applied sciences

MDPI

Article

Solving Load Balancing Problems in Routing and Limiting Traffic at the Network Edge

Alexander Barkalov [1,2], Oleksandr Lemeshko [3], Oleksandra Yeremenko [3,*], Larysa Titarenko [1,3] and Maryna Yevdokymenko [3]

[1] Institute of Metrology, Electronics and Computer Science, University of Zielona Góra, ul. Licealna 9, 65-417 Zielona Góra, Poland; a.barkalov@imei.uz.zgora.pl (A.B.); l.titarenko@imei.uz.zgora.pl (L.T.)
[2] Department of Computer Science and Information Technology, Vasyl Stus' Donetsk National University, 600-Richchia Str. 21, 21021 Vinnytsia, Ukraine
[3] V.V. Popovskyy Department of Infocommunication Engineering, Kharkiv National University of Radio Electronics, Nauky Ave. 14, 61166 Kharkiv, Ukraine; oleksandr.lemeshko@nure.ua (O.L.); marina.ievdokymenko@nure.ua (M.Y.)
* Correspondence: oleksandra.yeremenko@nure.ua

Abstract: This study focuses on creating and investigating models that optimize load balancing in communication networks by managing routing and traffic limitations. The purpose is to use these models to optimize the network's routing and traffic limitations while ensuring predictable quality of service levels, and adhering to traffic engineering requirements for routing and limiting traffic at the network edge. In order to achieve this aim, a mathematical optimization model was developed based on a chosen optimality criterion. Two modifications of the traffic engineering routing were created, including the linear limitation model (TER-LLM) and traffic engineering limitation (TER-TEL), each considering the main features of packet flow: intensity and priority. The proposed solutions were compared by analyzing various data inputs, including the ratio of flow parameters and the intensity with which packets will be limited at the border router. The study presents recommendations on the optimal use of the proposed solutions based on their respective features and advantages.

Keywords: routing; load balancing; traffic limiting; traffic engineering

Citation: Barkalov, A.; Lemeshko, O.; Yeremenko, O.; Titarenko, L.; Yevdokymenko, M. Solving Load Balancing Problems in Routing and Limiting Traffic at the Network Edge. *Appl. Sci.* **2023**, *13*, 9489. https://doi.org/10.3390/app13179489

Academic Editors: Panagiotis Sarigiannidis, Thomas Lagkas, Alexandros-Apostolos Boulogeorgos, Vasileios Argyriou and Pantelis Angelidis

Received: 22 June 2023
Revised: 19 August 2023
Accepted: 20 August 2023
Published: 22 August 2023

1. Introduction

Ensuring high quality of service (QoS) was and remains the most important task entrusted to modern communication networks. In the increasing network load, a growing variety of contemporary traffic, and QoS requirements, the problem involves improving the efficiency of using available network resources, such as the bandwidth of communication links, queue buffers, and processing time of routers [1,2].

Quality of service is a key characteristic of modern networks, which can be evaluated using a variety of indicators (metrics). Traditional QoS indicators in solving OSI network layer problems are network performance indicators: bandwidth, network delay, jitter, and packet loss probability [1,2]. The nature of the calculated routes largely determines the values of end-to-end QoS indicators and the characteristics of the communication links included in the routes: bandwidth and its utilization. Thus, according to the listed indicators, the current research developed an approach where a balanced communication link utilization improves the QoS level.

As the analysis [3–9] showed, an effective mechanism for solving the given task is the implementation of traffic engineering (TE) principles to ensure the balanced loading of available network resources. For example, routing protocols (OSPF-TE, IS-IS-TE), signaling, and resource reservation, as well as RSVP-TE, modified for traffic engineering requirements, have found their use in MPLS networks [2]. Great attention is also being paid to the development of TE solutions by scientists working on optimizing network solutions [10–22].

Their proposed findings apply to load balancing in software-defined networks (SDNs), cloud environments, mobile networks, queue management on routers and switches, etc. Another group of TE solutions involves developments based on using artificial intelligence to make network decisions [23–27].

The shortcomings of known solutions include implementing a heuristic approach to solving traffic limitation problems based on the token bucket and leaky bucket algorithms [1,2]. In addition, these algorithms do not consider the results of solving routing problems. Although these solutions' goal is common—to combat congestion and ensure the quality of service—new solutions aimed at the coordinated solution of routing and traffic limitation problems are subject to the following requirements: optimality and adaptability to changes in the network state (its structure, load, etc.) and packet flow characteristics (intensity, priority, etc.), as well as ensuring load balancing based on TE principles to guarantee a given QoS level.

Nevertheless, modern communication networks often work in overload conditions when the amount of available network resources is insufficient, which negatively affects the QoS level, and is accompanied by an increase in service failures and traffic limitation at the network edge. However, the delays and packet loss increases are unacceptable for specific flows. Therefore, it is essential to ensure the predictability of the QoS level and controllability of limiting traffic entering the network, which should be implemented through maximum coordination by routing tools and traffic limitations. The desired effect is possible using unified optimization mathematical models that describe the relation of such critical network processes. Specific solutions in this direction were proposed in studies [21,22] and modified, for example, under the problem of fault-tolerant routing. However, already-known mathematical models need further improvement to increase the adaptability of load balancing solutions to changes in the network state, flow parameters, requirements for the level of differentiated quality of service, etc.

Therefore, in the current study, a mathematical optimization model is proposed, which received two modifications depending on the chosen type of optimality criterion. Each of these modifications aims to provide solutions for load balancing within the routing, and for traffic limitation problems based on different principles relating to the packet flow parameter consideration. The primary purpose of the research is to ensure the predictability of the QoS level and manageability under traffic engineering requirements by routing and load limiting at the network edge.

The remainder of this study is structured as follows. Section 2 is devoted to related work analysis. Section 3 defines a basic traffic engineering routing and limiting model in a communication network. Section 4 presents the optimality criteria selection for load balancing problem solutions. Section 5 contains the numerical research of studies on network topology to evaluate and compare the effectiveness of network load balancing solutions based on various mathematical models and optimality criteria. In turn, Section 6 discusses the obtained research results, and offers recommendations regarding applying the proposed load balancing models inside and at the edge of a communication network. Section 7 concludes the study.

2. Related Research Analysis

The most current TE applications are intended for software-defined networking [3,5,7–16,18,21,23,26–28]. While a significant number of solutions found their field of use in wireless networks [4,6,19,20,23], Internet of Things technology [18,24,28], smart cities [8,18,23], edge, and fog computing [8,19,28]. Particularly noteworthy are the developments of complex (joint) solutions for TE-based load balancing problems and ensuring network fault tolerance [8,17,21,22].

The following overview was performed based on analyzing existing TE solutions research results in promising telecommunication networks (Table 1).

Table 1. Existing TE applications overview.

Ref.	Contribution	Underlying Approach	Field of Application
[10]	Legacy networking devices' gradual deployment to SDN switches, using traffic engineering measures such as minimizing the highest link utilization. Doing so can identify the most appropriate devices to migrate first, thereby determining an optimal deployment sequence	Optimization	Hybrid IP/SDN
[11]	Simultaneous optimization of both traffic matrix measurement (TMM) and traffic engineering (TE) while considering the constraints posed by TCAM capacity and flow aggregation, enabling a substantial improvement in TMM accuracy and TE effectiveness	Optimization	SDN
[12]	Development of traffic engineering-aware distributed routing (TEDR) algorithm, which maximizes link utilization comparable to full SDN considering TCAM resource limitation	Optimization	Hybrid SDN
[13]	Joint mathematical formulation to solve load balancing challenges in cloud computing; two multi-objective particle swarm optimization (MP) models: distance angle multi-objective particle swarm optimization (DAMP) and angle multi-objective particle swarm optimization (AMP); incorporation of meta-heuristic in the cloud networks management layer	Optimization	SDN, Cloud Computing
[14]	Dynamic load balancing (LB) scheme integrating genetic algorithm (GA) and ant colony optimization (ACO); fast global search of GA and efficient search of an optimal solution of ACO are achieved; improvement of the round robin and ACO algorithm with optimal path search rate, round trip time, and packet loss rate	Optimization	SDN
[15]	Development and implementation of the energy-aware routing multi-level and mapping problem (EARMLP) algorithm to minimize the overall power consumption; optimal routing strategy that considers system configuration and traffic demand between the data and control planes; establishment of the controller placement problem (CPP) to select the optimal locations and controller numbers	Optimization	SDN-based core networks
[16]	Mixed integer programming algorithm designed to optimize the network devices' power consumption utilizing energy-aware traffic engineering	Optimization	SDN Data Centers
[17]	Load-balanced and fast failure recovery solution to provision routing paths so a Fibbing network can apply loop-free alternate (LFA) in the network in a case of a single node or single link failure	Optimization	SDN
[18]	Multiple distributed controller load balancing (MDCLB) algorithm on an immense-scale SDN-IoT for smart cities	Optimization	IoT, SDN, Smart City
[19]	Task allocation in the mobile edge computing (MEC) scenario of ultra-dense network based on routing between MEC servers; load balancing algorithm based on load estimation by user load prediction	Optimization, Genetic Algorithm	SDN, Edge Computing
[20]	Joint optimization algorithm for enhancing the performance and traffic load balancing of the wireless network; the objectives are data transmission latency, the energy consumption of wireless microbase stations, and links' throughput	Optimization, Heuristic Scheme	Wireless Network
[21]	Traffic Engineering Fast ReRoute with support of Traffic Policing (TE-FRR-TP); optimality criterion focuses on minimizing the dynamically controlled upper bound of links utilization and the intensity of flows that receive a denial of service at the network edge weighted with the priority of serving	Optimization	SD-WAN
[23]	Load balancing mechanism based on SDWSN (Software-Defined Wireless Sensor Network); load-balanced routing improvement based on Elman neural network—SDSNLB (Software-Defined Sensor Network Load Balancing) routing algorithm	Optimization, Elman neural network	SDWSN, Smart City
[24]	Optimization model with mobile edges for multimedia sensors using artificial intelligence of things (AIoT), which aims to maintain real-time data collection with low resource consumption	Optimization, AI	SDN, AIoT, MIoT
[25]	Development of the deep reinforcement learning (DRL)-based TE scheme of multipath transmission to dynamically adjust the traffic splitting ratio among different paths based on the network traffic distribution in IP and segment routing (SR) hybrid network	Deep Reinforcement Learning	Hybrid IP/SR network

Table 1. *Cont.*

Ref.	Contribution	Underlying Approach	Field of Application
[26]	Proposal of a reinforcement learning (RL) based switch and controller selection scheme for switch migration, switch-aware RL load balancing (SAR-LB) under the utilization ratio of various resource types in controllers and switches as the inputs of the neural network	Reinforcement Learning	SDN
[27]	QoS-aware adaptive routing protocol (SQAR) based on reinforcement learning, which can intelligently select routes to satisfy the QoS requirements of multiple Internet of Underwater Things (IoUT) services	Reinforcement Learning	IoUT
[28]	Development of secure and energy-aware fog computing architecture; load balancing technique improving the complete resources utilization in SDN-based fog environment; implementation of deep belief network (DBN)-based intrusion detection method reducing workload communication delays in the fog layer	Heuristic algorithms	SDN, IoT, Fog Computing

An analysis of current research regarding TE implementation in different types of networks shows that the most promising approaches are grounded in optimization methods. Particular attention should be paid to solutions combining optimization and artificial intelligence [23,24]. Moreover, one of the features of implementing load balancing based on TE is its use in networks with edge/fog computing. Consequently, the proposed approach in this study can be effectively applied to managing routing and load limitation at the network's edge in different types of networks, mainly based on Software-Defined Networking architectures.

