# Two Distributed Reactive MPLS Traffic Engineering Mechanisms for Throughput Optimization in Best Effort MPLS Networks

Stefan Butenweg
*TU München*
*Institute of Communication Networks*
*80290 Munich, Germany*
*stefan.butenweg@ei.tum.de*

## Abstract

*This paper describes a scalable distributed reactive MPLS traffic engineering system, which can be used for throughput optimization of Best Effort traffic in IP networks. The load rebalancing in the system is realized either by path rerouting or by multi path balancing. To prevent the system from routing oscillations, the distributed traffic engineering units coordinate the load rebalancing actions by using routing update messages. It is shown by simulation, that the reactive MPLS traffic engineering system performs well in great networks and optimizes the network throughput dramatically compared to shortest path routing. The path rerouting and multi path balancing approach reach comparable results while the multi path balancing performs slightly better in the presented scenario. To evaluate the reactive MPLS traffic engineering approaches in realistic network scenarios, a scalable rate based simulation environment is used.*

## 1. Introduction

Reactive traffic engineering reacts to undesired load distributions in a network and optimizes the network performance by rebalancing load during network operation. Unexpected traffic variations are caused e.g. by link failures in the own or in neighboring networks, by changes of the BGP metrics or by new web content.

A basic technology to perform traffic engineering in IP networks is Multiprotocol Label Switching. MPLS establishes tunnels (Label Switched Paths LSP) in an IP network and transports an arbitrary aggregate of IP flows over an LSP. Reactive MPLS traffic engineering is realized either by rerouting LSPs or by redistributing the load over two or more LSPs, which connect the same Label Switched Routers LSRs.

To be able to react to unexpected traffic variations, a reactive traffic engineering system monitors the network load. Basing on this information, the traffic engineering units calculate a load rebalancing. The task of the rebalancing calculation algorithm is to improve an undesired load distribution towards an acceptable load distribution in a minimum amount of time and with a minimum amount of rebalanced data. Additionally the rebalancing algorithm has to be simple and fast due to online calculation during network operation. A special requirement of distributed reactive traffic engineering is to prevent the system from routing oscillations due to uncoordinated concurrent load rebalancing actions.

In this paper a scalable reactive distributed MPLS traffic engineering system is presented, which performs path rerouting as well as multi path balancing. Section 2 gives an overview over related work and describes the innovation of this proposal. Section 3 presents the reactive MPLS traffic engineering system with the integrated path rerouting and multi path balancing algorithm and discusses the system stability. Section 4 defines the simulation environment for the performance investigation and section 5 shows the simulation results. Section 6 summarizes the paper.

## 2. Related work

There are several proposals, which discuss reactive MPLS traffic engineering. In [1] the MPLS Adaptive Traffic Engineering system MATE is described. MATE is based on a distributed multi path balancing approach. The source routers perform an active measurement of each LSP by sending probing packets and measuring the delay jitter and the loss of the packets. The calculation of the new load distribution relies on the optimal routing with the gradient projection algorithm [6]. The traffic engineering capable routers perform rebalancing actions without coordination. To prevent the system from routing

Proceedings of the Eighth IEEE International Symposium on Computers and Communication (ISCC'03)
1530-1346/03 $17.00 © 2003 IEEE

oscillations due to concurrent rebalancing actions, each LSR adapts the load distribution of its LSPs with a limited step size. After each load rebalancing action, the LSPs are measured again. Because the step size decreases with an increasing network size and the load has to be measured after each rebalancing action over a certain period of time, MATE converges slowly in great networks.

Within the TEQUILA project [2] solutions on dynamic route management are discussed. The investigations made in TEQUILA consider MPLS-based networks as well as IP-based networks. In [3] the concept of the dynamic route management is sketched. The concept for MPLS is based on a multi path balancing approach, in which the dynamic route management is distributed over each edge router. With the monitoring of the performance parameters of the LSPs and each single link the load distribution of the LSPs is calculated. Neither the algorithm for the calculation of the load distribution weights nor the coordination of the rebalancing actions is described in detail.

