Internet Technology

The "inner network" view, part 2 (C): MPLS

Michael Welzl http://www.welzl.at

DPS NSG Team http://dps.uibk.ac.at/nsg
Institute of Computer Science
University of Innsbruck, Austria

TE Deployment considerations

- Scalability: how many LSPs possible / needed / reasonable?
  - one of the most important deployment considerations; hard to determine
  - limited by connectivity requirements: any-to-any connectivity needs O(n^2) LSPs – hence normally only deployed in the core, where scalability issues can be solved with LSP hierarchy
  - limited by bandwidth ("size") of traffic trunk: if capacity exceeded, load balance via multiple LSPs
  - Max. no. of supported LSPs normally provided by vendors
    - range of several tens of thousand LSPs
    - often different numbers given fro head end and transit (middle LSR)
- Reservation granularity: size of individual reservations
  - limited by bottleneck capacity
  - limited by number of LSPs (see above)

Using TE for resource optimization

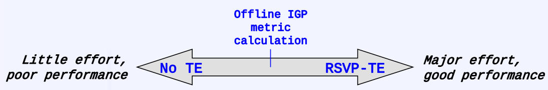
- TE deployment in parts of network for routing traffic away from congested link
  - tactical application: for quickly solving an immediate resource problem
  - e.g. fix problems that occur as scheduled link upgrade is delayed, or optimize usage of a particularly expensive link
- TE deployment throughout entire network for improving overall bandwidth utilization
  - strategic application: for long-term benefit
  - e.g. delay costly link upgrades by applying TE in network core
- In any case, TE based on knowledge about bandwidth requirement for LSP at its head end, available bandwidth at network nodes
  - but this information is not always available…

Autobandwidth

- How much bandwidth to reserve for an LSP?
  - based on knowledge about available bandwidth, i.e. traffic patterns
  - Manual estimations can be difficult (usually fluctuates with time of day)
- Wrong estimations possible:
  - estimate too high ⇒ waste of bandwidth
  - estimate too low ⇒ LSP cannot accommodate traffic
    - worse (packet drops), so usually estimated conservatively
    - note: RSVP only operates in control plane – traffic shaping needed to ensure conformance
- Solution: Autobandwidth
  - Ingress of an LSP monitors traffic statistics and periodically adjusts LSP's bandwidth reservation to traffic demand
  - Done by setting up new LSP and switching in make-before-break fashion
  - Proprietary technology (no IETF standards)

Offline optimization

- Possible to add offline optimization loop:
  - measure traffic, simulate the network, derive settings, adjust if necessary, repeat
- Was shown to enable traffic engineering in LDP based networks by manipulating IGP link metrics
  - less overhead and easier maintenance than RSVP-TE (at the cost of reduced control of network elements)
  - normally not advisable: influencing IGP can affect the whole network
  - test results show worse results than with explicit routing, but much better results than without any TE



Offline path computation

- Remember CSPF / multiple paths example: suboptimal performance because future reservations unknown
  - no optimal strategy; can only be solved with offline path computation
- Several other practical advantages
  - global view of reservations
  - no surprises from dynamic computation
  - ability to traverse AS boundaries (information for calculation not necessarily limited to TED)
  - can calculate normal and failure cases, take both into account
  - can use more sophisticated algorithms than CSPF
    - CSPF only takes calculating head end's LSPs into account, offline path computation can use view of the whole network

Offline path computation difficulties

- Volume of necessary data for calculation
- Changing network conditions can lead to large number of LSP configuration changes
  - may be impractical
  - solution: incorporate performance vs. configuration effort trade-off in calculation
- Result must contain order of upgrade
  - configurations cannot be changed simultaneously on all routers
  - during update, problems can arise
- typically slow calculation; impractical for quick temporary fixes

Protection and Restoration

The problem

- Remember: MPLS enables convergence of services
  - e.g., send best-effort IP + voice + video + ATM CBR over the same net
  - some of this traffic is "fragile": users do not accept phone interruptions (but requirement slightly relaxed for cell phones ⇒ different levels of loss tolerance)

⇒ Fast recovery from failures = key functionality of multiservice nets
  - IGP reconvergence speed may not be fast enough