3. Basic Model of Traffic Engineering Routing and Limiting

We describe the network model and its structural and functional parameters according to the notations introduced in [21,22], and the new ones used in the current research. The primary notations used in the models are presented in Table 2.

Table 2. Notation summary.

Symbol	Meaning
$G = (R, E)$	Network structure graph
$R = \{R_i;\ i = \overline{1,m}\}$	Set of vertices (routers)
$E = \left\{E_{i,j};\ i,j = \overline{1,m};\ i \neq j\right\}$	Set of edges (network links)
n	Number of links
m	Number of nodes
K	Set of packet flows in the network
s_k	Source node of the kth packet flow ($k \in K$)
d_k	Destination node of the kth packet flow ($k \in K$)
λ^k	Average packet intensity of the kth flow (packets per second, pps)
$\varphi_{i,j}$	Link bandwidth (packets per second, pps) between the ith and jth nodes ($i,j = \overline{1,m};\ i \neq j$)
β^k	Proportion of the kth flow intensity that receives a denial of service when using the multipath
b^k	kth flow intensity, with which packets will be limited on the border router
α	Upper bound of the network links utilization
α_{TH}	Threshold of the network links utilization upper bound α
v^k	Weighting coefficient responsible for the packet flows limitation at the network edge

Table 2. *Cont.*

Symbol	Meaning
c	Weighting coefficient responsible for the load balancing in the network
PR^k	kth packet flow priority
γ	Threshold for flow rate limiting that enters the network via border routers
w	Weighting coefficient ($w > 1$), which determines how many times the threshold γ is more important than the utilization upper bound α

Thus, the graph $G = (R, E)$ describes the network structure under the sets of vertices (routers) R and edges (network links) E. Then, K packet flows circulate in the network. Namely, the kth packet flow ($k \in K$) is transmitted from the source node s_k to the destination node d_k, with an average packet intensity λ^k.

In general, the result of solving the routing problem is the determination of the routing variables $x_{i,j}^k$ that characterize the fraction of the kth flow intensity in the communication links belonging to the multipath. When using multipath routing, the following restrictions are imposed on these variables [21,22]:

$$0 \le x_{i,j}^k \le 1. \tag{1}$$

The following flow conservation conditions are introduced to ensure the routes' connectivity in the whole network, which differ from the previously known ones that allow description of traffic policing at the network edge under conditions of the overload [21,22]:

$$\begin{cases} \sum\limits_{j:E_{i,j}\in E} x_{i,j}^k - \sum\limits_{j:E_{j,i}\in E} x_{j,i}^k = 0;\ k \in K,\ R_i \ne s_k, d_k; \\ \sum\limits_{j:E_{i,j}\in E} x_{i,j}^k - \sum\limits_{j:E_{j,i}\in E} x_{j,i}^k = 1 - \beta^k;\ k \in K,\ R_i = s_k; \\ \sum\limits_{j:E_{i,j}\in E} x_{i,j}^k - \sum\limits_{j:E_{j,i}\in E} x_{j,i}^k = \beta^k - 1;\ k \in K,\ R_i = d_k. \end{cases} \tag{2}$$

The following restrictions are imposed on variables β^k according to their physical content:

$$0 \le \beta^k \le 1. \tag{3}$$

We denote by $b^k = \lambda^k \cdot \beta^k$ the intensity (rate) of the kth flow, with which packets will be limited (rejected) on the border router.

The conditions for preventing overloading and providing load balancing in the network [21] have the following form:

$$\sum_{k\in K} \lambda^k \cdot x_{i,j}^k \le \alpha \cdot \varphi_{i,j}\ \left(E_{i,j} \in E\right), \tag{4}$$

where α is the additional control variable that obeys the following conditions:

$$0 \le \alpha \le \alpha_{TH}, \tag{5}$$

where α_{TH} is the α threshold. The network requirements for the quality of service level determine its value.

4. Optimality Criteria for Load Balancing Problem Solutions

The calculation of the control variables responsible for routing (1), load balancing based on the principles of traffic engineering (5), and limiting traffic at the network edge (3)

is proposed in the process of solving the optimization problem, as it is carried out in studies [21,22]. However, the quality of applied solutions traditionally depends on the form and content of the selected optimality criterion. In this research, various optimality criteria are proposed and investigated, considering the relation between control variables and packet flows parameters (priorities and intensities) in different ways.

The first type of optimality criterion is based on the linear objective function minimum as follows:

$$J = \sum_{k \in K} v^k \cdot \beta^k + c \cdot \alpha \to \textbf{min}, \quad (6)$$

where v^k and c are weighting coefficients that should consider the packet flows' parameters routed and limited at the network edge. For traffic limitation at the network edge to occur only after the bound reaches its maximum value, i.e., at $\alpha = \alpha_{TH}$, the next conditions should be met when choosing weighting coefficients:

$$v^k > c \ (k \in K). \quad (7)$$

The solution, which is based on the use of criterion (6) and models (1)–(5), will be called "***Traffic Engineering routing, linear limitation model***" (***TER-LLM***) later in the study.

This study will consider several options for forming the weighting coefficients v_k. The ***first option*** is based on the use of such weighting coefficients:

$$v^k = PR^k + 1 \ (k \in K), \quad (8)$$

where PR^k is the kth packet flow priority. In an IP network using the 3 bits of IP precedence in the packet header for prioritization, the priority value ranges from 0 to 7. In contrast, for the DSCP (differentiated services code point), the priorities vary from 0 to 63 [29].

Let us consider in more detail the physical sense of the optimality criterion (6) when coefficients (8) are presented in the following form:

$$v^k = \frac{(PR^k + 1) \cdot \lambda^k}{\lambda^k}.$$

Then, the optimality criterion (6) can be rewritten as follows:

$$J = \sum_{k \in K} \frac{(PR^k + 1) \cdot \lambda^k \cdot \beta^k}{\lambda^k} + c \cdot \alpha \to \textbf{min}.$$

Given that $b^k = \lambda^k \cdot \beta^k$, we have the following:

$$J = \sum_{k \in K} \frac{(PR^k + 1) \cdot b^k}{\lambda^k} + c \cdot \alpha \to \textbf{min}.$$

Therefore, when using (6) and (8), flow limitation will be carried out according to their intensity (b^k) weighted directly and proportionally relative to the packet flows' priorities (PR^k), and inversely proportional relative to the kth flow intensity (λ^k). On the other hand, we can say that the flow limitation will be carried out according to its fraction (β^k) weighted in direct proportion to the packet flow priorities (PR^k).

Within the scope of the ***second option***, weighting coefficients of the following type are used:

$$v^k = (PR^k + 1) \cdot \lambda^k \ (k \in K). \quad (9)$$

From (9), it becomes clear that, in this case, the limitation in the set of flows will be carried out according to their intensity (b^k) weighted in direct proportion to the priorities of the packet flows (PR^k). That is, priority flows should be limited less intensively. On the other hand, it can be said that the limitation of the flows will be carried out according to their fraction (β^k) weighted directly and proportionally relative to the priorities of packet

flows (PR^k), and to the intensity of kth flow packets (λ^k). That is, the share of service denials will be lower for those flows with a higher priority and intensity.

The peculiarity of optimality criterion (6) applications, which operate under the constraints and conditions (1)–(5), is that load balancing is ensured within the network based on traffic engineering principles. At the same time, traffic limitation at the network edge is carried out within a linear model, considering the intensities and packet flow priorities. As shown in studies [21,22], with the help of these solutions, it is possible to provide a more intensive limitation of flows that were a source of overload in terms of their intensity, and that had the lowest priority.

In the model proposed in this study and named "***Traffic Engineering routing, Traffic Engineering limitation***" (***TER-TEL***), expressions (1)–(5) should be supplemented with the following condition:

$$v^k \cdot \beta^k \leq \gamma, \tag{10}$$

where γ is the additional control variable that characterizes the load limitation threshold entering the network via border routers.

The introduction of expression (10) makes it possible to give the process of the load limitation at the network edge a balanced character under TE principles. Furthermore, depending on the chosen expression (8) or (9) for determining the v^k coefficients, it is possible to ensure that the packet flows' parameters are considered, namely, their intensities and priorities. At the same time, it is advisable to choose the following condition as an optimality criterion for the decisions regarding routing and traffic limitation based on the TE principles:

$$J = w \cdot \gamma + \alpha \rightarrow \mathbf{min}, \tag{11}$$

where w is the weighting coefficient ($w > 1$) that determines how much the load limitation threshold γ is more important than the upper bound of the network links utilization α.

Consequently, when using the optimality criterion (11) in the presence of restrictions (1)–(5) and (10), the traffic management process will be balanced both from the point of view of the TE routing within the network, and the application of TE traffic limitation on the network edge.

