MPLS Based Best Effort Traffic Engineering

Jerapong Rojanarowan, Bernd G. Koehler, and Henry L. Owen
Department of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332, USA
jerapong@ece.gatech.edu, bernd.koehler@usma.edu, henry.owen@ece.gatech.edu

Abstract

The advent of Multi-protocol Label Switching (MPLS) enables traffic engineering by introducing connection-oriented features of forwarding packets over arbitrary non-shortest paths. Our goal in this research is to improve the network utilization for best effort traffic in IP networks. By examining the best effort traffic class, we assume The large volume of research that has been conducted on traffic engineering for assured forwarding and expedited forwarding with admission control allows traffic engineering on those traffic classes. We examine a different and in some sense a "more complex" problem (traffic demands are not known a priori) of traffic engineering for the remaining network bandwidth which is utilized by best effort traffic. We present a generic traffic engineering framework and four specific algorithms. This framework has two prominent features 1) it uses MPLS to encapsulate source-destination aggregate flows within a Label Switched Path (LSP), with multiple LSPs per source-destination. 2) The framework is "stateless", only the topology is used to determine the traffic routing.

1. Introduction

As the Internet becomes more popular, there is more traffic in the network. Growing network traffic can cause network congestion. Essentially network congestion may result from a shortage of bandwidth or inefficient traffic management that causes uneven traffic distribution. In particular, the forwarding paradigm in IP networks based on the destination address maps the traffic onto the shortest path whereas the capacity of the links not on the shortest path is underutilized. Therefore, there is a possibility for traffic engineering to improve the performance of IP networks.

In order to provide such capability, the basic IP forwarding paradigm of present-day IP networks must be enhanced to support traffic engineering. The advent of Multi-protocol Label Switching (MPLS) made this feasible by introducing the connection-oriented features of forwarding packets over arbitrary non-shortest paths.

1.1. MPLS and Traffic Engineering

In this section we describe MPLS application to traffic engineering. MPLS was introduced by the Internet Engineering Task Force (IETF) as a novel forwarding paradigm in IP networks based on labels [1]. It is aimed to accommodate IP Quality of Service (QoS), forwarding speed, and traffic engineering [2].

MPLS displaces the hop-by-hop forwarding paradigm with a label swapping forwarding paradigm. A label is a short, fixed length, locally significant identifier assigned to a packet at the ingress router of an MPLS domain corresponding to its Forwarding Equivalence Class (FEC). A FEC is considered as a group of packets that has the same path forwarding requirements. A benefit of the FEC is that it can support a wide range of forwarding granularities, ranging from per-destination to per-application [3]. MPLS directs the flows of packets along the predetermined Label Switched Paths (LSPs) across the network based on labels.

In MPLS, there are two alternative LSP selection mechanisms: hop-by-hop and explicit routing. A hop-by-hop path is calculated based on the normal layer 3 routing information. With an explicit routing mechanism, the path is completely assigned by the originator independent of layer 3 routing. There are two methods to set up an explicit route: Constraint-based Routed Label Distribution Protocol (CR-LDP) [4] and extensions to Resource Reservation Protocol (RSVP) [5].

Traffic engineering is the process of optimizing the network utilization or fulfilling some policy objectives. The limitation of traffic engineering in IP networks is due to the destination based forwarding scheme. While this mechanism is scalable, it unevenly distributes traffic using only the shortest paths which in turn leads to inefficient use of network resources. The Explicitly Routed LSP (ER-LSP) with optimization objectives is the primary tool for traffic

0-7803-8367-2/04/$20.00

engineering in MPLS. With ER-LSP, the ingress routers are able to control the traffic distribution in the network to meet the performance objectives. However, the exact algorithm for determining the ER-LSP is not specified in the IETF.

