Network Working Group
Request for Comments: 3272
Category: Informational

D. Awduche
Movaz Networks
A. Chiu
Celion Networks
A. Elwalid
I. Widjaja
Lucent Technologies
X. Xiao
Redback Networks

May 2002

Overview and Principles of Internet Traffic Engineering

Status of this Memo

Copyright Notice

Abstract

This memo describes the principles of Traffic Engineering (TE) in the Internet. The document is intended to promote better understanding of the issues surrounding traffic engineering in IP networks, and to provide a common basis for the development of traffic engineering capabilities for the Internet. The principles, architectures, and methodologies for performance evaluation and performance optimization of operational IP networks are discussed throughout this document.

Table of Contents

1.0 Introduction

This memo describes the principles of Internet traffic engineering. The objective of the document is to articulate the general issues and principles for Internet traffic engineering; and where appropriate to provide recommendations, guidelines, and options for the development of online and offline Internet traffic engineering capabilities and support systems.

This document can aid service providers in devising and implementing traffic engineering solutions for their networks. Networking hardware and software vendors will also find this document helpful in the development of mechanisms and support systems for the Internet environment that support the traffic engineering function.

This document provides a terminology for describing and understanding common Internet traffic engineering concepts. This document also provides a taxonomy of known traffic engineering styles. In this context, a traffic engineering style abstracts important aspects from a traffic engineering methodology. Traffic engineering styles can be viewed in different ways depending upon the specific context in which they are used and the specific purpose which they serve. The combination of styles and views results in a natural taxonomy of traffic engineering systems.

Even though Internet traffic engineering is most effective when applied end-to-end, the initial focus of this document document is intra-domain traffic engineering (that is, traffic engineering within a given autonomous system). However, because a preponderance of Internet traffic tends to be inter-domain (originating in one autonomous system and terminating in another), this document provides an overview of aspects pertaining to inter-domain traffic engineering.

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119.

1.1. What is Internet Traffic Engineering?

Internet traffic engineering is defined as that aspect of Internet network engineering dealing with the issue of performance evaluation and performance optimization of operational IP networks. Traffic Engineering encompasses the application of technology and scientific principles to the measurement, characterization, modeling, and control of Internet traffic [RFC-2702, AWD2].

Enhancing the performance of an operational network, at both the traffic and resource levels, are major objectives of Internet traffic engineering. This is accomplished by addressing traffic oriented performance requirements, while utilizing network resources economically and reliably. Traffic oriented performance measures include delay, delay variation, packet loss, and throughput.

An important objective of Internet traffic engineering is to facilitate reliable network operations [RFC-2702]. Reliable network operations can be facilitated by providing mechanisms that enhance network integrity and by embracing policies emphasizing network survivability. This results in a minimization of the vulnerability of the network to service outages arising from errors, faults, and failures occurring within the infrastructure.

The Internet exists in order to transfer information from source nodes to destination nodes. Accordingly, one of the most significant functions performed by the Internet is the routing of traffic from ingress nodes to egress nodes. Therefore, one of the most distinctive functions performed by Internet traffic engineering is the control and optimization of the routing function, to steer traffic through the network in the most effective way.

Ultimately, it is the performance of the network as seen by end users of network services that is truly paramount. This crucial point should be considered throughout the development of traffic engineering mechanisms and policies. The characteristics visible to end users are the emergent properties of the network, which are the characteristics of the network when viewed as a whole. A central goal of the service provider, therefore, is to enhance the emergent properties of the network while taking economic considerations into account.

The importance of the above observation regarding the emergent properties of networks is that special care must be taken when choosing network performance measures to optimize. Optimizing the wrong measures may achieve certain local objectives, but may have

disastrous consequences on the emergent properties of the network and thereby on the quality of service perceived by end-users of network services.

A subtle, but practical advantage of the systematic application of traffic engineering concepts to operational networks is that it helps to identify and structure goals and priorities in terms of enhancing the quality of service delivered to end-users of network services. The application of traffic engineering concepts also aids in the measurement and analysis of the achievement of these goals.

The optimization aspects of traffic engineering can be achieved through capacity management and traffic management. As used in this document, capacity management includes capacity planning, routing control, and resource management. Network resources of particular interest include link bandwidth, buffer space, and computational resources. Likewise, as used in this document, traffic management includes (1) nodal traffic control functions such as traffic conditioning, queue management, scheduling, and (2) other functions that regulate traffic flow through the network or that arbitrate access to network resources between different packets or between different traffic streams.

The optimization objectives of Internet traffic engineering should be viewed as a continual and iterative process of network performance improvement and not simply as a one time goal. Traffic engineering also demands continual development of new technologies and new methodologies for network performance enhancement.

The optimization objectives of Internet traffic engineering may change over time as new requirements are imposed, as new technologies emerge, or as new insights are brought to bear on the underlying problems. Moreover, different networks may have different optimization objectives, depending upon their business models, capabilities, and operating constraints. The optimization aspects of traffic engineering are ultimately concerned with network control regardless of the specific optimization goals in any particular environment.

Thus, the optimization aspects of traffic engineering can be viewed from a control perspective. The aspect of control within the Internet traffic engineering arena can be pro-active and/or reactive. In the pro-active case, the traffic engineering control system takes preventive action to obviate predicted unfavorable future network states. It may also take perfective action to induce a more desirable state in the future. In the reactive case, the control system responds correctively and perhaps adaptively to events that have already transpired in the network.

The control dimension of Internet traffic engineering responds at multiple levels of temporal resolution to network events. Certain aspects of capacity management, such as capacity planning, respond at very coarse temporal levels, ranging from days to possibly years. The introduction of automatically switched optical transport networks (e.g., based on the Multi-protocol Lambda Switching concepts) could significantly reduce the lifecycle for capacity planning by expediting provisioning of optical bandwidth. Routing control functions operate at intermediate levels of temporal resolution, ranging from milliseconds to days. Finally, the packet level processing functions (e.g., rate shaping, queue management, and scheduling) operate at very fine levels of temporal resolution, ranging from picoseconds to milliseconds while responding to the real-time statistical behavior of traffic. The subsystems of Internet traffic engineering control include: capacity augmentation, routing control, traffic control, and resource control (including control of service policies at network elements). When capacity is to be augmented for tactical purposes, it may be desirable to devise a deployment plan that expedites bandwidth provisioning while minimizing installation costs.

Inputs into the traffic engineering control system include network state variables, policy variables, and decision variables.

One major challenge of Internet traffic engineering is the realization of automated control capabilities that adapt quickly and cost effectively to significant changes in a network's state, while still maintaining stability.

Another critical dimension of Internet traffic engineering is network performance evaluation, which is important for assessing the effectiveness of traffic engineering methods, and for monitoring and verifying compliance with network performance goals. Results from performance evaluation can be used to identify existing problems, guide network re-optimization, and aid in the prediction of potential future problems.

Performance evaluation can be achieved in many different ways. The most notable techniques include analytical methods, simulation, and empirical methods based on measurements. When analytical methods or simulation are used, network nodes and links can be modeled to capture relevant operational features such as topology, bandwidth, buffer space, and nodal service policies (link scheduling, packet prioritization, buffer management, etc.). Analytical traffic models can be used to depict dynamic and behavioral traffic characteristics, such as burstiness, statistical distributions, and dependence.

Performance evaluation can be quite complicated in practical network contexts. A number of techniques can be used to simplify the analysis, such as abstraction, decomposition, and approximation. For example, simplifying concepts such as effective bandwidth and effective buffer [Elwalid] may be used to approximate nodal behaviors at the packet level and simplify the analysis at the connection level. Network analysis techniques using, for example, queuing models and approximation schemes based on asymptotic and decomposition techniques can render the analysis even more tractable. In particular, an emerging set of concepts known as network calculus [CRUZ] based on deterministic bounds may simplify network analysis relative to classical stochastic techniques. When using analytical techniques, care should be taken to ensure that the models faithfully reflect the relevant operational characteristics of the modeled network entities.

Simulation can be used to evaluate network performance or to verify and validate analytical approximations. Simulation can, however, be computationally costly and may not always provide sufficient insights. An appropriate approach to a given network performance evaluation problem may involve a hybrid combination of analytical techniques, simulation, and empirical methods.

As a general rule, traffic engineering concepts and mechanisms must be sufficiently specific and well defined to address known requirements, but simultaneously flexible and extensible to accommodate unforeseen future demands.

1.2. Scope

The scope of this document is intra-domain traffic engineering; that is, traffic engineering within a given autonomous system in the Internet. This document will discuss concepts pertaining to intra-domain traffic control, including such issues as routing control, micro and macro resource allocation, and the control coordination problems that arise consequently.

This document will describe and characterize techniques already in use or in advanced development for Internet traffic engineering. The way these techniques fit together will be discussed and scenarios in which they are useful will be identified.

While this document considers various intra-domain traffic engineering approaches, it focuses more on traffic engineering with MPLS. Traffic engineering based upon manipulation of IGP metrics is not addressed in detail. This topic may be addressed by other working group document(s).

Although the emphasis is on intra-domain traffic engineering, in Section 7.0, an overview of the high level considerations pertaining to inter-domain traffic engineering will be provided. Inter-domain Internet traffic engineering is crucial to the performance enhancement of the global Internet infrastructure.

Whenever possible, relevant requirements from existing IETF documents and other sources will be incorporated by reference.

1.3 Terminology

This subsection provides terminology which is useful for Internet traffic engineering. The definitions presented apply to this document. These terms may have other meanings elsewhere.

- Baseline analysis:
  A study conducted to serve as a baseline for comparison to the actual behavior of the network.

- Busy hour:
  A one hour period within a specified interval of time (typically 24 hours) in which the traffic load in a network or sub-network is greatest.

- Bottleneck:
  A network element whose input traffic rate tends to be greater than its output rate.

- Congestion:
  A state of a network resource in which the traffic incident on the resource exceeds its output capacity over an interval of time.

- Congestion avoidance:
  An approach to congestion management that attempts to obviate the occurrence of congestion.

- Congestion control:
  An approach to congestion management that attempts to remedy congestion problems that have already occurred.

- Constraint-based routing:
  A class of routing protocols that take specified traffic attributes, network constraints, and policy constraints into account when making routing decisions. Constraint-based routing is applicable to traffic aggregates as well as flows. It is a generalization of QoS routing.

- Demand side congestion management:
  A congestion management scheme that addresses congestion problems by regulating or conditioning offered load.

- Effective bandwidth:
  The minimum amount of bandwidth that can be assigned to a flow or traffic aggregate in order to deliver 'acceptable service quality' to the flow or traffic aggregate.

- Egress traffic:
  Traffic exiting a network or network element.

- Hot-spot:
  A network element or subsystem which is in a state of congestion.

- Ingress traffic:
  Traffic entering a network or network element.

- Inter-domain traffic:
  Traffic that originates in one Autonomous system and terminates in another.

- Loss network:
  A network that does not provide adequate buffering for traffic, so that traffic entering a busy resource within the network will be dropped rather than queued.

- Metric:
  A parameter defined in terms of standard units of measurement.

- Measurement Methodology:
  A repeatable measurement technique used to derive one or more metrics of interest.

- Network Survivability:
  The capability to provide a prescribed level of QoS for existing services after a given number of failures occur within the network.

- Offline traffic engineering:
  A traffic engineering system that exists outside of the network.

- Online traffic engineering:
  A traffic engineering system that exists within the network, typically implemented on or as adjuncts to operational network elements.

- Performance measures:
  Metrics that provide quantitative or qualitative measures of the performance of systems or subsystems of interest.

- Performance management:
  A systematic approach to improving effectiveness in the accomplishment of specific networking goals related to performance improvement.

- Performance Metric:
  A performance parameter defined in terms of standard units of measurement.

- Provisioning:
  The process of assigning or configuring network resources to meet certain requests.

- QoS routing:
  Class of routing systems that selects paths to be used by a flow based on the QoS requirements of the flow.

- Service Level Agreement:
  A contract between a provider and a customer that guarantees specific levels of performance and reliability at a certain cost.

- Stability:
  An operational state in which a network does not oscillate in a disruptive manner from one mode to another mode.

- Supply side congestion management:
  A congestion management scheme that provisions additional network resources to address existing and/or anticipated congestion problems.

- Transit traffic:
  Traffic whose origin and destination are both outside of the network under consideration.

- Traffic characteristic:
  A description of the temporal behavior or a description of the attributes of a given traffic flow or traffic aggregate.

- Traffic engineering system:
  A collection of objects, mechanisms, and protocols that are used conjunctively to accomplish traffic engineering objectives.

- Traffic flow:
  A stream of packets between two end-points that can be characterized in a certain way. A micro-flow has a more specific definition: A micro-flow is a stream of packets with the same source and destination addresses, source and destination ports, and protocol ID.

- Traffic intensity:
  A measure of traffic loading with respect to a resource capacity over a specified period of time. In classical telephony systems, traffic intensity is measured in units of Erlang.

- Traffic matrix:
  A representation of the traffic demand between a set of origin and destination abstract nodes. An abstract node can consist of one or more network elements.

- Traffic monitoring:
  The process of observing traffic characteristics at a given point in a network and collecting the traffic information for analysis and further action.

- Traffic trunk:
  An aggregation of traffic flows belonging to the same class which are forwarded through a common path. A traffic trunk may be characterized by an ingress and egress node, and a set of attributes which determine its behavioral characteristics and requirements from the network.

2.0 Background

The Internet has quickly evolved into a very critical communications infrastructure, supporting significant economic, educational, and social activities. Simultaneously, the delivery of Internet communications services has become very competitive and end-users are demanding very high quality service from their service providers. Consequently, performance optimization of large scale IP networks, especially public Internet backbones, have become an important problem. Network performance requirements are multi-dimensional, complex, and sometimes contradictory; making the traffic engineering problem very challenging.

The network must convey IP packets from ingress nodes to egress nodes efficiently, expeditiously, and economically. Furthermore, in a multiclass service environment (e.g., Diffserv capable networks), the resource sharing parameters of the network must be appropriately determined and configured according to prevailing policies and service models to resolve resource contention issues arising from mutual interference between packets traversing through the network. Thus, consideration must be given to resolving competition for network resources between traffic streams belonging to the same service class (intra-class contention resolution) and traffic streams belonging to different classes (inter-class contention resolution).

2.1 Context of Internet Traffic Engineering

The context of Internet traffic engineering pertains to the scenarios where traffic engineering is used. A traffic engineering methodology establishes appropriate rules to resolve traffic performance issues occurring in a specific context. The context of Internet traffic engineering includes:

(1) A network context defining the universe of discourse, and in particular the situations in which the traffic engineering problems occur. The network context includes network structure, network policies, network characteristics, network constraints, network quality attributes, and network optimization criteria.

(2) A problem context defining the general and concrete issues that traffic engineering addresses. The problem context includes identification, abstraction of relevant features, representation, formulation, specification of the requirements on the solution space, and specification of the desirable features of acceptable solutions.

(3) A solution context suggesting how to address the issues identified by the problem context. The solution context includes analysis, evaluation of alternatives, prescription, and resolution.