In [4] a multi path approach for distributed reactive MPLS traffic engineering is specified. The system measures the end-to-end packet delay by sending active probing packets and equalizes the mean delay of the LSPs of the network ingress-egress pairs. The coordination of the rebalancing actions is not in the focus of the paper.

Dinan et all [5] describes an analytical model of a multi path balancing algorithm for MPLS traffic engineering. The coordination of the rebalancing actions is also not in the focus of the paper.

In comparison to the mentioned proposals this paper presents a distributed MPLS traffic engineering system, which is also scalable. Additionally it investigates also the path rerouting approach, which is not mentioned in the other proposals. Finally the behavior of the reactive MPLS traffic engineering system is investigated in realistic networks. The other proposals present simulation results only for small networks.

## 3. The MPLS traffic engineering system

Figure 1 shows the components of the distributed reactive Traffic Engineering system. The network consists of MPLS capable Label Switched Routers, which are connected via bi-directional links. At the network ingress an LSR maps the incoming IP traffic to an LSP and adds an MPLS header to an IP packet. The intermediate router forwards a packet to the next neighbor router according to the MPLS label. The egress LSR strips the MPLS header of the IP packet and sends it out as a regular IP packet.

In the case of path rerouting, each ingress LSR is connected to each egress LSR with one LSP. In the case of multi path balancing there are two or more LSPs, which connect an LSR pair.

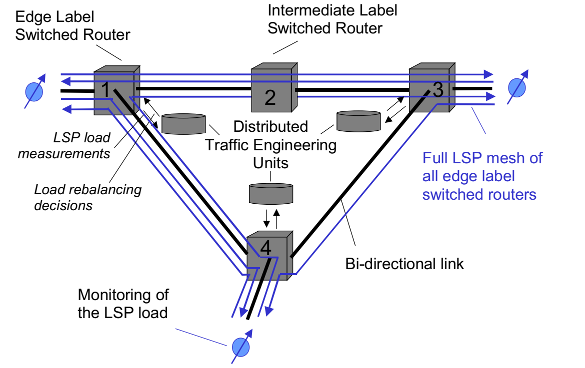


**Figure 1: Components of the distributed reactive MPLS traffic engineering system**

All ingress LSRs recalculate the link loads in the network basing on the knowledge of the mean load and the link path of all LSPs in the network. Therefore each LSR distributes the path information of its LSPs once during the system initialization and after each rerouting. Additionally each LSR passively monitors the ingress LSP load and periodically distributes the LSP loads. The monitor interval is in the range of several minutes. For the load and path distribution the flooding mechanism of OSPF/ISIS and their extensions [8] can be used.

The MPLS traffic engineering functionality is distributed over the ingress LSRs. The MPLS TE units perform the rebalancing coordination and calculation.

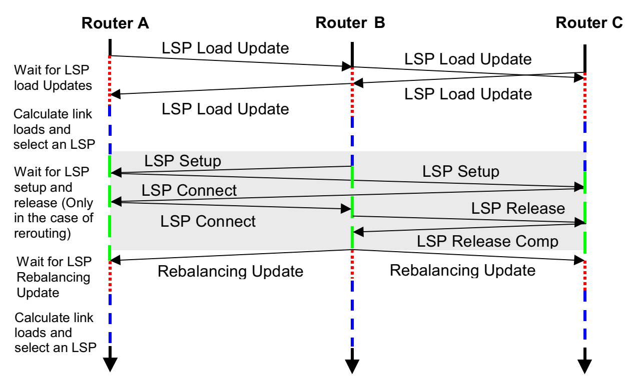


**Figure 2: LSR coordination with topology updates**

The coordination of the rebalancing actions requires, that all MPLS TE units have the same link load view. The MPLS TE units realize it by using LSP load update messages and rebalancing update messages (see Figure 2). After sending or receiving an LSP load update message, an LSR changes into the rebalancing state and waits until each LSR has received all currently flooded LSP load updates. In the rebalancing state an LSR is not allowed to send further LSP load updates. The LSRs recalculate the current link loads and check, if the link loads exceed the rebalancing threshold. The rebalancing threshold defines the maximum tolerable link load.