5. Numerical Research

During the preliminary research, various topologies were used that concerned the number and order of routers' connections, different sets of characteristics of communication links, and packet flows. A common feature of many studies is the use of mesh class network topologies [30], which include well-known topologies such as NSF, German network, GEANT2, and other similar structures. In this research, we have chosen the mesh topology (Figure 1), which is generally more complex, as it contains even more routing options, and supports more network load balancing options.

Several studies on different network topologies were conducted to evaluate and compare the effectiveness of network load balancing solutions based on various mathematical models and optimality criteria. One example of such a network topology is shown in Figure 1. The network consists of 16 routers and 24 communication links, in the gaps of which their bandwidths are indicated. Let packets of two flows be transmitted in the network from the first router to the sixteenth. The first packet flow had the fifth IP priority ($PR^1 = 5$), and the second one had the second IP priority ($PR^2 = 2$). The intensity of each packet flow varied from 10 to 900 pps.

The simulation was carried out in the MATLAB R2020b environment. The formulated problem was solved on a real-time scale.

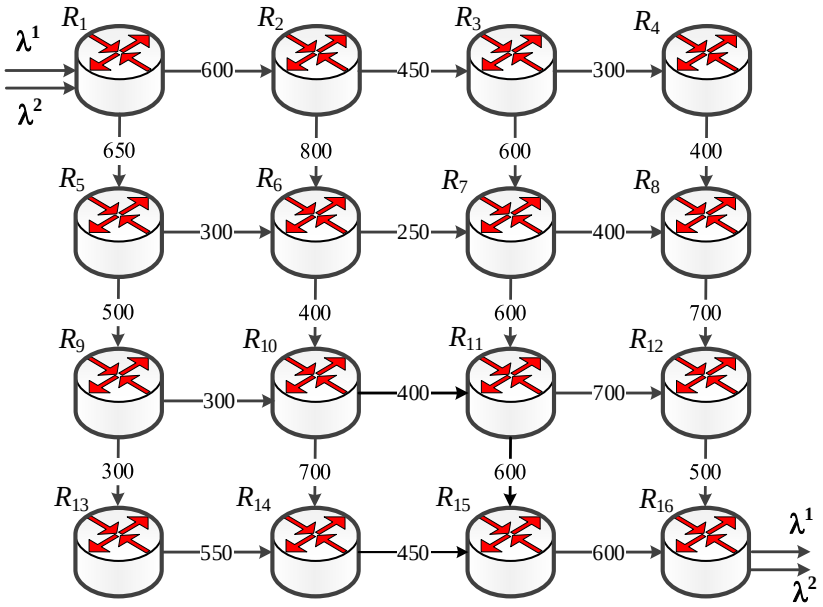

Figure 1. Network structure under the numerical study.

5.1. Investigation of the TER-LLM and TER-TEL Models

At the first stage of research, the order of load balancing was compared, which was provided by the use of the two proposed optimality criteria (6) and (11) with the weighting coefficients (8) in the corresponding objective function. Thus, Figure 2 shows the dependence of the upper bound of the network link utilization α on the packet flow intensities when using different models—TER-LLM and TER-TEL when $\alpha_{TH} = 0.7$.

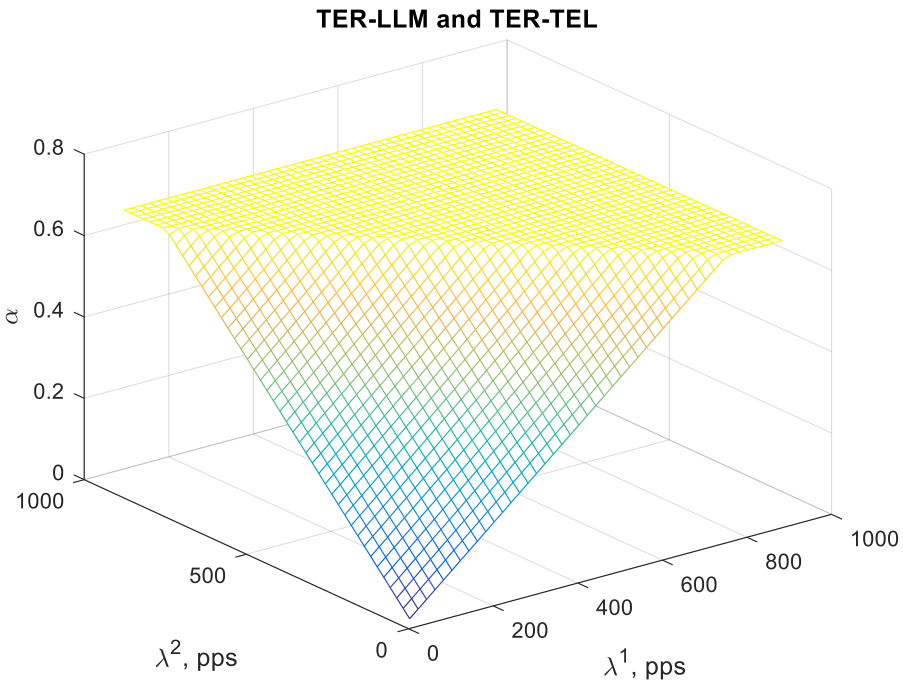

Figure 2. Dependencies of the network links utilization upper bound α on the packet flow intensities when using weighting coefficients (8) in different models—TER-LLM and TER-TEL when $\alpha_{TH} = 0.7$, $PR^1 = 5$, and $PR^2 = 2$.

From Figure 2, it can be seen that at high values of flow intensities λ^1 and λ^2, the upper bound of the network links utilization stabilized at the level $\alpha = \alpha_{TH} = 0.7$, which is a

common pattern for all types of studied models. However, the processes of limiting various types of packet flows, which differed in IP priorities and intensities, diverged significantly.

Hence, Figure 3 for the TER-LLM model shows the dependencies of the fractions (β^1 and β^2) and flow intensities (b^1 and b^2) with which these packet flows will be limited (rejected) at the border router. If the fractions and flow intensities (β^k and b^k) were equal to zero, it means that the TE routing tools coped with the load entering the network. With a further increase in the network load, its resources, first of all, the bandwidth, were insufficient to ensure that constraint (5) was met at $\alpha_{TH} = 0.7$, so the load limitation process was initiated on the border routers.

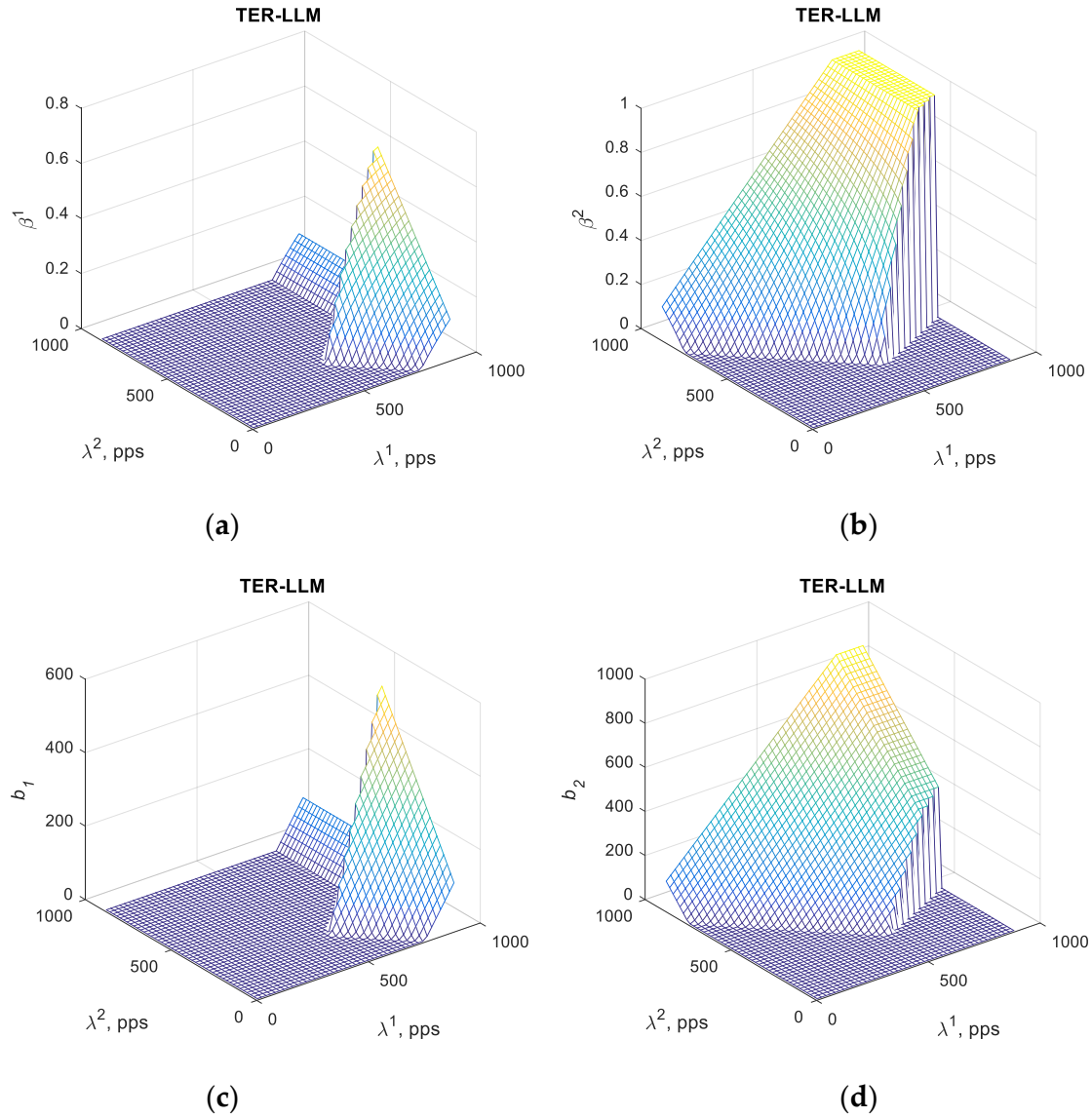

Figure 3. The procedure for limiting different packet flow types when using the TER-LLM model and weighting coefficients (8), when $\alpha_{TH} = 0.7$, $PR^1 = 5$ and $PR^2 = 2$: (**a**) the proportion of the first flow intensity that receives a denial of service; (**b**) the proportion of the second flow intensity that receives a denial of service; (**c**) the limitation intensity of the first flow at the border router; (**d**) the limitation intensity of the second flow at the border router.

