Multi Protocol Label Switching
Motivation

- IP
  - connectionless
  - IP routing is based on shortest-path routing
    - Shortest path can become the bottleneck
    - Limited QoS support in IP network
    - Routing in layer 3 is slower than layer 2 switching
Motivation (cont.)

- ATM
  - connection oriented
  - Supports QoS
  - fast packet switching with fixed length packets (cells)
  - integration of different traffic types (voice, data, video)
  - Complex technology
Motivation (cont.)

- Combine the forwarding algorithm used in ATM with IP.
Motivation (cont..)

- Multiprotocol Label Switching (MPLS) came about from two different paradigms in data networks: datagram (IP) and virtual circuit (ATM).
- IP over ATM solution (overlay model) was not a viable solution due to scaling issues.
- This lead to label-switching solution
  - IP and ATM are tightly integrated
IP over ATM problems

- Nodes within a LIS have to be connected in a partial or full mesh \((O(n^2))\) connections required.
- When two IP subnets are physically on the same ATM network, packets cannot be switched all the way between the two subnets
  - Needs Next Hop Resolution Protocol (NHRP)
- ARP: ATM does not support broadcast
  - Hence ATMARP introduced ARP server where each IP device has to register its ATM address
- This led to label-switching solution
  - IP and ATM are tightly integrated
Advantages of Label Switching

- Simpler packet forwarding paradigm
  - IP lookup involves longest-prefix match, which requires extensive preprocessing and multiple memory access. With label switching, packets are forwarded by doing an exact match against a short label by looking up the label switching table
  - Makes forwarding independent of routing architectures. Once Label Switching Paths (LSP) are established packet forwarding is always the same. Thus new routing schemes can be developed without changes in the forwarding logic
Advantages of Label Switching (cont.)

- Better forwarding granularity. For current IP based routing granularity is destination-based. But sometimes more granularity is desired e.g. an ISP may want to know from which interface or customer a particular packet came from. Label switching allows multiple granularities e.g. packets from a particular ingress interface can be put into an LSP.

- Can be used for traffic engineering
Label Switching Proposals

• Tag Switching (from Cisco)
  - Control-driven approach: setup of LSPs closely follows IP control msgs such as routing updates
  - Tag bindings are carried by tag distribution protocol (TDP)
  - Each tag switch router (TSR) distributes tag binding to all its neighbors for all FEC (forwarding equivalence class) (independent distribution).
Label Switching Proposals (cont.)

- Aggregate Route-Based IP switching (ARIS) (IBM)
  - Control-driven approach like tag switch
  - LSP setup starts at the egress node for a FEC (egress node is the last node along the path of FEC the next hop of which is a non-ARIS node)
  - Provides mechanism for loop prevention and detection and hop count facility
Label Switching Proposals (cont.)

- IP switching (IPSILON)
  - Traffic driven protocol that sets up shortcut paths on the fly for long lasting flows
  - Initially packets from the flow is sent to controller of IP switch and are forwarded with normal destination based IP forwarding.
  - When number of packets from a flow exceeds a threshold, controller uses IFMP (Ipsilon Flow Management Protocol) to set up LSP for the flow
  - Once LSP is established, packets from the flow start to follow shortcut LSP, rather than going through the controller
Label Switching Proposals (cont.)

- Cell Switch Router (CSR) (from Toshiba)
  - Similar to IP switching in many respects
  - Primarily designed as a device for interconnecting ATM clouds
  - Multiple ATM logical IP subnets (LIS) are interconnected with CSRs that are capable of both IP and cell forwarding
  - CSRs can choose to set up LSPs for long lasting flows (based on port # of the packets)
  - Setup of LSPs is data driven for best effort traffic and RSVP driven for flows that require resource reservation
MPLS Architecture

- Multi Protocol Label Switching is arranged between Layer 2 and Layer 3
MPLS Architecture (cont..)

