguration is necessary. This is accomplished by distributing the same information across multiple routes so that in the event of a failure, only the end node needs to respond. Conversely, *reactive* methods involve altering the network configuration once a failure has been detected. Note that while reactive strategies can conserve a considerable amount of bandwidth compared to proactive ones, they require immediate network reconfiguration after a failure occurs, which can be relatively slow in practice. This makes reactive approaches unsuitable for protecting latency-sensitive data flows. This leaves us with proactive protection schemes. Among these, for achieving the prescribed availabilities, current best practice is using solutions offering *dedicated protection*, due to their simplicity, robustness, and flexibility.

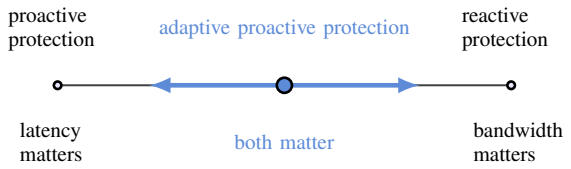


Fig. 1. Proactive protection schemes waste bandwidths, while reactive protection schemes are too slow for latency-sensitive applications. An ideal protection scheme re-introduces adaptiveness to a proactive approach by letting it prepare (react) to *predicted* bandwidth usages instead of peak rates.

Being stuck with a single static dedicated protection scheme, however, would waste a significant amount of bandwidth. Also, it could cause unnecessary bandwidth restrictions on the working paths (WPs). This is because the capacity of the WP is often over-provisioned for peak rates; however, a significant share of the bandwidths is unused most of the time [6]. To put this issue in a slightly different way, the drawback of using a single static dedicated protection scheme is the lack of *adaptability*. We argue that an ideal protection scheme would reintroduce adaptiveness to a proactive approach by letting it prepare (react) to *predicted* bandwidth usages instead of peak rates (see Fig. 1). Here the idea is that the proactive scheme could utilize the predicted spare bandwidth. The drawback of the approach is that if the bandwidth used by the actual traffic demands exceeds the predictions, some transient bandwidth limitations should be imposed.[1]

This issue should be seen as a trade-off between adaptability to changing traffic patterns and occasional bandwidth limi-

Attila Dobai-Pataky, Lehel Csató and Balázs Vass are with the Babeș-Bolyai University of Cluj Napoca, Romania. Balázs Vass is also with the Budapest University of Technology and Economics (BME). László Németh is with J. Selye University, Komárno, Slovakia. Email: attila.dobai@stud.ubbcluj.ro, lehel.csato@ubbcluj.ro, nemethl@ujs.sk, balazs.vass@ubbcluj.ro.

An earlier version of this paper appeared on IEEE RNDM 2025 [1].

[1]Transient bandwidth limitations may appear when applying reactive schemes too. Reactive schemes inevitably lose some packets when failure hits. In the case of ESnet, e.g., the performance impact of losing only .0046% of the packets can cause the throughput to become almost 17 times smaller compared to the loss-free state [7].

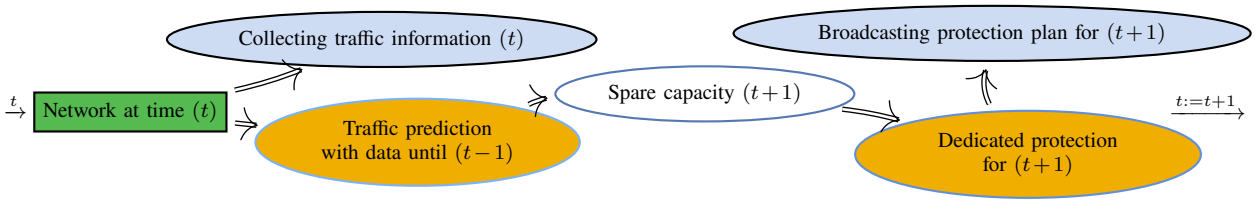


Fig. 2. Scheme of an adaptive proactive protection approach. Time slots could be as short as a few seconds [8].

tations. No adaptivity, with a basic ability of the proactive protection to enhance availability, and no bandwidth limitations should be considered as the baseline. More opportunistic adaptive schemes utilizing traffic predictions would achieve a higher bandwidth utilization of protection via trading a(n ideally) small timeshare with minor bandwidth limitations. Note that here, the experienced performance of the traffic prediction used plays a key role.

Summarized, the envisioned adaptive proactive protection schemes have two main components:

1) *Prediction:* forecasting the spare bandwidth in an upcoming time period, and
2) *Routing:* computing a best-fit dedicated protection for the period based on the estimated spare bandwidth.

The current study aims to understand the key possibilities and limitations of such approaches by investigating their two key components. To do so, we follow up on recent article [8] and an earlier version of our current study, presented at RNDM [1], which, to the best of our knowledge, represent the current state-of-the-art among adaptive proactive protection schemes. These articles focused on the prediction of the *mean* traffic volume. If the future demand's distribution is symmetric, this coincides with the median, which means that traffic volume is underestimated in as many cases as it is overestimated. In the adaptive protection framework, however, underestimation has severe consequences compared to overestimation. The former leads to congestion, possibly packet loss, while the latter leads to a waste of the spare capacity. This issue is acknowledged by [8]; they propose two different solutions: the first is a "buffer" that is 10% of the mean prediction, regardless of the demand distribution; the second is an asymmetric loss-function to be minimized by the prediction models, with aims to achieve overestimation of the traffic in 90% of the cases. However, a 90% overestimation ratio is still low when the relative disadvantages of over- and underestimation are weighed against each other. Neural network-based prediction methods, like those used in [8] and [1] struggle to reliably achieve higher overestimation ratios while staying accurate, due to the sparsity of extreme values in their training sets. We propose a method that works more reliably for target overestimation ratios as high as 99% and 99.9%, using Extreme Value Theory (EVT).

Our main contributions are as follows:

- Analyzing the components of an ideal adaptive proactive protection framework, we propose refinements to the state-of-art model [8]. Namely, we argue that with practical discretization of time, a lookahead of as low as 2 can be applied for prediction (instead of the baseline value of 8). We also hint at alternative scalable routing algorithms to be used to compute the dedicated protection routes.
- We argue that predicting a *high quantile* of the future demands' distribution is superior to classical methods that simply predict the *mean* (as done in [1], [8]) in terms of the ability to be recycled to securely and 'slightly' overestimate the traffic.
- For the prediction of high quantiles, we describe a simple neural network-based architecture, and an improved version that is based on neural networks and extreme value theory.
- We evaluate the performance of these traffic prediction approaches for high quantiles on real-world data (collected from the ESNet). We use asymmetric error metrics that penalize *underestimation* more severely than *overestimation*, accounting for the consequences of underestimation in a real network setting.

Relative to our RNDM 2025 paper [1], this version extends the mean-focused analysis to high-quantile forecasting, introduces the TCN+POT construction and its EVT treatment, and evaluates these methods on a new ESnet period across 73 links. The rest of the paper is structured as follows: §II presents our model, and background in both prediction and routing. §III presents our test data and methodology for predicting bandwidth demand, §IV discusses our evaluation results, and finally, §V concludes our work.