Figure 3 demonstrates that when using the TER-LLM model and weighting coefficients (8), the lower-priority second packet flow is limited more intensively. However, if the high-priority first flow is the source of the overload in terms of its intensity, then it will

be limited in the first place without balancing. Therefore, both flow priorities and their intensities affect traffic limitations.

Then, Figure 4 demonstrates the procedure for limiting different types of packet flows when using the TER-LLM model and weighting coefficients (8) when the difference in the IP priorities of the flows was maximal: $PR^1 = 7$ and $PR^2 = 0$. Figure 4 reveals that when the difference in IP priorities of flows increases, their limitation will minimally depend on flow intensities, and will be determined entirely by the ratio of IP priorities. The low-priority packet flow ($PR^2 = 0$) was more intensively limited in this case. The high-priority (first) packet flow was limited after the low-priority flow was completely limited, or when the intensity of the high-priority flow was a maximum (close to 900 pps) and the intensity of the low-priority flow was a minimum (close to 10 pps).

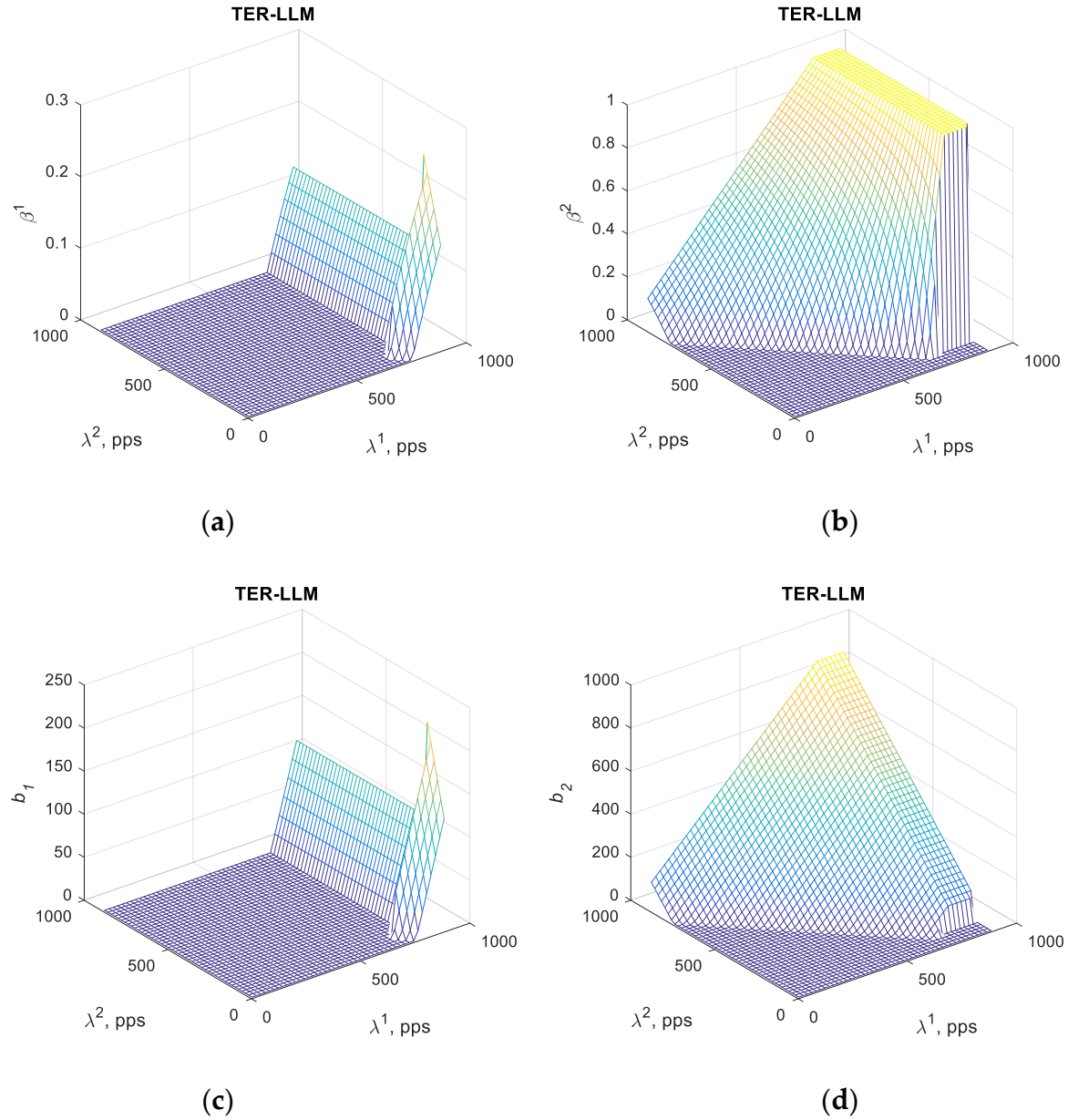

Figure 4. The procedure for limiting different packet flow types when using the TER-LLM model and weighting coefficients (8), when $\alpha_{TH} = 0.7$, $PR^1 = 7$, and $PR^2 = 0$: (**a**) the proportion of the first flow intensity that receives a denial of service; (**b**) the proportion of the second flow intensity that receives a denial of service; (**c**) the limitation intensity of the first flow at the border router; (**d**) the limitation intensity of the second flow at the border router.

Conversely, when the flows' IP priorities were almost the same ($PR^1 = 4$ and $PR^2 = 3$), the differentiation in limitation was determined only by their intensities (Figure 5), i.e., whichever flow had a higher intensity, it was restricted more strongly.

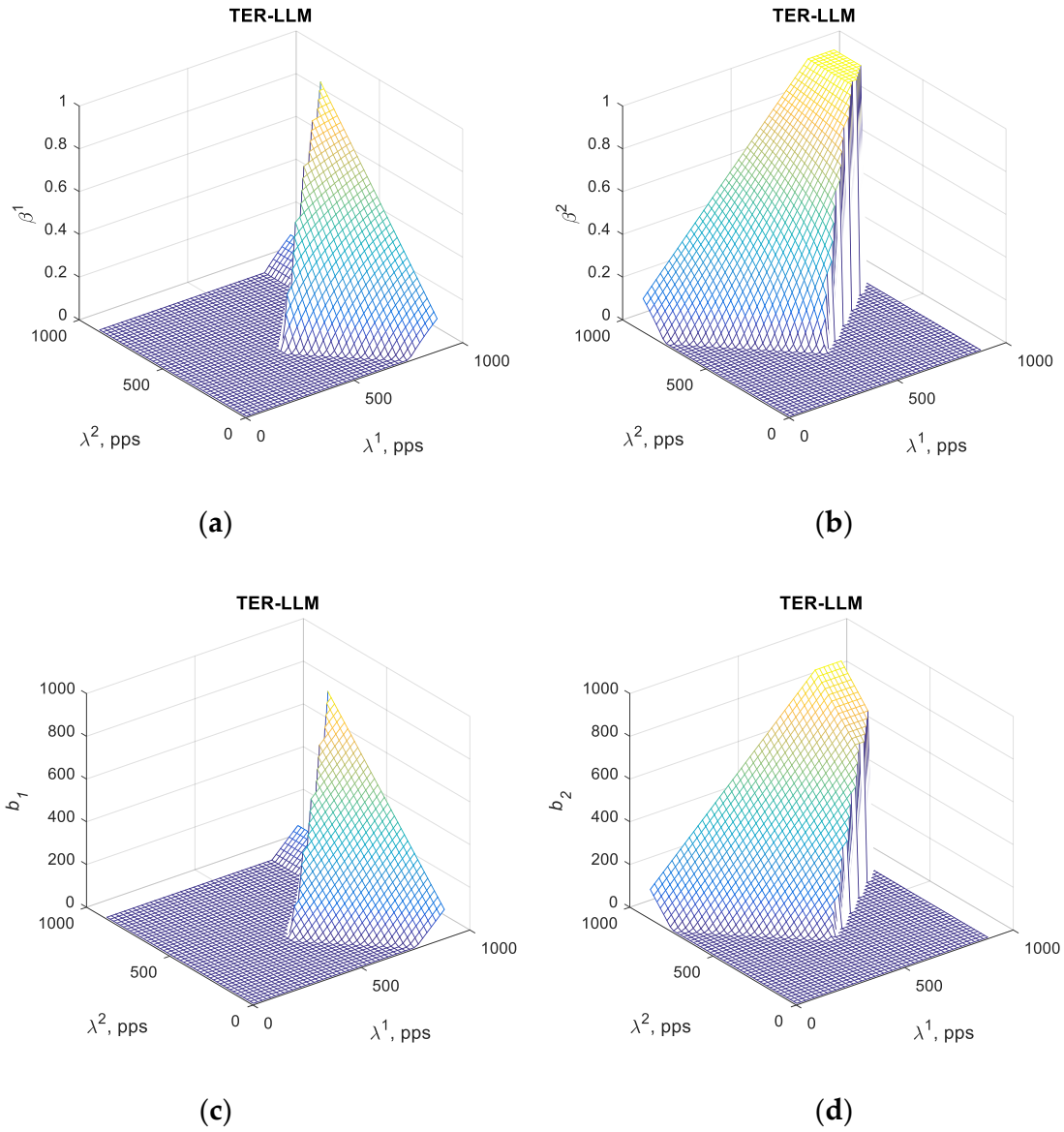

Figure 5. The procedure for limiting different packet flow types when using the TER-LLM model and weighting coefficients (8), when $\alpha_{TH} = 0.7$, $PR^1 = 4$, and $PR^2 = 3$: (**a**) the proportion of the first flow intensity that receives a denial of service; (**b**) the proportion of the second flow intensity that receives a denial of service; (**c**) the limitation intensity of the first flow at the border router; (**d**) the limitation intensity of the second flow at the border router.