(4) An implementation and operational context in which the solutions are methodologically instantiated. The implementation and operational context includes planning, organization, and execution.

The context of Internet traffic engineering and the different problem scenarios are discussed in the following subsections.

2.2 Network Context

IP networks range in size from small clusters of routers situated within a given location, to thousands of interconnected routers, switches, and other components distributed all over the world.

Conceptually, at the most basic level of abstraction, an IP network can be represented as a distributed dynamical system consisting of: (1) a set of interconnected resources which provide transport services for IP traffic subject to certain constraints, (2) a demand system representing the offered load to be transported through the network, and (3) a response system consisting of network processes, protocols, and related mechanisms which facilitate the movement of traffic through the network [see also AWD2].

The network elements and resources may have specific characteristics restricting the manner in which the demand is handled. Additionally, network resources may be equipped with traffic control mechanisms superintending the way in which the demand is serviced. Traffic control mechanisms may, for example, be used to control various packet processing activities within a given resource, arbitrate contention for access to the resource by different packets, and regulate traffic behavior through the resource. A configuration management and provisioning system may allow the settings of the traffic control mechanisms to be manipulated by external or internal entities in order to exercise control over the way in which the network elements respond to internal and external stimuli.

The details of how the network provides transport services for packets are specified in the policies of the network administrators and are installed through network configuration management and policy based provisioning systems. Generally, the types of services provided by the network also depends upon the technology and characteristics of the network elements and protocols, the prevailing service and utility models, and the ability of the network administrators to translate policies into network configurations.

Contemporary Internet networks have three significant characteristics: (1) they provide real-time services, (2) they have become mission critical, and (3) their operating environments are very dynamic. The dynamic characteristics of IP networks can be attributed in part to fluctuations in demand, to the interaction between various network protocols and processes, to the rapid evolution of the infrastructure which demands the constant inclusion of new technologies and new network elements, and to transient and persistent impairments which occur within the system.

Packets contend for the use of network resources as they are conveyed through the network. A network resource is considered to be congested if the arrival rate of packets exceed the output capacity of the resource over an interval of time. Congestion may result in some of the arrival packets being delayed or even dropped.

Congestion increases transit delays, delay variation, packet loss, and reduces the predictability of network services. Clearly, congestion is a highly undesirable phenomenon.

Combating congestion at a reasonable cost is a major objective of Internet traffic engineering.

Efficient sharing of network resources by multiple traffic streams is a basic economic premise for packet switched networks in general and for the Internet in particular. A fundamental challenge in network operation, especially in a large scale public IP network, is to increase the efficiency of resource utilization while minimizing the possibility of congestion.

Increasingly, the Internet will have to function in the presence of different classes of traffic with different service requirements. The advent of Differentiated Services [RFC-2475] makes this requirement particularly acute. Thus, packets may be grouped into behavior aggregates such that each behavior aggregate may have a common set of behavioral characteristics or a common set of delivery requirements. In practice, the delivery requirements of a specific set of packets may be specified explicitly or implicitly. Two of the most important traffic delivery requirements are capacity constraints and QoS constraints.

Capacity constraints can be expressed statistically as peak rates, mean rates, burst sizes, or as some deterministic notion of effective bandwidth. QoS requirements can be expressed in terms of (1) integrity constraints such as packet loss and (2) in terms of temporal constraints such as timing restrictions for the delivery of each packet (delay) and timing restrictions for the delivery of consecutive packets belonging to the same traffic stream (delay variation).

2.3 Problem Context

Fundamental problems exist in association with the operation of a network described by the simple model of the previous subsection. This subsection reviews the problem context in relation to the traffic engineering function.

The identification, abstraction, representation, and measurement of network features relevant to traffic engineering is a significant issue.

One particularly important class of problems concerns how to explicitly formulate the problems that traffic engineering attempts to solve, how to identify the requirements on the solution space, how to specify the desirable features of good solutions, how to actually solve the problems, and how to measure and characterize the effectiveness of the solutions.

Another class of problems concerns how to measure and estimate relevant network state parameters. Effective traffic engineering relies on a good estimate of the offered traffic load as well as a view of the underlying topology and associated resource constraints. A network-wide view of the topology is also a must for offline planning.

Still another class of problems concerns how to characterize the state of the network and how to evaluate its performance under a variety of scenarios. The performance evaluation problem is two-fold. One aspect of this problem relates to the evaluation of the system level performance of the network. The other aspect relates to the evaluation of the resource level performance, which restricts attention to the performance analysis of individual network resources. In this memo, we refer to the system level characteristics of the network as the "macro-states" and the resource level characteristics as the "micro-states." The system level characteristics are also known as the emergent properties of the network as noted earlier. Correspondingly, we shall refer to the traffic engineering schemes dealing with network performance optimization at the systems level as "macro-TE" and the schemes that optimize at the individual resource level as "micro-TE." Under certain circumstances, the system level performance can be derived from the resource level performance using appropriate rules of composition, depending upon the particular performance measures of interest.

Another fundamental class of problems concerns how to effectively optimize network performance. Performance optimization may entail translating solutions to specific traffic engineering problems into network configurations. Optimization may also entail some degree of resource management control, routing control, and/or capacity augmentation.

As noted previously, congestion is an undesirable phenomena in operational networks. Therefore, the next subsection addresses the issue of congestion and its ramifications within the problem context of Internet traffic engineering.

2.3.1 Congestion and its Ramifications

Congestion is one of the most significant problems in an operational IP context. A network element is said to be congested if it experiences sustained overload over an interval of time. Congestion almost always results in degradation of service quality to end users. Congestion control schemes can include demand side policies and supply side policies. Demand side policies may restrict access to congested resources and/or dynamically regulate the demand to alleviate the overload situation. Supply side policies may expand or augment network capacity to better accommodate offered traffic. Supply side policies may also re-allocate network resources by redistributing traffic over the infrastructure. Traffic redistribution and resource re-allocation serve to increase the 'effective capacity' seen by the demand.

The emphasis of this memo is primarily on congestion management schemes falling within the scope of the network, rather than on congestion management systems dependent upon sensitivity and adaptivity from end-systems. That is, the aspects that are considered in this memo with respect to congestion management are those solutions that can be provided by control entities operating on the network and by the actions of network administrators and network operations systems.

2.4 Solution Context

The solution context for Internet traffic engineering involves analysis, evaluation of alternatives, and choice between alternative courses of action. Generally the solution context is predicated on making reasonable inferences about the current or future state of the network, and subsequently making appropriate decisions that may involve a preference between alternative sets of action. More specifically, the solution context demands reasonable estimates of traffic workload, characterization of network state, deriving solutions to traffic engineering problems which may be implicitly or explicitly formulated, and possibly instantiating a set of control actions. Control actions may involve the manipulation of parameters associated with routing, control over tactical capacity acquisition, and control over the traffic management functions.

The following list of instruments may be applicable to the solution context of Internet traffic engineering.

(1) A set of policies, objectives, and requirements (which may be context dependent) for network performance evaluation and performance optimization.

(2) A collection of online and possibly offline tools and mechanisms for measurement, characterization, modeling, and control of Internet traffic and control over the placement and allocation of network resources, as well as control over the mapping or distribution of traffic onto the infrastructure.

(3) A set of constraints on the operating environment, the network protocols, and the traffic engineering system itself.

(4) A set of quantitative and qualitative techniques and methodologies for abstracting, formulating, and solving traffic engineering problems.

(5) A set of administrative control parameters which may be manipulated through a Configuration Management (CM) system. The CM system itself may include a configuration control subsystem, a configuration repository, a configuration accounting subsystem, and a configuration auditing subsystem.

(6) A set of guidelines for network performance evaluation, performance optimization, and performance improvement.

Derivation of traffic characteristics through measurement and/or estimation is very useful within the realm of the solution space for traffic engineering. Traffic estimates can be derived from customer subscription information, traffic projections, traffic models, and from actual empirical measurements. The empirical measurements may be performed at the traffic aggregate level or at the flow level in order to derive traffic statistics at various levels of detail. Measurements at the flow level or on small traffic aggregates may be performed at edge nodes, where traffic enters and leaves the network. Measurements at large traffic aggregate levels may be performed within the core of the network where potentially numerous traffic flows may be in transit concurrently.

To conduct performance studies and to support planning of existing and future networks, a routing analysis may be performed to determine the path(s) the routing protocols will choose for various traffic demands, and to ascertain the utilization of network resources as traffic is routed through the network. The routing analysis should capture the selection of paths through the network, the assignment of

traffic across multiple feasible routes, and the multiplexing of IP traffic over traffic trunks (if such constructs exists) and over the underlying network infrastructure. A network topology model is a necessity for routing analysis. A network topology model may be extracted from network architecture documents, from network designs, from information contained in router configuration files, from routing databases, from routing tables, or from automated tools that discover and depict network topology information. Topology information may also be derived from servers that monitor network state, and from servers that perform provisioning functions.

Routing in operational IP networks can be administratively controlled at various levels of abstraction including the manipulation of BGP attributes and manipulation of IGP metrics. For path oriented technologies such as MPLS, routing can be further controlled by the manipulation of relevant traffic engineering parameters, resource parameters, and administrative policy constraints. Within the context of MPLS, the path of an explicit label switched path (LSP) can be computed and established in various ways including: (1) manually, (2) automatically online using constraint-based routing processes implemented on label switching routers, and (3) automatically offline using constraint-based routing entities implemented on external traffic engineering support systems.

2.4.1 Combating the Congestion Problem

Minimizing congestion is a significant aspect of Internet traffic engineering. This subsection gives an overview of the general approaches that have been used or proposed to combat congestion problems.

Congestion management policies can be categorized based upon the following criteria (see e.g., [YARE95] for a more detailed taxonomy of congestion control schemes): (1) Response time scale which can be characterized as long, medium, or short; (2) reactive versus preventive which relates to congestion control and congestion avoidance; and (3) supply side versus demand side congestion management schemes. These aspects are discussed in the following paragraphs.

(1) Congestion Management based on Response Time Scales

- Long (weeks to months): Capacity planning works over a relatively long time scale to expand network capacity based on estimates or forecasts of future traffic demand and traffic distribution. Since router and link provisioning take time and are generally expensive, these upgrades are typically carried out in the weeks-to-months or even years time scale.

- Medium (minutes to days): Several control policies fall within the medium time scale category. Examples include: (1) Adjusting IGP and/or BGP parameters to route traffic away or towards certain segments of the network; (2) Setting up and/or adjusting some explicitly routed label switched paths (ER-LSPs) in MPLS networks to route some traffic trunks away from possibly congested resources or towards possibly more favorable routes; (3) re-configuring the logical topology of the network to make it correlate more closely with the spatial traffic distribution using for example some underlying path-oriented technology such as MPLS LSPs, ATM PVCs, or optical channel trails. Many of these adaptive medium time scale response schemes rely on a measurement system that monitors changes in traffic distribution, traffic shifts, and network resource utilization and subsequently provides feedback to the online and/or offline traffic engineering mechanisms and tools which employ this feedback information to trigger certain control actions to occur within the network. The traffic engineering mechanisms and tools can be implemented in a distributed fashion or in a centralized fashion, and may have a hierarchical structure or a flat structure. The comparative merits of distributed and centralized control structures for networks are well known. A centralized scheme may have global visibility into the network state and may produce potentially more optimal solutions. However, centralized schemes are prone to single points of failure and may not scale as well as distributed schemes. Moreover, the information utilized by a centralized scheme may be stale and may not reflect the actual state of the network. It is not an objective of this memo to make a recommendation between distributed and centralized schemes. This is a choice that network administrators must make based on their specific needs.

- Short (picoseconds to minutes): This category includes packet level processing functions and events on the order of several round trip times. It includes router mechanisms such as passive and active buffer management. These mechanisms are used to control congestion and/or signal congestion to end systems so that they can adaptively regulate the rate at which traffic is injected into the network. One of the most popular active queue management schemes, especially for TCP traffic, is Random Early Detection (RED) [FLJA93], which supports congestion avoidance by controlling the average queue size. During congestion (but before the queue is filled), the RED scheme chooses arriving packets to "mark" according to a probabilistic algorithm which takes into account the average queue size. For a router that does not utilize explicit congestion notification (ECN) see e.g., [FLOY94], the marked packets can simply be dropped to signal the inception of congestion to end systems. On the other hand, if the router supports ECN, then it can set the ECN field in the packet header. Several variations of RED have been proposed to support different drop precedence levels in multi-class environments [RFC-

2597], e.g., RED with In and Out (RIO) and Weighted RED. There is general consensus that RED provides congestion avoidance performance which is not worse than traditional Tail-Drop (TD) queue management (drop arriving packets only when the queue is full). Importantly, however, RED reduces the possibility of global synchronization and improves fairness among different TCP sessions. However, RED by itself can not prevent congestion and unfairness caused by sources unresponsive to RED, e.g., UDP traffic and some misbehaved greedy connections. Other schemes have been proposed to improve the performance and fairness in the presence of unresponsive traffic. Some of these schemes were proposed as theoretical frameworks and are typically not available in existing commercial products. Two such schemes are Longest Queue Drop (LQD) and Dynamic Soft Partitioning with Random Drop (RND) [SLDC98].

(2) Congestion Management: Reactive versus Preventive Schemes

- Reactive: reactive (recovery) congestion management policies react to existing congestion problems to improve it. All the policies described in the long and medium time scales above can be categorized as being reactive especially if the policies are based on monitoring and identifying existing congestion problems, and on the initiation of relevant actions to ease a situation.

- Preventive: preventive (predictive/avoidance) policies take proactive action to prevent congestion based on estimates and predictions of future potential congestion problems. Some of the policies described in the long and medium time scales fall into this category. They do not necessarily respond immediately to existing congestion problems. Instead forecasts of traffic demand and workload distribution are considered and action may be taken to prevent potential congestion problems in the future. The schemes described in the short time scale (e.g., RED and its variations, ECN, LQD, and RND) are also used for congestion avoidance since dropping or marking packets before queues actually overflow would trigger corresponding TCP sources to slow down.

(3) Congestion Management: Supply Side versus Demand Side Schemes

- Supply side: supply side congestion management policies increase the effective capacity available to traffic in order to control or obviate congestion. This can be accomplished by augmenting capacity. Another way to accomplish this is to minimize congestion by having a relatively balanced distribution of traffic over the network. For example, capacity planning should aim to provide a physical topology and associated link bandwidths that match estimated traffic workload and traffic distribution based on forecasting (subject to budgetary and other constraints). However, if actual traffic distribution does

not match the topology derived from capacity panning (due to forecasting errors or facility constraints for example), then the traffic can be mapped onto the existing topology using routing control mechanisms, using path oriented technologies (e.g., MPLS LSPs and optical channel trails) to modify the logical topology, or by using some other load redistribution mechanisms.