II. Model, Assumptions, and Background

The network is modeled as a directed graph $D=(V,A)$, with each link $a \in A$ having capacity $c_a > 0$, and a related time series $\{X_t^a\}$ $(t \in \mathbb{N})$ describing the bandwidth usage on a at time t. If known in advance, spare bandwidths $s_t^a := c_a - x_t^a$ could be used to host backup paths for protected data flows, computed by highly scalable resilient routing algorithms.[2] Unfortunately, future demands are not known. Let f denote our desired prediction distance. Thus, we aim to predict the future demand $\widehat{x}_{t+f}^a$. Here, two conflicting challenges arise: 1) Overestimation of x_{t+f}^a (equating to underestimation of spare bandwidth s_{t+f}^a) leads to less capacity to be used as a backup resource, and 2) an eventual underestimation of x_{t+f}^a may yield bandwidth limitations. We consider a forecasting distance of $f=2$ (amounting to 60 `[sec]` in our use case) to be sufficient to perform all necessary preparations (prediction, routing, updating). This is in contrast to [8], which supposes a distance of 240 `[sec]` is necessary, making prediction a

[2]Following standard statistical convention, we denote random variables with uppercase letters and their specific realizations with lowercase letters.

more challenging task. Hereafter, we will refer to cases where $f = 2$ as *short predictions*.

Since an underestimation of x^a_{t+f} poses a higher risk to service availability, $\widehat{x}^a_{t+f}$ is defined to be the prediction of the τ^{th} order quantile of the (unknown) future demand distribution X^a_{t+f}. Hyperparameter $\tau \in (0,1)$ is the target non-exceedance probability, defining the ratio in which we prefer to err on the side of overestimation of traffic. Note that a target τ does not directly define end-to-end service availability, which depends on several factors, such as the network's routing, topology, and response to capacity exceedance. In practice, τ should be chosen close to 1 to prevent frequent capacity violations.

A. Dedicated Protection Approaches

Nowadays, the so-called 1+1 is the most widespread dedicated protection approach. With 1+1, the data can be sent parallel on edge- or node-disjoint working and backup paths (WP and BP), ensuring instantaneous recovery against single link or node failures in a simple manner. Such a path pair can be quickly found, via, e.g., Suurballe's algorithm [9] that computes a WP-BP path pair of minimum total length in $O(|E|+|V|\log|V|)$. Being such a computationally efficient subroutine, Suurballe's algorithm is very useful for preparing the network for single element failures with short predictions.

To exceed a certain availability threshold, however, protection against the simultaneous failure of multiple elements becomes inevitable. Such failures can be caused by natural or man-made disasters (such as earthquakes, electromagnetic pulses, etc.). Such failures are modeled as *Shared Risk Group*s (SRG), which are sets of network elements that are expected to fail simultaneously. For routing purposes, the failure of a node v can be modeled as the failure of all the links incident to v. Thus, it is enough to take into count Shared Risk *Link* Groups (SRLGS), which are simply just SRGs containing only links. To tackle the issue of SRLG failures, the framework of [8] used the GDP-R routing of [10]. Unfortunately, since their problem formulation is NP-hard, the routing of GDP-R is calculated via the help of an ILP formulation - that can take an infeasibly long time to be optimized, especially if we want to make it work together with short predictions.

Studying SRLG-disjoint path computing problems has a long history [11]. Papers [12]–[14] prove the NP-hardness of the problem in various settings. [15]–[17] rely at least partly on ILP/MILP formulations to solve or approximate the weighted version of the SRLG-disjoint paths problem. Heuristics were also investigated [18], [19], unfortunately, with issues like possibly non-polynomial runtime or occasional forwarding loops under disasters. On the positive side, resulting from the chain of studies [20]–[23], the current state-of-art algorithm published in [24], [25] called *DateLine* is highly scalable (with a near-linear computational complexity in function of $|V|$ in practice), and has a low worst-case time complexity for solving the SRLG-disjoint paths problem supposing the topology is planar and the paths should be node-disjoint. While the DateLine algorithm on its own does not take into account the spare bandwidths of links, being computationally efficient, it can be used for preparing the network for SRLG failures with short predictions. In this sense, it can be seen as the best known counterpart of Suurballe's algorithm for SRLG failures.

In conclusion, we believe the current state of SRLG-disjoint routing algorithms allows replacing the ILP-based GDP-R routing algorithm used in [8] with more scalable, efficient algorithms, enabling the choice of a lookahead time of as low as $f = 2$ for prediction of the network traffic.

B. Mean and Median Traffic Prediction Techniques

In the area of time series forecasting, the most widespread objectives are the prediction of the *mean* and the *median* of the distribution of future values [26]. The techniques mentioned in this subsection are mostly used for these objectives, yet most of them could easily be adapted for our purpose of predicting a high quantile.

A classical, linear forecasting model is the Autoregressive Integrated Moving Average (ARIMA) [27]. The ARIMA is interpretable due to its linear nature and simple structure; it performs well on periodic, linearly dependent data. ARIMA models were evaluated by both [8] and [1] for the purpose of traffic prediction, and they have both found that, on average, these models have worse performance than their nonlinear counterparts. Another limitation of the ARIMA is that it predicts the mean by design. Even if an accompanying GARCH [28] model were used to estimate the variance, the prediction would be bound by the underlying assumption of both these models, namely, that the future value distribution is *gaussian*.

Given the noisiness and nonlinear behavior of real-world bandwidth usage data, nonlinear models, especially neural networks, are also applied in this domain. Examples include feed-forward neural networks (more specifically, Convolutional Neural Networks – CNNS), such as the

- Temporal Convolutional Networks (TCNS) [29], devised for general-purpose time series forecasting, employing residual connections known from ResNet [30], and
- WaveNet [31], an architecture similar to TCNS, originally used for voice generation, but applicable to other time series prediction problems [32]. Wavenet uses *local* and *global conditioning*, based on metadata provided as input (for example, linguistic features or the speaker's identity). In the case of traffic volume forecasting, such metadata could be the time-of-day or day-of-week.

Recurrent Neural Networks (RNNS) are another family of neural networks suited for processing sequential data [33]. Although RNNS are structurally designed for processing sequential data, they are generally harder to train. (The sequential computation of data inherent in RNNS limits parallelization during training, as opposed to CNNS, which may process a long input sequence as a whole.) Furthermore, TCNS were shown to outperform RNNS (including specialized architectures like LSTMS and GRUS) on multiple sequential tasks [29].

A third, novel family of neural networks, namely Transformer-based architectures have outstanding performance in the domains of natural language processing and speech

recognition, yet they are harder to apply to time series forecasting problems, due to their permutation-invariant self-attention mechanism. Study [34] shows that, in many cases, Transformers specialized for time series forecasting have poorer performance than a simple linear model, on various datasets. The original Transformer architecture also has $O(N^2)$ time and memory complexity in terms of input sequence length N, which may be prohibitively expensive in the case of high-volume time series data (whereas the TCN scales linearly). Although some specialized Transformer models reduce this theoretical complexity to $O(N\log N)$, or even $O(N)$, [34] also shows that it is unclear whether this improves the actual inference time and memory cost. A more recent architecture, iTransformer [35], achieves superior performance on multivariate time series prediction tasks compared to the previous models assessed in [34].

Despite the recent emphasis on deep learning models, new linear models are still being developed; Random Convolutional Kernel Transform (ROCKET) [36], for example, is essentially a convolutional neural network with no hidden layers and a very high number of filters. ROCKET was demonstrated to be on par with deep learning architectures like ResNet in terms of accuracy on certain time series benchmarks.