- Some layer 1 technology can do this (but need to use such layers)
  - e.g. SONET Automatic Protection Switching (APS)
- MPLS can help, but only with RSVP-TE

Failure detection

- Automatic indication hardware dependent (e.g. provided in packet-over-SONET/SDH, not provided in Ethernet)
  - need a general solution
- IGP can detect failure - but inefficient
  - message frequency = (CPU + network) load vs. detection speed trade-off
- Solution: Bidirectional Forwarding Detection (BFD) protocol
  - fast low-layer per-link ping
- BSD works well, so fast failure detection assumed to be available and work in upcoming slides

End-to-end protection

- Set up two LSPs: primary and secondary (also called "protection path")
  - primary used; switch to secondary upon failure
  - setting secondary up in advance helps ensure
    - fast switchover
    - conformance of secondary path to traffic requirements
    - path diversity (shared links increase chance of double failure)
- Switching to secondary path done by LSP head end
  - upon reception of RSVP error message
- Issues
  - Secondary LSP resource reservation usually similar to primary
    - total reservations = 2 x necessary reservations under normal operation
    - wasted bandwidth can be prevented by assigning better priorities to primary LSPs
  - Unnecessary protection for some links (e.g. when they have SONET APS)
  - Delay until arrival of RSVP error message nondeterministic

Local protection

- Problems with end-to-end protection partially due to LSP head end being in control
- Hence, solution: protect as close as possible to point of failure
  - Use alternate sub-path (called "detour" or "bypass") within LSP
  - consider cars on highway: bypass problem by using a country road for a while, but not all the way
- Faster reaction possible ⇒ Fast Reroute (FRR)
  - Only done until head end acts
    - head end's secondary path can be better
    - interior LSRs have different shortest paths to dest. than head end
    - not feasible to require interior LSRs to additionally maintain shortest paths from head end's point of view

Local protection /2

- Distinguish:
  - Resource that is protected: link or node (influences placement of detour)
  - Number of LSPs protected: 1 ("one-to-one backup") or N ("facility backup") (both cases protected with only 1 detour)
- Some terminology
  - backup path called detour in case of 1:1 backup, bypass in case of N:1
  - head end of backup path (router upstream of failure) called Point of Local Repair (PLR)
  - tail end of backup path (where traffic merges into normal path again) called Merge Point (MP)
  - "normal path" = LSP receiving protection called "protected path"

Link protection: control plane before failure

- Backup path established around link
  - need to compute path (CSPF) + install state in PLR, MP and transit nodes
- PLR must learn that it should do this
  - for a certain link + for certain LSP(s)
  - may not be necessary for all LSPs (e.g. voice vs. best effort IP)
- Choice of link configured at PLR, but LSP configured at LSP head end
  - information propagated from head end to PLR via RSVP Path messages ("local protection desired" flag + optional Fast Reroute Object for telling PLR about constraints to be used in CSPF)

Data plane

- One-to-one backup: can use alternate path in a "normal" way
  - labels are swapped by all LSPs including PLR and MP, additional state necessary for alternate path at PLR and MP
- Facility backup: additional label state necessary PLR and MP per LSP
  - may not be feasible
- Solution for facility backup: stack labels
  - PLR pushes backup path label on top of existing label
  - Penultimate hop popping used
  - traffic arrives at MP with the same label as if it would arrive via the failed link
    ⇒ no per-LSP state necessary at PLR (just push) or MP (just forward)

Link protection: control plane after failure

- Error messages (e.g. IGP) leading to LSP teardown must be suppressed
- LSP head end must be notified about failure
  - now is the time for the RSVP error message
  - contains "Notify" error code + "Tunnel locally repaired" subcode + flag in Record Route Object
  - Could theoretically be omitted, head end could rely on IGP messages – but this would not work across multiple AS'es
- LSP head end now switches to secondary LSP (make-before-break)
  - because it learns what happened via "Tunnel locally repaired" in RSVP
  - new path could be the path that is already used
  - so why bother switching? depends on policy / implementation at head end
  - note: if path is kept, must ensure that RSVP messages are correctly forwarded over backup path to avoid timeouts