- Label
  - A short, fixed length, locally significant identifier that is used for label switching
  - Label can be encoded into a packet (shim header)
  - Label may be mapped to some field in an existing data link or network layer protocol (e.g. in ATM network vpi/vci fields are used as label)
  - Each label is associated with a FEC (Forwarding Equivalence Class). FEC defines a group of IP packets that are forwarded over the same LSP with same treatment
Hierarchical Label Stack
- MPLS allows more than one label to be encoded in a packet: referred to as label stack
- Used to support nested tunnels.
- When a new label is pushed on the top of current label, packet will follow the tunnel due to the new label
- When packet reaches the end of the tunnel, LSR at the end of the tunnel discards the top level label and the label below pops up.
**MPLS Architecture (cont.)**

- **Label-Switching Table**
  - Also called incoming label map (ILM), maintains mapping between an incoming label to outgoing interface and outgoing label

<table>
<thead>
<tr>
<th>Incoming Label</th>
<th>First Subentry</th>
<th>Second Subentry</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Outgoing Label</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Outgoing</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Interface</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Next hop address</td>
<td></td>
</tr>
</tbody>
</table>

- Multiple subentries are useful for multicasting
MPLS Architecture (cont.)

- The entry that incoming label points to is called next-hop label-forwarding entry (NHLFE)
- When a packet arrives, the LSR finds the corresponding NHFLE for incoming label by performing a lookup in the label switching table
- Incoming label is replaced with outgoing label and forwards the packets along the interface specified in the NHFLE
• Label Distribution Protocols
  – Before LSPs can be used, label switching table at each LSR must be populated with mapping from incoming label to outgoing label
  – So a label distribution protocol is required so that two LSRs learn each other’s MPLS capabilities and exchange label-mapping information
  – Sometimes called signaling protocols for MPLS networks
MPLS Architecture (cont.)

- Label Assignment
  - Decision to bind a particular label to a particular FEC is made by the downstream LSR.
  - Downstream LSR informs the upstream LSR of the binding
  - Thus, data traffic and control traffic flow in different directions
  - Two different modes of downstream label distribution: downstream on demand and unsolicited downstream
MPLS Architecture (cont.)

- Downstream on demand: LSR explicitly requests a neighbor for a label binding for a particular FEC.
- Unsolicited downstream: LSRs can distribute label bindings to its neighbors that have not requested them.
MPLS Architecture (cont.)

- Label Merging
  - Two or more LSPs may be merged into one
  - When an LSR has bound multiple incoming labels to a particular FEC, an LSR may have single outgoing label to all packets in the same FEC
  - Once packets are forwarded with the same outgoing label, the information that they came with different incoming label is lost.
  - Reduces the requirement on label space
  - May not be possible in some technologies
    - If ATM based MPLS network does label merging, then cells from different packets will be interleaved.
Label Edge Router - LER

- Resides at the edge of an MPLS network and assigns and removes the labels from the packets.
- Support multiple ports connected to dissimilar networks (such as frame relay, ATM, and Ethernet).
Label Switching Router - LSR

- Is a high speed router in the core on an MPLS network.
- ATM switches can be used as LSRs without changing their hardware. Label switching is equivalent to VP/VC switching.
Positions of LERs & LSRs