1.2. Best Effort Traffic Class

Currently IP networks can support only a single best effort service class. Recent research has begun examining how to deliver QoS for best effort traffic [6]. Many traditional applications such as file transfer, remote terminal, and electronic mail have been served sufficiently because of their elasticity. These applications can tolerate performance variations during the presence of congestion in the network. The majority of the best effort traffic over the Internet is Transmission Control Protocol (TCP). The close-loop congestion control mechanisms in TCP regulate the transmission rate of the source. The offered load presented by a TCP connection varies with the network condition along with the complex interaction between routing, queueing and flow control. This causes the offered traffic to be dynamic and elastic in nature. Typically, the size of elastic flows is variable and exhibits a heavy-tailed distribution [7]. In addition, best effort traffic is not subject to admission control. As a result, it is very difficult to describe its aggregate bandwidth requirements. These characteristics make traffic engineering of best effort traffic challenging.

2. Best Effort Framework

One constraint in the examination of traffic engineering applied to best effort traffic is that traffic demands are not well-defined. Consequently traditional optimization techniques do not directly apply. Given this difficult constraint, this research examines how well we can do with traffic engineering on best effort traffic without a set of well-defined best effort demands.

We propose a *stateless* framework where the input to the traffic engineering algorithm is restricted to the network topology. We assume that our techniques are applied only to the remaining best effort bandwidth in a network with multiple traffic classes. The other priority classes use admission control and traffic engineering based on other more typical traffic engineering algorithms. These more traditional traffic engineering algorithms do not apply to best effort traffic because the exact amount of best effort traffic is not known a priori. We examine the remaining network capacity and attempt to maximize the capability to deliver best effort traffic with this limited available capacity.

The motivation for exploring this approach is the following observation: current IP routing protocols are state-independent - they utilize metrics that only depend upon the topology to calculate the shortest path. Therefore, paths do not change in response to network congestion or demand variation. Despite the non-adaptive nature of this scheme, it consistently provides an adequate level of performance over a broad range of demand patterns. One can do better, but the trade-off is in the additional complexity required for state-sensitive routes and metrics. The objective of this work is to develop a framework that, although static and stateless, improves upon the current paradigm and provides an acceptable level of performance over a broad range of network conditions as applied to best effort traffic.

The general operation of the proposed framework consists of the following steps [8]:

1. For each source-destination pair $w \in \mathscr{W}$ a set of paths $\mathscr{P}_w$ is computed based on the network topology $\mathscr{G}(\mathscr{V}, \mathscr{A})$.

2. Each path $p \in \mathscr{P}$ is instantiated as a permanent LSP.

3. For each path, an optimal rate r_p is computed. Each path is guaranteed this rate along each link in the path through link sharing discipline such as Weighted Fair Queueing (WFQ).

4. For each path, $\mathscr{P}_w$ assigned to a source-destination pair, a fraction $\phi_p \in [0,1]$ of the total traffic is calculated.

5. Each ingress node i of a source-destination pair $(i,j) \equiv w$ splits j-bound traffic according to the assigned path fractions $\phi_p, \forall p \in \mathscr{P}_w$.

Barring any topology changes (e.g., link failure/recovery), path, rate, and fraction computation is only executed once, during the system initialization. A potential performance enhancement is to periodically recalculate the assigned rates and fractions based on aggregate flow measurements. We do not examine that enhancement in this paper.

The primary design criteria for the proposed framework is that the input is restricted to the network topology. The primary advantage of this restriction is inherent simplicity - there is no additional overhead for the exchange of network state. Clearly, the framework should

1. Split aggregate source-destination traffic over multiple paths.

2. Account for the elastic nature of best effort traffic.

3. Restrict the input to the network topology.

4. Provide a level of service better than that of shortest-path routing.

In fact, there are exactly two design unknowns: 1) the path set $\mathscr{P}_k$, and 2) the fractions ϕ_p, $\forall p \in \mathscr{P}_k$. The primary difference between the proposed framework and an adaptive scheme is there is no periodic exchange of network state and adjustment of path flow parameters. Instead, the objective is by choosing a "good" path set $\mathscr{P}_k$ and fractions ϕ_p, improvements over shortest-path routing can be achieved over a wide range of traffic demands. While the results may not be as good as an adaptive scheme, the payoff is in a much simpler framework. The proposed framework at present incorporates four algorithms for calculating optimal paths and flow rates.