- Demand side: demand side congestion management policies control or regulate the offered traffic to alleviate congestion problems. For example, some of the short time scale mechanisms described earlier (such as RED and its variations, ECN, LQD, and RND) as well as policing and rate shaping mechanisms attempt to regulate the offered load in various ways. Tariffs may also be applied as a demand side instrument. To date, however, tariffs have not been used as a means of demand side congestion management within the Internet.

In summary, a variety of mechanisms can be used to address congestion problems in IP networks. These mechanisms may operate at multiple time-scales.

2.5 Implementation and Operational Context

The operational context of Internet traffic engineering is characterized by constant change which occur at multiple levels of abstraction. The implementation context demands effective planning, organization, and execution. The planning aspects may involve determining prior sets of actions to achieve desired objectives. Organizing involves arranging and assigning responsibility to the various components of the traffic engineering system and coordinating the activities to accomplish the desired TE objectives. Execution involves measuring and applying corrective or perfective actions to attain and maintain desired TE goals.

3.0 Traffic Engineering Process Model(s)

This section describes a generic process model that captures the high level practical aspects of Internet traffic engineering in an operational context. The process model is described as a sequence of actions that a traffic engineer, or more generally a traffic engineering system, must perform to optimize the performance of an operational network (see also [RFC-2702, AWD2]). The process model described here represents the broad activities common to most traffic engineering methodologies although the details regarding how traffic engineering is executed may differ from network to network. This process model may be enacted explicitly or implicitly, by an automaton and/or by a human.

The traffic engineering process model is iterative [AWD2]. The four phases of the process model described below are repeated continually.

The first phase of the TE process model is to define the relevant control policies that govern the operation of the network. These policies may depend upon many factors including the prevailing business model, the network cost structure, the operating constraints, the utility model, and optimization criteria.

The second phase of the process model is a feedback mechanism involving the acquisition of measurement data from the operational network. If empirical data is not readily available from the network, then synthetic workloads may be used instead which reflect either the prevailing or the expected workload of the network. Synthetic workloads may be derived by estimation or extrapolation using prior empirical data. Their derivation may also be obtained using mathematical models of traffic characteristics or other means.

The third phase of the process model is to analyze the network state and to characterize traffic workload. Performance analysis may be proactive and/or reactive. Proactive performance analysis identifies potential problems that do not exist, but could manifest in the future. Reactive performance analysis identifies existing problems, determines their cause through diagnosis, and evaluates alternative approaches to remedy the problem, if necessary. A number of quantitative and qualitative techniques may be used in the analysis process, including modeling based analysis and simulation. The analysis phase of the process model may involve investigating the concentration and distribution of traffic across the network or relevant subsets of the network, identifying the characteristics of the offered traffic workload, identifying existing or potential bottlenecks, and identifying network pathologies such as ineffective link placement, single points of failures, etc. Network pathologies may result from many factors including inferior network architecture, inferior network design, and configuration problems. A traffic matrix may be constructed as part of the analysis process. Network analysis may also be descriptive or prescriptive.

The fourth phase of the TE process model is the performance optimization of the network. The performance optimization phase involves a decision process which selects and implements a set of actions from a set of alternatives. Optimization actions may include the use of appropriate techniques to either control the offered traffic or to control the distribution of traffic across the network. Optimization actions may also involve adding additional links or increasing link capacity, deploying additional hardware such as routers and switches, systematically adjusting parameters associated with routing such as IGP metrics and BGP attributes, and adjusting

traffic management parameters. Network performance optimization may also involve starting a network planning process to improve the network architecture, network design, network capacity, network technology, and the configuration of network elements to accommodate current and future growth.

3.1 Components of the Traffic Engineering Process Model

The key components of the traffic engineering process model include a measurement subsystem, a modeling and analysis subsystem, and an optimization subsystem. The following subsections examine these components as they apply to the traffic engineering process model.

3.2 Measurement

Measurement is crucial to the traffic engineering function. The operational state of a network can be conclusively determined only through measurement. Measurement is also critical to the optimization function because it provides feedback data which is used by traffic engineering control subsystems. This data is used to adaptively optimize network performance in response to events and stimuli originating within and outside the network. Measurement is also needed to determine the quality of network services and to evaluate the effectiveness of traffic engineering policies. Experience suggests that measurement is most effective when acquired and applied systematically.

When developing a measurement system to support the traffic engineering function in IP networks, the following questions should be carefully considered: Why is measurement needed in this particular context? What parameters are to be measured? How should the measurement be accomplished? Where should the measurement be performed? When should the measurement be performed? How frequently should the monitored variables be measured? What level of measurement accuracy and reliability is desirable? What level of measurement accuracy and reliability is realistically attainable? To what extent can the measurement system permissibly interfere with the monitored network components and variables? What is the acceptable cost of measurement? The answers to these questions will determine the measurement tools and methodologies appropriate in any given traffic engineering context.

It should also be noted that there is a distinction between measurement and evaluation. Measurement provides raw data concerning state parameters and variables of monitored network elements. Evaluation utilizes the raw data to make inferences regarding the monitored system.

Measurement in support of the TE function can occur at different levels of abstraction. For example, measurement can be used to derive packet level characteristics, flow level characteristics, user or customer level characteristics, traffic aggregate characteristics, component level characteristics, and network wide characteristics.

3.3 Modeling, Analysis, and Simulation

Modeling and analysis are important aspects of Internet traffic engineering. Modeling involves constructing an abstract or physical representation which depicts relevant traffic characteristics and network attributes.

A network model is an abstract representation of the network which captures relevant network features, attributes, and characteristics, such as link and nodal attributes and constraints. A network model may facilitate analysis and/or simulation which can be used to predict network performance under various conditions as well as to guide network expansion plans.

In general, Internet traffic engineering models can be classified as either structural or behavioral. Structural models focus on the organization of the network and its components. Behavioral models focus on the dynamics of the network and the traffic workload. Modeling for Internet traffic engineering may also be formal or informal.

Accurate behavioral models for traffic sources are particularly useful for analysis. Development of behavioral traffic source models that are consistent with empirical data obtained from operational networks is a major research topic in Internet traffic engineering. These source models should also be tractable and amenable to analysis. The topic of source models for IP traffic is a research topic and is therefore outside the scope of this document. Its importance, however, must be emphasized.

Network simulation tools are extremely useful for traffic engineering. Because of the complexity of realistic quantitative analysis of network behavior, certain aspects of network performance studies can only be conducted effectively using simulation. A good network simulator can be used to mimic and visualize network characteristics under various conditions in a safe and non-disruptive manner. For example, a network simulator may be used to depict congested resources and hot spots, and to provide hints regarding possible solutions to network performance problems. A good simulator may also be used to validate the effectiveness of planned solutions to network issues without the need to tamper with the operational network, or to commence an expensive network upgrade which may not

achieve the desired objectives. Furthermore, during the process of network planning, a network simulator may reveal pathologies such as single points of failure which may require additional redundancy, and potential bottlenecks and hot spots which may require additional capacity.

Routing simulators are especially useful in large networks. A routing simulator may identify planned links which may not actually be used to route traffic by the existing routing protocols. Simulators can also be used to conduct scenario based and perturbation based analysis, as well as sensitivity studies. Simulation results can be used to initiate appropriate actions in various ways. For example, an important application of network simulation tools is to investigate and identify how best to make the network evolve and grow, in order to accommodate projected future demands.

3.4 Optimization

Network performance optimization involves resolving network issues by transforming such issues into concepts that enable a solution, identification of a solution, and implementation of the solution. Network performance optimization can be corrective or perfective. In corrective optimization, the goal is to remedy a problem that has occurred or that is incipient. In perfective optimization, the goal is to improve network performance even when explicit problems do not exist and are not anticipated.

Network performance optimization is a continual process, as noted previously. Performance optimization iterations may consist of real-time optimization sub-processes and non-real-time network planning sub-processes. The difference between real-time optimization and network planning is primarily in the relative time-scale in which they operate and in the granularity of actions. One of the objectives of a real-time optimization sub-process is to control the mapping and distribution of traffic over the existing network infrastructure to avoid and/or relieve congestion, to assure satisfactory service delivery, and to optimize resource utilization. Real-time optimization is needed because random incidents such as fiber cuts or shifts in traffic demand will occur irrespective of how well a network is designed. These incidents can cause congestion and other problems to manifest in an operational network. Real-time optimization must solve such problems in small to medium time-scales ranging from micro-seconds to minutes or hours. Examples of real-time optimization include queue management, IGP/BGP metric tuning, and using technologies such as MPLS explicit LSPs to change the paths of some traffic trunks [XIAO].

One of the functions of the network planning sub-process is to initiate actions to systematically evolve the architecture, technology, topology, and capacity of a network. When a problem exists in the network, real-time optimization should provide an immediate remedy. Because a prompt response is necessary, the real-time solution may not be the best possible solution. Network planning may subsequently be needed to refine the solution and improve the situation. Network planning is also required to expand the network to support traffic growth and changes in traffic distribution over time. As previously noted, a change in the topology and/or capacity of the network may be the outcome of network planning.

Clearly, network planning and real-time performance optimization are mutually complementary activities. A well-planned and designed network makes real-time optimization easier, while a systematic approach to real-time network performance optimization allows network planning to focus on long term issues rather than tactical considerations. Systematic real-time network performance optimization also provides valuable inputs and insights toward network planning.

Stability is an important consideration in real-time network performance optimization. This aspect will be repeatedly addressed throughout this memo.

4.0 Historical Review and Recent Developments

This section briefly reviews different traffic engineering approaches proposed and implemented in telecommunications and computer networks. The discussion is not intended to be comprehensive. It is primarily intended to illuminate pre-existing perspectives and prior art concerning traffic engineering in the Internet and in legacy telecommunications networks.

4.1 Traffic Engineering in Classical Telephone Networks

This subsection presents a brief overview of traffic engineering in telephone networks which often relates to the way user traffic is steered from an originating node to the terminating node. This subsection presents a brief overview of this topic. A detailed description of the various routing strategies applied in telephone networks is included in the book by G. Ash [ASH2].

The early telephone network relied on static hierarchical routing, whereby routing patterns remained fixed independent of the state of the network or time of day. The hierarchy was intended to accommodate overflow traffic, improve network reliability via

alternate routes, and prevent call looping by employing strict hierarchical rules. The network was typically over-provisioned since a given fixed route had to be dimensioned so that it could carry user traffic during a busy hour of any busy day. Hierarchical routing in the telephony network was found to be too rigid upon the advent of digital switches and stored program control which were able to manage more complicated traffic engineering rules.

Dynamic routing was introduced to alleviate the routing inflexibility in the static hierarchical routing so that the network would operate more efficiently. This resulted in significant economic gains [HUSS87]. Dynamic routing typically reduces the overall loss probability by 10 to 20 percent (compared to static hierarchical routing). Dynamic routing can also improve network resilience by recalculating routes on a per-call basis and periodically updating routes.

There are three main types of dynamic routing in the telephone network. They are time-dependent routing, state-dependent routing (SDR), and event dependent routing (EDR).

In time-dependent routing, regular variations in traffic loads (such as time of day or day of week) are exploited in pre-planned routing tables. In state-dependent routing, routing tables are updated online according to the current state of the network (e.g., traffic demand, utilization, etc.). In event dependent routing, routing changes are incepted by events (such as call setups encountering congested or blocked links) whereupon new paths are searched out using learning models. EDR methods are real-time adaptive, but they do not require global state information as does SDR. Examples of EDR schemes include the dynamic alternate routing (DAR) from BT, the state-and-time dependent routing (STR) from NTT, and the success-to-the-top (STT) routing from AT&T.

Dynamic non-hierarchical routing (DNHR) is an example of dynamic routing that was introduced in the AT&T toll network in the 1980's to respond to time-dependent information such as regular load variations as a function of time. Time-dependent information in terms of load may be divided into three time scales: hourly, weekly, and yearly. Correspondingly, three algorithms are defined to pre-plan the routing tables. The network design algorithm operates over a year-long interval while the demand servicing algorithm operates on a weekly basis to fine tune link sizes and routing tables to correct forecast errors on the yearly basis. At the smallest time scale, the routing algorithm is used to make limited adjustments based on daily traffic variations. Network design and demand servicing are computed using offline calculations. Typically, the calculations require extensive searches on possible routes. On the other hand, routing may need

online calculations to handle crankback. DNHR adopts a "two-link" approach whereby a path can consist of two links at most. The routing algorithm presents an ordered list of route choices between an originating switch and a terminating switch. If a call overflows, a via switch (a tandem exchange between the originating switch and the terminating switch) would send a crankback signal to the originating switch. This switch would then select the next route, and so on, until there are no alternative routes available in which the call is blocked.

4.2 Evolution of Traffic Engineering in Packet Networks

This subsection reviews related prior work that was intended to improve the performance of data networks. Indeed, optimization of the performance of data networks started in the early days of the ARPANET. Other early commercial networks such as SNA also recognized the importance of performance optimization and service differentiation.

In terms of traffic management, the Internet has been a best effort service environment until recently. In particular, very limited traffic management capabilities existed in IP networks to provide differentiated queue management and scheduling services to packets belonging to different classes.

In terms of routing control, the Internet has employed distributed protocols for intra-domain routing. These protocols are highly scalable and resilient. However, they are based on simple algorithms for path selection which have very limited functionality to allow flexible control of the path selection process.

In the following subsections, the evolution of practical traffic engineering mechanisms in IP networks and its predecessors are reviewed.

4.2.1 Adaptive Routing in the ARPANET

The early ARPANET recognized the importance of adaptive routing where routing decisions were based on the current state of the network [MCQ80]. Early minimum delay routing approaches forwarded each packet to its destination along a path for which the total estimated transit time was the smallest. Each node maintained a table of network delays, representing the estimated delay that a packet would experience along a given path toward its destination. The minimum delay table was periodically transmitted by a node to its neighbors. The shortest path, in terms of hop count, was also propagated to give the connectivity information.

One drawback to this approach is that dynamic link metrics tend to create "traffic magnets" causing congestion to be shifted from one location of a network to another location, resulting in oscillation and network instability.

4.2.2 Dynamic Routing in the Internet

The Internet evolved from the APARNET and adopted dynamic routing algorithms with distributed control to determine the paths that packets should take en-route to their destinations. The routing algorithms are adaptations of shortest path algorithms where costs are based on link metrics. The link metric can be based on static or dynamic quantities. The link metric based on static quantities may be assigned administratively according to local criteria. The link metric based on dynamic quantities may be a function of a network congestion measure such as delay or packet loss.

It was apparent early that static link metric assignment was inadequate because it can easily lead to unfavorable scenarios in which some links become congested while others remain lightly loaded. One of the many reasons for the inadequacy of static link metrics is that link metric assignment was often done without considering the traffic matrix in the network. Also, the routing protocols did not take traffic attributes and capacity constraints into account when making routing decisions. This results in traffic concentration being localized in subsets of the network infrastructure and potentially causing congestion. Even if link metrics are assigned in accordance with the traffic matrix, unbalanced loads in the network can still occur due to a number factors including:

- Resources may not be deployed in the most optimal locations from a routing perspective.

- Forecasting errors in traffic volume and/or traffic distribution.