Node protection

- If downstream end of link fails, must bypass the node (two links)
- PLR can only establish backup path if it knows the address of the downstream node after the failed node (and the label it expects)
- Address available in RRO of RSVP Path message, but not label
  ⇒ flag "label recording desired" was added to RRO
  - normally, LSRs only learn about immediate downstream labels
- Forwarding done as with facility backup (label stacking), but PLR must swap label with the one expected by the correct MP before pushing

Fate sharing

- If primary and secondary path use the same optical fiber, a bulldozer can eliminate both at the same time
  - this is called fate sharing
  - the paths are said to be in the same Shared Risk Link Group (SRLG) or fate sharing group
- Avoiding SRLG = constraint for calculating the protection path
  - user-defined; like link colors, but can be dynamic: models dependencies between links, and link usage depends on routing changes
  - not a very strict constraint
    - e.g. increase link costs to make creating a SRLG less likely
    - but generally better to have a SRLG than to have no protection path
- How to learn about SRLGs?
  - knowledge comes from network operator's database
  - either manually configure routers or use IGP (GMPLS extensions for OSPF)

Bandwidth protection

- 100% working protection paths for all LSPs in the whole network without packet loss only possible if total network capacity is doubled
  - trade-off: (overprovisioning + better protection) vs. bandwidth costs
  - common rule of thumb: upgrade when average load exceeds 50%
- Bandwidth protection: other methods for guaranteeing that enough bandwidth will be available
  - makes sense for local protection (FRR): traffic will only use backup path for a few seconds, there should be little packet loss during this interval
  - PLR can announce this capability with flag in Record Route Object
  - head can then request its usage for the LSP using flag in Session Attribute and Fast Reroute Objects
    - LSPs where PLRs cannot do this can be made less attractive, e.g. by increasing their metric if incorporated as link in IGP

Bandwidth protection /2

- One-to-one backup: upon request for bandwidth protection, PLR only establishes backup path with enough bandwidth
- Facility backup: establishing appropriate backup path per LSP impossible by design - but bandwidth on single path may not suffice for all LSPs
  - Solution: reserve fixed bandwidth + perform admission control
- So far, assumed that backup path is idle unless a failure happens
  - That's a waste
  - But if we let "normal" traffic share links with the backup path, failure can affect this traffic
  - Solution: apply DiffServ; map DSCP onto EXP bits in label, give protected traffic higher drop precedence (i.e. preferably dropped during congestion)

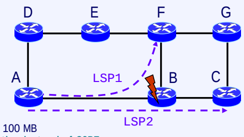
FRR deployment considerations

- Scalability
  - Problem complexity
    - Local protection said to be difficult to configure, but up-to-date implementations make it easier (dynamic calculation and establishment of protection paths)
    - Still, number of resources that can be protected limited by complexity
  - Number of LSPs
    - 1:1 protection: # backup paths depends on # protected resources and # LSPs
    - N:1 protection: # backup paths only depends on # protected resources
      - Only true for link protection
      - node protection: also depends on topology (different MPs for LSPs possible)
  - Forwarding state
    - Depends on topology (e.g. length of protection path), protection type (1:1 or facility)
    - Make-before-break temporarily consumes resources

FRR deployment considerations /2

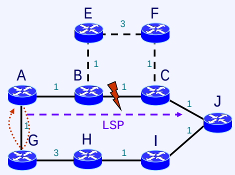
- Recovery speed (influences number of lost packets)
  - Detection time: hardware detection vs. BFD support and operation speed
  - Switchover time: how fast to switch from one LSP to protection path
  - Number of LSPs switched over in a certain amount of time
  - IP routing forwarding state update speed: relevant when problem happens at head end ⇒ LSP failure can influence IP routing
- Cost of bandwidth protection
  - Overall amount of bandwidth reserved for protection should be minimized; the longer the path, the more resources are kept idle
  - Example on the right: assume all link capacities = 100 MB, LSP1 and LSP 2 need up to 100 MB bandwidth
  - Failure at B:
    - protection path of LSP1: reserve 100 MB along A-D-E-F
    - Protection path of LSP2: reserve 100 MB along A-D-E-F-G-C