**LSP**

**LER**

**LSR**

**LSR**

**LER**

```
<table>
<thead>
<tr>
<th>IP Addr</th>
<th>Out Label</th>
<th>In Label</th>
<th>Out Label</th>
<th>In Label</th>
<th>Out Label</th>
<th>In Label</th>
<th>Next Hop</th>
</tr>
</thead>
<tbody>
<tr>
<td>192.4/16</td>
<td>5</td>
<td>5</td>
<td>9</td>
<td>9</td>
<td>2</td>
<td>2</td>
<td>192.4/16</td>
</tr>
<tr>
<td>Layer 2 Transport</td>
<td>Assign init label</td>
<td>Label Swapping</td>
<td>Label Swapping</td>
<td>Remove Label</td>
<td>Layer 2 Transport</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

“ROUTE AT EDGE, SWITCH IN CORE”
MPLS Architecture (cont.)

- Route Selection and Explicit routing
  - During label distribution, LSR needs to determine which is the next hop for the LSP.
  - Two approaches: hop-by-hop routing and explicit routing
  - Hop-by-hop approach relies on IP routing to set up LSPs
  - Explicit routing, a single LSR (usually ingress or egress of the LSP) specifies the entire route for the LSP. Could be strictly explicit routed or loosely explicit routed.
MPLS Architecture (cont.)

• Forwarding Equivalence Class
  - A group of IP packets which are forwarded in the same manner (same next hop, same path, same forwarding treatment, etc.)
  - Current IP based networks support one kind of FEC i.e. destination-based FEC. So packets destined for the same subnet belong to the same FEC
  - MPLS can easily support multiple FECs, since classification of packets is done in control path
MPLS Architecture (cont.)

- More types of FEC support means finer forwarding granularity
- MPLS supports the following common types of FECs
  - IP prefix: packets that match an IP destination prefix in routing table are considered as one FEC. This is a direct mapping from routing table to label-switching table. Advantage of such FECs is that the label distribution can be closely coupled with IP routing
  - Egress router: All packets that go out on a particular egress router can make a useful FEC (for supporting traffic engineering). LSPs can be set up based on information from BGP next hop in a BGP update msg or from router id in OSPF environment
MPLS Architecture (cont.)

- Application flow: each application flow can be put into a FEC. This is the finest granularity of packet forwarding. Provides end-to-end switching and allows maximum control of traffic flows.
MPLS Architecture (cont.)

- Label Stack encoding
  - MPLS works over many different link layer technologies. So exact encoding of a label depends on the type of layer-2 technology used.
  - For packet based technologies such as ethernet, MPLS label is inserted between layer-2 and layer-3 (shim header)
  - For ATM or FR which are inherently label switching, the label is mapped to VPI/VCI in ATM cell header. Thus label switching is performed with native header of link-layer protocol
MPLS Architecture (cont.)

PPP/HDLC header | MPLS header | IP packet | HDLC trailer

PPP frame

AAL5 PDU

Label stack | IP packet | AAL5 pad and trailer

48 bytes | 48 bytes | 48 bytes

ATM cell headers carry top label in VPI/VCI
MPLS Architecture (cont.)

- **MPLS header**
  - Label value: 20 bits
  - Experimental use: reserved for experimental use: can be used for drop priorities similar to diffserv

<table>
<thead>
<tr>
<th>Label</th>
<th>Exp</th>
<th>S</th>
<th>TTL</th>
</tr>
</thead>
<tbody>
<tr>
<td>(20)</td>
<td>(3)</td>
<td>(1)</td>
<td>(8)</td>
</tr>
</tbody>
</table>

Label stack format
MPLS Architecture (cont.)

- Bottom of stack: S bit is used to indicate the bottom of label stack. Bit is set to 1 for the last entry in the label stack
- TTL: for avoiding loops
Label-Switched Paths - LSPs

- A path is established before the data transmission starts.
- A path is a representation of a FEC.
Loop Detection

• Loops can affect performance severely in MPLS network: traffic remains in loop as long as the LSP exists

• When IP packet is labeled at the ingress, TTL field is copied into the label stack header. This TTL is decremented at each LSRs

• When the last label is popped, TTL value of the label stack is copied back to IP packet
Loop Detection

- MPLS packets in ATM network do not have such mechanism since ATM does not have TTL field; so loop detection is taken care of during LSP set up phase
- A path vector field is carried in the LDP (Label Distribution Protocol) msg
  - It contains a list of LSRs that the msg has traversed