In the case of link overload, the LSRs have to pick an LSP for rebalancing. They choose the LSP from the highest loaded link, which fits best to reduce the link load below the rebalancing threshold. All LSRs, which are not the source of the chosen LSP, wait for a rebalancing update of the source LSR. The source LSR performs the LSP rebalancing. In the case of path rerouting it calculates a new path (see subsection 3.1), triggers the setup of a new LSP, switches the load from the old LSP to the new LSP and releases the old LSP. In the case of multi path balancing, it recalculates the load distribution weights and redistributes the traffic according to these weights (see subsection 3.2). After a successful rebalancing, the source LSR floods the rebalancing update to all ingress LSRs. Then the rebalancing process repeats, until either no link exceeds the rebalancing threshold or no rebalancing increases the network performance. To guarantee the functionality of the system, the waiting times have to be adapted to the maximum message exchange time between two LSRs.

## 3.1 Path rerouting algorithm

The path rerouting calculation relies on the shortest distance path routing [7]. In this case the shortest path is calculated basing on additive load depending link costs. The link cost $C_{link}$ increases in a convex manner with the link load. A qualitative run of a cost function is shown in Figure 3. In this proposal the cost function is derived from the used bandwidth on an outgoing link $L_{out}$ and the maximum link bandwidth $B_{max}$ as shown in Equation 1.

$$C_{link} = \frac{B_{\max}}{B_{\max} - L_{out}}$$

**Equation 1: Link cost function**

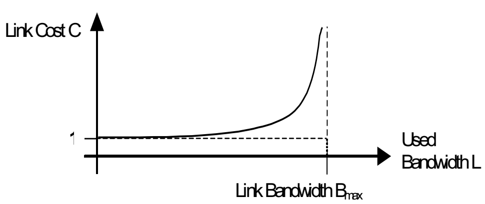


**Figure 3: Link cost as a function of the link load $L_{out}$**

To calculate a new LSP path, the load of this LSP is removed from the current link load view. Basing on the resulting link load view, the shortest distance path is calculated. If the new path differs from the old path, the total link costs with the old and the new path are calculated and compared to each other. The total link cost is the sum of the cost of each link in the network (see Equation 2). An LSP is rerouted, if either the total link cost decreases or the new path has a smaller hop count with the same total link cost.

$$C_{tot} = \sum_{i}^{all\_links} C(L_i)$$

**Equation 2: Total link cost**

This guarantees, that without a change of the ingress traffic the network will not reach the same load distribution after one or several rerouting actions. This prevents the rerouting algorithm from routing oscillations.

## 3.2 Multi path balancing algorithm

The load redistribution is based on the optimal routing, which is described in [6]. It says: "The load of a set of paths is distributed optimal over the paths if and only if the path flow is positive only on paths with a minimum first derivative of the cost function".

$$C_{tot\,new} = C_{tot\,old} + \frac{\partial C_B(L)}{\partial L} \cdot \Delta L - \frac{\partial C_A(L)}{\partial L} \cdot \Delta L < C_{tot\,old}$$
$$with \frac{\partial C_A(L)}{\partial L} > \frac{\partial_B(L)}{\partial L}$$

**Equation 3: Reduction of the total link cost by shifting load to a path with a lower first derivative of the cost**

Equation 3 explains this statement for the case, in which an LSR ingress-egress pair is connected with two LSPs, LSP A and LSP B. LSP A carries load and the first derivative of the path cost is greater than of LSP B. Shifting a small load portion $\Delta L$ from LSP A to LSP B reduces the total link cost. If all load carrying LSPs have the same minimum first derivative of the path cost, the total link cost reaches its minimum. As within the rerouting approach, load is only rebalanced, if the total link cost is decreased. This prevents the load balancing from routing oscillation. The first derivative of the path cost is the sum of the first derivatives of the link costs.