When using the TER-TEL model and weighting coefficients (8), the order of traffic limitation is presented in Figure 6. In this way, limiting the load entering the network was more balanced than when using the TER-LLM model (Figure 6), but also considering both flows' IP priorities and intensities. In the present case, when the network was overloaded, both flows were limited at the network edge at once. However, the flow of packets with a higher intensity (i.e., its contribution to the overload was more significant) and a lower IP priority was limited more.

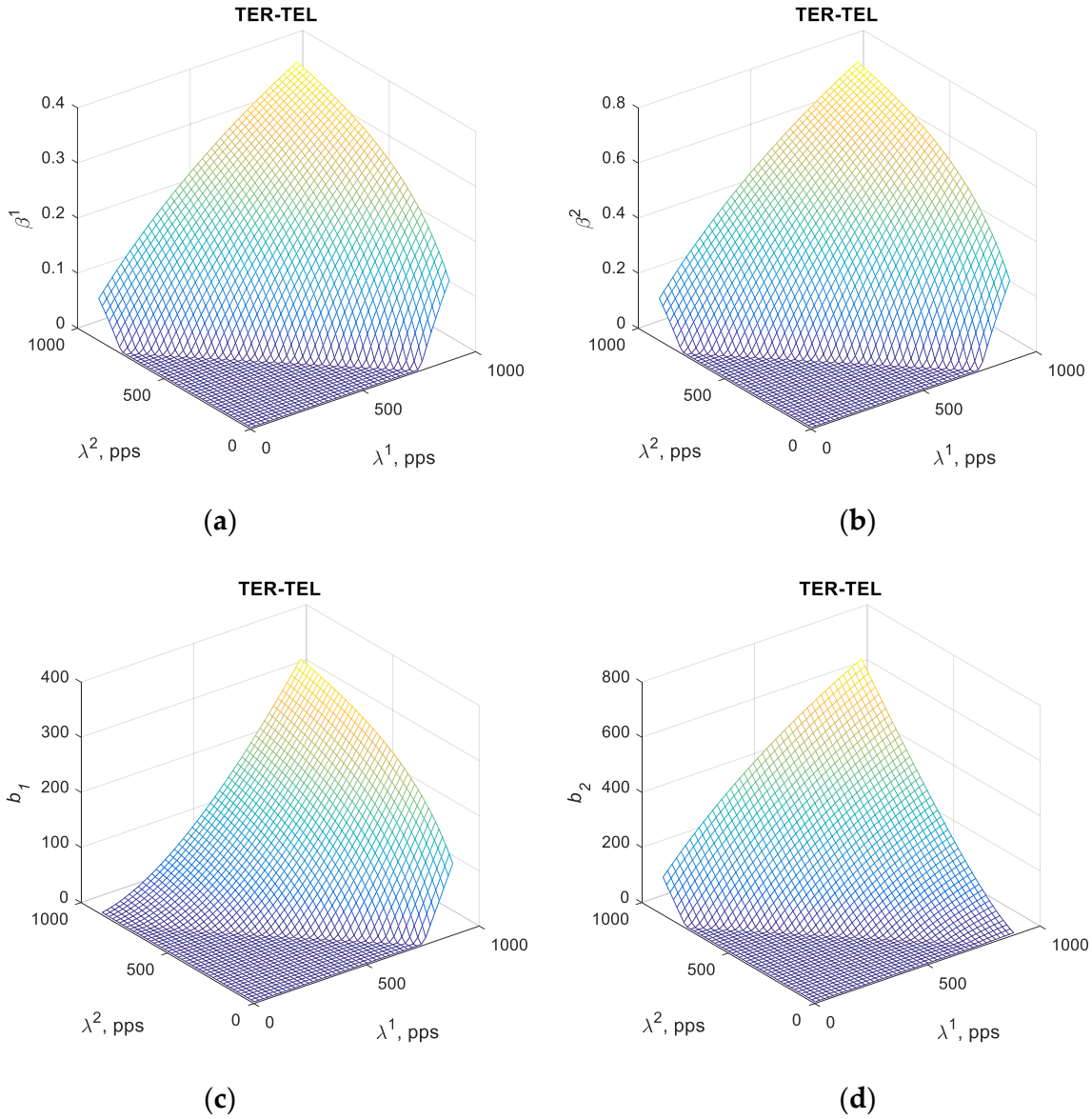

Figure 6. The procedure for limiting different packet flow types when using the TER-TEL model and weighting coefficients (8), when $\alpha_{TH} = 0.7$, $PR^1 = 5$, and $PR^2 = 2$: (**a**) the proportion of the first flow intensity that receives a denial of service; (**b**) the proportion of the second flow intensity that receives a denial of service; (**c**) the limitation intensity of the first flow at the border router; (**d**) the limitation intensity of the second flow at the border router.

The load balancing order was again compared at the second research stage. It was provided under the application of the two proposed optimality criteria (6) and (11), but with weighting coefficients (9) for the TER-LLM and TER-TEL solutions. At the same time, the dependence character of the upper bound of the network link utilization on the packet flows' intensities when using different models, TER-LLM and TER-TEL, with weighting coefficients (9), when $\alpha_{TH} = 0.7$, $PR^1 = 5$, and $PR^2 = 2$, entirely coincides with Figure 2. However, the nature of failures when using weighting coefficients (9) was significantly different (Figures 7 and 8) from the results obtained when using weighting coefficients (8) in optimality criteria (6) and (11).

Figure 7 for the TER-LLM model indicates the procedure for limiting the load in the network using weighting coefficients (9). Figure 7 shows that using the TER-LLM model and weighting coefficients (9) leads to the primary limitation of low-priority flows ($PR^2 = 2$). Only when the low-priority flow is wholly blocked will the higher-priority flow

($PR^1 = 5$) begin to be limited. Thus, within the TER-LLM solution, balancing is supported only within the network when solving routing problems, and balancing is not supported when limiting the load.

Figure 8, as an example, demonstrates the research results on the TER-TEL model.

As shown in Figure 8, the limitations of different types of flows are carried out in a balanced way, considering only their priorities. In the case of overload, the low-priority flow ($PR^2 = 2$) was more intensively limited, and the high-priority flow ($PR^1 = 5$) was restricted, but less intensively. In contrast to the use of weighting coefficients (8), in this case, the flow intensity does not significantly affect the order of traffic limitation. That is, the identification of the overload source is not supported.

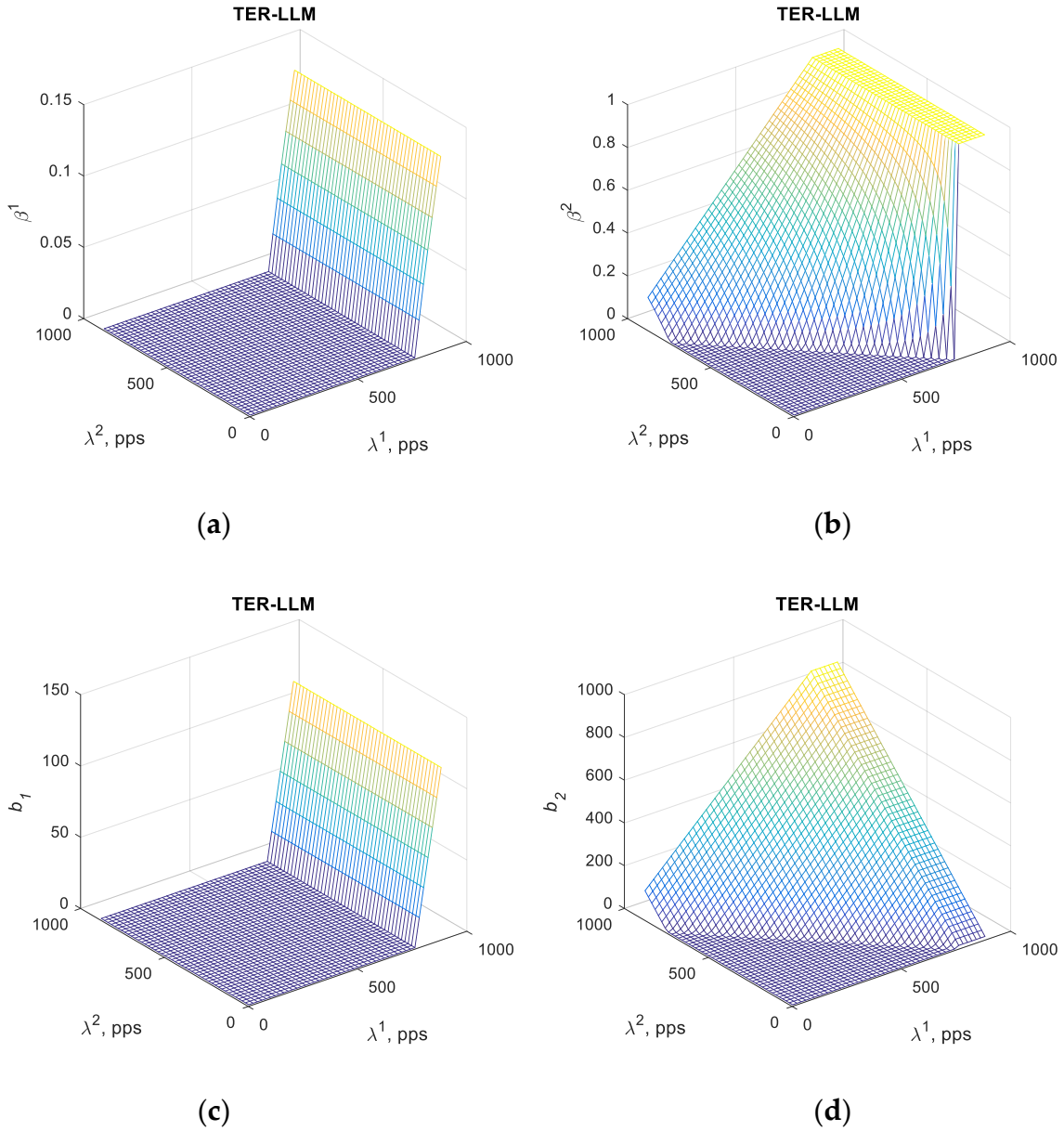

Figure 7. The procedure for limiting different packet flow types when using the TER-LLM model and weighting coefficients (9), when $\alpha_{TH} = 0.7$, $PR^1 = 5$, and $PR^2 = 2$: (**a**) the proportion of the first flow intensity that receives a denial of service; (**b**) the proportion of the second flow intensity that receives a denial of service; (**c**) the limitation intensity of the first flow at the border router; (**d**) the limitation intensity of the second flow at the border router.