  - Each LSR when transmitting LDP msg adds its ID to the msg
  - If LSR finds its ID already present in path vector, then a loop is detected
  - A hop count is also used in the msg. When this value reaches a configured max, a loop is detected
Label Distribution Protocols

• IETF initially considered only one label distribution protocol (LDP)
• Later on two different proposals were put forward
  – Constraint routing based LDP (CR-LDP)
  – RSVP with traffic engineering extension (RSVP-TE)
LDP

• Provides LSR discovery mechanism to enable LSR peers find each other

• Defines four classes of msgs:
  – DISCOVERY msg
  – ADJACENCY msgs
    • Deals with initialization, keepalive and shutdown of sessions
  – LABEL ADVERTISEMENT msgs
  – NOTIFICATION msgs
    • Provides advisory and error information
LDP (cont.)

- Runs over TCP to provide reliable delivery msg (except DISCOVERY msg)
- Designed to be easily extensible using msgs specified as collection of TLV (type, length, value) encoded objects
- The most commonly used msgs are
  - INITIALIZATION
  - KEEPALIVE
  - LABEL MAPPING
  - LABEL WITHDRAWAL
  - LABEL RELEASE
  - LABEL REQUEST
  - LABEL REQUEST ABORT
MPLS Operation Example
Label Distribution Modes

• Conservative label retention mode
  – Retains only those label-to-FEC mapping that it needs at the current time. Any other mapping are released
  • If LSR2 observes that LSR1 is not its next hop for a FEC that is being advertised by LSR1, it will throw away that label

• Liberal label retention mode
  – Retains all mappings that are advertised to it, even if some of them are not directly useful at the time of advertisement
  – Quicker response to changes in routing
Label Distribution Modes (cont.)

• Independent LSP control
  – LSR can advertise label mapping to its neighbors at any time it desires
  – Similar to IP datagram model
    • Each node independently decides how to forward a packet and relies on the underlying routing protocol to converge.

• Ordered LSP control
  – LSR will send a label mapping only for a FEC for which it has a label mapping for FEC next hop or for which the LSR is the egress
Label Distribution Using BGP

- MPLS defines a new address family in BGP which includes one or more labels (plain BGP supports address families such as IPv4 and IPv6)

<table>
<thead>
<tr>
<th>Length (1 byte)</th>
<th>Label (3 bytes)</th>
<th>Optional labels</th>
<th>Prefix (variable)</th>
</tr>
</thead>
</table>

(Label and address prefix encoding in BGP)
• Any BGP speaker that advertises a route may also advertise a label or label stack to be used for packets using that route
• The label is bound to the address prefix carried in the BGP msg.
• So the advertised label is used for packets using that route
• All normal BGP procedures such as withdrawal of routes can be used on these advertisements
Label Distribution Using BGP (cont.)

A sends

<table>
<thead>
<tr>
<th>y</th>
<th>x</th>
<th>payload</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>A sends</td>
</tr>
</tbody>
</table>

If A and B are not adjacent

If A and B are adjacent

Border gateway

Label x (via BGP)

Interior gateway
RSVP-TE

- Extends original RSVP to perform label distribution and support explicit routing
- Plain RSVP augmented by Label Object can establish MPLS forwarding state (LSP) only along the path computed by IP routing
- To steer the PATH msg along a particular explicit route, RSVP-TE introduces an Explicit Route Object (ERO).
• ERO contains explicit route that the PATH msg should take
• ERO could contain strict or loose explicit route
• Can be used for rerouting existing LSPs (to optimize resource utilization)
  – A replacement LSP is first set up
  – Then traffic is switched over to the new LSP
  – Finally the old LSP is torn down
RSVP-TE (cont.)

- During transition, old and new LSP may coexist and compete with each other for resources on network segment that they have in common.
- There may be race condition where new LSP may not get resources because old LSP is still holding resources.
- To resolve this problem, it is necessary to make sure that resource reservation is not counted twice: can be achieved by using SE-style reservation.
CR-LDP

• Plain LDP does not have ability to establish label forwarding state on all LSRs along an explicit route
• CR-LDP introduces a new object called Explicit Route (ER)