$$\frac{\partial C_{path}}{\partial L} = \sum_{i}^{links} \frac{\partial C_{link}}{\partial L_i}$$
$$with \frac{\partial C_{link}}{\partial L} = \frac{B_{\max}}{\left(B_{\max} - L_{out}\right)^2}$$

**Equation 4: First derivative of the path cost**

## 3.3 Stability of the reactive MPLS traffic engineering system

The stability of both rebalancing algorithms relies on a correct link load view for the rebalancing calculation. This cannot be always guaranteed. Concurrent

Proceedings of the Eighth IEEE International Symposium on Computers and Communication (ISCC'03)
1530-1346/03 $17.00 © 2003 IEEE

rebalancing actions and the impreciseness of the load monitoring can lead to incorrect link load views.

Due to the sequential coordination of the rebalancing actions and the flooding of rebalancing updates, concurrent rebalancing actions do not occur in the presented MPLS traffic engineering system.

Incorrect link load views due to the impreciseness of load monitoring cannot always be prevented. In the presented system the load is monitored over a time period of several minutes. During this time the load on an LSP may change dramatically. Until the monitor interval is not finished, the change of the LSP load is not considered in the following rebalancing decisions. To minimize the occurrence of these situations, the monitor interval should be much shorter than the mean time between two traffic variations.

The same problem occurs, if several rebalancing actions are performed in a row. Within this time, no LSP load updates are sent and current load changes are not considered within the rebalancing process. To minimize the number of these situations, the number of sequential rebalancing actions without LSP load updates is limited.

## 4. Simulation environment

To investigate the performance of the presented reactive MLPS traffic engineering system in realistic network scenarios, a scalable simulation environment is needed. Because the simulation of networks of typical size and traffic volume is not possible with the standard packet-based simulators like NS2 or OPNET, a rate-based simulation approach is used.

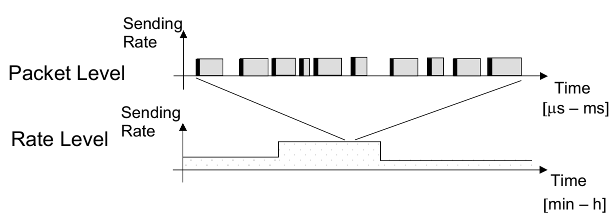


**Figure 4: Rate-based traffic model**

Within this simulation approach the traffic is modeled with piece-wise constant rates over time periods of varying lengths. An example of the rate-based traffic modeling is shown in Figure 4. This simulation approach increases the simulation speed, because only the rate changes instead of each data packet is modeled. Though this traffic model simplifies the real traffic behavior, it is sufficient for load distribution evaluation in networks.

Within the simulation of the reactive MPLS TE system, each aggregated traffic stream of a network ingress-egress pair is modeled with one rate-based source. An LSR is modeled in the simulation as an ideal output buffered node without internal blocking. The buffers are modeled with a M/M/1/S queuing system with a finite buffer space. S is the number of packets, which can be stored in the outgoing buffer. In the following simulations, the buffer space is set to 50 packets. The outgoing link load $L_{out}$ is calculated from the incoming link load of the link buffer $L_{in}$ and the packet loss probability $P_{loss}$. This is shown in Equation 5.

$$L_{out} = L_{in} \cdot (1 - P_{loss})$$

**Equation 5: Deriving the outgoing link load $L_{out}$ from the incoming link load $L_{in}$**

### 4.1 Simulation scenario

The presented MPLS Traffic Engineering system is evaluated with the network topology shown in Figure 5. Each link has a capacity of 2500MBit/s. All non-core routers are fully meshed. These are 870 LSPs within the path rerouting approach and 1740 LSPs within the multi path balancing approach, in which each ingress egress LSR pair is connected by two LSPs. A normal distributed random number generates the ingress traffic of each LSP. The mean and the variance of this traffic is the same for each LSP.