The neural network-based models described above are all aiming to minimize the error of their predictions conforming to a *loss function*. This loss function is typically the *mean squared error* (MSE) or the *mean average error* (MAE). Minimizing these functions leads to predictors of the mean and the median, respectively [27]. To create a predictor for a τ^{th} order quantile, we can use the *pinball loss* [37], [38], an asymmetric loss function that is a generalisation of the MAE:

$$L_{\text{pinball}}(x,\widehat{x};\tau) = \tau \sum_{x_i > \widehat{x}_i} (x_i - \widehat{x}_i) + (\tau - 1) \sum_{x_i < \widehat{x}_i} (x_i - \widehat{x}_i).$$

A similar loss function was employed by [8], although their primary focus was on mean prediction. While [8] aimed to achieve a 90% overestimation ratio (corresponding to $\tau = 0.9$), we argue that, under practical settings, τ should be at least 0.99 to keep the time share with possible congestion induced by the adaptive protection approach reasonably low. Since both [8] and [1] showed that neural networks perform better for mean prediction than the linear methods, and they are also easier to adapt for the prediction of high quantiles, we choose temporal convolutional networks (TCNs) as our base model. TCNs also have relatively low resource requirements compared to RNNs and Transformers [39].

The direct prediction of high quantiles with neural networks is, however, still a difficult task [40]. During training, the data points close in range to the high quantile should be guiding the network to learn a correct characterization of the distribution's tail. But these data points are, by definition, rare.

C. Peaks Over Threshold (POT) Method

An approach for dealing with the rare, extreme events that constitute the tail of the future value distribution is given by the statistical framework known as *extreme value theory* (EVT) [41], [42]. Let us define extreme values as observations exceeding $q_u = Q(x;u)$, a high quantile (u^{th} order quantile) of the past observations x. We refer to the practice of collecting and analyzing these values as the *Peaks Over Threshold* (POT) method. For a random variable X with distribution F_X, the distribution of the exceedances $y = x - q_u$ is described as [42]:

$$Pr(X > q_u + y | X > q_u) = \frac{1 - F_X(q_u + y)}{1 - F_X(q_u)}$$

A main result of EVT, the Pickands-Balkema-De Haan theorem, states that for a sufficiently high u value, the distribution of the exceedances of the extreme values can be approximated by the *Generalised Pareto Distribution* (GPD) [41], [42]:

$$G_{\xi,\sigma}(y) = \begin{cases} 1 - \left(1 + \dfrac{\xi y}{\sigma}\right)^{-1/\xi}, & \text{if } \xi \neq 0, \\ 1 - \exp\left(-\dfrac{y}{\sigma}\right), & \text{if } \xi = 0. \end{cases}$$

Parameter ξ determines the shape of the distribution. If $\xi < 0$, the upper bound of the distribution is $-\sigma/\xi$. If $\xi \geq 0$, the distribution has no upper bound. This result holds under the assumption that the data points come from independent and identically distributed (*iid*) distributions. This is not the case for bandwidth usage data by default. As a way to remove seasonality, trend and other time-dependent factors from the time series, we apply a "base model" to the data. The GPD is then fit using maximum likelihood estimation to the *residuals*, the difference between the base model's predictions and the realised values. We can thus model the tail behaviour of the residuals, which we will use in the final prediction of the high quantiles. A thorough description of the POT-based hybrid method for the prediction of high quantiles follows in §III-C, where Fig. 4 shows its application to a link from the evaluated traffic dataset. Note that EVT does not give a solution for the *prediction* of extreme events, rather a way to describe their *extent*, should they occur. This still proved useful for our objective of the prediction of a high quantile of the future value distribution.

III. Methodology for Traffic Prediction

A. Test Data

We analyzed bandwidth usage data obtained from the Energy Sciences Network (ESnet), covering the 4-week period of 2026-02-09 to 2026-03-09. The resolution of the time steps is 30 secs; this gives 80640 data points per direction per link. The tests were executed on 73 data links. We have analyzed a single data flow direction for each link. (In the context of traffic volume prediction, we may treat each direction as a separate link since we have per-direction bandwidth usage data on them.) The first 3 weeks of the analyzed interval were used as training data, while the last week was used to evaluate the prediction capabilities of the models. This results in a 75%-25% train-test split ratio.

For the sake of simplicity and computational efficiency, we make no assumptions about the correlation between traffic volumes on links, meaning all links are treated as separate

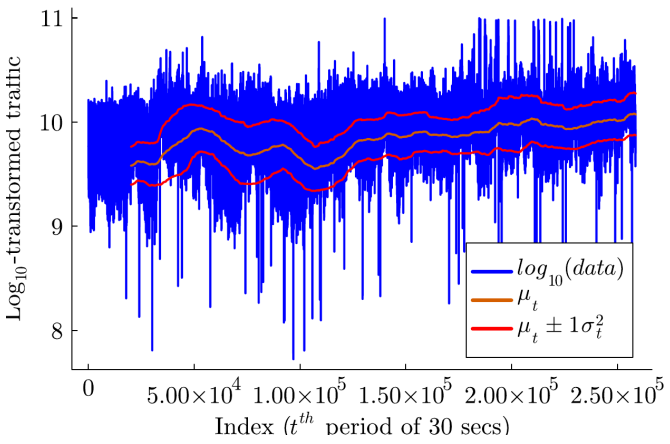


Fig. 3. Logarithm of traffic volume on link CERN513–WASH; mean μ_t, and variance σ_t^2 calculated with window size $w = 20160$, i.e., 1 week.

univariate time series prediction problems. Although each model evaluation happens with the same set of hyperparameters across links, model fitting / training happens separately for each link, since traffic volume, potential periodicity, and other patterns vary widely between links. The constant τ is the target quantile that we aim to predict; it is also treated as a hyperparameter of the models.

B. Temporal Convolutional Networks (TCN)

TCNs are convolutional neural networks specifically designed for sequential data processing. Their core component is the *causal convolution*, where the output corresponding to time step $t+f$ depends strictly on inputs from t and earlier, preventing future information leakage. TCNs employ *residual connections* [30] to mitigate the problem of vanishing gradients. Causal convolutions are *dilated*, in order to expand the receptive field: the convolutional kernels are applied to input elements spaced d positions apart, where d is the dilation rate. By choosing the dilation rates to be increasing powers of two in consecutive layers, an exponential growth in the size of the receptive field of the model is achieved with a linear increase in the number of parameters [29], [31]. Dropout layers are employed in order to prevent overfitting [29]. Another method applicable to reduce overfitting is the L2 regularization of weights, as used in the case of CNNs for time series forecasting in [32]. In order to detect overfitting, 10% of the training dataset was randomly selected to constitute the validation set.

We observed that, on most links, the marginal distribution of the data on the entire time interval is approximately log-normal, as is common for internet traffic [43]. The exceptions are the 0-values and, in some cases, a heavy tail induced by extreme events (bursts). The most significant outliers are the 0-values, comprising 0.47% of the data per link on average. We replaced these with a moving average, in order to be able to normalize the data. This leads to more inaccurate predictions corresponding to 0-values, but better performance in general. Occasional outages, although rare, were also represented with 0-values in the dataset.

Furthermore, any sufficiently long interval taken from a link's data also has an approximately log-normal marginal distribution. This, however, does not mean that the data is stationary: the means and variances of the log-transformed data vary significantly in function of the starting point and size of the interval. For an example, see Fig. 3.