- Dynamics in traffic matrix due to the temporal nature of traffic patterns, BGP policy change from peers, etc.

The inadequacy of the legacy Internet interior gateway routing system is one of the factors motivating the interest in path oriented technology with explicit routing and constraint-based routing capability such as MPLS.

4.2.3 ToS Routing

Type-of-Service (ToS) routing involves different routes going to the same destination with selection dependent upon the ToS field of an IP packet [RFC-2474]. The ToS classes may be classified as low delay and high throughput. Each link is associated with multiple link costs and each link cost is used to compute routes for a particular ToS. A separate shortest path tree is computed for each ToS. The shortest path algorithm must be run for each ToS resulting in very expensive computation. Classical ToS-based routing is now outdated as the IP header field has been replaced by a Diffserv field. Effective traffic engineering is difficult to perform in classical ToS-based routing because each class still relies exclusively on shortest path routing which results in localization of traffic concentration within the network.

4.2.4 Equal Cost Multi-Path

Equal Cost Multi-Path (ECMP) is another technique that attempts to address the deficiency in the Shortest Path First (SPF) interior gateway routing systems [RFC-2328]. In the classical SPF algorithm, if two or more shortest paths exist to a given destination, the algorithm will choose one of them. The algorithm is modified slightly in ECMP so that if two or more equal cost shortest paths exist between two nodes, the traffic between the nodes is distributed among the multiple equal-cost paths. Traffic distribution across the equal-cost paths is usually performed in one of two ways: (1) packet-based in a round-robin fashion, or (2) flow-based using hashing on source and destination IP addresses and possibly other fields of the IP header. The first approach can easily cause out-of-order packets while the second approach is dependent upon the number and distribution of flows. Flow-based load sharing may be unpredictable in an enterprise network where the number of flows is relatively small and less heterogeneous (for example, hashing may not be uniform), but it is generally effective in core public networks where the number of flows is large and heterogeneous.

In ECMP, link costs are static and bandwidth constraints are not considered, so ECMP attempts to distribute the traffic as equally as possible among the equal-cost paths independent of the congestion status of each path. As a result, given two equal-cost paths, it is possible that one of the paths will be more congested than the other. Another drawback of ECMP is that load sharing cannot be achieved on multiple paths which have non-identical costs.

4.2.5 Nimrod

Nimrod is a routing system developed to provide heterogeneous service specific routing in the Internet, while taking multiple constraints into account [RFC-1992]. Essentially, Nimrod is a link state routing protocol which supports path oriented packet forwarding. It uses the concept of maps to represent network connectivity and services at multiple levels of abstraction. Mechanisms are provided to allow restriction of the distribution of routing information.

Even though Nimrod did not enjoy deployment in the public Internet, a number of key concepts incorporated into the Nimrod architecture, such as explicit routing which allows selection of paths at originating nodes, are beginning to find applications in some recent constraint-based routing initiatives.

4.3 Overlay Model

In the overlay model, a virtual-circuit network, such as ATM, frame relay, or WDM, provides virtual-circuit connectivity between routers that are located at the edges of a virtual-circuit cloud. In this mode, two routers that are connected through a virtual circuit see a direct adjacency between themselves independent of the physical route taken by the virtual circuit through the ATM, frame relay, or WDM network. Thus, the overlay model essentially decouples the logical topology that routers see from the physical topology that the ATM, frame relay, or WDM network manages. The overlay model based on ATM or frame relay enables a network administrator or an automaton to employ traffic engineering concepts to perform path optimization by re-configuring or rearranging the virtual circuits so that a virtual circuit on a congested or sub-optimal physical link can be re-routed to a less congested or more optimal one. In the overlay model, traffic engineering is also employed to establish relationships between the traffic management parameters (e.g., PCR, SCR, and MBS for ATM) of the virtual-circuit technology and the actual traffic that traverses each circuit. These relationships can be established based upon known or projected traffic profiles, and some other factors.

The overlay model using IP over ATM requires the management of two separate networks with different technologies (IP and ATM) resulting in increased operational complexity and cost. In the fully-meshed overlay model, each router would peer to every other router in the network, so that the total number of adjacencies is a quadratic function of the number of routers. Some of the issues with the overlay model are discussed in [AWD2].

4.4 Constrained-Based Routing

Constraint-based routing refers to a class of routing systems that compute routes through a network subject to the satisfaction of a set of constraints and requirements. In the most general setting, constraint-based routing may also seek to optimize overall network performance while minimizing costs.

The constraints and requirements may be imposed by the network itself or by administrative policies. Constraints may include bandwidth, hop count, delay, and policy instruments such as resource class attributes. Constraints may also include domain specific attributes of certain network technologies and contexts which impose restrictions on the solution space of the routing function. Path oriented technologies such as MPLS have made constraint-based routing feasible and attractive in public IP networks.

The concept of constraint-based routing within the context of MPLS traffic engineering requirements in IP networks was first defined in [RFC-2702].

Unlike QoS routing (for example, see [RFC-2386] and [MA]) which generally addresses the issue of routing individual traffic flows to satisfy prescribed flow based QoS requirements subject to network resource availability, constraint-based routing is applicable to traffic aggregates as well as flows and may be subject to a wide variety of constraints which may include policy restrictions.

4.5 Overview of Other IETF Projects Related to Traffic Engineering

This subsection reviews a number of IETF activities pertinent to Internet traffic engineering. These activities are primarily intended to evolve the IP architecture to support new service definitions which allow preferential or differentiated treatment to be accorded to certain types of traffic.

4.5.1 Integrated Services

The IETF Integrated Services working group developed the integrated services (Intserv) model. This model requires resources, such as bandwidth and buffers, to be reserved a priori for a given traffic flow to ensure that the quality of service requested by the traffic flow is satisfied. The integrated services model includes additional components beyond those used in the best-effort model such as packet classifiers, packet schedulers, and admission control. A packet classifier is used to identify flows that are to receive a certain level of service. A packet scheduler handles the scheduling of

service to different packet flows to ensure that QoS commitments are met. Admission control is used to determine whether a router has the necessary resources to accept a new flow.

Two services have been defined under the Integrated Services model: guaranteed service [RFC-2212] and controlled-load service [RFC-2211].

The guaranteed service can be used for applications requiring bounded packet delivery time. For this type of application, data that is delivered to the application after a pre-defined amount of time has elapsed is usually considered worthless. Therefore, guaranteed service was intended to provide a firm quantitative bound on the end-to-end packet delay for a flow. This is accomplished by controlling the queuing delay on network elements along the data flow path. The guaranteed service model does not, however, provide bounds on jitter (inter-arrival times between consecutive packets).

The controlled-load service can be used for adaptive applications that can tolerate some delay but are sensitive to traffic overload conditions. This type of application typically functions satisfactorily when the network is lightly loaded but its performance degrades significantly when the network is heavily loaded. Controlled-load service, therefore, has been designed to provide approximately the same service as best-effort service in a lightly loaded network regardless of actual network conditions. Controlled-load service is described qualitatively in that no target values of delay or loss are specified.

The main issue with the Integrated Services model has been scalability [RFC-2998], especially in large public IP networks which may potentially have millions of active micro-flows in transit concurrently.

A notable feature of the Integrated Services model is that it requires explicit signaling of QoS requirements from end systems to routers [RFC-2753]. The Resource Reservation Protocol (RSVP) performs this signaling function and is a critical component of the Integrated Services model. The RSVP protocol is described next.

4.5.2 RSVP

RSVP is a soft state signaling protocol [RFC-2205]. It supports receiver initiated establishment of resource reservations for both multicast and unicast flows. RSVP was originally developed as a signaling protocol within the integrated services framework for applications to communicate QoS requirements to the network and for the network to reserve relevant resources to satisfy the QoS requirements [RFC-2205].

Under RSVP, the sender or source node sends a PATH message to the receiver with the same source and destination addresses as the traffic which the sender will generate. The PATH message contains: (1) a sender Tspec specifying the characteristics of the traffic, (2) a sender Template specifying the format of the traffic, and (3) an optional Adspec which is used to support the concept of one pass with advertising" (OPWA) [RFC-2205]. Every intermediate router along the path forwards the PATH Message to the next hop determined by the routing protocol. Upon receiving a PATH Message, the receiver responds with a RESV message which includes a flow descriptor used to request resource reservations. The RESV message travels to the sender or source node in the opposite direction along the path that the PATH message traversed. Every intermediate router along the path can reject or accept the reservation request of the RESV message. If the request is rejected, the rejecting router will send an error message to the receiver and the signaling process will terminate. If the request is accepted, link bandwidth and buffer space are allocated for the flow and the related flow state information is installed in the router.

One of the issues with the original RSVP specification was Scalability. This is because reservations were required for micro-flows, so that the amount of state maintained by network elements tends to increase linearly with the number of micro-flows. These issues are described in [RFC-2961].

Recently, RSVP has been modified and extended in several ways to mitigate the scaling problems. As a result, it is becoming a versatile signaling protocol for the Internet. For example, RSVP has been extended to reserve resources for aggregation of flows, to set up MPLS explicit label switched paths, and to perform other signaling functions within the Internet. There are also a number of proposals to reduce the amount of refresh messages required to maintain established RSVP sessions [RFC-2961].

A number of IETF working groups have been engaged in activities related to the RSVP protocol. These include the original RSVP working group, the MPLS working group, the Resource Allocation Protocol working group, and the Policy Framework working group.

4.5.3 Differentiated Services

The goal of the Differentiated Services (Diffserv) effort within the IETF is to devise scalable mechanisms for categorization of traffic into behavior aggregates, which ultimately allows each behavior aggregate to be treated differently, especially when there is a shortage of resources such as link bandwidth and buffer space [RFC-2475]. One of the primary motivations for the Diffserv effort was to

devise alternative mechanisms for service differentiation in the Internet that mitigate the scalability issues encountered with the Intserv model.

The IETF Diffserv working group has defined a Differentiated Services field in the IP header (DS field). The DS field consists of six bits of the part of the IP header formerly known as TOS octet. The DS field is used to indicate the forwarding treatment that a packet should receive at a node [RFC-2474]. The Diffserv working group has also standardized a number of Per-Hop Behavior (PHB) groups. Using the PHBs, several classes of services can be defined using different classification, policing, shaping, and scheduling rules.

For an end-user of network services to receive Differentiated Services from its Internet Service Provider (ISP), it may be necessary for the user to have a Service Level Agreement (SLA) with the ISP. An SLA may explicitly or implicitly specify a Traffic Conditioning Agreement (TCA) which defines classifier rules as well as metering, marking, discarding, and shaping rules.

Packets are classified, and possibly policed and shaped at the ingress to a Diffserv network. When a packet traverses the boundary between different Diffserv domains, the DS field of the packet may be re-marked according to existing agreements between the domains.

Differentiated Services allows only a finite number of service classes to be indicated by the DS field. The main advantage of the Diffserv approach relative to the Intserv model is scalability. Resources are allocated on a per-class basis and the amount of state information is proportional to the number of classes rather than to the number of application flows.

It should be obvious from the previous discussion that the Diffserv model essentially deals with traffic management issues on a per hop basis. The Diffserv control model consists of a collection of micro-TE control mechanisms. Other traffic engineering capabilities, such as capacity management (including routing control), are also required in order to deliver acceptable service quality in Diffserv networks. The concept of Per Domain Behaviors has been introduced to better capture the notion of differentiated services across a complete domain [RFC-3086].

4.5.4 MPLS

MPLS is an advanced forwarding scheme which also includes extensions to conventional IP control plane protocols. MPLS extends the Internet routing model and enhances packet forwarding and path control [RFC-3031].

At the ingress to an MPLS domain, label switching routers (LSRs) classify IP packets into forwarding equivalence classes (FECs) based on a variety of factors, including, e.g., a combination of the information carried in the IP header of the packets and the local routing information maintained by the LSRs. An MPLS label is then prepended to each packet according to their forwarding equivalence classes. In a non-ATM/FR environment, the label is 32 bits long and contains a 20-bit label field, a 3-bit experimental field (formerly known as Class-of-Service or CoS field), a 1-bit label stack indicator and an 8-bit TTL field. In an ATM (FR) environment, the label consists of information encoded in the VCI/VPI (DLCI) field. An MPLS capable router (an LSR) examines the label and possibly the experimental field and uses this information to make packet forwarding decisions.

An LSR makes forwarding decisions by using the label prepended to packets as the index into a local next hop label forwarding entry (NHLFE). The packet is then processed as specified in the NHLFE. The incoming label may be replaced by an outgoing label, and the packet may be switched to the next LSR. This label-switching process is very similar to the label (VCI/VPI) swapping process in ATM networks. Before a packet leaves an MPLS domain, its MPLS label may be removed. A Label Switched Path (LSP) is the path between an ingress LSRs and an egress LSRs through which a labeled packet traverses. The path of an explicit LSP is defined at the originating (ingress) node of the LSP. MPLS can use a signaling protocol such as RSVP or LDP to set up LSPs.

MPLS is a very powerful technology for Internet traffic engineering because it supports explicit LSPs which allow constraint-based routing to be implemented efficiently in IP networks [AWD2]. The requirements for traffic engineering over MPLS are described in [RFC-2702]. Extensions to RSVP to support instantiation of explicit LSP are discussed in [RFC-3209]. Extensions to LDP, known as CR-LDP, to support explicit LSPs are presented in [JAM].

4.5.5 IP Performance Metrics

The IETF IP Performance Metrics (IPPM) working group has been developing a set of standard metrics that can be used to monitor the quality, performance, and reliability of Internet services. These metrics can be applied by network operators, end-users, and independent testing groups to provide users and service providers with a common understanding of the performance and reliability of the Internet component 'clouds' they use/provide [RFC-2330]. The criteria for performance metrics developed by the IPPM WG are described in [RFC-2330]. Examples of performance metrics include one-way packet

loss [RFC-2680], one-way delay [RFC-2679], and connectivity measures between two nodes [RFC-2678]. Other metrics include second-order measures of packet loss and delay.

Some of the performance metrics specified by the IPPM WG are useful for specifying Service Level Agreements (SLAs). SLAs are sets of service level objectives negotiated between users and service providers, wherein each objective is a combination of one or more performance metrics, possibly subject to certain constraints.

4.5.6 Flow Measurement

The IETF Real Time Flow Measurement (RTFM) working group has produced an architecture document defining a method to specify traffic flows as well as a number of components for flow measurement (meters, meter readers, manager) [RFC-2722]. A flow measurement system enables network traffic flows to be measured and analyzed at the flow level for a variety of purposes. As noted in RFC 2722, a flow measurement system can be very useful in the following contexts: (1) understanding the behavior of existing networks, (2) planning for network development and expansion, (3) quantification of network performance, (4) verifying the quality of network service, and (5) attribution of network usage to users.