The initial routing of the LSPs is calculated with the shortest path for the path rerouting and a k-shortest path algorithm for multi path balancing. The k-shortest path calculates the shortest path, hides the edges of this path and calculates a shortest path on the residual graph.

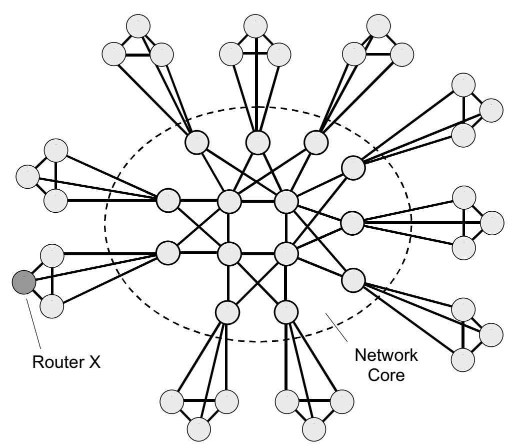


**Figure 5: Simulated network topology**

In a first simulation run the convergence behavior of the MPLS traffic engineering system is evaluated under idealized conditions. The traffic load of each LSP is set at the simulation start and does not vary. The rebalancing decisions base on the ideal knowledge of the current link loads. If the load distribution cannot be further optimized, the path rerouting and multi path balancing algorithms stop. The reached load distribution and the number of performed rebalancing actions are evaluated. For the convergence evaluation the mean LSP ingress traffic is varied between 30Mbit/s and 50Mbit/s. For each mean LSP load the convergence evaluation is repeated 1000 times. To evaluate the convergence behavior, the path rerouting and the multi path balancing algorithm are compared to the shortest path routing and to the optimal load distribution. The optimal load distribution is received by formulating the problem as a multi commodity linear program. The constraints of this program are to keep the link load below the maximum link capacity and to guarantee that each node receives the same load of an LSP as it sends out. In contrast to MPLS, this allows a splitting of an LSP load over several links at each node in the network. The objective of this linear program is to minimize the total link cost. To solve the problem with a linear program, the original cost function is approximated with a piecewise linear cost function. Straight lines connect the integer cost values (see Figure 6). The cost value of link load L=0 is set to zero.

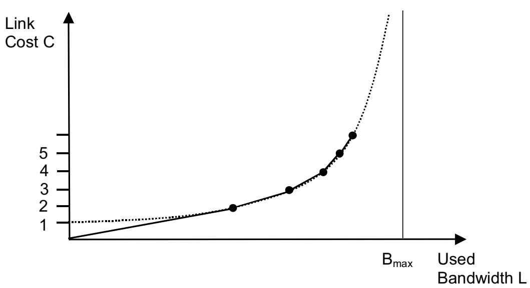


**Figure 6: Piecewise-linear cost function**

In a second simulation run the behavior of the reactive MPLS Traffic Engineering system is evaluated under varying traffic load. The LSP load with a mean value of 28Mbit/s changes after a negative exponential distributed time with a mean value of 1500sec. During the simulation, router X (see Figure 5) increases the load of each LSP with a factor of three. This models a traffic variation e.g. due to a link failure in a neighboring network. The rebalancing threshold is set to 1500Mbit/s and the monitoring time interval is set to 300sec. The monitoring intervals of the LSRs start randomly distributed within the first 300sec. The LSP setup time is set to 4sec and the rebalancing calculation time and the topology update distribution time are set to 2sec.

## 5. Simulation Results

The convergence evaluation of the reactive MPLS TE system shows an enormous decrease of the mean total link cost and the maximum link utilization compared to the shortest path routing (Figure 7 and Figure 8). Taking the results of the shortest path routing with source ingress traffic of 30Mbit/s as a reference, the ingress traffic can be increased with about 50% using the reactive MPLS TE system without increasing the maximum link utilization. Comparing the reactive MPLS TE system to the optimum load distribution, the achieved results of the total link costs and maximum link utilizations are close to the optimum load distribution.