These distribution properties are crucial for a meaningful standardization of the data. For each timestep t, the transformation $(\log(x_t) - \mu_t)/\sigma_t$ is applied to the input data. μ_t and σ_t are calculated from a running window $\log(x_{t-w}),...,\log(x_t)$. An additional clipping step is performed on the running window's values to lessen the disproportionate influence of outliers on the calculation of μ_t and σ_t. The running window length w is considered a hyperparameter of our model.

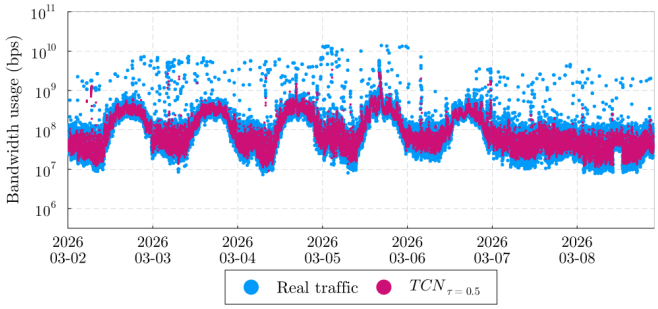
While the primary input channel is the normalized data described above, we apply local conditioning [31], [32] using *time encoding*. This is done by supplying two sine-cosine pairs as extra input channels: one of them has a period of 2880 units (1 day), while the other has a period of 20160 units (7 days). These channels provide the network with features that help it distinguish recurring traffic patterns based on time-of-day and time-of-week. Our preliminary tests showed that time encoding improved MAE loss on the links used for hyperparameter-tuning by about $\sim 1\%$. Additionally, a *differenced* version of the logarithmized time series is also supplied as an extra channel. This is analogous to the *log-returns* that are used in financial modeling [44].

The best-performing TCN model had 8 convolutional layers, with residual connections after every second layer. The 6 original input channels are projected to 32 channels internally. After the convolutional layers a fully connected layer with 50 neurons follows, implemented using 1x1 convolutions. This final hidden layer is summarised using a 1x1 convolution to give the final prediction.

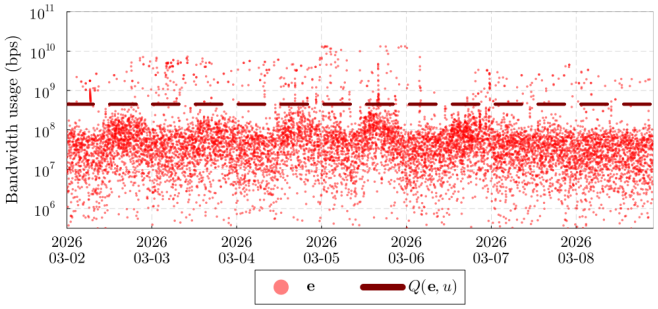
C. TCN-POT Hybrid Method

TCNs and other time series prediction techniques are useful to model temporal patterns; they assume temporal dependencies between data points. On the contrary, the POT method assumes the data points are sampled from *iid.* random variables. TCNs are useful for describing the most probable outcome, but we cannot expect them to foresee unpredictable events. On the other hand, the POT method is used exactly for the characterization of these extreme events. Thus the idea to combine these two methods arises naturally.

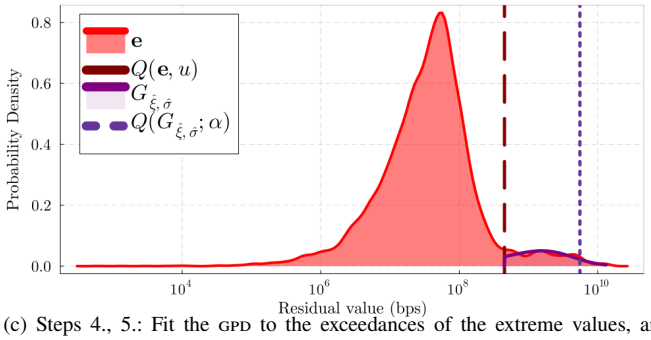
In this subsection we describe a way to chain these two methods. The idea of chaining a "base model" with the POT method has been previously explored for anomaly detection in computer networks [45], in vehicular traffic [46], and for the prediction of value-at-risk in financial assets [44]. This base model could be a simple moving average as in the case of [45], a linear method as in [44], or a RNN-LSTM as in [46]. However, to the best of our knowledge, this is the first instance of this hybrid method's use for the purpose of *prediction* of traffic in a network. The steps for this hybrid method are as follows.



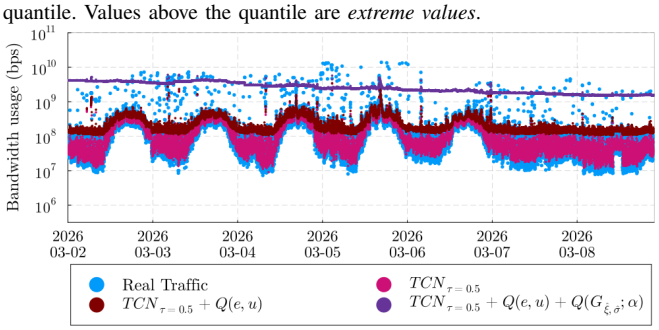
(a) Step 1.: Apply a TCN to the historical observations.



(b) Steps 2., 3.: Calculate the positive residuals and determine their u-th order quantile. Values above the quantile are *extreme values*.



(c) Steps 4., 5.: Fit the GPD to the exceedances of the extreme values, and determine the GPD's α-th order quantile.



(d) Step 6.: Calculate the final prediction based on the base model's prediction, the residuals' u-th quantile and the GPD's α-th quantile.

Fig. 4. Steps of the TCN+POT hybrid method, illustrated on the representative link PPPL–WASH for the one week test-period of 2026-03-02 to 2026-03-09.

1. Apply a TCN (trained to estimate the mean or median) to the historical observations:

$$\widehat{y}_{t_i+f} = \text{TCN}(x_{t_{i-l_{\text{TCN}}}},\ldots,x_{t_i}) \quad \forall l_{\text{TCN}} \le t_i \le t$$
$$\widehat{\mathbf{y}} = \{\widehat{y}_{t_i+f} \mid l_{\text{TCN}} < t_i < t\}.$$

2. Calculate the positive residuals (underestimation errors) by subtracting the forecasted values from the observed traffic volumes, retaining only those cases where demand exceeded the prediction:

$$\mathbf{e} = \{x_{t_i} - \widehat{y}_{t_i} \mid x_{t_i} > \widehat{y}_{t_i}\}.$$

3. Assuming the TCN effectively captures the temporal dependencies of the series, we treat the resulting residuals as a sample of approximately *iid.* random variables. We then select a threshold $u < \tau$ (representing a specific quantile of the residuals) above which observations are considered extreme. We keep the *exceedances* of the extreme events:

$$\mathbf{e}^+ = \{e_{t_i} - Q(\mathbf{e};u) \mid e_{t_i} \in \mathbf{e}, e_{t_i} > Q(\mathbf{e};u)\}.$$

4. Fit the Generalized Pareto Distribution (GPD) to the set of exceedances $\mathbf{e}^+$ using Maximum Likelihood Estimation:

$$\widehat{\xi},\widehat{\sigma} = \underset{\xi,\sigma}{\text{argmax}} \sum_i \ln g(e_i^+;\xi,\sigma)$$

5. Determine the specific α-th quantile of the GPD that corresponds to the target τ^{th} quantile of the original future demand distribution X_{t+f}. This is calculated as:

$$\alpha = 1 - \frac{1-\tau}{p_+(1-u)},$$

where p_+ is the probability of a residual being positive (approximated via the ratio of positive residuals in the historical observations).