    ⇒ Total 200 MB reservation doesn't match reality: bandwidth of LSP1 + LSP 2 cannot have exceeded 100 MB (A-B link capacity) ⇒ solution: offline computation instead of CSPF



LDP and IP FRR

- LDP
  - Attractive for operational simplicity, but does not support the mechanisms we've seen so far...
  - Possibility: use one-hop RSVP LSPs, tunnel LDP sessions through them
  - More attractive alternative: LDP based FRR
    - LDP uses IGP, hence LDP FRR = IP FRR
- IP FRR tunnel-based approach:
  - set up protection path with RSVP, tunnel through it only in case of failure; as with RSVP, PLR must learn MP's label
  - Microloops can happen due to IP routing
  - Example on the right:
    - all link costs 1, except G-H (3) and E-F (3) B-E-F-C = protection path
    - At A: costs of backup path > costs via G
    - Until IGP converged, G's shortest path to J is via A ⇒ for a while, traffic will loop!



IP FRR alternate path approach

- Maintain alternate path at head end
  - Example on previous slide: assume all link costs are 1
    - A-G-H-I-J calculated in addition to default path A-B-C-J
    - A forwards to G when B-C link fails
  - Link costs as in example: G would route back to A ("U-turn")
  - Prevention with U-turn alternates: let G detect that it sends traffic back via the incoming interface ⇒ use other (higher cost) path to destination (J) instead
    - Does not work in arbitrary topologies
    - Calculating alternate paths adds computational complexity and forwarding state (scalability concern)
    - No explicit control of path traffic will take upon failure

DiffServ Aware MPLS Traffic Engineering (MPLS DiffServ-TE)

About combining DiffServ and TE

- Complementary: each mechanism has benefits that the other doesn't
  - e.g. DiffServ can provide guarantees, but not resilience
- Convergence enabled by MPLS (carry IP + Ethernet, ATM, FR, ...) leads to strict SLA requirements
  - e.g. MPLS can provide resilience, but not prioritization via queuing
- Class-of-service (CoS) unknown to MPLS without DiffServ
  - Combining enables resource reservation with CoS granularity
  - Provide fault-tolerance properties of MPLS at a per-CoS level
- Reminder:
  - E-LSP (EXP-inferred LSP): map EXP ⇔ DSCP
  - L-LSP (Label-inferred LSP): map EXP+label ⇔ DSCP

Application scenarios

- Voice (delay-sensitive) and data
  - DiffServ can assign a priority to voice ⇒ queued on its own
  - Still, the voice-queues can grow ⇒ delay
  - Hence, amount of voice traffic per link should be kept small
  - ⇒ CoS becomes a (dynamic, i.e. depending on traffic amount) constraint when using an alternate path in case of failure
- Three classes (e.g. voice, video, data)
  - Queue sizes and scheduling policies should be configured for QoS
  - Should be a function of traffic load, which is a function of routing, LSP preemption, FRR, ..
  - ⇒ TE should enable fixing relative proportions of each traffic type on links
- Guaranteed bandwidth service and best effort service
  - How to do TE for best effort without violating guaranteed bandwidth service requirements?

Class Type

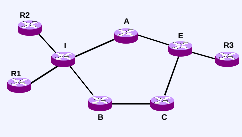
- Class Type (CT): can be thought of as queue and associated resources
  - 8 CTs supported: CT0 (best effort) – CT7
  - No predefined mappings; could be one or more PHBs
  - DiffServ-TE LSP
    - LSP which guarantees bandwidth for a particular CT
    - Carries one CT; non-DiffServ LSPs assumed to be mapped to CT0
- Voice / data example on previous slide
  - voice = EF PHB, mapped to CT1, data = BE PHB, mapped to CT0
  - Bandwidth available for CT1 limited to percentage of link required to ensure small queuing delays for voice traffic
  - Separate TE LSPs established for CT0 and CT1