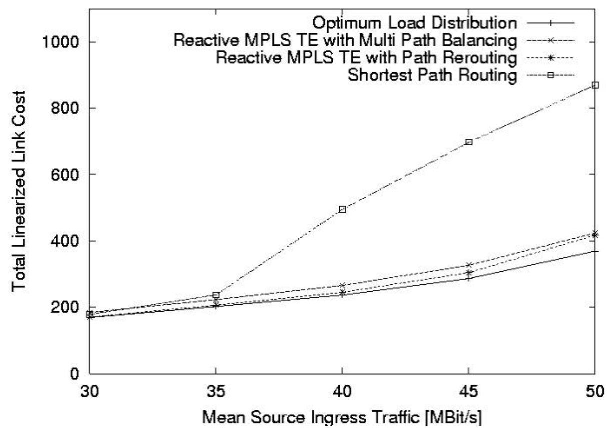


**Figure 7: Mean total link cost**

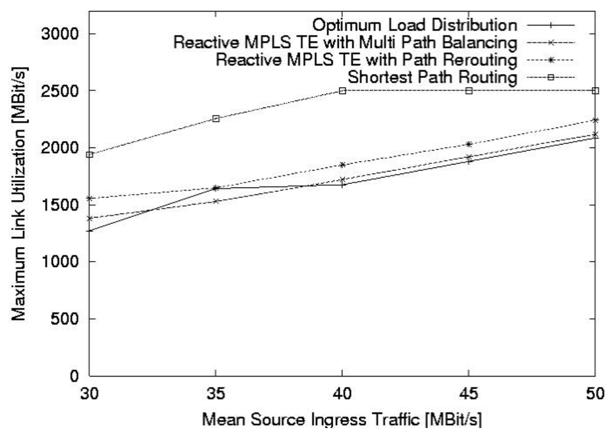


**Figure 8: Maximum link utilization**

Comparing the multi path balancing and the path rerouting, both approaches reaches the same total link cost. The multi path balancing reaches a lower maximum link utilization with a difference of about 100Mbit/s and approaches the optimal load distribution closer. The number of performed rebalancing actions of the two approaches is compared in Figure 9. It shows, that both approaches perform about 140 rebalancing actions with a source ingress rate of 30Mbit/s. Increasing the source ingress rate, the number of rebalancing actions of the path

Proceedings of the Eighth IEEE International Symposium on Computers and Communication (ISCC'03)
1530-1346/03 $17.00 © 2003 IEEE

rerouting approach is nearly constant, while it increases for the multi path balancing approach linearly up to 250.

To derive the convergence time from the number of rebalancing actions, it has to taken into account that due to the additional LSP setup time path rerouting actions are approximately twice as long as multi path balancing actions. It follows that the multi path balancing converges faster in the presented scenario. The amount of the switched data of both approaches is equal.

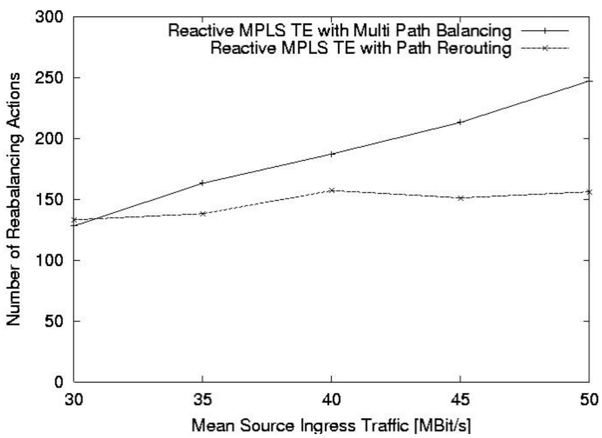


**Figure 9: Mean number of rebalancing actions**

The system behavior under varying traffic load is shown in Figure 10. It takes 300sec, that the load of each LSP is monitored and flooded through the network. Then the path rerouting approach and the multi path balancing approach starts to rebalance the load until no link load is above the rebalancing threshold. The multi path balancing converge faster than the path rerouting due to the approximation that a path rerouting action is twice as long as a multi path balancing action.