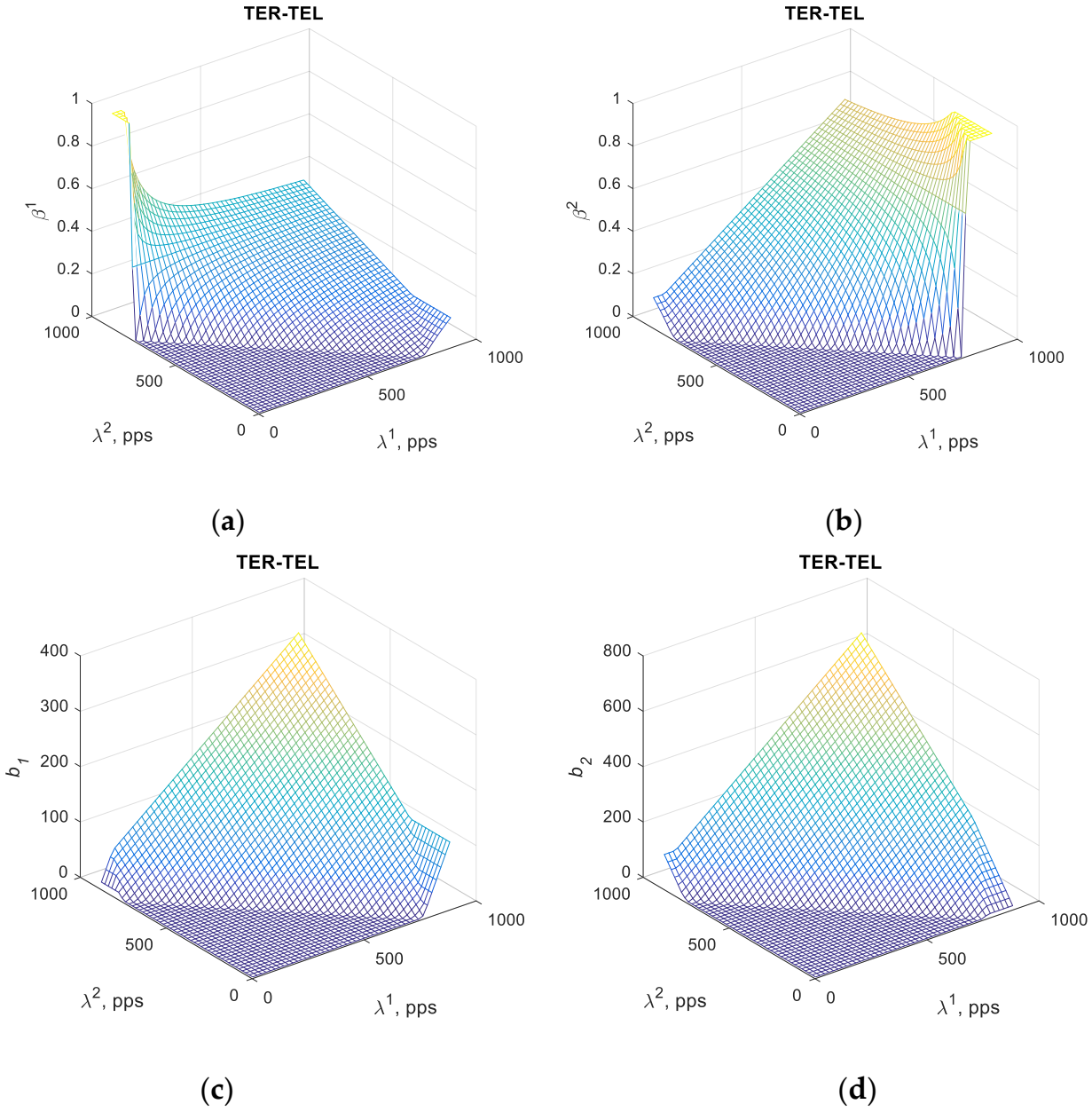

Figure 8. The procedure for limiting different packet flow types when using the TER-TEL model and weighting coefficients (9), when $\alpha_{TH} = 0.7$, $PR^1 = 5$, and $PR^2 = 2$: (**a**) the proportion of the first flow intensity that receives a denial of service; (**b**) the proportion of the second flow intensity that receives a denial of service; (**c**) the limitation intensity of the first flow at the border router; (**d**) the limitation intensity of the second flow at the border router.

5.2. Investigation of the OSPF-TS Model

To analyze the proposed models' effectiveness, we compared them with a model based on known technological solutions. It was taken into account that firstly, in practice, for example, in IP networks, decisions to limit traffic on border routers, unfortunately, are not coordinated with routing decisions in any way. Secondly, IP routing protocols cannot guarantee the maximum level of the network and its links utilization. Thirdly, traffic shaping (TS) mechanisms equalize and limit traffic according to the previously agreed rate of packets entering the network through mostly static settings.

Therefore, the ***OSPF-TS*** model was chosen for the compared technological solution, based on reasonably common traffic management tools in IP networks—the OSPF routing

protocol and the traffic shaping mechanism at the network edge [1]. The OSPF IP routing protocol supported uniform load balancing along paths with the same metric, influenced by the number of links in the route and their bandwidth. The traffic shaping mechanism limited the intensity of the aggregated flow to ensure that condition (5) was met.

As shown by the study results of the selected network topology (Figure 1), the use of the OSPF protocol functionality, its routing metrics, and balancing schemes has led to a change in the utilization of communication links and the network as a whole (Figure 9). Compared to Figure 2, the changes concerned the value of the total load when the link utilization reached the threshold level $\alpha_{TH} = 0.7$. When using the TER-LLM and TER-TEL models, traffic limitation began at a network load of 760 pps, whereas when using the OSPF-TS model, the limitation began much earlier, at a lower network load of 320 pps. Consequently, balancing based on the principles of traffic engineering within the TER-LLM and TER-TEL models was more effective than load balancing using the OSPF protocol. This made it possible to increase the network performance (the amount of load served without failures) by almost 2.4 times, subject to condition (5).

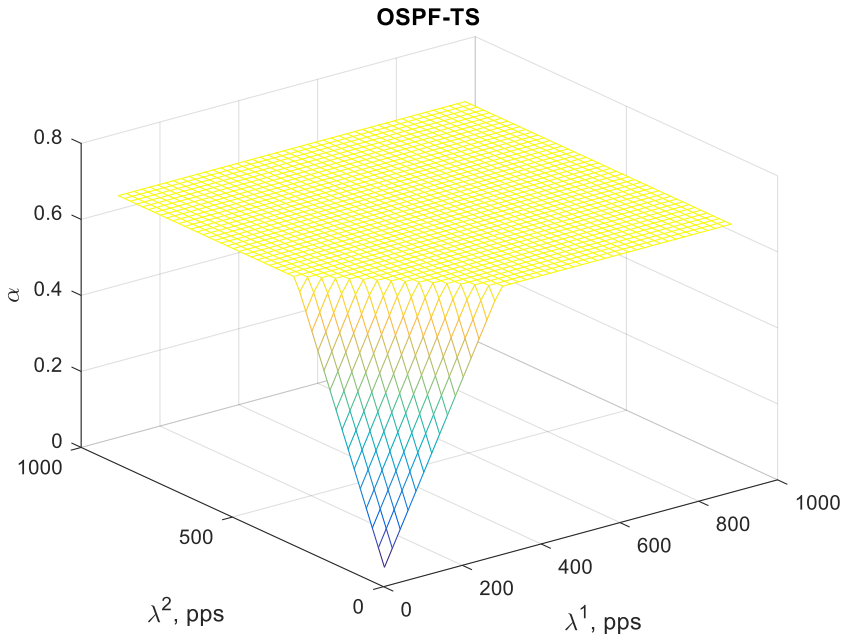

Figure 9. Dependencies of the network links utilization upper bound α on the packet flow intensities when using OSPF-TS model, when $\alpha_{TH} = 0.7$, $PR^1 = 5$, and $PR^2 = 2$.

The application of traffic shaping led to a uniform intensity limitation of the analyzed packet flows entering the network (Figure 10). The limitation uniformity was manifested in the fact that for both flows, the proportions of their intensities that received a denial of service (β^1 and β^2) were always the same (Figure 10a,b) because traffic shaping worked with the aggregated packet flow, and the limitation rates differed (Figure 10c,d). Thus, it can be concluded that when using the OSPF-TS model for load balancing, its key characteristics (neither priority nor intensity) were taken into account, which in practice leads to ignoring the requirements for DiffServ and localization of the overload source.