6. The final high-quantile prediction is obtained by combining the TCN's forecast, the chosen threshold, and the EVT-based safety margin:

$$y^*_{t+f} = \widehat{y}_{t+f} + Q(\mathbf{e};u) + Q(G_{\widehat{\xi},\widehat{\sigma}};\alpha).$$

See Fig. 4 for an illustration of the process. Instead of the TCN, any sufficiently accurate prediction method could be used. To demonstrate the importance of a good base model, in §IV we also evaluate the hybrid method using a naive lag-based model ($\widehat{x}_{t+f} = x_t$) and a rolling window-based median-predicting model ($\widehat{x}_{t+f} = Q(x_{t-w},\ldots,x_t;0.5)$).

For the hyperparameter u, we considered four natural high thresholds: $u \in \{0.85, 0.9, 0.95, 0.99\}$, where $\alpha > 0$. Theoretically, all four approaches may yield similar estimates. However, for different sample sizes and underlying distributions, one threshold may produce slightly more accurate estimates than the others. To assess their relative performance, we applied all of them in our analysis. The evaluation of the residuals for different u-levels is relatively inexpensive compared to the TCN-prediction step.

The question arises: why employ a theoretical distribution like GPD rather than a simpler heuristic, such as adding the maximum observed residual to the base forecast? Our method offers two primary advantages. First, the GPD allows for *extrapolation*: our model can estimate quantiles for extreme events that have not yet occurred within the limited observation window. This is critical for network reliability, where the goal

is to prepare for spikes that may not have occurred during the observation period. Second, relying on the highest seen error makes the protection scheme highly sensitive to individual noisy outliers or measurement errors. By contrast, the GPD integrates information from the entire set of exceedances above the threshold u. This results in a smoother and more robust estimate, preventing the inefficient bandwidth over-provisioning that would otherwise be triggered by a single, unrepresentative extreme value.

IV. Evaluation

For conducting our evaluations, we implemented the prediction module[3] in the Julia programming language, using the Flux.jl library [47] for the TCNs and Extremes.jl [48] for the EVT-analysis. Our goal is to achieve predictions that err on the side of overestimation of the traffic, since the resulting bandwidth waste has less negative consequences than the underestimation-induced congestion. The exact ratio of penalization for over- and underestimation should be determined in the context of the risk entailed for the specific network. We chose to evaluate our models for $\tau \in \{0.9, 0.99, 0.999\}$.

Our primary method is the hybrid approach described in Section §III-C, using a median-predicting TCN as the base model. As a simple baseline, we take the τ^{th} quantile from a running window of past observations. We also compare against a standard TCN conditioned on the pinball loss to directly predict the high quantile τ, as well as variants of the hybrid method using naive lag-based and median-based predictors.

The candidate u-levels were fixed before test set evaluation and are reported as separate configurations; the test set was not used to tune u. Hyperparameters of the TCN models were chosen based on their respective performance on the validation set.

In order to be able to aggregate the short term prediction performances of different models across links, we use a capacity-normalized version of pinball loss:

$$L_{\text{pinballCap}}(x, \widehat{x}; \tau, c_{\text{link}}) = \frac{L_{\text{pinball}}(x, \widehat{x}; \tau)}{c_{\text{link}}}.$$

We have also evaluated the models' underestimation ratio. In certain cases, a model's underestimation ratio may be close to the objective determined by τ, while its $L_{\text{pinballCap}}$ value is high. This means that, although the underestimation ratio is correct, the model's predictions are far from the actual observations. In practice, the most useful models are those that minimize $L_{\text{pinballCap}}$, this is our primary objective. The models' $L_{\text{pinballCap}}$ losses are illustrated on Fig. 5. For detailed results, see Tables I, II, III, for $\tau \in \{0.9, 0.99, 0.999\}$, respectively. The results with the best mean and variance are marked in bold.

The standard deviation of the $L_{\text{pinballCap}}$ results across links is relatively high, generally approaching the magnitude of the mean. This variance is attributable to the high heterogeneity of the 73 analyzed links; for instance, the daily and weekly periodicities typical of many backbone networks are not universally present.

[3]Implementation available: https://gitlab.com/dobaipatakyattila/qoserm-tsa

TABLE I
MODEL PERFORMANCE FOR $\tau = 0.9$, BEST RESULTS IN BOLD. $\text{TCN}_{\tau=0.9}$ AND TCN+POT HYBRID PERFORMANCE ALMOST IDENTICAL.

Model	Params.	Underestim. %		PinballCap ($\times 10^2$)	
		mean	std	mean	std
TCN	$\tau = 0.9$	8.571	2.311	**10.838**	**11.011**
TCN+POT	$u = 0.85$	**10.084**	**0.478**	11.805	11.935
Quantile	$\tau = 0.9$	10.359	1.216	34.074	28.697
Median+POT	$u = 0.85$	11.619	1.432	37.260	29.644
Lag+POT	$u = 0.85$	10.351	0.499	15.790	13.481

TABLE II
MODEL PERFORMANCE FOR $\tau = 0.99$, BEST RESULTS IN BOLD. TCN+POT PERFORMS BEST, BUT $\text{TCN}_{\tau=0.99}$ IS A CLOSE SECOND.

Model	Params.	Underestim. %		PinballCap ($\times 10^2$)	
		mean	std	mean	std
TCN	$\tau = 0.99$	1.575	5.645	3.416	3.405
TCN +POT	$u = 0.85$	1.058	0.218	3.439	3.674
	$u = 0.90$	1.108	0.157	3.294	3.280
	$u = 0.95$	1.104	**0.127**	**3.281**	**3.255**
Quantile	$\tau = 0.99$	1.253	0.369	8.199	6.425
Median +POT	$u = 0.85$	1.465	0.670	10.150	7.425
	$u = 0.90$	1.530	0.685	9.988	7.314
	$u = 0.95$	1.603	0.717	9.603	6.941
Lag +POT	$u = 0.85$	**0.995**	0.361	5.747	6.312
	$u = 0.90$	1.023	0.286	5.634	6.083
	$u = 0.95$	1.093	0.232	5.194	5.178

For the lowest investigated target of $\tau = 0.9$, the direct quantile-forecasting model ($\text{TCN}_{\tau=0.9}$) achieved the best performance in terms of $L_{\text{pinballCap}}$. This suggests that at this level, observations do not yet qualify as "extreme," allowing conventional forecasting methods to perform effectively. Nevertheless, the EVT-based hybrid method achieved results comparable to the top-performing model. Furthermore, the hybrid approach approximated the target underestimation rate (which is 10% for $\tau = 0.9$) most accurately.

In contrast, while the naive moving-window quantile calculation was relatively accurate regarding the underestimation rate, exhibiting lower standard deviation and staying closer to the target than the $\text{TCN}_{\tau=0.9}$, its $L_{\text{pinballCap}}$ loss was more than triple that of the TCN-based models.

