An overview of MPLS Traffic Engineering

JP Vasseur – CISCO SYSTEMS

Technical Leader, ITD engineering, co-chair of the IETF PCE Working Group - jpv@cisco.com
Agenda

• Traffic Engineering
• Overview of MPLS Traffic Engineering
• Typical deployment of MPLS Traffic Engineering
• A quick overview of some more advanced MPLS Traffic Engineering topics
Agenda

• Traffic Engineering
• Overview of MPLS Traffic Engineering
• Typical deployment of MPLS Traffic Engineering
• A quick overview of some more advanced MPLS Traffic Engineering topics
Traffic Engineering

• Traffic engineering is the “art” of efficiently route the traffic so as to make an efficient use of the network resources and meet specific traffic requirements.

• Traffic engineering has been used for decades in telecommunication networks:
  • Telephony,
  • Private data networks,
  • Internet.

• Using a wide set of techniques …
Traffic Engineering

- Traffic engineering can be performed using several technologies …
  - An underlying layer-2 technology such as ATM (Well-known overlay model)
  - IP with metric optimization
  - MPLS Traffic Engineering

- Objective functions …
  - Do not exceed the link capacity,
  - Do not exceed a fraction of the link capacity (e.g. for sensitive traffic in conjunction with queuing techniques to bound queuing delay and jitter)
  - Additional constraints (delay, …)
Traffic Engineering

• The challenges of Traffic Engineering …
  • Requires to know the traffic matrix
  • Some traffic engineering techniques are relatively traffic disruptive and cannot efficiently cope with rapid traffic growth, sporadic traffic, multiple failures scenario, …
  • Multi-constraints objective functions are NP-Complete

• Various models of Traffic Engineering exists …
  • Centralized: efficient to solve multi-constraints problems (NP-Complete) *but* poorly scale and quite slow …
  • Distributed: highly scalable and dynamic, sometimes more complex to troubleshoot
Motivation for Traffic Engineering

• Increase efficiency of bandwidth resources
  Prevent over-utilized (congested) links whilst other links are under-utilized

• Ensure the most desirable/appropriate path for some/all traffic
  Override the shortest path selected by the IGP

• Replace ATM/FR cores
  PVC-like traffic placement without IGP full mesh and associated $O(N^2)$ flooding

• An important goal is COST SAVING
  Service development also progressing
Agenda

• Traffic Engineering

• Overview of MPLS Traffic Engineering

• Typical deployment of MPLS Traffic Engineering

• A quick overview of some more advanced MPLS Traffic Engineering topics
Shortest Path and Congestion

20Mbps traffic to R5
60Mbps aggregate
26Mbps drops!

R1
R2
R3
R4
R5
R6
R7

IP uses destination based routing …
Techniques exist to optimize IP metrics (knowing the traffic matrix) … but not very flexible/granular
MPLS TE Solution: based on admission control

- MPLS Labels can be used to engineer explicit paths
- Tunnels are UNI-DIRECTIONAL

Normal path: R8 ➔ R2 ➔ R3 ➔ R4 ➔ R5
Tunnel path: R1 ➔ R2 ➔ R6 ➔ R7 ➔ R4

20Mbps traffic to R5
40Mbps traffic to R5
20Mbps traffic to R5 from R8
40Mbps traffic to R1 from R8
MPLS Traffic Engineering Components

- Information distribution (topology and resource)
- Path computation
- Path setup
- Admission control
- Forwarding traffic on to tunnel
- Path maintenance
Topology and resource distribution

• Need to flood TE information (Resource Attributes) across the network
  Available bandwidth per priority level, links attributes (coloring), …

• IGP extensions flood this information
  OSPF uses Type 10 (area-local) Opaque LSAs
  ISIS uses new TLVs

• MPLS Traffic Engineering is a control-plane function
  … Data plane (queuing, congestion avoidance algorithms, …) is orthogonal. Can be used in combination!
Path Computation

• Once available bandwidth information and attributes are flooded, router may calculate a path from head (source) to tail (destination)

• The Head-end performs a “Constrained SPF” (CSPF) calculation to find the best path:
  - Constraints include available bandwidth, link affinity, delays
  - An objective function (reduce the path cost with respect to a specific metrics).

*Many* possible implementation of CSPF algorithms.