2.1. Defining the Path Set

The initial step of the proposed framework is to determine the path set $\mathscr{P}_k$ for each SD pair. Paths with smaller hop-counts are more efficient in terms of bandwidth utilization. We also expect diminishing returns as the path set size grows large, especially for sparse networks. Large path sets frequently include paths with high hop-counts, and these extremely long paths contribute little to the optimal solution. They also preclude traffic on shorter paths contained within them. A small path set also has the added benefit of reducing the complexity of the rate computation problem.

Two alternative path sets are proposed. The first set consists of the k-shortest paths between source and destination nodes. The second set consists of k-disjoint paths between source and destination nodes. If there are no disjoint paths, $\mathscr{P}_k$ consists of the shortest path. If there are only $n \leq k$ disjoint paths, then only the n disjoint paths are included in $\mathscr{P}_k$.

2.2. Calculating Path Fractions ϕ_k and Rates r_k

Two alternative methods for calculating r_k are proposed. The first method stems from classical flow control theory and is based on a "max-min fair" rate allocation [9].

A rate allocation vector $\mathbf{R}$ is *max-min fair* if it is feasible and for each $p \in \mathscr{P}$, r_p cannot be increased without decreasing some r'_p for which $r'_p \leq r_p$. In other words, it is impossible to increase a given path flow without taking bandwidth away from a path less well off. To derive r_k, we first determine a path set $\mathscr{P}_k$ for each source-destination pair. We then calculate a max-min fair allocation to each $\{r_p : p \in \mathscr{P}\}$. Individual path rates are given by $x_p = r_p$, and the total rate assigned to each source-destination pair $k \in \mathscr{K}$ is

$$r_k = \sum_{p \in \mathscr{P}_k} r_p \tag{1}$$

Table 1. Proposed framework algorithms

Algorithm	Paths	Rates
SMM	shortest	max-min
SOP	shortest	optimal
DMM	disjoint	max-min
DOP	disjoint	optimal

The path fractions ϕ_p are calculated by

$$\phi_p = \frac{x_p}{r_k} = \frac{x_p}{\sum_{p \in \mathscr{P}_k} r_p} \tag{2}$$

The second method solves a combined rate allocation/delay minimization network optimization problem. Recall that in general, solving a multi-commodity flow problem requires a finite demand set to constrain the feasible region of the flow vector $\mathbf{x}$. In the absence of any demand constraints, the optimal solution tends to $\mathbf{x} = \mathbf{0}$. Zero flow entails zero (or minimum) delay. However, for best effort traffic source-destination demands are not well defined, and the best that can be said is that each $d_k \in [0, d_k^*]$ where d_k^* is the intrinsic demand defined as the demand in the absence of competing traffic. We propose to solve an *unconstrained* optimization problem with an objective function comprised of two components. One component minimizes the total aggregate delay, while the other is a penalty term for making source-destination flows too small. The objective function is given by

$$f(\mathbf{x}) = \sum_{(i,j) \in A} D_{ij}(\mathbf{x}) + \sum_{k \in \mathscr{K}} \frac{a}{r_k} \tag{3}$$

$$D_{ij}(\mathbf{x}) = \frac{f_{ij}}{c_{ij} - f_{ij}}, \;\; f_{ij} = \sum_{p \ni (i,j)} x_p \tag{4}$$

where D_{ij} represents the delay of each link $(i,j) \in \mathscr{A}$, and $\sum_{k \in \mathscr{K}} \frac{a}{r_k}$ is a penalty term for making source-destination flow r_k too small. By adding a penalty term the optimal solution finds a balance between congestion and throughput, and at a coarse level models the behavior of a best effort network flow control - flows increase the offered load until throttled by network congestion.

We have proposed two methods for selecting the path set $\mathscr{P}_w$ and two methods for calculating the allocation rates r_w and fractions ϕ_p. This leads to four different candidate traffic engineering algorithms labeled SMM, SOP, DMM and DOP. They are summarized in Table 1. We compare and contrast these four algorithms with each other, as well as the currently deployed shortest-path algorithm (SPF).