For the higher target quantiles of $\tau \in \{0.99, 0.999\}$, the TCN+POT hybrid method proved optimal in terms of both the mean and standard deviation of $L_{\text{pinballCap}}$, as well as the stability of the underestimation rate:

- For $\tau = 0.99$, the $L_{\text{pinballCap}}$ performance of the TCN+POT hybrid and the direct $\text{TCN}_{\tau=0.99}$ model were relatively similar. However, the standard deviation of the underestimation rate for the $\text{TCN}_{\tau=0.99}$ model was extremely high; 44 times higher than that of the optimal TCN+POT model.
- For $\tau = 0.999$, the average $L_{\text{pinballCap}}$ loss of the TCN+POT hybrid model was 1.7 times lower than the next best performers (the $\text{TCN}_{\tau=0.999}$ and the Lag+POT hybrid models).

At these high τ levels, the Median+POT and Lag+POT models are largely unreliable, except when $u = 0.99$. With relatively low u values, the corresponding α parameter becomes high, causing the α-quantile of the fitted GPD to produce excessively large positive-direction estimates. This is further evidenced by

TABLE III

MODEL PERFORMANCE FOR $\tau = 0.999$, BEST RESULTS IN BOLD. TCN+POT OUTPERFORMS THE OTHER MODELS BY A LARGE MARGIN.

Model	Params.	Underestim. %		PinballCap ($\times 10^2$)	
		mean	std	mean	std
TCN	$\tau = 0.999$	**0.096**	0.145	0.954	1.148
TCN +POT	$u = 0.85$	0.092	0.054	2.632	10.100
	$u = 0.90$	0.089	0.043	1.592	4.467
	$u = 0.95$	0.096	0.033	0.661	0.773
	$u = 0.99$	0.119	**0.022**	**0.542**	**0.492**
Quantile	$\tau = 0.999$	0.162	0.075	1.325	1.139
Median +POT	$u = 0.85$	0.220	0.249	93.989	> 100.0
	$u = 0.90$	0.247	0.270	> 100.0	> 100.0
	$u = 0.95$	0.264	0.255	> 100.0	> 100.0
	$u = 0.99$	0.293	0.281	1.970	1.957
Lag +POT	$u = 0.85$	0.097	0.081	> 100.0	> 100.0
	$u = 0.90$	0.093	0.065	> 100.0	> 100.0
	$u = 0.95$	0.093	0.053	> 100.0	> 100.0
	$u = 0.99$	0.112	0.044	0.944	0.895

the observation that the underestimation rates for these models are less catastrophic than their absolute $L_{\text{pinballCap}}$ values.

A general observation is that POT-based models perform best at the highest permissible u values; that is, they are most accurate when operating on fewer, more extreme data points. In summary, when underestimation is penalized at least 100 times more severely than over-estimation, the TCN+POT hybrid model described in §III-C provides significantly superior predictions compared to direct quantile-forecasting TCNs or other hybrid models utilizing simpler base techniques. This is of paramount importance for forecasting spare capacity in internet networks, where traffic underestimation can lead to severe performance degradation. For less extreme requirements, the standard TCN model remains a sufficient solution.

V. CONCLUSIONS

In this paper, we investigated the key components of an adaptive proactive protection framework for backbone networks, focusing on the critical trade-off between bandwidth efficiency and service availability. Our work motivates that the advent of highly scalable routing algorithms, such as DateLine, could enable a significant reduction in prediction lookahead; from the 8-slot baseline established in prior literature to just 2 slots (60 seconds). This reduction in horizon is a crucial enabler for more accurate forecasting and more responsive network management.

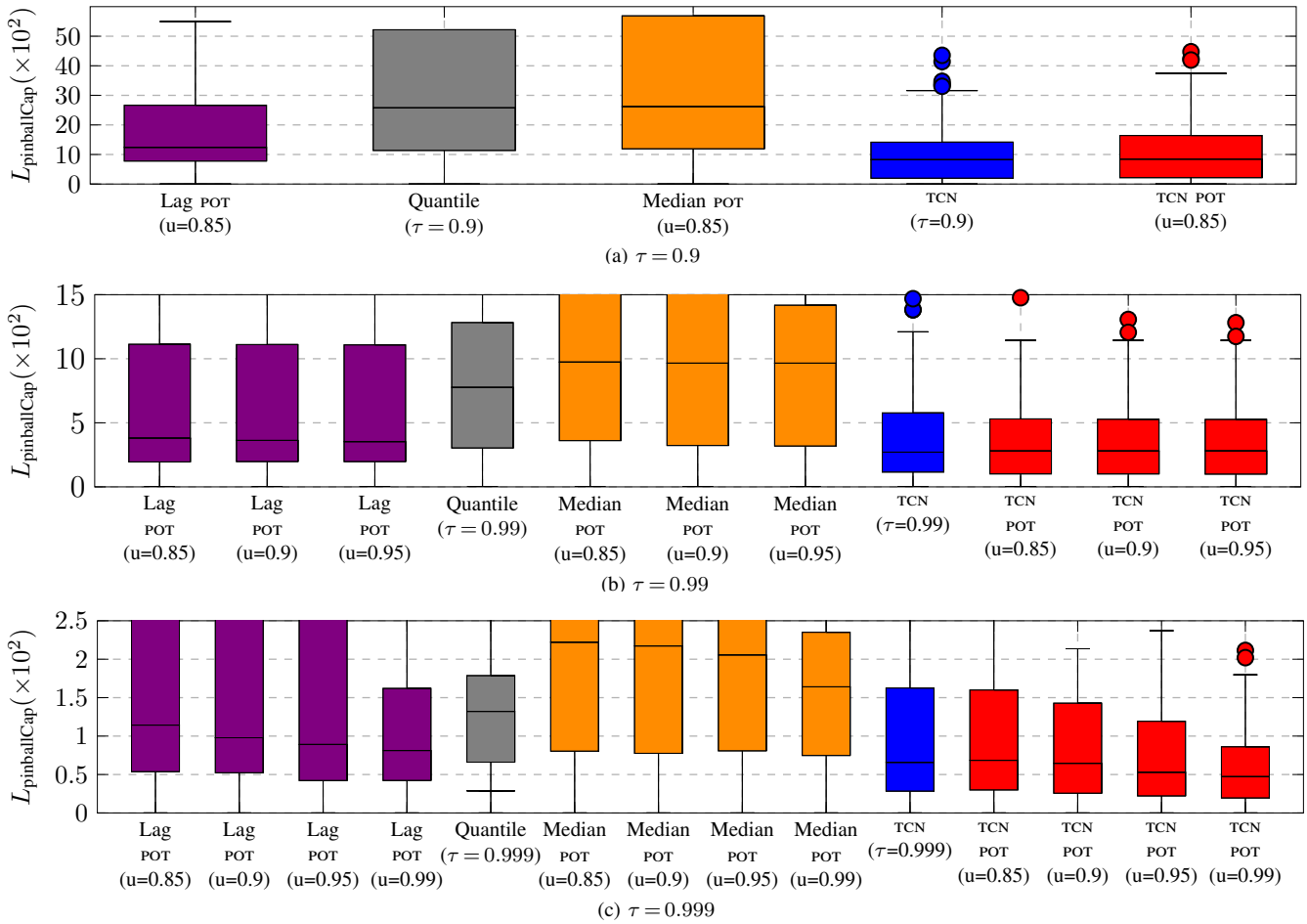


Fig. 5. $L_{\text{pinballCap}}$ loss values for different τ targets. The TCN and TCN+POT models outperform the naive baseline methods. For $\tau \in \{0.99, 0.999\}$, the hybrid model with the highest u parameter achieves the best performance. The boxplots of the poorly performing naive methods are not shown in full range, as they are less relevant.