A flow measurement system consists of meters, meter readers, and managers. A meter observes packets passing through a measurement point, classifies them into certain groups, accumulates certain usage data (such as the number of packets and bytes for each group), and stores the usage data in a flow table. A group may represent a user application, a host, a network, a group of networks, etc. A meter reader gathers usage data from various meters so it can be made available for analysis. A manager is responsible for configuring and controlling meters and meter readers. The instructions received by a meter from a manager include flow specification, meter control parameters, and sampling techniques. The instructions received by a meter reader from a manager include the address of the meter whose date is to be collected, the frequency of data collection, and the types of flows to be collected.

4.5.7 Endpoint Congestion Management

[RFC-3124] is intended to provide a set of congestion control mechanisms that transport protocols can use. It is also intended to develop mechanisms for unifying congestion control across a subset of an endpoint’s active unicast connections (called a congestion group). A congestion manager continuously monitors the state of the path for

each congestion group under its control. The manager uses that information to instruct a scheduler on how to partition bandwidth among the connections of that congestion group.

4.6 Overview of ITU Activities Related to Traffic Engineering

This section provides an overview of prior work within the ITU-T pertaining to traffic engineering in traditional telecommunications networks.

ITU-T Recommendations E.600 [ITU-E600], E.701 [ITU-E701], and E.801 [ITU-E801] address traffic engineering issues in traditional telecommunications networks. Recommendation E.600 provides a vocabulary for describing traffic engineering concepts, while E.701 defines reference connections, Grade of Service (GOS), and traffic parameters for ISDN. Recommendation E.701 uses the concept of a reference connection to identify representative cases of different types of connections without describing the specifics of their actual realizations by different physical means. As defined in Recommendation E.600, "a connection is an association of resources providing means for communication between two or more devices in, or attached to, a telecommunication network." Also, E.600 defines "a resource as any set of physically or conceptually identifiable entities within a telecommunication network, the use of which can be unambiguously determined" [ITU-E600]. There can be different types of connections as the number and types of resources in a connection may vary.

Typically, different network segments are involved in the path of a connection. For example, a connection may be local, national, or international. The purposes of reference connections are to clarify and specify traffic performance issues at various interfaces between different network domains. Each domain may consist of one or more service provider networks.

Reference connections provide a basis to define grade of service (GoS) parameters related to traffic engineering within the ITU-T framework. As defined in E.600, "GoS refers to a number of traffic engineering variables which are used to provide a measure of the adequacy of a group of resources under specified conditions." These GoS variables may be probability of loss, dial tone, delay, etc. They are essential for network internal design and operation as well as for component performance specification.

GoS is different from quality of service (QoS) in the ITU framework. QoS is the performance perceivable by a telecommunication service user and expresses the user’s degree of satisfaction of the service. QoS parameters focus on performance aspects observable at the service

access points and network interfaces, rather than their causes within the network. GoS, on the other hand, is a set of network oriented measures which characterize the adequacy of a group of resources under specified conditions. For a network to be effective in serving its users, the values of both GoS and QoS parameters must be related, with GoS parameters typically making a major contribution to the QoS.

Recommendation E.600 stipulates that a set of GoS parameters must be selected and defined on an end-to-end basis for each major service category provided by a network to assist the network provider with improving efficiency and effectiveness of the network. Based on a selected set of reference connections, suitable target values are assigned to the selected GoS parameters under normal and high load conditions. These end-to-end GoS target values are then apportioned to individual resource components of the reference connections for dimensioning purposes.

4.7 Content Distribution

The Internet is dominated by client-server interactions, especially Web traffic (in the future, more sophisticated media servers may become dominant). The location and performance of major information servers has a significant impact on the traffic patterns within the Internet as well as on the perception of service quality by end users.

A number of dynamic load balancing techniques have been devised to improve the performance of replicated information servers. These techniques can cause spatial traffic characteristics to become more dynamic in the Internet because information servers can be dynamically picked based upon the location of the clients, the location of the servers, the relative utilization of the servers, the relative performance of different networks, and the relative performance of different parts of a network. This process of assignment of distributed servers to clients is called Traffic Directing. It functions at the application layer.

Traffic Directing schemes that allocate servers in multiple geographically dispersed locations to clients may require empirical network performance statistics to make more effective decisions. In the future, network measurement systems may need to provide this type of information. The exact parameters needed are not yet defined.

When congestion exists in the network, Traffic Directing and Traffic Engineering systems should act in a coordinated manner. This topic is for further study.

The issues related to location and replication of information servers, particularly web servers, are important for Internet traffic engineering because these servers contribute a substantial proportion of Internet traffic.

5.0 Taxonomy of Traffic Engineering Systems

This section presents a short taxonomy of traffic engineering systems. A taxonomy of traffic engineering systems can be constructed based on traffic engineering styles and views as listed below:

- Time-dependent vs State-dependent vs Event-dependent
- Offline vs Online
- Centralized vs Distributed
- Local vs Global Information
- Prescriptive vs Descriptive
- Open Loop vs Closed Loop
- Tactical vs Strategic

These classification systems are described in greater detail in the following subsections of this document.

5.1 Time-Dependent Versus State-Dependent Versus Event Dependent

Traffic engineering methodologies can be classified as time-dependent, or state-dependent, or event-dependent. All TE schemes are considered to be dynamic in this document. Static TE implies that no traffic engineering methodology or algorithm is being applied.

In the time-dependent TE, historical information based on periodic variations in traffic, (such as time of day), is used to pre-program routing plans and other TE control mechanisms. Additionally, customer subscription or traffic projection may be used. Pre-programmed routing plans typically change on a relatively long time scale (e.g., diurnal). Time-dependent algorithms do not attempt to adapt to random variations in traffic or changing network conditions. An example of a time-dependent algorithm is a global centralized optimizer where the input to the system is a traffic matrix and multi-class QoS requirements as described [MR99].

State-dependent TE adapts the routing plans for packets based on the current state of the network. The current state of the network provides additional information on variations in actual traffic (i.e., perturbations from regular variations) that could not be predicted using historical information. Constraint-based routing is

an example of state-dependent TE operating in a relatively long time scale. An example operating in a relatively short time scale is a load-balancing algorithm described in [MATE].

The state of the network can be based on parameters such as utilization, packet delay, packet loss, etc. These parameters can be obtained in several ways. For example, each router may flood these parameters periodically or by means of some kind of trigger to other routers. Another approach is for a particular router performing adaptive TE to send probe packets along a path to gather the state of that path. Still another approach is for a management system to gather relevant information from network elements.

Expeditious and accurate gathering and distribution of state information is critical for adaptive TE due to the dynamic nature of network conditions. State-dependent algorithms may be applied to increase network efficiency and resilience. Time-dependent algorithms are more suitable for predictable traffic variations. On the other hand, state-dependent algorithms are more suitable for adapting to the prevailing network state.

Event-dependent TE methods can also be used for TE path selection. Event-dependent TE methods are distinct from time-dependent and state-dependent TE methods in the manner in which paths are selected. These algorithms are adaptive and distributed in nature and typically use learning models to find good paths for TE in a network. While state-dependent TE models typically use available-link-bandwidth (ALB) flooding for TE path selection, event-dependent TE methods do not require ALB flooding. Rather, event-dependent TE methods typically search out capacity by learning models, as in the success-to-the-top (STT) method. ALB flooding can be resource intensive, since it requires link bandwidth to carry LSAs, processor capacity to process LSAs, and the overhead can limit area/autonomous system (AS) size. Modeling results suggest that event-dependent TE methods could lead to a reduction in ALB flooding overhead without loss of network throughput performance [ASH3].

5.2 Offline Versus Online

Traffic engineering requires the computation of routing plans. The computation may be performed offline or online. The computation can be done offline for scenarios where routing plans need not be executed in real-time. For example, routing plans computed from forecast information may be computed offline. Typically, offline computation is also used to perform extensive searches on multi-dimensional solution spaces.

Online computation is required when the routing plans must adapt to changing network conditions as in state-dependent algorithms. Unlike offline computation (which can be computationally demanding), online computation is geared toward relative simple and fast calculations to select routes, fine-tune the allocations of resources, and perform load balancing.

5.3 Centralized Versus Distributed

Centralized control has a central authority which determines routing plans and perhaps other TE control parameters on behalf of each router. The central authority collects the network-state information from all routers periodically and returns the routing information to the routers. The routing update cycle is a critical parameter directly impacting the performance of the network being controlled. Centralized control may need high processing power and high bandwidth control channels.

Distributed control determines route selection by each router autonomously based on the routers view of the state of the network. The network state information may be obtained by the router using a probing method or distributed by other routers on a periodic basis using link state advertisements. Network state information may also be disseminated under exceptional conditions.

5.4 Local Versus Global

Traffic engineering algorithms may require local or global network-state information.

Local information pertains to the state of a portion of the domain. Examples include the bandwidth and packet loss rate of a particular path. Local state information may be sufficient for certain instances of distributed-controlled TEs.

Global information pertains to the state of the entire domain undergoing traffic engineering. Examples include a global traffic matrix and loading information on each link throughout the domain of interest. Global state information is typically required with centralized control. Distributed TE systems may also need global information in some cases.

5.5 Prescriptive Versus Descriptive

TE systems may also be classified as prescriptive or descriptive.

Prescriptive traffic engineering evaluates alternatives and recommends a course of action. Prescriptive traffic engineering can be further categorized as either corrective or perfective. Corrective TE prescribes a course of action to address an existing or predicted anomaly. Perfective TE prescribes a course of action to evolve and improve network performance even when no anomalies are evident.

Descriptive traffic engineering, on the other hand, characterizes the state of the network and assesses the impact of various policies without recommending any particular course of action.

5.6 Open-Loop Versus Closed-Loop

Open-loop traffic engineering control is where control action does not use feedback information from the current network state. The control action may use its own local information for accounting purposes, however.

Closed-loop traffic engineering control is where control action utilizes feedback information from the network state. The feedback information may be in the form of historical information or current measurement.

5.7 Tactical vs Strategic

Tactical traffic engineering aims to address specific performance problems (such as hot-spots) that occur in the network from a tactical perspective, without consideration of overall strategic imperatives. Without proper planning and insights, tactical TE tends to be ad hoc in nature.

Strategic traffic engineering approaches the TE problem from a more organized and systematic perspective, taking into consideration the immediate and longer term consequences of specific policies and actions.

6.0 Recommendations for Internet Traffic Engineering

This section describes high level recommendations for traffic engineering in the Internet. These recommendations are presented in general terms.

The recommendations describe the capabilities needed to solve a traffic engineering problem or to achieve a traffic engineering objective. Broadly speaking, these recommendations can be categorized as either functional and non-functional recommendations.

Functional recommendations for Internet traffic engineering describe the functions that a traffic engineering system should perform. These functions are needed to realize traffic engineering objectives by addressing traffic engineering problems.

Non-functional recommendations for Internet traffic engineering relate to the quality attributes or state characteristics of a traffic engineering system. These recommendations may contain conflicting assertions and may sometimes be difficult to quantify precisely.

6.1 Generic Non-functional Recommendations

The generic non-functional recommendations for Internet traffic engineering include: usability, automation, scalability, stability, visibility, simplicity, efficiency, reliability, correctness, maintainability, extensibility, interoperability, and security. In a given context, some of these recommendations may be critical while others may be optional. Therefore, prioritization may be required during the development phase of a traffic engineering system (or components thereof) to tailor it to a specific operational context.

In the following paragraphs, some of the aspects of the non-functional recommendations for Internet traffic engineering are summarized.

Usability: Usability is a human factor aspect of traffic engineering systems. Usability refers to the ease with which a traffic engineering system can be deployed and operated. In general, it is desirable to have a TE system that can be readily deployed in an existing network. It is also desirable to have a TE system that is easy to operate and maintain.

Automation: Whenever feasible, a traffic engineering system should automate as many traffic engineering functions as possible to minimize the amount of human effort needed to control and analyze operational networks. Automation is particularly imperative in large scale public networks because of the high cost of the human aspects of network operations and the high risk of network problems caused by human errors. Automation may entail the incorporation of automatic feedback and intelligence into some components of the traffic engineering system.

Scalability: Contemporary public networks are growing very fast with respect to network size and traffic volume. Therefore, a TE system should be scalable to remain applicable as the network evolves. In particular, a TE system should remain functional as the network expands with regard to the number of routers and links, and with

respect to the traffic volume. A TE system should have a scalable architecture, should not adversely impair other functions and processes in a network element, and should not consume too much network resources when collecting and distributing state information or when exerting control.

Stability: Stability is a very important consideration in traffic engineering systems that respond to changes in the state of the network. State-dependent traffic engineering methodologies typically mandate a tradeoff between responsiveness and stability. It is strongly recommended that when tradeoffs are warranted between responsiveness and stability, that the tradeoff should be made in favor of stability (especially in public IP backbone networks).

Flexibility: A TE system should be flexible to allow for changes in optimization policy. In particular, a TE system should provide sufficient configuration options so that a network administrator can tailor the TE system to a particular environment. It may also be desirable to have both online and offline TE subsystems which can be independently enabled and disabled. TE systems that are used in multi-class networks should also have options to support class based performance evaluation and optimization.

Visibility: As part of the TE system, mechanisms should exist to collect statistics from the network and to analyze these statistics to determine how well the network is functioning. Derived statistics such as traffic matrices, link utilization, latency, packet loss, and other performance measures of interest which are determined from network measurements can be used as indicators of prevailing network conditions. Other examples of status information which should be observed include existing functional routing information (additionally, in the context of MPLS existing LSP routes), etc.

Simplicity: Generally, a TE system should be as simple as possible. More importantly, the TE system should be relatively easy to use (i.e., clean, convenient, and intuitive user interfaces). Simplicity in user interface does not necessarily imply that the TE system will use naive algorithms. When complex algorithms and internal structures are used, such complexities should be hidden as much as possible from the network administrator through the user interface.

Interoperability: Whenever feasible, traffic engineering systems and their components should be developed with open standards based interfaces to allow interoperation with other systems and components.

Security: Security is a critical consideration in traffic engineering systems. Such traffic engineering systems typically exert control over certain functional aspects of the network to achieve the desired

performance objectives. Therefore, adequate measures must be taken to safeguard the integrity of the traffic engineering system. Adequate measures must also be taken to protect the network from vulnerabilities that originate from security breaches and other impairments within the traffic engineering system.

The remainder of this section will focus on some of the high level functional recommendations for traffic engineering.

6.2 Routing Recommendations

Routing control is a significant aspect of Internet traffic engineering. Routing impacts many of the key performance measures associated with networks, such as throughput, delay, and utilization. Generally, it is very difficult to provide good service quality in a wide area network without effective routing control. A desirable routing system is one that takes traffic characteristics and network constraints into account during route selection while maintaining stability.

Traditional shortest path first (SPF) interior gateway protocols are based on shortest path algorithms and have limited control capabilities for traffic engineering [RFC-2702, AWD2]. These limitations include :

1. The well known issues with pure SPF protocols, which do not take network constraints and traffic characteristics into account during route selection. For example, since IGPs always use the shortest paths (based on administratively assigned link metrics) to forward traffic, load sharing cannot be accomplished among paths of different costs. Using shortest paths to forward traffic conserves network resources, but may cause the following problems: 1) If traffic from a source to a destination exceeds the capacity of a link along the shortest path, the link (hence the shortest path) becomes congested while a longer path between these two nodes may be under-utilized; 2) the shortest paths from different sources can overlap at some links. If the total traffic from the sources exceeds the capacity of any of these links, congestion will occur. Problems can also occur because traffic demand changes over time but network topology and routing configuration cannot be changed as rapidly. This causes the network topology and routing configuration to become sub-optimal over time, which may result in persistent congestion problems.