• Minimal impact on CPU utilization using CSPF

• Path can also be explicitly configured
Path (Traffic Engineering LSP) Setup

- Once the path is calculated, it must be signaled across the network
  Reserve any bandwidth to avoid “double booking” from other TE reservations
  Priority can be used to pre-empt low priority existing tunnels - MULTIPLE preemption algorithms exist.
- RSVP used to set up TE LSP
  PATH messages (from head to tail) carries LABEL_REQUEST
  RESV messages (from tail to head) carries LABEL
- When RESV reaches headend, tunnel interface = UP
- RSVP messages exist for LSP teardown and error sig
Path Setup Example

RSVP PATH: R1 → R2 → R6 → R7 → R4 → R9
RSVP RESV: Returns labels and reserves bandwidth on each link
Bandwidth available
Returned label via RESV message

PATH Message 20Mbps
Admission Control

• On receipt of PATH message
  
  Router will check there is bandwidth available to honour the reservation
  
  If bandwidth available then RSVP accepted

• On receipt of a RESV message
  
  Router actually reserves the bandwidth for the TE LSP
  
  If pre-emption is required lower priority LSP are torn down

• OSPF/ISIS updates are triggered if threshold is crossed
Load Balancing

Many types of load balancing:

• 1) Load balancing of tunnels across multiple possible paths

• 2) Load balancing of traffic over multiple tunnels (TE LSPs)

IP routing has equal-cost load balancing, but not unequal cost* load balancing (difficult to do while guaranteeing a loop-free topology)

MPLS Traffic Engineering support unequal load balancing with configurable share.
Path Maintenance

• RSVP states must be refreshed (Soft state protocol) but in Steady-state information load is low
  Especially with refresh reduction (RFC2961)

• Path re-optimization
  Process where some traffic trunks are rerouted to new paths so as to improve the overall efficiency in bandwidth utilization
  For example, traffic may be moved to an alternate path during failure; when more optimal path is restored traffic moved back

• Network recovery
  Comprised of two techniques; local protection (link and node) and path (Global) protection
Diff-Serv-Aware TE (DS-TE)

- DS-TE is more than MPLS TE + MPLS Diff-Serv
- DS-TE makes MPLS TE aware of Diff-Serv:
  - DS-TE engineers separate LSPs for different QoS classes
  - DS-TE takes into account the bandwidth available to each class (e.g. to queue)
  - DS-TE takes into account separate engineering constraints for each class
    - e.g. I want to limit voice traffic to 70% of link max, but I don’t mind having up to 100% of BE traffic
    - e.g. I want overbooking ratio of 1 for voice but 3 for BE
  - DS-TE may take into account different metrics (e.g. delay)
- DS-TE ensures specific QoS level of each Diff-Serv class is achieved
If I Can Keep EF Traffic $< \alpha \%$, I Will Keep EF Delay Under $M_1$ ms
If I Can Keep AF1 Traffic $< \beta \%$, I Will Keep AF1 Delay Under $M_2$ ms
Bandwidth Pools
Maximum Allocation Model

- Each pool can have its own overbooking ratio
- Unused BW in one pool is not available for other pools
Bandwidth Pools: Russian Dolls

- Each pool has its own overbooking ratio
- Unused BW in one pool is available to others
Why the Russian Doll Model?

- It is a good match to the way many network operators manage QoS in the data plane
  
  e.g. Voice in the LLQ, business data in a high weight CBWFQ class, best effort gets whatever is left
  
  VOIP in LLQ is unaffected by other traffic: use BC2
  
  Business data gets preferential access to link vs. BE: use BC1
  
  Best effort can use most of the link if the other classes are not fully used, but should be reduced if lots of VOIP or business data is present: use BC0

- Good isolation between classes, efficient use of bandwidth
Routing/Forwarding on DS-TE Tunnels

• **<Destination/COS>-based Routing/Forwarding:**
  
  Needed when Different COS destined to same Prefixes (eg voice/data within a VPN Site)

  NextHop Tunnel depends on Destination AND on COS

  Dynamic Routing (i.e. no manual configuration of routes, rerouting,..)