Our empirical evaluation on real-world ESnet traffic demonstrates that while Temporal Convolutional Networks (TCNs) are effective at capturing the non-linear temporal patterns required for mean-based forecasting, they struggle to directly predict the extreme tail behavior necessary for high-availability protection.

To address this limitation, we proposed a hybrid TCN+POT framework. By utilizing the TCN to filter predictable dependencies and fitting the Generalized Pareto Distribution (GPD) to the resulting residuals, we achieved superior stability and accuracy for high reliability targets ($\tau \geq 0.99$). The hybrid method's ability to extrapolate beyond observed data while maintaining statistical stability against noisy outliers makes it a more robust choice for mission-critical infrastructures compared to purely empirical or mean-centric approaches.

Acknowledgements

This study has received funding from the European Union's Horizon Europe research and innovation programme under the Marie Sk lodowska-Curie grant agreement No. 101155116 and from the "Romanian Hub for Artificial Intelligence - HRIA", Smart Growth, Digitization and Financial Instruments Program, 2021-2027, MySMIS no. 334906.

References

[1] A. Dobai-Pataky, B. Vass, and L. Csató, "On traffic prediction in backbone networks for adaptive proactive protection," in *Int. Workshop on Resilient Networks Design and Modeling (RNDM)*, Trondheim, Norway, Jun. 2025. DOI: 10.1109/RNDM66856.2025.11073795 .

[2] A Review of Critical Infrastructure Domains in Europe. SPEAR, www.spear2020.eu/News/Details?id=120. Accessed: 2025.

[3] Z. Wang, *Internet QoS: Architectures and Mechanisms for Quality of Service*, 1st ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001. DOI: 10.1016/B978-1-55860-608-1.X5000-5 .

[4] F. Tariq, M. R. A. Khandaker, K.-K. Wong, M. A. Imran, M. Bennis, and M. Debbah, "A speculative study on 6g," *IEEE Wireless Communications*, vol. 27, no. 4, pp. 118–125, 2020. DOI: 10.1109/MWC.001.1900488 .

[5] D. Feng, L. Lai, J. Luo, Y. Zhong, C. Zheng, and K. Ying, "Ultra-reliable and low-latency communications: applications, opportunities and challenges," *Science China Information Sciences*, vol. 64, pp. 1–12, 2021. DOI: 10.1007/s11432-020-2852-1 .

[6] Y. Huang and R. Guerin, "Does over-provisioning become more or less efficient as networks grow larger?" in *13TH IEEE International Conference on Network Protocols (ICNP'05)*, 2005, pp. 11 pp.–235. DOI: 10.1109/ICNP.2005.12 .

[7] On packet loss in ESNet. fasterdata.es.net/network-tuning/tcp-issues-explained/packet-loss/. Accessed: 2025.

[8] F. Mogyorósi, A. Pašić, R. Cziva, P. Revisnyei, Z. Kenesi, and J. Tapolcai, "Adaptive protection of scientific backbone networks using machine learning," *IEEE Transactions on Network and Service Management*, vol. 18, no. 1, pp. 1064–1076, 2021. DOI: 10.1109/TNSM.2021.3050964 .

[9] J. W. Suurballe and R. E. Tarjan, "A quick method for finding shortest pairs of disjoint paths," *Networks*, vol. 14, no. 2, pp. 325–336, 1984. DOI: 10.1002/net.3230140209 .

[10] P. Babarczi, A. Pašić, J. Tapolcai, F. Németh, and B. Ladóczki, "Instantaneous recovery of unicast connections in transport networks: Routing versus coding," *Computer Networks*, vol. 82, pp. 68–80, 2015. DOI: 10.1016/j.comnet.2015.02.010 ., robust and Fault-Tolerant Communication Networks.

[11] P. Datta, M. T. Frederick, and A. K. Somani, "Sub-graph routing : A novel fault-tolerant architecture for shared-risk link group failures in wdm optical networks," in *Design of Reliable Communication Networks (DRCN)*, 2003, pp. 296–303. DOI: 10.1109/DRCN.2003.1275369 .

[12] J.-Q. Hu, "Diverse routing in optical mesh networks," *IEEE Trans. Communications*, vol. 51, pp. 489–494, 2003. DOI: 10.1109/TCOMM.2003.809779 .

[13] J.-C. Bermond, D. Coudert, G. D'Angelo, and F. Z. Moataz, "SRLG-diverse routing with the star property," in *Design of Reliable Communication Networks (DRCN)*. IEEE, 2013, pp. 163–170.

[14] X. Luo and B. Wang, "Diverse routing in WDM optical networks with shared risk link group (SLRG) failures," in *Proceedings of 5th International Workshop on the Design of Reliable Communication Networks (DRCN 2005)*, Oct. 16-19 2005. DOI: 10.1109/DRCN.2005.1563905 .

[15] D. Xu, G. Li, B. Ramamurthy, A. Chiu, D. Wang, and R. Doverspike, "SRLG-diverse routing of multiple circuits in a heterogeneous optical transport network," in *8th International Workshop on the Design of Reliable Communication Networks (DRCN 2011)*, 2011, pp. 180–187. DOI: 10.1109/DRCN.2011.6076901 .

[16] T. Gomes, M. Soares, J. Craveirinha, P. Melo, L. Jorge, V. Mirones, and A. é. Brízido, "Two heuristics for calculating a shared risk link group disjoint set of paths of min-sum cost," *Journal of Network and Systems Management*, vol. 23, no. 4, pp. 1067–1103, 2015. DOI: 10.1007/s10922-014-9332-6 .

[17] A. de Sousa, D. Santos, and P. Monteiro, "Determination of the minimum cost pair of D-geodiverse paths," in *DRCN 2017 - 13th International Conference on the Design of Reliable Communication Networks*, Munich, Germany, March 8-10 2017.

[18] K. Xie, H. Tao, X. Wang, G. Xie, J. Wen, J. Cao, and Z. Qin, "Divide and conquer for fast SRLG disjoint routing," in *2018 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, 2018, pp. 622–633. DOI: 10.1109/DSN.2018.00069 .

[19] B. Yang, J. Liu, S. Shenker, J. Li, and K. Zheng, "Keep forwarding: Towards k-link failure resilient routing," in *IEEE INFOCOM 2014-IEEE Conference on Computer Communications*. IEEE, 2014, pp. 1617–1625. DOI: 10.1109/INFOCOM.2014.6848098 .

[20] C. McDiarmid, B. Reed, A. Schrijver, and B. Shepherd, "Non-interfering network flows," in *Algorithm Theory—SWAT'92: Third Scandinavian Workshop on Algorithm Theory Helsinki, Finland, July 8–10, 1992 Proceedings 3*. Springer, 1992, pp. 245–257. DOI: 10.1007/3-540-55706-7 _21 .

[21] S. Neumayer, A. Efrat, and E. Modiano, "Geographic max-flow and min-cut under a circular disk failure model," in *IEEE INFOCOM*, 2012, pp. 2736–2740. DOI: 10.1109/INFCOM.2012.6195690 .