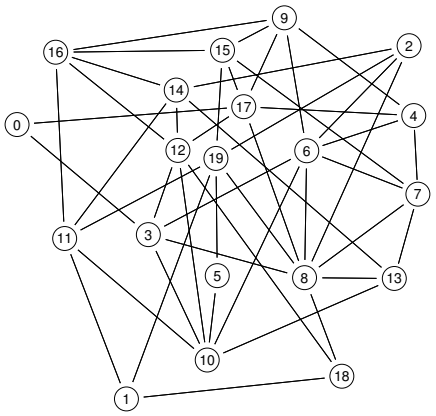

Figure 1. Random Network Topology

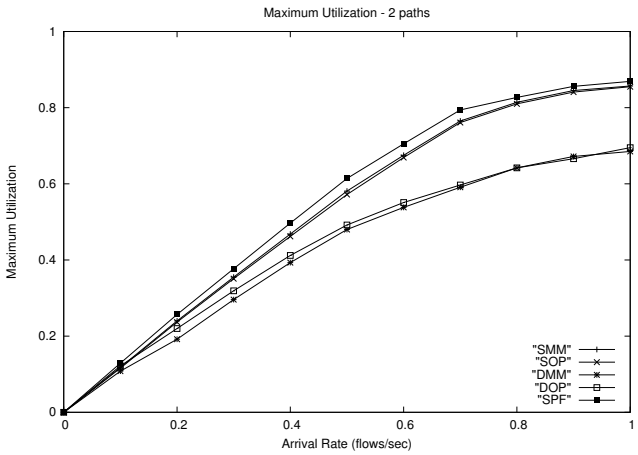

Figure 2. Maximum network utilization for TCP traffic: 2-path case

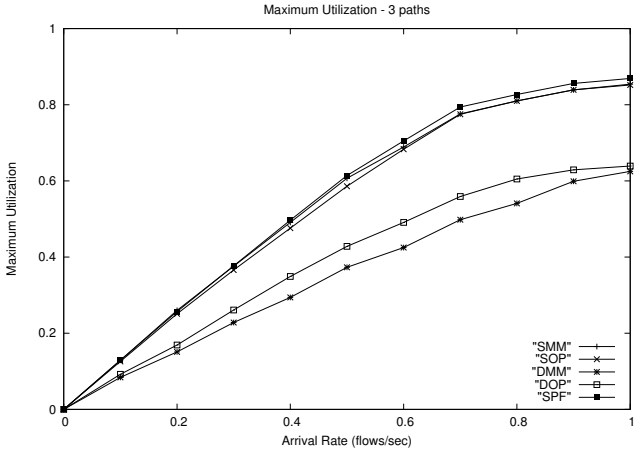

Figure 3. Maximum network utilization for TCP traffic: 3-path case

3. Simulation Scenarios

The simulations in this paper are conducted using the NS simulator with MPLS Traffic Engineering Extensions. MplsWF2Q is the modified version of Worst-case Fair WFQ (WF2Q+) [10], and it is used to guarantee that each LSP receives the assigned flow rate.

Fig. 1 shows the random network topology used in this paper to obtain the results. The random network topology is generated by the Georgia Tech Internetworking Topology Models (GT-ITM) [11]. The network consists of 20 nodes and 96 bi-directional links with capacity of 45 Mbps. There are 12 nodes generating traffic to all other nodes. We simulated two different types of traffic, User Datagram Protocol (UDP) and TCP, with packet sizes of 1000 bytes. The traffic flows generated in each node are exponentially distributed. We simulate each algorithm with the number of desired paths, in k-shortest and k-disjoint paths, set to 2 to 5 paths.