  This is “COS Based TE Tunnel Selection (CBTS)”
An efficient MPLS network recovery mechanism: MPLS Traffic Engineering Fast Reroute

- MPLS TE Fast Reroute link/node protection is:
  - a **Local** recovery technique (compared to IGP, MPLS TE LSP reroute, or Path protection (1:1) which are global protection/restoration mechanisms)
  - Using **Protection**

→ **Allows to achieve a O (10s msecs) convergence time**

- Temporary mechanism followed by reoptimization with “Make before break” in order to find a more optimal path

“Network Recovery” - JP Vasseur, Mario Pickavet and Piet Demeester - Morgan Kaufmann - July 2004
Fast Reroute Basics

- Reroutable LSP
- NNHOP Backup LSP
- Protected LSP
- Point of Local Repair (PLR)
- Merge Point (MP)
- NHOP Backup LSP
The Path messages are sent onto the backup tunnel to refresh the downstream states
MPLS TE FRR Local repair

- **MPLS TE FRR BYPASS** make use of nested LSPs (stack of labels)

**Convergence Time**

FRR LSPs: $O(50\text{ms})$

Non FRR LSPs: $O(s)$

- Fast reroutable LSPs
- NON Fast reroutable LSPs are rerouted using restoration
SRLG aware backup tunnel path computation: an example with Link Protection

• Step 1: on each link, configure the set of SRLGs the link belongs to:

[no] mpls traffic-eng srlg <0-4294967295>

• The SRLG membership is then passed to OSPF/ISIS and flooded throughout the area/level

• ex (OSPF): draft-ietf-ccamp-ospf-gmpls-extensions defines several new sub-TLVs carried in the link TLV (TLV type 2)

Sub-TLV Type Length Name
16 variable Shared Risk Link Group

• The topology, resources information along with the SRLG membership are stored in the TE database.
SRLG aware backup tunnel path computation: an example with Link Protection

- Step 2: on each LSR, one single command allows to automatically configure the set of NHOP and/or NNHOP backup tunnels:

  [no] mpls traffic-eng auto-tunnel backup srlg exclude {force | preferred }

  “force” indicates that no backup tunnel must be created if the SRLG diversity constraint cannot be satisfied.

  “Preferred”: if no SRLG diverse path can be found the constraint is dynamically relaxed.

Example with NHOP:

For each link (with an IGP adjacency):

→ If there is no Fast Reroute TE LSP traversing the link, do nothing.

→ If there is at least one Fast reroutable TE LSP traversing the link, then compute an SRLG diverse path for the NHOP backup tunnel and establish the NHOP backup tunnel.
Introduction

• Bandwidth Protection mechanisms is required:
  – For some types of traffic
    → Not all the traffic types require bandwidth protection
    Ex: Voice, AToM traffics typically require bandwidth protection.
  – In some networks
    → Not all the networks require bandwidth protection
Bandwidth/Jitter protection during failure

- Simple approaches can be used for the backup tunnel resource reservation but expensive...

- More sophisticated algorithms have been designed to minimize the required backup capacity thanks to Bandwidth Sharing while being capable of enforcing other constraints.
MPLS TE Fast reroute is a temporary mechanism (backup tunnels are used for a short period of time, until the TE LSPs are rerouted/reoptimized by head-ends and traverse a protected path),

Therefore in practice, during that period:

\[ P_b \text{ (multiple failures)} \ll 1 \]

That makes the single simultaneous failure assumption valid
Bandwidth sharing

Objectives

→ for each LSR, find a set of backup tunnels (to the NHOP for link P and NNHOP for Node P) with capacity constraints,

→ Two backup tunnels protecting the same facility cannot share bandwidth,

→ Two backup tunnel protecting independent facilities can share bandwidth (as they are not simultaneously active under the assumption of single failure)
Agenda

• Traffic Engineering
• Overview of MPLS Traffic Engineering
• Typical deployment of MPLS Traffic Engineering
• A quick overview of some more advanced MPLS Traffic Engineering topics
Coupling Layer-2 Services with MPLS

- **TE - AToM Tunnel Selection**

  - Static mapping between pseudo-wire and TE Tunnel on **PE**
  - Implies PE-to-PE TE deployment
  - TE tunnel defined as preferred path for pseudo-wire
  - Traffic will fall back to peer LSP if tunnel goes down
Agenda

• Traffic Engineering
• Overview of MPLS Traffic Engineering
• Typical deployment of MPLS Traffic Engineering
• A quick overview of some more advanced MPLS Traffic Engineering topics
Inter-domain MPLS Traffic Engineering

• A challenging problem ...
  • By definition, network topology and resource information is limited to the domain (domain=IGP area or an Autonomous System)
  • Path computation involves sophisticated techniques ...
  • Path reoptimization is another challenge: how does the head-end detect a more optimal path in a downstream domain?
  • When domain belong to different administrative authority, confidentiality becomes an issue!