[22] Y. Kobayashi and K. Otsuki, "Max-flow min-cut theorem and faster algorithms in a circular disk failure model," in *IEEE INFOCOM 2014 - IEEE Conference on Computer Communications*, April 2014, pp. 1635–1643. DOI: 10.1109/INFOCOM.2014.6848100 .

[23] B. Vass, E. Bérczi-Kovács, A. Barabás, Z. L. Hajdú, and J. Tapolcai, "A whirling dervish: Polynomial-time algorithm for the regional srlg-disjoint paths problem," *IEEE/ACM Transactions on Networking*, 2023. DOI: 10.1109/TNET.2023.3276815 .

[24] E. Bérczi-Kovács, P. Gyimesi, B. Vass, and J. Tapolcai, "Efficient algorithm for Region-Disjoint survivable routing in backbone networks," in *IEEE INFOCOM 2024*, Vancouver, Canada, 2024. DOI: 10.1109/INFOCOM52122.2024.10621256 .

[25] ——, "DateLine: efficient algorithm for computing region disjoint paths in backbone networks," *IEEE JSAC Advances in Internet Routing and Addressing 2025*, pp. 1–14, Jan. 2025. DOI: 10.1109/JSAC.2025.3528810 .

[26] T. Gneiting, "Making and evaluating point forecasts," 2010. DOI: 10.1198/jasa.2011.r10138 .

[27] R. Hyndman, *Forecasting: principles and practice*. OTexts, 2018.

[28] T. Bollerslev, "Generalized autoregressive conditional heteroskedasticity," *Journal of Econometrics*, vol. 31, no. 3, p. 307–327, Apr. 1986. DOI: 10.1016/0304-4076(86)90063-1 .

[29] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," 2018.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015. DOI: 10.48550/ARXIV.1512.03385 .

[31] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," in *Arxiv*, 2016. DOI: 10.48550/arXiv.1609.03499 .

[32] A. Borovykh, S. Bohte, and C. W. Oosterlee, "Conditional time series forecasting with convolutional neural networks," 2018. DOI: 10.48550/arXiv.1703.04691 .

[33] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.
[34] A. Zeng, M. Chen, L. Zhang, and Q. Xu, "Are transformers effective for time series forecasting?" in *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, ser. AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. DOI: 10.1609/aaai.v37i9.26317.
[35] Y. Liu, T. Hu, H. Zhang, H. Wu, S. Wang, L. Ma, and M. Long, "itransformer: Inverted transformers are effective for time series forecasting," 2024. DOI: 10.48550/arXiv.2310.06625.
[36] A. Dempster, F. Petitjean, and G. I. Webb, "Rocket: exceptionally fast and accurate time series classification using random convolutional kernels," *Data Mining and Knowledge Discovery*, vol. 34, no. 5, p. 1454–1495, Jul. 2020. DOI: 10.1007/s10618-020-00701-z.
[37] I. Steinwart and A. Christmann, "Estimating conditional quantiles with the help of the pinball loss," *Bernoulli*, vol. 17, no. 1, Feb. 2011. DOI: 10.3150/10-bej267.
[38] G. Petneházi, "Quantile convolutional neural networks for value at risk forecasting," *Machine Learning with Applications*, vol. 6, p. 100096, 2021. DOI: 10.1016/j.mlwa.2021.100096.
[39] J. Kim, H. Kim, H. Kim, D. Lee, and S. Yoon, "A comprehensive survey of deep learning for time series forecasting: Architectural diversity and open challenges," 2024. DOI: 10.48550/ARXIV.2411.05793.
[40] O. C. Pasche and S. Engelke, "Neural networks for extreme quantile regression with an application to forecasting of flood risk," *The Annals of Applied Statistics*, vol. 18, no. 4, Dec. 2024. DOI: 10.1214/24-aoas1907.
[41] P. Embrechts, C. Klüppelberg, and T. Mikosch, *Modelling Extremal Events*. Springer Berlin Heidelberg, 1997. DOI: 10.1007/978-3-642-33483-2.
[42] S. Coles, *An Introduction to Statistical Modeling of Extreme Values*. Springer London, 2001. DOI: 10.1007/978-1-4471-3675-0.
[43] M. Alasmar, R. Clegg, N. Zakhleniuk, and G. Parisis, "Internet traffic volumes are not gaussian—they are log-normal: An 18-year longitudinal study with implications for modelling and prediction," *IEEE/ACM Transactions on Networking*, vol. 29, no. 3, p. 1266–1279, 2021. DOI: 10.1109/tnet.2021.3059542.
[44] A. J. McNeil and R. Frey, "Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach," *Journal of Empirical Finance*, vol. 7, no. 3-4, pp. 271–300, November 2000. DOI: 10.1016/S0927-5398(00)00012-8.
[45] A. Siffer, P.-A. Fouque, A. Termier, and C. Largouet, "Anomaly detection in streams with extreme value theory," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '17. ACM, Aug. 2017, p. 1067–1075. DOI: 10.1145/3097983.3098144.
[46] N. Davis, G. Raina, and K. Jagannathan, "A framework for end-to-end deep learning-based anomaly detection in transportation networks," 2019. DOI: 10.1016/j.trip.2020.100112.
[47] M. Innes, E. Saba, K. Fischer, D. Gandhi, M. C. Rudilosso, N. M. Joy, T. Karmali, A. Pal, and V. Shah, "Fashionable modelling with flux," *CoRR*, vol. abs/1811.01457, 2018.
[48] J. Jalbert, M. Farmer, G. Gobeil, and P. Roy, "Extremes.jl: Extreme value analysis in julia," *Journal of Statistical Software*, vol. 109, no. 6, p. 1–35, 2024. DOI: 10.18637/jss.v109.i06.

Attila Dobai-Pataky received his MScs in data analysis and modeling and didactics of computer science from Babeș-Bolyai University, Cluj, in 2026. He is set to begin his PhD in computer science at Babeș-Bolyai University in the autumn of 2026. His research interests include artificial intelligence, data analysis, computer networks, and computer graphics.

Lehel Csató received his PhD from Aston University, UK; focusing on inference using probabilistic non-parametric methods. He is a professor at the Faculty of Mathematics and Informatics, Babeş-Bolyai University, Romania. He is interested in the mathematical aspects of machine learning, sparse inference, classification of complex data, inverse modelling. With the onset of deep learning, he tries to understand the working of "deep" systems; and the development of explainable deep learning.

László Németh received his MSc in applied mathematics from Eötvös Loránd University (ELTE), Budapest, in 2016, and his PhD in mathematics and computer sciences (applied mathematics) also from Eötvös Loránd University (ELTE) in 2022. He is currently an assistant professor at J. Selye University (Komárno). His research focuses on extreme values, statistics, biostatistics and machine learning.

Balázs Vass received his MSc in applied mathematics from Eötvös Loránd University (ELTE), Budapest, in 2016, and his PhD in informatics from the Budapest University of Technology and Economics (BME) in 2022. He is currently an associate professor at Babeș-Bolyai University. His research focuses on network survivability, combinatorial optimization, and graph algorithms. He has served on the Technical Program Committee of IEEE INFOCOM since 2023 and is a recipient of a 2023 MSCA Postdoctoral Fellowship.