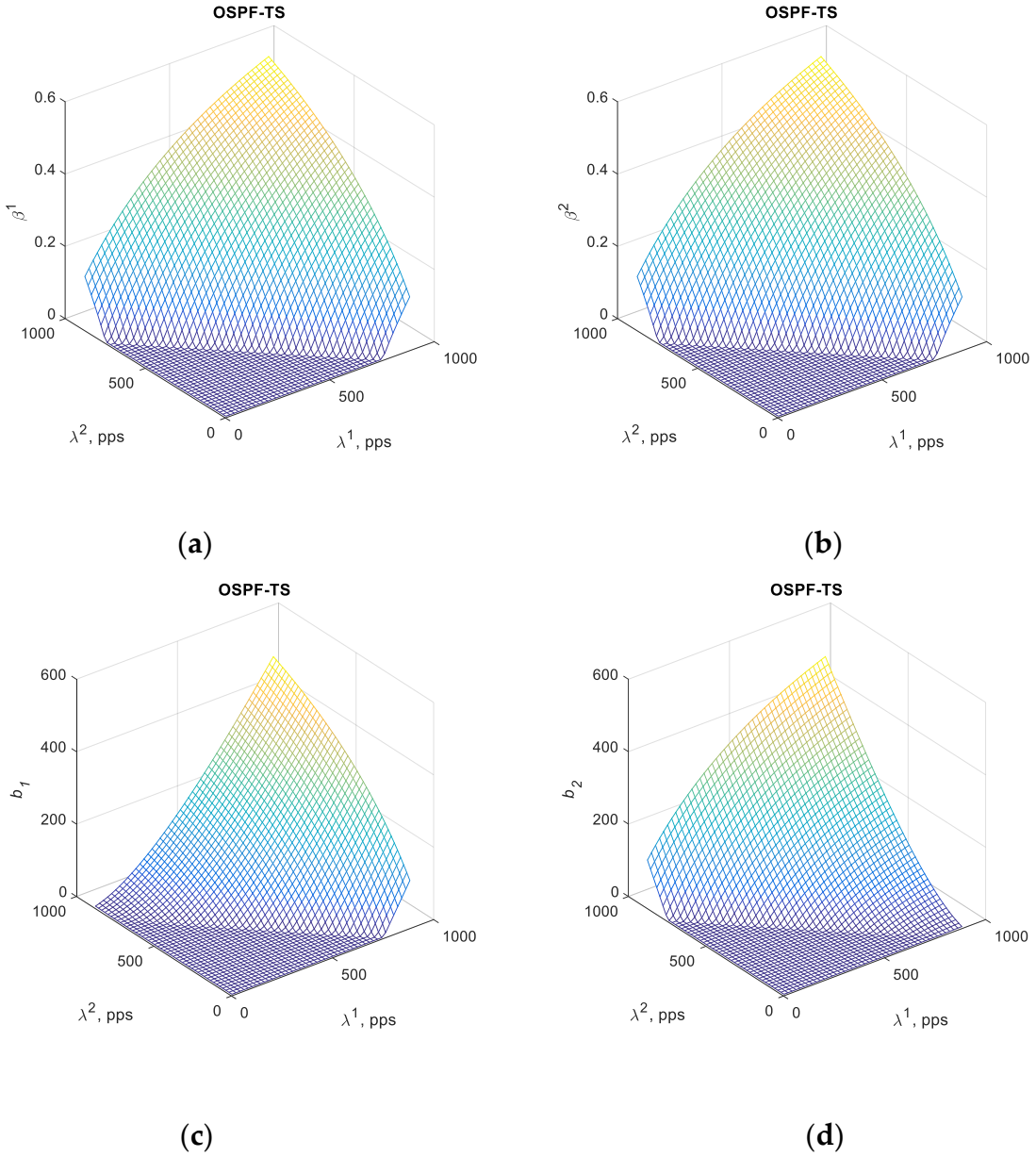

Figure 10. The procedure for limiting different packet flow types when using the OSPF-TS model, when $\alpha_{TH} = 0.7$, $PR^1 = 5$, and $PR^2 = 2$: (**a**) the proportion of the first flow intensity that receives a denial of service; (**b**) the proportion of the second flow intensity that receives a denial of service; (**c**) the limitation intensity of the first flow at the border router; (**d**) the limitation intensity of the second flow at the border router.

6. Discussion

The following recommendations regarding applying the proposed load balancing models inside the network and at its edge can be offered based on the research results. The general recommendations include considering various traffic parameters when limiting it systematically, including the IP priorities of packet flows and their intensities. This is important from the point of view of ensuring DiffServ QoS and combating network overload.

Thus, when using the technology of absolute priorities, when it is necessary to guarantee the service of high-priority traffic, it is appropriate to use the TER-LLM solution with weighting coefficients (9) in optimality criterion (6). During the implementation of this solution, as the research results showed (Figure 7), that packet flows with the highest IP priority are limited by their rate at the network edge last.

A more adaptive solution to limit the load at the network edge is the TER-LLM model with weighting coefficients (8) in optimality criterion (6). This solution generally supports the previous version's functionality, but also considers the packet flow rate arriving on the network. As the research results showed (Figures 3–5), when the upper bound of the network link utilization reaches the threshold value $\alpha = \alpha_{TH}$, this solution implements sequential load balancing in the event of overloading. The packet flow that caused the overload was primarily limited in two prominent cases: first, in the load areas where the packet flow rates differed quite a lot (Figure 3); and second, in the case where the IP priorities of the flows were almost the same (Figure 5). When the flow rates were equalized, the restriction was implemented based on analyzing their IP priorities (Figure 3). On the other hand, the influence of flow rates on the limitation process was almost not felt at the maximum difference in the values of their IP priorities (Figure 4), which reflected the principle of the TER-LLM model with weighting coefficients (9).

As the research results showed, the TER-TEL model was confirmed to have the highest level of adaptability when balancing the load in limiting it at the network edge. A common feature of these solutions is that IP priorities and packet flow rates are comprehensively considered when limiting the load. However, relative priorities are implemented when the network is overloaded, and all flows are limited; however, those with a higher IP priority are limited less intensively than lower-priority flows. Using weighting coefficients (8) in optimality criterion (11) allows, in addition to IP priorities, to consider flow rates in order to more intensively limit precisely those flows that cause overload.

Then, Tables 3 and 4 demonstrate the results of traffic limitation for four characteristic points for comparison concerning the intensities of two flows under $\alpha_{TH} = 0.7$, $PR^1 = 5$, and $PR^2 = 2$ for the first stage of research when the two proposed optimality criteria (6) and (11) were used with weighting coefficients (8) in the corresponding objective function.

Table 3. Traffic limitation under $\alpha_{TH} = 0.7$, $PR^1 = 5$, and $PR^2 = 2$ for TER-LLM model.

Flow Intensity		**TER-LLM**			
λ^1	λ^2	β^1	b^1	β^2	b^2
50	850	0	0	0.153	130.05
850	50	0.153	130.05	0	0
450	450	0	0	0.29	130.5
900	900	0.14	130	1	900

Table 4. Traffic limitation under $\alpha_{TH} = 0.7$, $PR^1 = 5$, and $PR^2 = 2$ for TER-TEL model.

Flow Intensity		**TER-TEL**			
λ^1	λ^2	β^1	b^1	β^2	b^2
50	850	0.074	3.7	0.149	126.35
850	50	0.137	116.4	0.274	13.65
450	450	0.096	43.2	0.19	85.5
900	900	0.38	344	0.76	686

Using the TER-TEL model and weighting coefficients (8), the order of traffic limitation (Figure 6) and the load entering the network was balanced, considering both the IP priorities and intensities of the packet flows. As mentioned above, using the TER-LLM model and weighting coefficients (9) leads to limiting low-priority flows, even without considering the intensity of flows. That is, when limiting the load, balancing is not supported (Figure 7).

Let us note how the results will differ if you choose a different network topology. It has been found that the number of links, their bandwidths, and the order in which routers are connected determine the amount of available network resources. With they are deficient, the load limitation begins, which the authors in this research try to make manageable by considering the results of solving routing problems, flow priorities, and

their intensities; this was the study's primary purpose. With an increase in the available network resources (structural and link), the level of utilization of communication links will naturally decrease, and load limitation at the network edge will begin at higher packet flow rates. The qualitative nature of the dependencies presented in Figures 3–8, regarding the impact of priorities and flow intensities on load balancing, will not significantly change.

By analogy with Tables 3 and 4, Table 5 was formed based on Figure 10. Table 5 shows the data for the same four characteristic points in the plots (Figure 10) for which the proportions of the limited packet flows were equal in their values. Table 5 confirms the uniform load limitation of two packet flows at the network edge, without considering their priority and intensity.

Table 5. Traffic limitation under $\alpha_{TH} = 0.7$, $PR^1 = 5$, and $PR^2 = 2$ for OSPF-TS model.

Flow Intensity		**OSPF-TS**			
λ^1	λ^2	β^1	b^1	β^2	b^2
50	850	0.1444	7222	0.1444	122,778
850	50	0.1444	122,778	0.1444	7222
450	450	0.1444	65	0.1444	65
900	900	0.5722	515	0.5722	515

7. Conclusions

The study substantiates the current scientific problem of load balancing process optimization in the network. It is noted that an effective mechanism for balancing the load in the network is the implementation of traffic engineering concept principles to ensure the effective use of the available network resource, first of all, the bandwidth of communication links. That is why, in theory and practice, mathematical and protocol solutions related to implementing traffic engineering routing, queuing, etc., are increasingly proposed and implemented.

It is essential to ensure the predictability of the QoS level and manageability following traffic engineering requirements, routing processes, and load limitation at the network edge. Therefore, appropriate mathematical models and methods should be developed, which are the basis of any communication protocol, to adequately describe the abovementioned network processes. Thus, this research proposed mathematical optimization models (1)–(5) and (10), which, depending on the selected type of optimality criterion (6) or (11), received two modifications: TER-LLM and TER-TEL. Each of these modifications aims to provide solutions for load balancing within the traffic routing and rate limitation problems, based on different principles related to the flow parameters consideration.

During the research, a comparative analysis of the proposed solutions TER-LLM and TER-TEL (Figures 2–8) was carried out for different variants of the initial data regarding the ratio of the flows' parameters served by the network. A system of recommendations is presented related to the preferential use of the proposed solutions under their features and advantages. Thus, it was established that the TER-LLM solution is better for servicing packet flows based on the so-called absolute priorities when low-priority flows are first limited. The TER-TEL solution is recommended for balanced flow limitation based on flow priorities and their rates. Thus, considering IP priorities, precisely those flows that directly create overload in the network are more intensively limited.

The comparison results indicate that the solution of routing and load balancing problems coordinated on the principles of traffic engineering, represented by the TER-LLM and TER-TEL models, will significantly improve network performance. In the example shown in Figure 1, compared to the OSPF-TS model, it was possible to increase the performance (the amount of traffic served without limitations) by almost 2.4 times. In contrast to the OSPF-TS model, the proposed solutions for network overload allowed differential traffic limitation, considering such key packet flow characteristics as their intensities and priorities.

This fits into the DiffServ paradigm and ensures the localization of the overload source, if necessary.