4. Results

Fig. 2 to Fig. 5 show the maximum utilization of any link in the network on the y axis and the incoming flow rate in unit of flows per second on the x axis, with the number of paths ranging from 2 to 5 respectively. Lower values of maximum utilization are better. We notice performance improvements over SPF from the results when multi-path best effort traffic engineering algorithms are emplyed. The DMM algorithm is the best candidate to reduce the maximum network utilization. In Fig. 2 where incoming traffic is routed over 2 paths, the performance gain of the algorithms that use k-disjoint paths (DOP and DMM) is considerably better than those algorithms that use k-shortest paths (SOP and SMM). This results from the fact that a set of disjoint paths will not direct traffic over the same link whereas a set of shortest paths may share some links in common. Therefore, algorithms that use k-shortest paths do not produce improvements as large as k-disjoint paths. When the number of paths increases to 4, the performance improvements of DOP and DMM algorithms are most visible. However, when the number of paths increases to 5, algorithms (SOP and SMM) obtain only a slight enhancement whereas algorithms (DOP and DMM) stay nearly the same. This is because the average number of paths per source-destination pair of the k-disjoint paths is limited when k increases because of mutually exclusive path requirement whereas the k-shortest paths have no such constraint.

Fig. 6 to Fig. 9 repeat the same situations in Fig. 2 to Fig. 5 except that UDP traffic is used instead of TCP traffic. The best choices are still DOP and DMM which use a disjoint path set. As in the case of TCP traffic, increas-