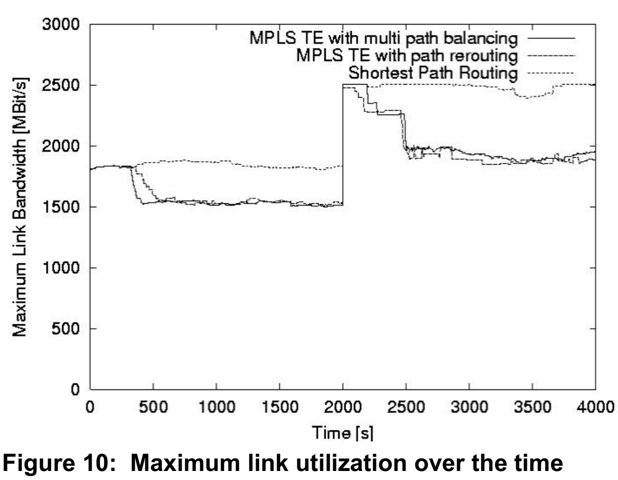


**Figure 10: Maximum link utilization over the time**

After 2000sec router X increases its ingress LSP loads with the factor of three and the maximum link load increases up to 100% of the link capacity. At the end of the next monitor interval of router X, the load change is recognized and the rebalancing process starts. The finally reached maximum link utilization is below 80% of the link capacity. In the presented scenario the monitoring interval length mainly influences the convergence speed.

## 6. Conclusion

In this paper a scalable distributed reactive MPLS traffic engineering system is presented. The distributed TE units exchange periodically routing update messages to actualize their link load view. The exchanged messages are also used to coordinate the load rebalancing actions of the TE units, which prevents the system from routing oscillations. The reactive MPLS TE system is realized with path rerouting and with multi path balancing. To evaluate the performance of the system in realistic network scenarios, a scalable rate-based simulation environment is used. For the investigated scenario, the simulation results show an enormous performance increase compared to the shortest path routing. It is also shown, that the both the path rerouting and the multi path balancing produce results close to the optimum load distribution. The results of both approaches are comparable, while the multi path is considered to reach lower maximum link utilizations and converges faster.

In future work the convergence speed of the system is further improved. Therefore the gradient projection algorithm for the multi path balancing is integrated in the system. Additionally the monitor interval length is dynamically adapted due to changes of the LSP load.

### REFERENCES


*[1] A. Elwalid, C. Jin, S. Low, I. Widjaja, "MATE: MPLS Adaptive Traffic Engineering", Infocom 2001*

*[2] www.ist-tequila.org*

*[3] P. Trimintzios, L. Georgiadis, G. Pavlou, D. Griffin, C.F. Cavalcanti, P. Georgatsos, C. Jacquenet, "Engineering the Multi-Service Internet: MPLS and IP-based Techniques", ICT2001*

*[4] D. Gao, Y. Shu, S. Liu, "Delay-based adaptive load balancing in MPLS networks", ICC2002*

*[5] E. Dinan, D. Awduche, B. Jabbari, „Optimal Traffic Partitioning in MPLS Networks", NETWORKING 2000*

*[6] D. Bertsekas, R. Gallager, „Data Networks", Prentice-Hall, 1991*

*[7] Q. Ma, P. Steenkiste, H. Zhang, „Routing High-bandwidth Traffic in Max-min Fair Share Networks", SIGCOMM 1996*

*[8] D. Katz, D. Yeung, K. Kompella, „Traffic Engineering Extensions to OSPF Version 2", Internet Draft <draft-katz-yeung-ospf-traffic-09.txt>, 2002*