2. The Equal-Cost Multi-Path (ECMP) capability of SPF IGPs supports sharing of traffic among equal cost paths between two nodes. However, ECMP attempts to divide the traffic as equally as possible among the equal cost shortest paths. Generally, ECMP

does not support configurable load sharing ratios among equal cost paths. The result is that one of the paths may carry significantly more traffic than other paths because it may also carry traffic from other sources. This situation can result in congestion along the path that carries more traffic.

3. Modifying IGP metrics to control traffic routing tends to have network-wide effect. Consequently, undesirable and unanticipated traffic shifts can be triggered as a result. Recent work described in Section 8.0 may be capable of better control [FT00, FT01].

Because of these limitations, new capabilities are needed to enhance the routing function in IP networks. Some of these capabilities have been described elsewhere and are summarized below.

Constraint-based routing is desirable to evolve the routing architecture of IP networks, especially public IP backbones with complex topologies [RFC-2702]. Constraint-based routing computes routes to fulfill requirements subject to constraints. Constraints may include bandwidth, hop count, delay, and administrative policy instruments such as resource class attributes [RFC-2702, RFC-2386]. This makes it possible to select routes that satisfy a given set of requirements subject to network and administrative policy constraints. Routes computed through constraint-based routing are not necessarily the shortest paths. Constraint-based routing works best with path oriented technologies that support explicit routing, such as MPLS.

Constraint-based routing can also be used as a way to redistribute traffic onto the infrastructure (even for best effort traffic). For example, if the bandwidth requirements for path selection and reservable bandwidth attributes of network links are appropriately defined and configured, then congestion problems caused by uneven traffic distribution may be avoided or reduced. In this way, the performance and efficiency of the network can be improved.

A number of enhancements are needed to conventional link state IGPs, such as OSPF and IS-IS, to allow them to distribute additional state information required for constraint-based routing. These extensions to OSPF were described in [KATZ] and to IS-IS in [SMIT]. Essentially, these enhancements require the propagation of additional information in link state advertisements. Specifically, in addition to normal link-state information, an enhanced IGP is required to propagate topology state information needed for constraint-based routing. Some of the additional topology state information include link attributes such as reservable bandwidth and link resource class attribute (an administratively specified property of the link). The

resource class attribute concept was defined in [RFC-2702]. The additional topology state information is carried in new TLVs and sub-TLVs in IS-IS, or in the Opaque LSA in OSPF [SMIT, KATZ].

An enhanced link-state IGP may flood information more frequently than a normal IGP. This is because even without changes in topology, changes in reservable bandwidth or link affinity can trigger the enhanced IGP to initiate flooding. A tradeoff is typically required between the timeliness of the information flooded and the flooding frequency to avoid excessive consumption of link bandwidth and computational resources, and more importantly, to avoid instability.

In a TE system, it is also desirable for the routing subsystem to make the load splitting ratio among multiple paths (with equal cost or different cost) configurable. This capability gives network administrators more flexibility in the control of traffic distribution across the network. It can be very useful for avoiding/relieving congestion in certain situations. Examples can be found in [XIAO].

The routing system should also have the capability to control the routes of subsets of traffic without affecting the routes of other traffic if sufficient resources exist for this purpose. This capability allows a more refined control over the distribution of traffic across the network. For example, the ability to move traffic from a source to a destination away from its original path to another path (without affecting other traffic paths) allows traffic to be moved from resource-poor network segments to resource-rich segments. Path oriented technologies such as MPLS inherently support this capability as discussed in [AWD2].

Additionally, the routing subsystem should be able to select different paths for different classes of traffic (or for different traffic behavior aggregates) if the network supports multiple classes of service (different behavior aggregates).

6.3 Traffic Mapping Recommendations

Traffic mapping pertains to the assignment of traffic workload onto pre-established paths to meet certain requirements. Thus, while constraint-based routing deals with path selection, traffic mapping deals with the assignment of traffic to established paths which may have been selected by constraint-based routing or by some other means. Traffic mapping can be performed by time-dependent or state-dependent mechanisms, as described in Section 5.1.

An important aspect of the traffic mapping function is the ability to establish multiple paths between an originating node and a destination node, and the capability to distribute the traffic between the two nodes across the paths according to some policies. A pre-condition for this scheme is the existence of flexible mechanisms to partition traffic and then assign the traffic partitions onto the parallel paths. This requirement was noted in [RFC-2702]. When traffic is assigned to multiple parallel paths, it is recommended that special care should be taken to ensure proper ordering of packets belonging to the same application (or micro-flow) at the destination node of the parallel paths.

As a general rule, mechanisms that perform the traffic mapping functions should aim to map the traffic onto the network infrastructure to minimize congestion. If the total traffic load cannot be accommodated, or if the routing and mapping functions cannot react fast enough to changing traffic conditions, then a traffic mapping system may rely on short time scale congestion control mechanisms (such as queue management, scheduling, etc.) to mitigate congestion. Thus, mechanisms that perform the traffic mapping functions should complement existing congestion control mechanisms. In an operational network, it is generally desirable to map the traffic onto the infrastructure such that intra-class and inter-class resource contention are minimized.

When traffic mapping techniques that depend on dynamic state feedback (e.g., MATE and such like) are used, special care must be taken to guarantee network stability.

6.4 Measurement Recommendations

The importance of measurement in traffic engineering has been discussed throughout this document. Mechanisms should be provided to measure and collect statistics from the network to support the traffic engineering function. Additional capabilities may be needed to help in the analysis of the statistics. The actions of these mechanisms should not adversely affect the accuracy and integrity of the statistics collected. The mechanisms for statistical data acquisition should also be able to scale as the network evolves.

Traffic statistics may be classified according to long-term or short-term time scales. Long-term time scale traffic statistics are very useful for traffic engineering. Long-term time scale traffic statistics may capture or reflect periodicity in network workload (such as hourly, daily, and weekly variations in traffic profiles) as well as traffic trends. Aspects of the monitored traffic statistics may also depict class of service characteristics for a network supporting multiple classes of service. Analysis of the long-term

traffic statistics MAY yield secondary statistics such as busy hour characteristics, traffic growth patterns, persistent congestion problems, hot-spot, and imbalances in link utilization caused by routing anomalies.

A mechanism for constructing traffic matrices for both long-term and short-term traffic statistics should be in place. In multi-service IP networks, the traffic matrices may be constructed for different service classes. Each element of a traffic matrix represents a statistic of traffic flow between a pair of abstract nodes. An abstract node may represent a router, a collection of routers, or a site in a VPN.

Measured traffic statistics should provide reasonable and reliable indicators of the current state of the network on the short-term scale. Some short term traffic statistics may reflect link utilization and link congestion status. Examples of congestion indicators include excessive packet delay, packet loss, and high resource utilization. Examples of mechanisms for distributing this kind of information include SNMP, probing techniques, FTP, IGP link state advertisements, etc.

6.5 Network Survivability

Network survivability refers to the capability of a network to maintain service continuity in the presence of faults. This can be accomplished by promptly recovering from network impairments and maintaining the required QoS for existing services after recovery. Survivability has become an issue of great concern within the Internet community due to the increasing demands to carry mission critical traffic, real-time traffic, and other high priority traffic over the Internet. Survivability can be addressed at the device level by developing network elements that are more reliable; and at the network level by incorporating redundancy into the architecture, design, and operation of networks. It is recommended that a philosophy of robustness and survivability should be adopted in the architecture, design, and operation of traffic engineering that control IP networks (especially public IP networks). Because different contexts may demand different levels of survivability, the mechanisms developed to support network survivability should be flexible so that they can be tailored to different needs.

Failure protection and restoration capabilities have become available from multiple layers as network technologies have continued to improve. At the bottom of the layered stack, optical networks are now capable of providing dynamic ring and mesh restoration functionality at the wavelength level as well as traditional protection functionality. At the SONET/SDH layer survivability

capability is provided with Automatic Protection Switching (APS) as well as self-healing ring and mesh architectures. Similar functionality is provided by layer 2 technologies such as ATM (generally with slower mean restoration times). Rerouting is traditionally used at the IP layer to restore service following link and node outages. Rerouting at the IP layer occurs after a period of routing convergence which may require seconds to minutes to complete. Some new developments in the MPLS context make it possible to achieve recovery at the IP layer prior to convergence [SHAR].

To support advanced survivability requirements, path-oriented technologies such a MPLS can be used to enhance the survivability of IP networks in a potentially cost effective manner. The advantages of path oriented technologies such as MPLS for IP restoration becomes even more evident when class based protection and restoration capabilities are required.

Recently, a common suite of control plane protocols has been proposed for both MPLS and optical transport networks under the acronym Multi-protocol Lambda Switching [AWD1]. This new paradigm of Multi-protocol Lambda Switching will support even more sophisticated mesh restoration capabilities at the optical layer for the emerging IP over WDM network architectures.

Another important aspect regarding multi-layer survivability is that technologies at different layers provide protection and restoration capabilities at different temporal granularities (in terms of time scales) and at different bandwidth granularity (from packet-level to wavelength level). Protection and restoration capabilities can also be sensitive to different service classes and different network utility models.

The impact of service outages varies significantly for different service classes depending upon the effective duration of the outage. The duration of an outage can vary from milliseconds (with minor service impact) to seconds (with possible call drops for IP telephony and session time-outs for connection oriented transactions) to minutes and hours (with potentially considerable social and business impact).

Coordinating different protection and restoration capabilities across multiple layers in a cohesive manner to ensure network survivability is maintained at reasonable cost is a challenging task. Protection and restoration coordination across layers may not always be feasible, because networks at different layers may belong to different administrative domains.

The following paragraphs present some of the general recommendations for protection and restoration coordination.

- Protection and restoration capabilities from different layers should be coordinated whenever feasible and appropriate to provide network survivability in a flexible and cost effective manner. Minimization of function duplication across layers is one way to achieve the coordination. Escalation of alarms and other fault indicators from lower to higher layers may also be performed in a coordinated manner. A temporal order of restoration trigger timing at different layers is another way to coordinate multi-layer protection/restoration.

- Spare capacity at higher layers is often regarded as working traffic at lower layers. Placing protection/restoration functions in many layers may increase redundancy and robustness, but it should not result in significant and avoidable inefficiencies in network resource utilization.

- It is generally desirable to have protection and restoration schemes that are bandwidth efficient.

- Failure notification throughout the network should be timely and reliable.

- Alarms and other fault monitoring and reporting capabilities should be provided at appropriate layers.

6.5.1 Survivability in MPLS Based Networks

MPLS is an important emerging technology that enhances IP networks in terms of features, capabilities, and services. Because MPLS is path-oriented, it can potentially provide faster and more predictable protection and restoration capabilities than conventional hop by hop routed IP systems. This subsection describes some of the basic aspects and recommendations for MPLS networks regarding protection and restoration. See [SHAR] for a more comprehensive discussion on MPLS based recovery.

Protection types for MPLS networks can be categorized as link protection, node protection, path protection, and segment protection.

- Link Protection: The objective for link protection is to protect an LSP from a given link failure. Under link protection, the path of the protection or backup LSP (the secondary LSP) is disjoint from the path of the working or operational LSP at the particular link over which protection is required. When the protected link fails, traffic on the working LSP is switched over to the

protection LSP at the head-end of the failed link. This is a local repair method which can be fast. It might be more appropriate in situations where some network elements along a given path are less reliable than others.

- Node Protection: The objective of LSP node protection is to protect an LSP from a given node failure. Under node protection, the path of the protection LSP is disjoint from the path of the working LSP at the particular node to be protected. The secondary path is also disjoint from the primary path at all links associated with the node to be protected. When the node fails, traffic on the working LSP is switched over to the protection LSP at the upstream LSR directly connected to the failed node.

- Path Protection: The goal of LSP path protection is to protect an LSP from failure at any point along its routed path. Under path protection, the path of the protection LSP is completely disjoint from the path of the working LSP. The advantage of path protection is that the backup LSP protects the working LSP from all possible link and node failures along the path, except for failures that might occur at the ingress and egress LSRs, or for correlated failures that might impact both working and backup paths simultaneously. Additionally, since the path selection is end-to-end, path protection might be more efficient in terms of resource usage than link or node protection. However, path protection may be slower than link and node protection in general.

- Segment Protection: An MPLS domain may be partitioned into multiple protection domains whereby a failure in a protection domain is rectified within that domain. In cases where an LSP traverses multiple protection domains, a protection mechanism within a domain only needs to protect the segment of the LSP that lies within the domain. Segment protection will generally be faster than path protection because recovery generally occurs closer to the fault.

6.5.2 Protection Option

Another issue to consider is the concept of protection options. The protection option uses the notation m:n protection, where m is the number of protection LSPs used to protect n working LSPs. Feasible protection options follow.

- 1:1: one working LSP is protected/restored by one protection LSP.

- 1:n: one protection LSP is used to protect/restore n working LSPs.

- n:1: one working LSP is protected/restored by n protection LSPs, possibly with configurable load splitting ratio. When more than one protection LSP is used, it may be desirable to share the traffic across the protection LSPs when the working LSP fails to satisfy the bandwidth requirement of the traffic trunk associated with the working LSP. This may be especially useful when it is not feasible to find one path that can satisfy the bandwidth requirement of the primary LSP.

- 1+1: traffic is sent concurrently on both the working LSP and the protection LSP. In this case, the egress LSR selects one of the two LSPs based on a local traffic integrity decision process, which compares the traffic received from both the working and the protection LSP and identifies discrepancies. It is unlikely that this option would be used extensively in IP networks due to its resource utilization inefficiency. However, if bandwidth becomes plentiful and cheap, then this option might become quite viable and attractive in IP networks.

6.6 Traffic Engineering in Diffserv Environments

This section provides an overview of the traffic engineering features and recommendations that are specifically pertinent to Differentiated Services (Diffserv) [RFC-2475] capable IP networks.

Increasing requirements to support multiple classes of traffic, such as best effort and mission critical data, in the Internet calls for IP networks to differentiate traffic according to some criteria, and to accord preferential treatment to certain types of traffic. Large numbers of flows can be aggregated into a few behavior aggregates based on some criteria in terms of common performance requirements in terms of packet loss ratio, delay, and jitter; or in terms of common fields within the IP packet headers.

As Diffserv evolves and becomes deployed in operational networks, traffic engineering will be critical to ensuring that SLAs defined within a given Diffserv service model are met. Classes of service (CoS) can be supported in a Diffserv environment by concatenating per-hop behaviors (PHBs) along the routing path, using service provisioning mechanisms, and by appropriately configuring edge functionality such as traffic classification, marking, policing, and shaping. PHB is the forwarding behavior that a packet receives at a DS node (a Diffserv-compliant node). This is accomplished by means of buffer management and packet scheduling mechanisms. In this context, packets belonging to a class are those that are members of a corresponding ordering aggregate.