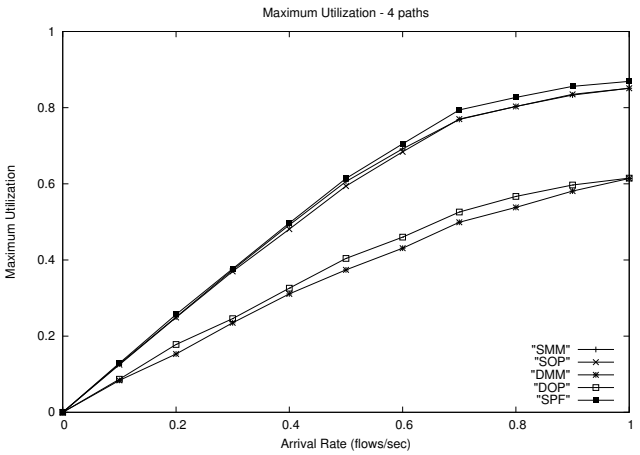

Figure 4. Maximum network utilization for TCP traffic: 4-path case

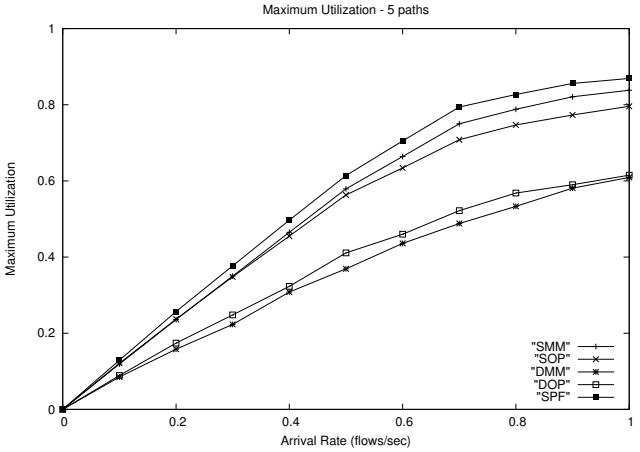

Figure 5. Maximum network utilization for TCP traffic: 5-path case

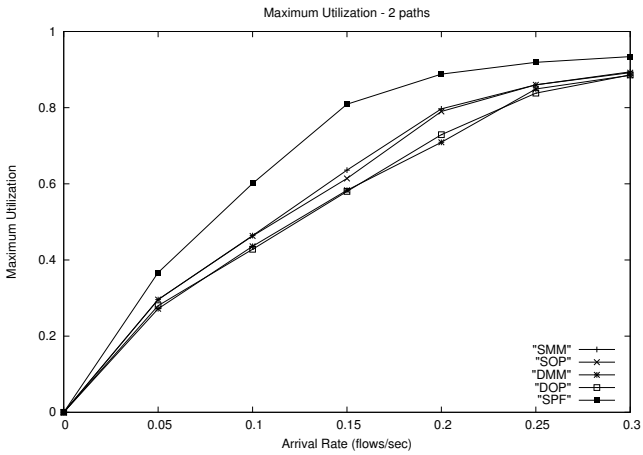

Figure 6. Maximum network utilization for UDP traffic: 2-path case

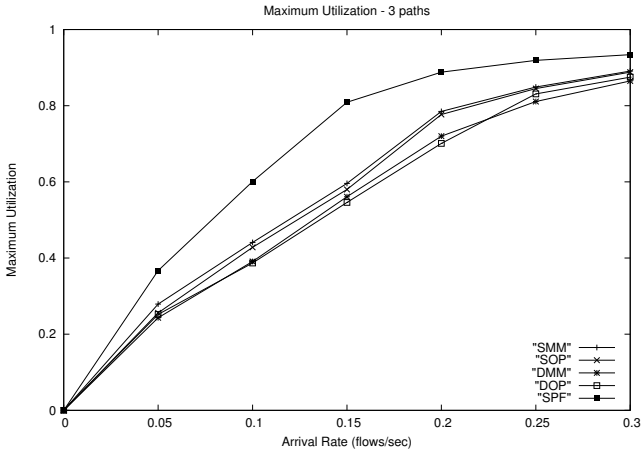

Figure 7. Maximum network utilization for UDP traffic: 3-path case

ing number of paths to 4 for UDP traffic does not result in a dramatic improvement. Only a slightly improvement is obtained. Also going to 5 paths improves the performance of k-shortest path algorithms. All of our best effort traffic engineering algorithms perform comparably when 5 paths are used. In general, there is a trade-off between performance gain and number of paths used. A higher number of paths used means more complexity in the routers and higher network resource usage because the non-minimum hop paths are used. If increasing number of paths does not result in a good improvement, keeping number of paths as low as possible while maintaining good performance is recommended.

We also simulated the transit-stub topology as illustrated in Fig. 10 but the results are not shown here. Stub domains correspond to the customer networks connected to transit domains whereas transit domains illustrate the backbone networks. The performance of DOP and DMM algorithms are basically the same as that of SPF. This is due to the limited number of disjoint paths available in the network. As one can see from Fig. 10 the number of disjoint paths from node 11 to node 19, or node 9 to node 10 consist of only one path. In particular, the average numbers of paths used per source-destination pair in 2, 3, 4, and 5-disjoint paths in the transit-stub network are 1.03, 1.04, 1.04, 1.04, respectively. Therefore, we cannot gain the performance improvement over the SPF when the number of available paths is limited to only one path. However, as stated earlier, the availability of the k-shortest path set is generally higher than that of the k-disjoint path set. Hence, the algorithms that use k-shortest paths (SOP and SMM) perform best in transit-stub networks. When the availability of alternative paths is very limited, these algorithms normally provide better performance gain than those that use disjoint