DiffServ-TE CSPF and path signaling

- CSPF with DiffServ constraints
  - Goal: serve requests like "LSP to destination X, using CT1 at (preemption) priority 3, bandwidth 30 Mbit/s"
  - Available bandwidths per CT must be known for each link
  - 8 priorities x 8 CTs = 64 values per link ("TE class matrix")
    - Limited to a choice of 8 by the IETF in RFC 4124 for practical reasons
  - Must be advertised by IGP; also specified in RFC 4124
- Path signaling
  - Class Type (CT) object of RSVP Path message specifies associated CT
  - Only used for CT1 – CT7 (CT0 = default when CT object is missing)
  - Incremental deployment: nodes which don't understand CT object must reject request
    - DiffServ-TE LSPs can only be established through LSPs which can accordingly serve the request
  - Side note: Constraint-based Routing LDP (CR-LDP) was also specified but eventually abandoned by the IETF

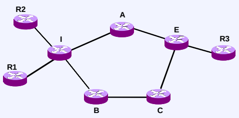
Bandwidth constraint models

- Bandwidth Constraint (BC): percentage of link's bandwidth that CT(s) can take up
- Maximum allocation model (MAM)
  - Map one BC to one CT; link bandwidth is divided among CTs
  - Completely isolates CTs ⇒ LSP priorities between different CTs irrelevant
  - Disadvantage: inflexible; bandwidth can be wasted
  - Topology on the right: assume all link capacities are 10 Mbit/s, 9 reserved for CT0, 1 for CT1
  - Establish LSP1, 9Mbit/s, R2-R3, CT0: I-A-E
  - Establish LSP2, 1Mbit/s, R1-R3, CT0: I-B-C-E (cannot use I-A-E anymore because 0 Mbit/s left for CT0)



Russian Dolls Model (RDM)

- Better bandwidth usage by allowing CTs to share bandwidth
- Different levels of strictness: CT7 > CT6 > ... CT0
  - BC7 is mapped to CT7 only
  - BC6 accommodates traffic from CT7 and CT6
  - BC5 accommodates traffic from CT7, CT6 and CT5
  - ... BC0 accommodates traffic from all classes
- Topology on the right:
  - All link capacities are 10 Mbit/s, CT0 for data, CT1 for voice, BC1 = CT1 = 1 Mbit/s, BC0 = CT0 + CT1 = 10 Mbit/s
  - Establish LSP1, 9Mbit/s, R2-R3, CT0: I-A-E
  - 1 Mbit/s left on I-A-E for either CT1 or CT0



- Guarantees in RDM: must use priorities (preemption)
  - Makes available bandwidth calculation (configuration) more complicated

Overbooking

- Reserve X Mbit/s for N LSPs along link of capacity X Mbit/s: some bandwidth will remain unused
  ⇒ Overbooking: reserve more than available
  - Several methods
- LSP size overbooking
  - reserve lower bandwidth value than the maximum traffic that will be mapped to the LSP
- Link size overbooking
  - Artificially raise max. reservable link bandwidth, work with these values
- Local Overbooking Multipliers (LOM)
  - Link size overbooking with different values for different CTs (e.g. 3:1 overbooking for CT0 but 1:1 overbooking for CT1)
- Manual BC configuration
  - User specifies bandwidth constraints, can overbook a class

Protection

- Backup path must reserve bandwidth for the same class type as protected path
  - No problem in 1:1 backup case
  - Facility backup: two options
    1. Single backup: all classes mapped onto single backup and treated as best-effort
    2. Separate backup per CT: one backup for each class type, admission control of LSPs into appropriate backup based on bandwidth request and class type
- Traffic must be kept within reservation limits
  - Police traffic at network edge (ingress) or use LSP policer (per-CT granularity; drop or mark out-of-profile traffic at head end)
  - Admission control: only admit connections that can be accommodated

Multiclass LSPs

- Mapping traffic with different DiffServ behaviors onto the same LSP
  - This LSP must satisfy the bandwidth constraints for each of these classes
  - Does not yield new functionality, but can reduce state (increase scalability) and facilitate configuration
- Application scenario: ATM trunk emulation
  - All traffic classes should follow the same path, exhibit the same behavior in case of failure
    - if EF class fails, BE should fail too
    - Otherwise, the same protection path should be used for all classes
  - Can also be achieved with separate LSPs, but more cumbersome

References

- Ina Minei, Julian Lucek: "MPLS-Enabled Applications", John Wiley & Sons, 2005, ISBN: 0-470-01453-9
- Slides from Dimitri Papadimitriou
  - Thanks!!!