Traffic engineering can be used as a compliment to Diffserv mechanisms to improve utilization of network resources, but not as a necessary element in general. When traffic engineering is used, it can be operated on an aggregated basis across all service classes [RFC-3270] or on a per service class basis. The former is used to provide better distribution of the aggregate traffic load over the network resources. (See [RFC-3270] for detailed mechanisms to support aggregate traffic engineering.) The latter case is discussed below since it is specific to the Diffserv environment, with so called Diffserv-aware traffic engineering [DIFF_TE].

For some Diffserv networks, it may be desirable to control the performance of some service classes by enforcing certain relationships between the traffic workload contributed by each service class and the amount of network resources allocated or provisioned for that service class. Such relationships between demand and resource allocation can be enforced using a combination of, for example: (1) traffic engineering mechanisms on a per service class basis that enforce the desired relationship between the amount of traffic contributed by a given service class and the resources allocated to that class, and (2) mechanisms that dynamically adjust the resources allocated to a given service class to relate to the amount of traffic contributed by that service class.

It may also be desirable to limit the performance impact of high priority traffic on relatively low priority traffic. This can be achieved by, for example, controlling the percentage of high priority traffic that is routed through a given link. Another way to accomplish this is to increase link capacities appropriately so that lower priority traffic can still enjoy adequate service quality. When the ratio of traffic workload contributed by different service classes vary significantly from router to router, it may not suffice to rely exclusively on conventional IGP routing protocols or on traffic engineering mechanisms that are insensitive to different service classes. Instead, it may be desirable to perform traffic engineering, especially routing control and mapping functions, on a per service class basis. One way to accomplish this in a domain that supports both MPLS and Diffserv is to define class specific LSPs and to map traffic from each class onto one or more LSPs that correspond to that service class. An LSP corresponding to a given service class can then be routed and protected/restored in a class dependent manner, according to specific policies.

Performing traffic engineering on a per class basis may require certain per-class parameters to be distributed. Note that it is common to have some classes share some aggregate constraint (e.g., maximum bandwidth requirement) without enforcing the constraint on each individual class. These classes then can be grouped into a

class-type and per-class-type parameters can be distributed instead to improve scalability. It also allows better bandwidth sharing between classes in the same class-type. A class-type is a set of classes that satisfy the following two conditions:

1) Classes in the same class-type have common aggregate requirements to satisfy required performance levels.

2) There is no requirement to be enforced at the level of individual class in the class-type. Note that it is still possible, nevertheless, to implement some priority policies for classes in the same class-type to permit preferential access to the class-type bandwidth through the use of preemption priorities.

An example of the class-type can be a low-loss class-type that includes both AF1-based and AF2-based Ordering Aggregates. With such a class-type, one may implement some priority policy which assigns higher preemption priority to AF1-based traffic trunks over AF2-based ones, vice versa, or the same priority.

See [DIFF-TE] for detailed requirements on Diffserv-aware traffic engineering.

6.7 Network Controllability

Off-line (and on-line) traffic engineering considerations would be of limited utility if the network could not be controlled effectively to implement the results of TE decisions and to achieve desired network performance objectives. Capacity augmentation is a coarse grained solution to traffic engineering issues. However, it is simple and may be advantageous if bandwidth is abundant and cheap or if the current or expected network workload demands it. However, bandwidth is not always abundant and cheap, and the workload may not always demand additional capacity. Adjustments of administrative weights and other parameters associated with routing protocols provide finer grained control, but is difficult to use and imprecise because of the routing interactions that occur across the network. In certain network contexts, more flexible, finer grained approaches which provide more precise control over the mapping of traffic to routes and over the selection and placement of routes may be appropriate and useful.

Control mechanisms can be manual (e.g., administrative configuration), partially-automated (e.g., scripts) or fully-automated (e.g., policy based management systems). Automated mechanisms are particularly required in large scale networks. Multi-vendor interoperability can be facilitated by developing and deploying standardized management

systems (e.g., standard MIBs) and policies (PIBs) to support the control functions required to address traffic engineering objectives such as load distribution and protection/restoration.

Network control functions should be secure, reliable, and stable as these are often needed to operate correctly in times of network impairments (e.g., during network congestion or security attacks).

7.0 Inter-Domain Considerations

Inter-domain traffic engineering is concerned with the performance optimization for traffic that originates in one administrative domain and terminates in a different one.

Traffic exchange between autonomous systems in the Internet occurs through exterior gateway protocols. Currently, BGP [BGP4] is the standard exterior gateway protocol for the Internet. BGP provides a number of attributes and capabilities (e.g., route filtering) that can be used for inter-domain traffic engineering. More specifically, BGP permits the control of routing information and traffic exchange between Autonomous Systems (AS's) in the Internet. BGP incorporates a sequential decision process which calculates the degree of preference for various routes to a given destination network. There are two fundamental aspects to inter-domain traffic engineering using BGP:

- Route Redistribution: controlling the import and export of routes between AS's, and controlling the redistribution of routes between BGP and other protocols within an AS.

- Best path selection: selecting the best path when there are multiple candidate paths to a given destination network. Best path selection is performed by the BGP decision process based on a sequential procedure, taking a number of different considerations into account. Ultimately, best path selection under BGP boils down to selecting preferred exit points out of an AS towards specific destination networks. The BGP path selection process can be influenced by manipulating the attributes associated with the BGP decision process. These attributes include: NEXT-HOP, WEIGHT (Cisco proprietary which is also implemented by some other vendors), LOCAL-PREFERENCE, AS-PATH, ROUTE-ORIGIN, MULTI-EXIT-DESCRIMINATOR (MED), IGP METRIC, etc.

Route-maps provide the flexibility to implement complex BGP policies based on pre-configured logical conditions. In particular, Route-maps can be used to control import and export policies for incoming and outgoing routes, control the redistribution of routes between BGP and other protocols, and influence the selection of best paths by

manipulating the attributes associated with the BGP decision process. Very complex logical expressions that implement various types of policies can be implemented using a combination of Route-maps, BGP-attributes, Access-lists, and Community attributes.

When looking at possible strategies for inter-domain TE with BGP, it must be noted that the outbound traffic exit point is controllable, whereas the interconnection point where inbound traffic is received from an EBGP peer typically is not, unless a special arrangement is made with the peer sending the traffic. Therefore, it is up to each individual network to implement sound TE strategies that deal with the efficient delivery of outbound traffic from one's customers to one's peering points. The vast majority of TE policy is based upon a "closest exit" strategy, which offloads interdomain traffic at the nearest outbound peer point towards the destination autonomous system. Most methods of manipulating the point at which inbound traffic enters a network from an EBGP peer (inconsistent route announcements between peering points, AS pre-pending, and sending MEDs) are either ineffective, or not accepted in the peering community.

Inter-domain TE with BGP is generally effective, but it is usually applied in a trial-and-error fashion. A systematic approach for inter-domain traffic engineering is yet to be devised.

Inter-domain TE is inherently more difficult than intra-domain TE under the current Internet architecture. The reasons for this are both technical and administrative. Technically, while topology and link state information are helpful for mapping traffic more effectively, BGP does not propagate such information across domain boundaries for stability and scalability reasons. Administratively, there are differences in operating costs and network capacities between domains. Generally, what may be considered a good solution in one domain may not necessarily be a good solution in another domain. Moreover, it would generally be considered inadvisable for one domain to permit another domain to influence the routing and management of traffic in its network.

MPLS TE-tunnels (explicit LSPs) can potentially add a degree of flexibility in the selection of exit points for inter-domain routing. The concept of relative and absolute metrics can be applied to this purpose. The idea is that if BGP attributes are defined such that the BGP decision process depends on IGP metrics to select exit points for inter-domain traffic, then some inter-domain traffic destined to a given peer network can be made to prefer a specific exit point by establishing a TE-tunnel between the router making the selection to the peering point via a TE-tunnel and assigning the TE-tunnel a metric which is smaller than the IGP cost to all other peering

points. If a peer accepts and processes MEDs, then a similar MPLS TE-tunnel based scheme can be applied to cause certain entrance points to be preferred by setting MED to be an IGP cost, which has been modified by the tunnel metric.

Similar to intra-domain TE, inter-domain TE is best accomplished when a traffic matrix can be derived to depict the volume of traffic from one autonomous system to another.

Generally, redistribution of inter-domain traffic requires coordination between peering partners. An export policy in one domain that results in load redistribution across peer points with another domain can significantly affect the local traffic matrix inside the domain of the peering partner. This, in turn, will affect the intra-domain TE due to changes in the spatial distribution of traffic. Therefore, it is mutually beneficial for peering partners to coordinate with each other before attempting any policy changes that may result in significant shifts in inter-domain traffic. In certain contexts, this coordination can be quite challenging due to technical and non- technical reasons.

It is a matter of speculation as to whether MPLS, or similar technologies, can be extended to allow selection of constrained paths across domain boundaries.

8.0 Overview of Contemporary TE Practices in Operational IP Networks

This section provides an overview of some contemporary traffic engineering practices in IP networks. The focus is primarily on the aspects that pertain to the control of the routing function in operational contexts. The intent here is to provide an overview of the commonly used practices. The discussion is not intended to be exhaustive.

Currently, service providers apply many of the traffic engineering mechanisms discussed in this document to optimize the performance of their IP networks. These techniques include capacity planning for long time scales, routing control using IGP metrics and MPLS for medium time scales, the overlay model also for medium time scales, and traffic management mechanisms for short time scale.

When a service provider plans to build an IP network, or expand the capacity of an existing network, effective capacity planning should be an important component of the process. Such plans may take the following aspects into account: location of new nodes if any, existing and predicted traffic patterns, costs, link capacity, topology, routing design, and survivability.

Performance optimization of operational networks is usually an ongoing process in which traffic statistics, performance parameters, and fault indicators are continually collected from the network. This empirical data is then analyzed and used to trigger various traffic engineering mechanisms. Tools that perform what-if analysis can also be used to assist the TE process by allowing various scenarios to be reviewed before a new set of configurations are implemented in the operational network.

Traditionally, intra-domain real-time TE with IGP is done by increasing the OSPF or IS-IS metric of a congested link until enough traffic has been diverted from that link. This approach has some limitations as discussed in Section 6.2. Recently, some new intra-domain TE approaches/tools have been proposed [RR94][FT00][FT01][WANG]. Such approaches/tools take traffic matrix, network topology, and network performance objective(s) as input, and produce some link metrics and possibly some unequal load-sharing ratios to be set at the head-end routers of some ECMPs as output. These new progresses open new possibility for intra-domain TE with IGP to be done in a more systematic way.

The overlay model (IP over ATM or IP over Frame relay) is another approach which is commonly used in practice [AWD2]. The IP over ATM technique is no longer viewed favorably due to recent advances in MPLS and router hardware technology.

Deployment of MPLS for traffic engineering applications has commenced in some service provider networks. One operational scenario is to deploy MPLS in conjunction with an IGP (IS-IS-TE or OSPF-TE) that supports the traffic engineering extensions, in conjunction with constraint-based routing for explicit route computations, and a signaling protocol (e.g., RSVP-TE or CRLDP) for LSP instantiation.

In contemporary MPLS traffic engineering contexts, network administrators specify and configure link attributes and resource constraints such as maximum reservable bandwidth and resource class attributes for links (interfaces) within the MPLS domain. A link state protocol that supports TE extensions (IS-IS-TE or OSPF-TE) is used to propagate information about network topology and link attribute to all routers in the routing area. Network administrators also specify all the LSPs that are to originate each router. For each LSP, the network administrator specifies the destination node and the attributes of the LSP which indicate the requirements that to be satisfied during the path selection process. Each router then uses a local constraint-based routing process to compute explicit paths for all LSPs originating from it. Subsequently, a signaling

protocol is used to instantiate the LSPs. By assigning proper bandwidth values to links and LSPs, congestion caused by uneven traffic distribution can generally be avoided or mitigated.

The bandwidth attributes of LSPs used for traffic engineering can be updated periodically. The basic concept is that the bandwidth assigned to an LSP should relate in some manner to the bandwidth requirements of traffic that actually flows through the LSP. The traffic attribute of an LSP can be modified to accommodate traffic growth and persistent traffic shifts. If network congestion occurs due to some unexpected events, existing LSPs can be rerouted to alleviate the situation or network administrator can configure new LSPs to divert some traffic to alternative paths. The reservable bandwidth of the congested links can also be reduced to force some LSPs to be rerouted to other paths.

In an MPLS domain, a traffic matrix can also be estimated by monitoring the traffic on LSPs. Such traffic statistics can be used for a variety of purposes including network planning and network optimization. Current practice suggests that deploying an MPLS network consisting of hundreds of routers and thousands of LSPs is feasible. In summary, recent deployment experience suggests that MPLS approach is very effective for traffic engineering in IP networks [XIAO].

As mentioned previously in Section 7.0, one usually has no direct control over the distribution of inbound traffic. Therefore, the main goal of contemporary inter-domain TE is to optimize the distribution of outbound traffic between multiple inter-domain links. When operating a global network, maintaining the ability to operate the network in a regional fashion where desired, while continuing to take advantage of the benefits of a global network, also becomes an important objective.

Inter-domain TE with BGP usually begins with the placement of multiple peering interconnection points in locations that have high peer density, are in close proximity to originating/terminating traffic locations on one’s own network, and are lowest in cost. There are generally several locations in each region of the world where the vast majority of major networks congregate and interconnect. Some location-decision problems that arise in association with inter-domain routing are discussed in [AWD5].

Once the locations of the interconnects are determined, and circuits are implemented, one decides how best to handle the routes heard from the peer, as well as how to propagate the peers’ routes within one’s own network. One way to engineer outbound traffic flows on a network with many EBGP peers is to create a hierarchy of peers. Generally,

the Local Preferences of all peers are set to the same value so that the shortest AS paths will be chosen to forward traffic. Then, by over-writing the inbound MED metric (Multi-exit-discriminator metric, also referred to as "BGP metric". Both terms are used interchangeably in this document) with BGP metrics to routes received at different peers, the hierarchy can be formed. For example, all Local Preferences can be set to 200, preferred private peers can be assigned a BGP metric of 50, the rest of the private peers can be assigned a BGP metric of 100, and public peers can be assigned a BGP metric of 600. "Preferred" peers might be defined as those peers with whom the most available capacity exists, whose customer base is larger in comparison to other peers, whose interconnection costs are the lowest, and with whom upgrading existing capacity is the easiest. In a network with low utilization at the edge, this works well. The same concept could be applied to a network with higher edge utilization by creating more levels of BGP metrics between peers, allowing for more granularity in selecting the exit points for traffic bound for a dual homed customer on a peer's network.