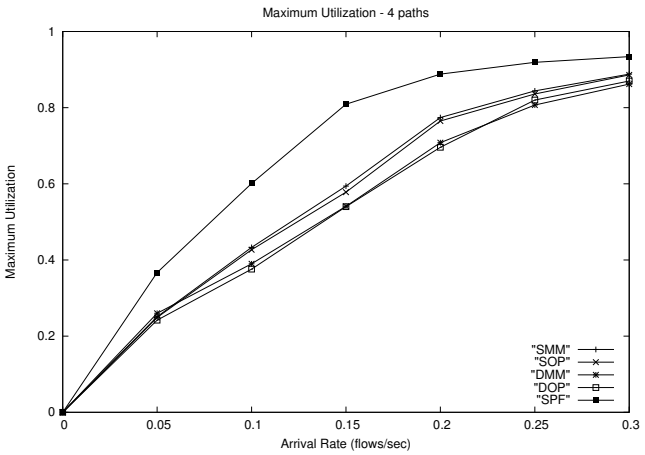

Figure 8. Maximum network utilization for UDP traffic: 4-path case

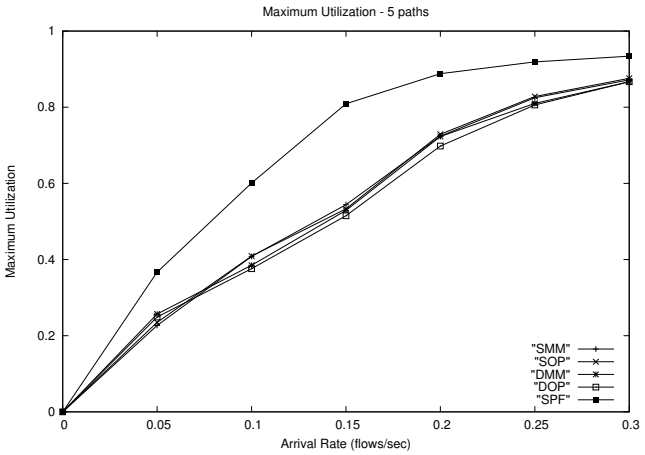

Figure 9. Maximum network utilization for UDP traffic: 5-path case

path sets.

Fig. 11 and Fig. 12 illustrate the maximum utilization of each link in the network sorted in ascending order. The x axis is the link index and the y axis shows the maximum link utilization. One can see from both figures that SPF algorithm causes uneven traffic distribution. There are more than 10 links not being utilized by the network (the low index links) while at the same time many links are highly congested (the high index links). We can see that all proposed best effort traffic engineering algorithms produce much better performance than SPF. By using these algorithms, we can both increase the utilization of under-utilized links and at the same time decrease the use of over-utilized ones. This will allow the network to support more best effort traffic in the network without investing in new high capacity links.

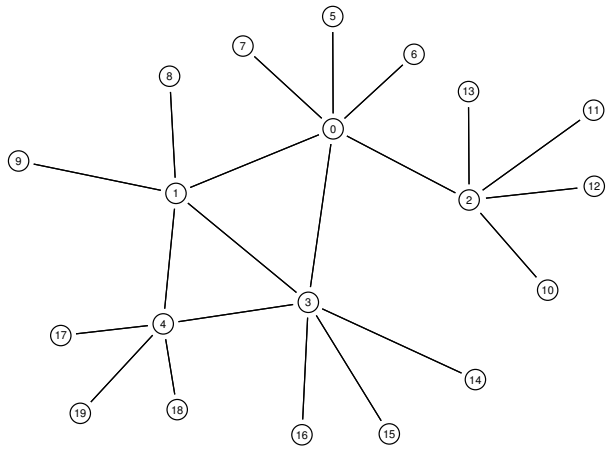

Figure 10. Transit-Stub Network Topology

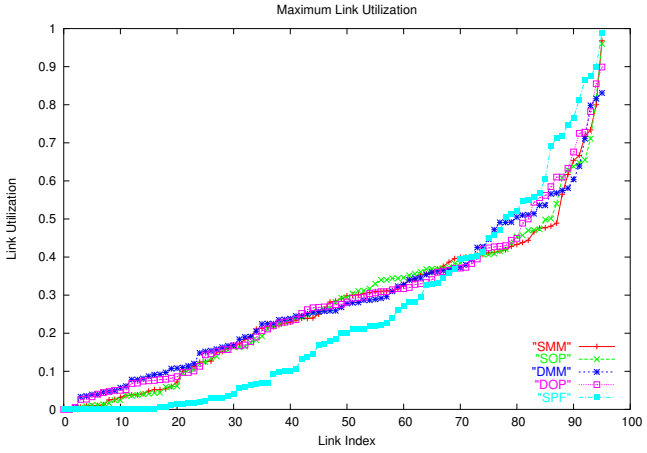

Figure 11. Maximum link utilization for TCP traffic: 3-path case, flow rate=0.7 flows/second