• Structure of this object and handling by LSRs is similar to ERO in RSVP-TE
• CR-LDP also has mechanism to reserve resources along a route: uses Traffic Parameter object for the purpose
CR-LDP (cont.)

- CR-LDP defines seven traffic parameters
  - Peak data rate (PDR)
  - Peak burst size (PBS)
  - Committed data rate (CDR)
  - Committed burst size (CBS)
  - Excess burst size (EBS)
  - Frequency
  - Weight

- PDR and PBS defines the token bucket params for maximum rate of traffic that can be sent down the LSP
CR-LDP (cont.)

- CDR and CBS similarly defines the average traffic for the LSP
- EBS is the amount by which bursts may exceed the committed burst size
- These params are used to allocate resources at an LSP as it is established or for policing
- Frequency is a very rough indication of the time interval over which an LSP is expected to provide available bw of CDR
- Weight is used to determine how extra bw above the total CDR of all LSPs should be divided among LSPs that share a common bottleneck link
Guaranteed Bandwidth LSPs

- RSVP-TE can be used to establish LSPs with reserved resources.
- Before sending PATH msg, sender node consults the link-state database and selects a path between sender and receiver that satisfies the constraint (e.g. enough bw along the path).
- ERO can then be used to establish LSP along such a path.
Fast Reroute with MPLS

- MPLS can be used to handle link failures by rerouting packets
- When a link fails, LSR attached to the failed link can use the label stacking capability of MPLS to “nest” (i.e. tunnel) all the LSPs that used to go over the failed link into the “protection” LSP
MPLS in DiffServ network

- Few issues to address to support MPLS in diffserv network
  - Provide some way to ensure that DSCP-marked packets receive appropriate QoS treatment at each LSR in the network
  - However, DSCP is carried in the IP header, and LSRs do not examine that header while forwarding packets
  - So there needs to be some way to determine appropriate PHB from label header
  - There are two ways to achieve this
MPLS in DiffServ network (cont.)

- Use the Exp field in MPLS header to encode DSCP
  - Since Exp field is 3 bits, only 8 PHBs can be supported in this scheme
    - Known as E-LSP
- If more than 8 PHBs are required, or in networks where shim header is not present (e.g. ATM) label has to be used in some way to convey the PHB
  - Use L-LSP (Label-LSP), where PHB is inferred from the label
MPLS in DiffServ network (cont.)

- Label distribution protocol needs to be enhanced to achieve this
- In basic LDP, advertisement binds labels and FECs such as address prefixes
- This has to be enhanced to such that a label can be bound to both FEC and PHB
  - Labels are bound to \( \langle \text{prefix}, \text{PHB} \rangle \) pairs.
- But to implement AF, it requires that packets belonging to a single microflow must not be misordered if they differ only in drop precedence
MPLS in DiffServ network (cont.)

Thus, packets of the same AF class go into a common queue.

Some LSRs implement a queue per LSP.

Two LSPs to a common destination may take different paths.

Thus, to meet ordering constraints for AF traffic, it will be required to send packets of the same AF class (AF11, AF12, AF13) on a common LSP.

For networks that supports shim header, Exp bits will be used to carry drop precedence.

For ATM that does not have shim header, the CLP bit in the cell header will be used for drop priority (only two drop priorities can be supported).
MPLS in DiffServ network (cont.)

- Label distribution protocol now has to advertise binding label to FEC and PHB
  - E.g. for LDP, msgs that request and advertise bindings of prefixes to labels are now extended to include a PHB. This allows labels to be bound to \(<\text{prefix}, \text{PHB}>\) pairs.

- To summarize
  - E-LSP does not require additional signaling, but not possible in network that cannot support shim header
    - Can only support up to eight PHBs
    - PHBs determined from Exp bits
  - In L-LSP PHB binding is signaled during LSP setup
    - Suitable for both ATM and non-ATM networks
    - PHBs determined from label plus Exp/CLP bits
References

- “Multiprotocol Label Switching Architecture” – RFC 3031
- “MPLS Technology and Applications” – Bruce Davie, Yakov Rekhter, Morgan Kaufmann Publishers
- “RSVP-TE: Extensions to RSVP for LSP Tunnels” – RFC 3209
- “LDP Specification “ – RFC 3036