When the upper bound of the network links utilization increases, it is worth switching to a nonlinear flow-based network model, for example, as described in [31]. Due to tensor generalization, it considers the packet loss probability at network routers. Such an approach analyzes and calculates network performance, average end-to-end delays, and packet loss probabilities. This option is proposed as a direction for further development of the solution obtained in this research. Furthermore, it is possible to develop the proposed solution to implement fault-tolerant routing, both at the level of fast rerouting (FRR) and first hop redundancy protocol (FHRP) enhancements.

Author Contributions: Conceptualization, A.B., O.L., O.Y., L.T. and M.Y.; methodology, O.L. and O.Y.; software, O.L., O.Y. and M.Y.; validation, A.B., O.L. and L.T.; formal analysis, A.B., O.L., O.Y., L.T. and M.Y.; investigation, O.L. and O.Y.; resources, O.L. and O.Y.; data curation, A.B. and L.T.; writing—original draft preparation, A.B., O.L., O.Y., L.T. and M.Y.; writing—review and editing, A.B., O.L., O.Y., L.T. and M.Y.; visualization, O.L., O.Y. and M.Y.; supervision, A.B., O.L., O.Y. and L.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DiffServ	Differentiated Services
FHRP	First Hop Redundancy Protocol
FRR	Fast ReRouting
IoT	Internet of Things
IP	Internet Protocol
IS-IS	Intermediate System to Intermediate System
LFA	Loop-Free Alternate
MPLS	Multiprotocol Label Switching
OSPF	Open Shortest Path First
QoS	Quality of Service
RSVP	Resource Reservation Protocol
SDN	Software-Defined Network
TE	Traffic Engineering
TER-LLM	Traffic Engineering Routing, Linear Limitation Model
TER-TEL	Traffic Engineering Routing, Traffic Engineering Limitation
TS	Traffic Shaping

References

1. Medhi, D.; Ramasamy, K. *Network Routing: Algorithms, Protocols, and Architectures*; Morgan Kaufmann: San Francisco, CA, USA, 2017.
2. Monge, A.S.; Szarkowicz, K.G. *MPLS in the SDN Era: Interoperable Scenarios to Make Networks Scale to New Services*; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2015.
3. Bhardwaj, S.; Girdhar, A. Software-Defined Networking: A Traffic Engineering Approach. In Proceedings of the 2021 IEEE 8th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON), Dehradun, India, 11–13 November 2021; pp. 1–5. [CrossRef]
4. Kim, G.T.; Lee, J.; Kant, L.; Poylisher, A. Utility-Driven Traffic Engineering via Joint Routing and Rate Control for 802.11 MANETs. In Proceedings of the 2021 International Conference on COMmunication Systems & NETworkS (COMSNETS), Bangalore, India, 5–9 January 2021; pp. 317–325. [CrossRef]
5. Chiang, S.H.; Wang, C.H.; Yang, D.N.; Liao, W.; Chen, W.T. Distributed Multicast Traffic Engineering for Multi-Domain Software-Defined Networks. *IEEE Trans. Parallel Distrib. Syst.* **2022**, *34*, 446–462. [CrossRef]
6. Zhang, Z.; Min, X.; Chen, Y. An Adaptive Control Scheme for Data-Driven Traffic Migration Engineering on 5G Network. *Symmetry* **2022**, *14*, 1105. [CrossRef]

7. Semong, T.; Maupong, T.; Anokye, S.; Kehulakae, K.; Dimakatso, S.; Boipelo, G.; Sarefo, S. Intelligent Load Balancing Techniques in Software Defined Networks: A Survey. *Electronics* **2020**, *9*, 1091. [CrossRef]
8. Ibrar, M.; Wang, L.; Shah, N.; Rottenstreich, O.; Muntean, G.M.; Akbar, A. Reliability-aware flow distribution algorithm in SDN-enabled fog computing for smart cities. *IEEE Trans. Veh. Technol.* **2023**, *72*, 573–588. [CrossRef]
9. Shahrbabaki, P.P.; Coutinho, R.W.; Shayan, Y.R. A Novel SDN-enabled Edge Computing Load Balancing Scheme for IoT Video Analytics. In Proceedings of the GLOBECOM 2022-2022 IEEE Global Communications Conference, Rio de Janeiro, Brazil, 4–8 December 2022; pp. 5025–5030. [CrossRef]
10. Kelkawi, A.; Mohammed, A.; Alyatama, A. Incremental deployment of hybrid IP/SDN network with optimized traffic engineering. In Proceedings of the 2020 IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN), Leganes, Spain, 10–12 November 2020; pp. 57–63. [CrossRef]
11. Wang, X.; Deng, Q.; Ren, J.; Malboubi, M.; Wang, S.; Xu, S.; Chuah, C.N. The joint optimization of online traffic matrix measurement and traffic engineering for software-defined networks. *IEEE/ACM Trans. Netw.* **2019**, *28*, 234–247. [CrossRef]
12. Ren, C.; Bai, S.; Wang, Y.; Li, Y. Achieving near-optimal traffic engineering using a distributed algorithm in hybrid SDN. *IEEE Access* **2020**, *8*, 29111–29124. [CrossRef]
13. Albowarab, M.H.; Zakaria, N.A.; Zainal Abidin, Z. Directionally-Enhanced Binary Multi-Objective Particle Swarm Optimisation for Load Balancing in Software Defined Networks. *Sensors* **2021**, *21*, 3356. [CrossRef] [PubMed]
14. Xue, H.; Kim, K.T.; Youn, H.Y. Dynamic Load Balancing of Software-Defined Networking Based on Genetic-Ant Colony Optimization. *Sensors* **2019**, *19*, 311. [CrossRef] [PubMed]
15. Ibrahim, A.A.; Hashim, F.; Sali, A.; Noordin, N.K.; Fadul, S.M. A Multi-Objective Routing Mechanism for Energy Management Optimization in SDN Multi-Control Architecture. *IEEE Access* **2022**, *10*, 20312–20327. [CrossRef]
16. Charalampou, P.; Sykas, E.D. An SDN Focused Approach for Energy Aware Traffic Engineering in Data Centers. *Sensors* **2019**, *19*, 3980. [CrossRef]
17. Ashraf, T.; Lee, S.S.W.; Iqbal, M.; Pan, J.-Y. Routing Path Assignment for Joint Load Balancing and Fast Failure Recovery in IP Network. *Appl. Sci.* **2021**, *11*, 10504. [CrossRef]
18. Babbar, H.; Rani, S.; Gupta, D.; Aljahdali, H.M.; Singh, A.; Al-Turjman, F. Load Balancing Algorithm on the Immense Scale of Internet of Things in SDN for Smart Cities. *Sustainability* **2021**, *13*, 9587. [CrossRef]
19. Chen, W.; Zhu, Y.; Liu, J.; Chen, Y. Enhancing Mobile Edge Computing with Efficient Load Balancing Using Load Estimation in Ultra-Dense Network. *Sensors* **2021**, *21*, 3135. [CrossRef]
20. Zhang, Z.; Duan, A. An Adaptive Data Traffic Control Scheme with Load Balancing in a Wireless Network. *Symmetry* **2022**, *14*, 2164. [CrossRef]
21. Lemeshko, O.; Yeremenko, O.; Hailan, A.M.; Yevdokymenko, M.; Shapovalova, A. Policing based traffic engineering fast reroute in SD-WAN architectures: Approach development and investigation. In *New Trends in Information and Communications Technology Applications, Proceedings of the 4th International Conference, NTICT 2020, Baghdad, Iraq, 15 June 2020*; Proceedings 4, Communications in Computer and Information Science; Springer: Cham, Switzerland, 2020; Volume 1183, pp. 29–43. [CrossRef]
22. Lemeshko, O.; Yevdokymenko, M.; Yeremenko, O.; Shapovalova, A. Investigation of Load-Balancing Fast ReRouting Model with Providing Fair Priority-Based Traffic Policing. In *Advances in Computer Science for Engineering and Education III*; Advances in Intelligent Systems and Computing; Springer: Cham, Switzerland, 2021; Volume 1247, pp. 108–119. [CrossRef]
23. Cui, X.; Huang, X.; Ma, Y.; Meng, Q. A Load Balancing Routing Mechanism Based on SDWSN in Smart City. *Electronics* **2019**, *8*, 273. [CrossRef]
24. Rehman, A.; Haseeb, K.; Saba, T.; Lloret, J.; Sendra, S. An Optimization Model with Network Edges for Multimedia Sensors Using Artificial Intelligence of Things. *Sensors* **2021**, *21*, 7103. [CrossRef] [PubMed]
25. Chen, B.; Sun, P.; Zhang, P.; Lan, J.; Bu, Y.; Shen, J. Traffic engineering based on deep reinforcement learning in hybrid IP/SR network. *China Commun.* **2021**, *18*, 204–213. [CrossRef]
26. Yeo, S.; Naing, Y.; Kim, T.; Oh, S. Achieving Balanced Load Distribution with Reinforcement Learning-Based Switch Migration in Distributed SDN Controllers. *Electronics* **2021**, *10*, 162. [CrossRef]
27. Shi, Y.; Yang, Q.; Huang, X.; Li, D.; Huang, X. An SDN-Enabled Framework for a Load-Balanced and QoS-Aware Internet of Underwater Things. *IEEE Internet Things J.* **2022**, *10*, 7824–7834. [CrossRef]
28. Singh, J.; Singh, P.; Amhoud, E.M.; Hedabou, M. Energy-Efficient and Secure Load Balancing Technique for SDN-Enabled Fog Computing. *Sustainability* **2022**, *14*, 12951. [CrossRef]
29. Nichols, K.; Blake, S.; Baker, F.; Black, D. RFC2474: Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers. 1998. Available online: https://www.rfc-editor.org/rfc/rfc2474 (accessed on 2 December 2022).
30. Ahmed, B.A.; Abdullah, N.S.; Rohani, M.F.; Waseem, S. Fractional Political Optimizer-Based Switch Migration in Software-Defined WAN for Load Balancing with Deep Q Network. *Cybern. Syst.* **2022**, 1–27. [CrossRef]
31. Lemeshko, O.; Papan, J.; Yeremenko, O.; Yevdokymenko, M.; Segec, P. Research and Development of Delay-Sensitive Routing Tensor Model in IoT Core Networks. *Sensors* **2021**, *21*, 3934. [CrossRef] [PubMed]