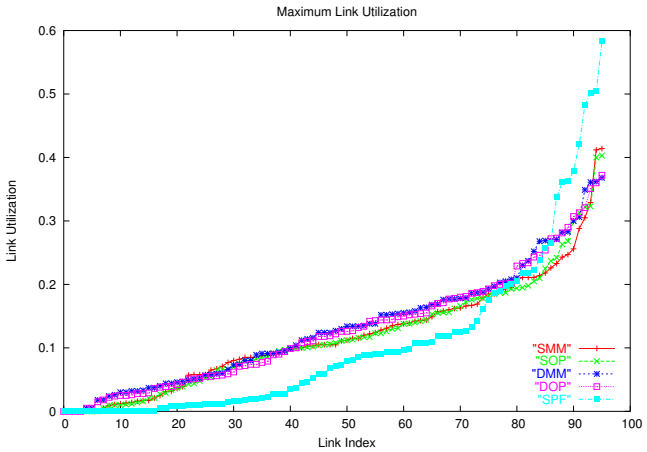

Figure 12. Maximum link utilization for UDP traffic: 3-path case, flow rate=0.1 flows/second

5. Conclusion

In this paper we have proposed a traffic engineering framework for best effort traffic. This framework is *"stateless"*, only the topology is used to determine the traffic routing. Four algorithms were investigated. The results show that there are performance improvements over SPF and traffic is more evenly distributed. Our proposed framework can improve the performace of both inelastic (e.g., UDP) and elastic (e.g., TCP) traffic. The algorithms that incorporate disjoint paths are our best candidates, DMM in TCP traffic case, and DOP and DMM in UDP traffic case.

We are presently examining additional new traffic engineering algorithms targeted for best effort traffic. Using these new algorithms we are comparing these more capable algorithms to our present set of algorithms to determine how to maximize the limited left over bandwidth which is available for best effort traffic. We believe that networks which use traditional traffic engineering algorithms coupled with admission control for the premium traffic classes and best effort traffic engineering algorithms with the remaining bandwidth left over for best effort traffic will better utilize limited network resources.

References

[1] E. Rosen, A. Viswanathan, and R. Callon, "Multiprotocol label switching architecture," RFC3031, Jan. 2001.

[2] D. Awduche, J. Malcolm, J. Agogbua, M. O'Dell, and J. McManus, "Requirements for traffic engineering over MPLS," RFC2702, Jan. 2002.

[3] Y.-D. Lin, N.-B. Hsu, and R.-H. Hwang, "QoS routing granularity in MPLS networks," *IEEE Commun. Mag.*, pp. 58–65, June 2002.

[4] B. Jamoussi, L. Andersson, R. Callon, R. Dantu, L. Wu, P. Doolan, T. Worster, N. Feldman, A. Fredette, M. Girish, E. Gray, J. Heinanen, T. Kilty, and A. Malis, "Constraint-based LSP setup using LDP," RFC3212, Jan. 2002.

[5] D. Awduche, L. Berger, D. Gan, T. Li, V. Srinivasan, and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP tunnels," RFC3209, Dec. 2001.

[6] B. Wydrowski and M. Zukerman, "QoS in best-effort networks," *IEEE Commun. Mag.*, pp. 44–49, Dec. 2002.

[7] J. W. Roberts, "Traffic theory and the internet," *IEEE Commun. Mag.*, pp. 94–99, Jan. 2001.

[8] B. Koehler, D. Barlow, H. Owen, and J. Sokol, "Traffic engieering communication protocols for best effort traffic," in *International Conference on Communication Systems and Networks CSN 2002*, Malaga, Spain, Sept. 2002, pp. 360–365.

[9] J. M. Jaffe, "Bottleneck flow control," *IEEE Trans. Commun.*, vol. 29, pp. 954–962, July 1981.

[10] J. C. R. Bennett and H. Zhang, "Hierachical packet fair queueing algorithms," *IEEE/ACM Trans. Networking*, vol. 5, pp. 675–689, Oct. 1997.

[11] E. W. Zegura, K. L. Calvert, and S. Bhattacharjee, "How to model an internetwork," in *Proc. IEEE INFOCOM'96*, Mar. 1996, pp. 594–602.