By only replacing inbound MED metrics with BGP metrics, only equal AS-Path length routes' exit points are being changed. (The BGP decision considers Local Preference first, then AS-Path length, and then BGP metric). For example, assume a network has two possible egress points, peer A and peer B. Each peer has 40% of the Internet's routes exclusively on its network, while the remaining 20% of the Internet's routes are from customers who dual home between A and B. Assume that both peers have a Local Preference of 200 and a BGP metric of 100. If the link to peer A is congested, increasing its BGP metric while leaving the Local Preference at 200 will ensure that the 20% of total routes belonging to dual homed customers will prefer peer B as the exit point. The previous example would be used in a situation where all exit points to a given peer were close to congestion levels, and traffic needed to be shifted away from that peer entirely.

When there are multiple exit points to a given peer, and only one of them is congested, it is not necessary to shift traffic away from the peer entirely, but only from the one congested circuit. This can be achieved by using passive IGP-metrics, AS-path filtering, or prefix filtering.

Occasionally, more drastic changes are needed, for example, in dealing with a "problem peer" who is difficult to work with on upgrades or is charging high prices for connectivity to their network. In that case, the Local Preference to that peer can be reduced below the level of other peers. This effectively reduces the amount of traffic sent to that peer to only originating traffic

(assuming no transit providers are involved). This type of change can affect a large amount of traffic, and is only used after other methods have failed to provide the desired results.

Although it is not much of an issue in regional networks, the propagation of a peer's routes back through the network must be considered when a network is peering on a global scale. Sometimes, business considerations can influence the choice of BGP policies in a given context. For example, it may be imprudent, from a business perspective, to operate a global network and provide full access to the global customer base to a small network in a particular country. However, for the purpose of providing one's own customers with quality service in a particular region, good connectivity to that in-country network may still be necessary. This can be achieved by assigning a set of communities at the edge of the network, which have a known behavior when routes tagged with those communities are propagating back through the core. Routes heard from local peers will be prevented from propagating back to the global network, whereas routes learned from larger peers may be allowed to propagate freely throughout the entire global network. By implementing a flexible community strategy, the benefits of using a single global AS Number (ASN) can be realized, while the benefits of operating regional networks can also be taken advantage of. An alternative to doing this is to use different ASNs in different regions, with the consequence that the AS path length for routes announced by that service provider will increase.

9.0 Conclusion

This document described principles for traffic engineering in the Internet. It presented an overview of some of the basic issues surrounding traffic engineering in IP networks. The context of TE was described, a TE process models and a taxonomy of TE styles were presented. A brief historical review of pertinent developments related to traffic engineering was provided. A survey of contemporary TE techniques in operational networks was presented. Additionally, the document specified a set of generic requirements, recommendations, and options for Internet traffic engineering.

10.0 Security Considerations

This document does not introduce new security issues.

11.0 Acknowledgments

The authors would like to thank Jim Boyle for inputs on the recommendations section, Francois Le Faucheur for inputs on Diffserv aspects, Blaine Christian for inputs on measurement, Gerald Ash for

inputs on routing in telephone networks and for text on event-dependent TE methods, Steven Wright for inputs on network controllability, and Jonathan Aufderheide for inputs on inter-domain TE with BGP. Special thanks to Randy Bush for proposing the TE taxonomy based on "tactical vs strategic" methods. The subsection describing an "Overview of ITU Activities Related to Traffic Engineering" was adapted from a contribution by Waisum Lai. Useful feedback and pointers to relevant materials were provided by J. Noel Chiappa. Additional comments were provided by Glenn Grotefeld during the working last call process. Finally, the authors would like to thank Ed Kern, the TEWG co-chair, for his comments and support.

12.0 References

[ASH2] J. Ash, Dynamic Routing in Telecommunications Networks, McGraw Hill, 1998.

[ASH3] Ash, J., "TE & QoS Methods for IP-, ATM-, & TDM-Based Networks", Work in Progress, March 2001.

[AWD1] D. Awduche and Y. Rekhter, "Multiprocotol Lambda Switching: Combining MPLS Traffic Engineering Control with Optical Crossconnects", IEEE Communications Magazine, March 2001.

[AWD2] D. Awduche, "MPLS and Traffic Engineering in IP Networks", IEEE Communications Magazine, Dec. 1999.

[AWD5] D. Awduche et al, "An Approach to Optimal Peering Between Autonomous Systems in the Internet", International Conference on Computer Communications and Networks (ICCCN’98), Oct. 1998.

[CRUZ] R. L. Cruz, "A Calculus for Network Delay, Part II: Network Analysis", IEEE Transactions on Information Theory, vol. 37, pp. 132-141, 1991.

[DIFF-TE] Le Faucheur, F., Nadeau, T., Tatham, M., Telkamp, T., Cooper, D., Boyle, J., Lai, W., Fang, L., Ash, J., Hicks, P., Chui, A., Townsend, W. and D. Skalecki, "Requirements for support of Diff-Serv-aware MPLS Traffic Engineering", Work in Progress, May 2001.

[ELW95] A. Elwalid, D. Mitra and R.H. Wentworth, "A New Approach for Allocating Buffers and Bandwidth to Heterogeneous, Regulated Traffic in an ATM Node", IEEE IEEE Journal on Selected Areas in Communications, 13:6, pp. 1115-1127, Aug. 1995.

[FGLR] A. Feldmann, A. Greenberg, C. Lund, N. Reingold, and J. Rexford, "NetScope: Traffic Engineering for IP Networks", IEEE Network Magazine, 2000.

[FLJA93] S. Floyd and V. Jacobson, "Random Early Detection Gateways for Congestion Avoidance", IEEE/ACM Transactions on Networking, Vol. 1 Nov. 4., p. 387-413, Aug. 1993.

[FLOY94] S. Floyd, "TCP and Explicit Congestion Notification", ACM Computer Communication Review, V. 24, No. 5, p. 10-23, Oct. 1994.

[FT00] B. Fortz and M. Thorup, "Internet Traffic Engineering by Optimizing OSPF Weights", IEEE INFOCOM 2000, Mar. 2000.

[FT01] B. Fortz and M. Thorup, "Optimizing OSPF/IS-IS Weights in a Changing World", www.research.att.com/~mthorup/PAPERS/papers.html.

[HUSS87] B.R. Hurley, C.J.R. Seidl and W.F. Sewel, "A Survey of Dynamic Routing Methods for Circuit-Switched Traffic", IEEE Communication Magazine, Sep. 1987.

[ITU-E600] ITU-T Recommendation E.600, "Terms and Definitions of Traffic Engineering", Mar. 1993.

[ITU-E701] ITU-T Recommendation E.701, "Reference Connections for Traffic Engineering", Oct. 1993.

[ITU-E801] ITU-T Recommendation E.801, "Framework for Service Quality Agreement", Oct. 1996.

[JAM] Jamoussi, B., Editior, Andersson, L., Collon, R. and R. Dantu, "Constraint-Based LSP Setup using LDP", RFC 3212, January 2002.

[KATZ] Katz, D., Yeung, D. and K. Kompella, "Traffic Engineering Extensions to OSPF", Work in Progress, February 2001.

[LNO96] T. Lakshman, A. Neidhardt, and T. Ott, "The Drop from Front Strategy in TCP over ATM and its Interworking with other Control Features", Proc. INFOCOM'96, p. 1242-1250, 1996.

[MA] Q. Ma, "Quality of Service Routing in Integrated Services Networks", PhD Dissertation, CMU-CS-98-138, CMU, 1998.

[MATE] A. Elwalid, C. Jin, S. Low, and I. Widjaja, "MATE: MPLS Adaptive Traffic Engineering", Proc. INFOCOM'01, Apr. 2001.

[MCQ80] J.M. McQuillan, I. Richer, and E.C. Rosen, "The New Routing Algorithm for the ARPANET", IEEE. Trans. on Communications, vol. 28, no. 5, pp. 711-719, May 1980.

[MR99] D. Mitra and K.G. Ramakrishnan, "A Case Study of Multiservice, Multipriority Traffic Engineering Design for Data Networks", Proc. Globecom'99, Dec 1999.

[RFC-1458] Braudes, R. and S. Zabele, "Requirements for Multicast Protocols", RFC 1458, May 1993.

[RFC-1771] Rekhter, Y. and T. Li, "A Border Gateway Protocol 4 (BGP-4)", RFC 1771, March 1995.

[RFC-1812] Baker, F., "Requirements for IP Version 4 Routers", STD 4, RFC 1812, June 1995.

[RFC-1992] Castineyra, I., Chiappa, N. and M. Steenstrup, "The Nimrod Routing Architecture", RFC 1992, August 1996.

[RFC-1997] Chandra, R., Traina, P. and T. Li, "BGP Community Attributes", RFC 1997, August 1996.

[RFC-1998] Chen, E. and T. Bates, "An Application of the BGP Community Attribute in Multi-home Routing", RFC 1998, August 1996.

[RFC-2205] Braden, R., Zhang, L., Berson, S., Herzog, S. and S. Jamin, "Resource Reservation Protocol (RSVP) - Version 1 Functional Specification", RFC 2205, September 1997.

[RFC-2211] Wroclawski, J., "Specification of the Controlled-Load Network Element Service", RFC 2211, September 1997.

[RFC-2212] Shenker, S., Partridge, C. and R. Guerin, "Specification of Guaranteed Quality of Service", RFC 2212, September 1997.

[RFC-2215] Shenker, S. and J. Wroclawski, "General Characterization Parameters for Integrated Service Network Elements", RFC 2215, September 1997.

[RFC-2216] Shenker, S. and J. Wroclawski, "Network Element Service Specification Template", RFC 2216, September 1997.

[RFC-2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, July 1997.

[RFC-2330] Paxson, V., Almes, G., Mahdavi, J. and M. Mathis, "Framework for IP Performance Metrics", RFC 2330, May 1998.

[RFC-2386] Crawley, E., Nair, R., Rajagopalan, B. and H. Sandick, "A Framework for QoS-based Routing in the Internet", RFC 2386, August 1998.

[RFC-2474] Nichols, K., Blake, S., Baker, F. and D. Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers", RFC 2474, December 1998.

[RFC-2475] Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z. and W. Weiss, "An Architecture for Differentiated Services", RFC 2475, December 1998.

[RFC-2597] Heinanen, J., Baker, F., Weiss, W. and J. Wroclawski, "Assured Forwarding PHB Group", RFC 2597, June 1999.

[RFC-2678] Mahdavi, J. and V. Paxson, "IPPM Metrics for Measuring Connectivity", RFC 2678, September 1999.

[RFC-2679] Almes, G., Kalidindi, S. and M. Zekauskas, "A One-way Delay Metric for IPPM", RFC 2679, September 1999.

[RFC-2680] Almes, G., Kalidindi, S. and M. Zekauskas, "A One-way Packet Loss Metric for IPPM", RFC 2680, September 1999.

[RFC-2702] Awduche, D., Malcolm, J., Agogbua, J., O'Dell, M. and J. McManus, "Requirements for Traffic Engineering over MPLS", RFC 2702, September 1999.

[RFC-2722] Brownlee, N., Mills, C. and G. Ruth, "Traffic Flow Measurement: Architecture", RFC 2722, October 1999.

[RFC-2753] Yavatkar, R., Pendarakis, D. and R. Guerin, "A Framework for Policy-based Admission Control", RFC 2753, January 2000.

[RFC-2961] Berger, L., Gan, D., Swallow, G., Pan, P., Tommasi, F. and S. Molendini, "RSVP Refresh Overhead Reduction Extensions", RFC 2961, April 2000.

[RFC-2998] Bernet, Y., Ford, P., Yavatkar, R., Baker, F., Zhang, L., Speer, M., Braden, R., Davie, B., Wroclawski, J. and E. Felstaine, "A Framework for Integrated Services Operation over Diffserv Networks", RFC 2998, November 2000.

[RFC-3031] Rosen, E., Viswanathan, A. and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.

[RFC-3086] Nichols, K. and B. Carpenter, "Definition of Differentiated Services Per Domain Behaviors and Rules for their Specification", RFC 3086, April 2001.

[RFC-3124] Balakrishnan, H. and S. Seshan, "The Congestion Manager", RFC 3124, June 2001.

[RFC-3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V. and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.

[RFC-3210] Awduche, D., Hannan, A. and X. Xiao, "Applicability Statement for Extensions to RSVP for LSP-Tunnels", RFC 3210, December 2001.

[RFC-3213] Ash, J., Girish, M., Gray, E., Jamoussi, B. and G. Wright, "Applicability Statement for CR-LDP", RFC 3213, January 2002.

[RFC-3270] Le Faucheur, F., Wu, L., Davie, B., Davari, S., Vaahanen, P., Krishnan, R., Cheval, P. and J. Heinanen, "Multi-Protocol Label Switching (MPLS) Support of Differentiated Services", RFC 3270, April 2002.

[RR94] M.A. Rodrigues and K.G. Ramakrishnan, "Optimal Routing in Shortest Path Networks", ITS'94, Rio de Janeiro, Brazil.

[SHAR] Sharma, V., Crane, B., Owens, K., Huang, C., Hellstrand, F., Weil, J., Anderson, L., Jamoussi, B., Cain, B., Civanlar, S. and A. Chui, "Framework for MPLS Based Recovery", Work in Progress.

[SLDC98] B. Suter, T. Lakshman, D. Stiliadis, and A. Choudhury, "Design Considerations for Supporting TCP with Per-flow Queueing", Proc. INFOCOM'98, p. 299-306, 1998.

[SMIT] Smit, H. and T. Li, "IS-IS extensions for Traffic Engineering", Work in Progress.

[WANG] Y. Wang, Z. Wang, L. Zhang, "Internet traffic engineering without full mesh overlaying", Proceedings of INFOCOM'2001, April 2001.

[XIAO] X. Xiao, A. Hannan, B. Bailey, L. Ni, "Traffic Engineering with MPLS in the Internet", IEEE Network magazine, Mar. 2000.

[YARE95] C. Yang and A. Reddy, "A Taxonomy for Congestion Control Algorithms in Packet Switching Networks", IEEE Network Magazine, p. 34-45, 1995.

13.0 Authors' Addresses

Daniel O. Awduche
Movaz Networks
7926 Jones Branch Drive, Suite 615
McLean, VA 22102

Phone: 703-298-5291
EMail: awduche@movaz.com

Angela Chiu
Celion Networks
1 Sheila Dr., Suite 2
Tinton Falls, NJ 07724

Phone: 732-747-9987
EMail: angela.chiu@celion.com

Anwar Elwalid
Lucent Technologies
Murray Hill, NJ 07974

Phone: 908 582-7589
EMail: anwar@lucent.com

Indra Widjaja
Bell Labs, Lucent Technologies
600 Mountain Avenue
Murray Hill, NJ 07974

Phone: 908 582-0435
EMail: iwidjaja@research.bell-labs.com

XiPeng Xiao
Redback Networks
300 Holger Way
San Jose, CA 95134

Phone: 408-750-5217
EMail: xipeng@redback.com

14.0 Full Copyright Statement

Acknowledgement

Funding for the RFC Editor function is currently provided by the Internet Society.