• Solutions?
  • Per-domain path computation
  • PCE-based path computation
Inter-AS MPLS Traffic Engineering

- One possible scenario: seamless TE plane for bandwidth optimisation, ASBR FRR node protection, strict QOS guaranties across ASes

AS1 SP1

ASBR

ASBR

ASBR

AS2 SP1

Bw optimisation

ASBR

ASBR

FRR

AS3 SP1

Strict QOS guaranties (DS-TE, ...)

VoIP
Per-domain path computation: Conclusion

A simple path computation model consisting of computing the TE LSP path on a per-domain basis, in a serialized fashion.

- **Pros**: simple

- **Cons**: increased probability of call set up failure (cranback :-( ), potentially suboptimal!

- **Path reoptimization** can be performed by in-band polling mechanism or explicit mid-point notification so as to reoptimize a TE LSP within a domain for local reoptimization within a domain (area or AS) (supported in IOS).
Inter-domain IETF Standardization

A new PCE Working group (Routing Area) has been formed:

• Chairs: JP Vasseur, A. Farrel / AD: A. Zinin, B. Fenner
• Mailing list: pce@ietf.org
• PCE Architecture ID published (draft-ash-pce-architecture)
• Various new drafts under work:
  • Generic PCE Communication protocol requirements
  • PCE discovery protocol requirements
  • …
• See http://www.ietf.org/html.charters/pce-charter.html
PCE-based path computation

- PCE architecture is defined in draft-ash-pce-architecture:

  "This document specifies the architecture for a Path Computation Element (PCE)-based model to address this problem space. This document does not attempt to provide a detailed description of all the architectural components, but rather it describes a set of building blocks for the PCE architecture from which solutions may be constructed."

- This presentation focuses on one model based on a distributed PCE-based computation of inter-domain TE LSP path where the ABR/ASBR act as PCE
Distributed PCE-based computation

AS1

ASBR1 build a virtual SPT (the shortest path is built using a backward recursive computation)

ASBR1 \* PCE

ASBR2 \* PCE

ASBR3 \* PCE

ASBR4 \* PCE

ASBR5

ASBR6 \* PCE

ASBR7

ASBR8 \* PCE

B

A

Path Computation Request

Path Computation reply

© 2001, Cisco Systems, Inc. All rights reserved.
Auto-bandwidth MPLS TE

• Estimating the TE LSP bandwidth can be challenging,

• Moreover, the traffic can be quite bursty and fluctuates within the day,

• International networks have multiple time zones,

• Auto-bandwidth TE LSP automatically resizes the TE LSP based on the traffic actually sent onto the TE LSP …

  → Adjustable sampling

  → Adjustable resizing frequency
Various possible resizing strategies.

---

**LSP26 Traffic Behavior, Mechanism 2**

- **Size**
- **Traffic**

---

**LSP26 Traffic Behavior, Mechanism 4**

- **Size**
- **Traffic**
Performance Comparison

Link Utilizations

Resizes per Mechanism

LSPs not on IGP path

Worst case LSP count for a node

- Mechanism 1
- Mechanism 2
- Mechanism 3
- Mechanism 4
Auto-mesh Traffic Engineering

• **Problem**: Deploying a full mesh of N router requires a significant configuration ➔ With 100 routers in a full mesh, 9900 TE LSPs must be configured!

• Moreover, adding Routers to an existing TE Full Mesh
  • TE Tunnels need to be built from the 101st router to every other 100 routers.
  • Since TE Tunnels are unidirectional, a TE Tunnel needs to be built from each of the 100 routers.
Auto-mesh Traffic Engineering

Mesh-group 1
=> LSA update

Mesh-group 1: no affinities, global-pool, IGP metric, preemption 7, ...

For example, Blue mesh = data – Red mesh = voice
Point to Multipoint MPLS Traffic Engineering

• Increasing demand for high rate source (Video/TV distribution) for network optimization and QoS guarantees for multicast flows.

Selection of an optimal (Steiner Tree) satisfying constraints by means of MPLS TE
Further reading

• IETF Specifications: [www.ietf.org](http://www.ietf.org) (MPLS, CCAMP, PCE and TE Working Groups)

• Books


Thank You !