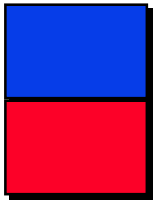
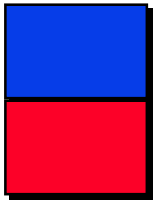


Tag Switching Architecture

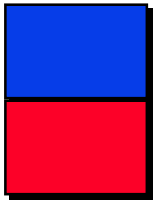
Cisco Systems, Inc.



Design Objectives

- **Address a broad range of open issues in Network Layer routing:**
  - **richer functionality**
    - **» destination-based forwarding is not enough**
  - **scalability**
  - **better performance**
  - **integration of cell-switching (ATM) and frame-switching technologies**
  - **ability to evolve routing system gracefully and in a timely fashion to meet new and emerging requirements**

CISCO SYSTEMS



Tag Switching

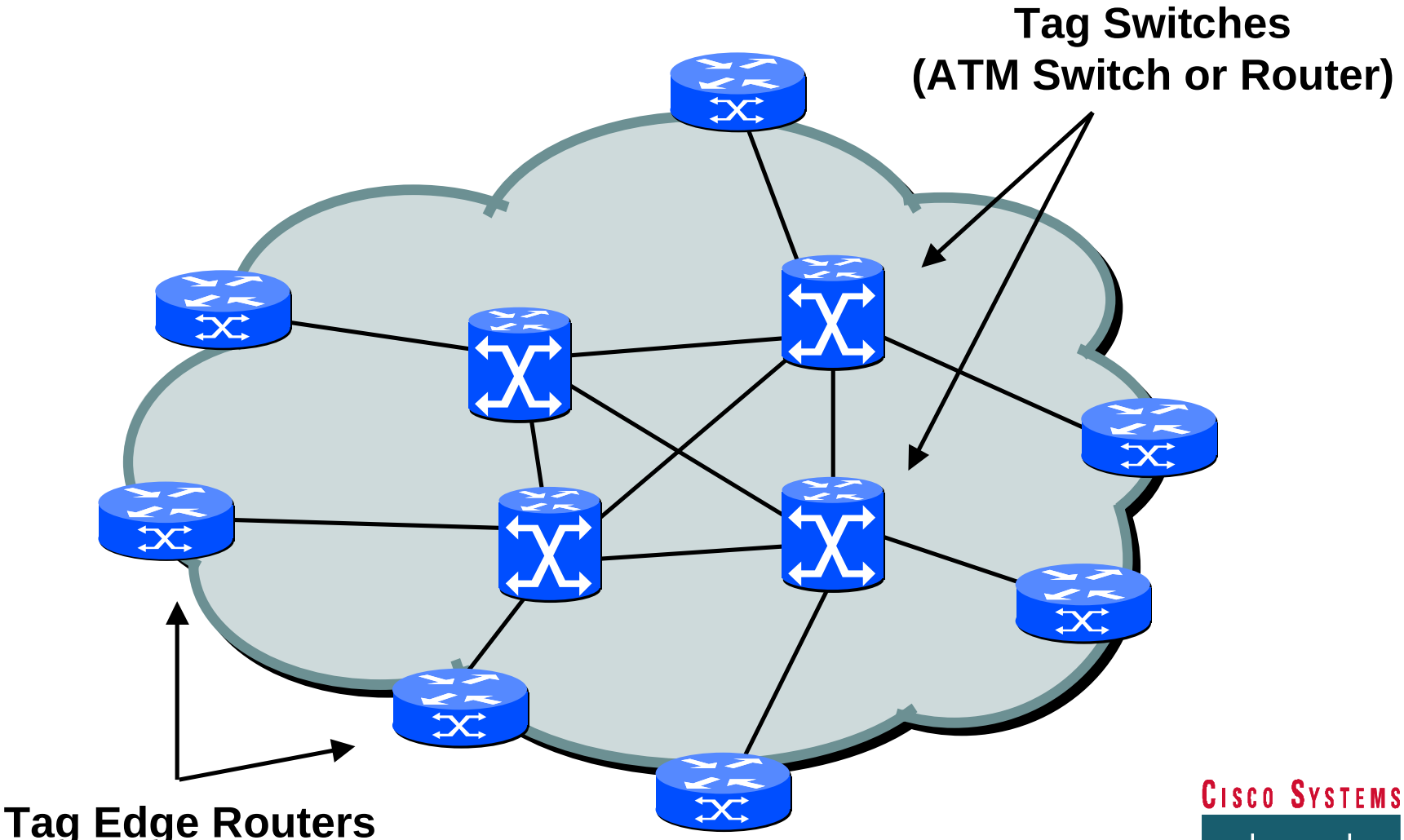
- **Blend of Network Layer routing with the label (tag) swapping forwarding paradigm**
  - **simple forwarding algorithm -> improved forwarding performance**
  - **wide range of forwarding granularities per tag + Network Layer routing -> wide variety of routing functions + good scaling properties**
  - **segregation of control from forwarding -> ability to evolve routing functionality while keeping forwarding paradigm intact -> support for graceful evolution of routing**

CISCO SYSTEMS

Unit Components

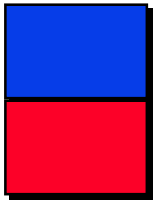
- **Tag Edge Routers**
  - **tag previously untagged packets**
    - **» at the beginning of a tag switched path**
  - **strip tags from tagged packets**
    - **» at the end of a tag switched path**
- **Tag Switches**
  - **forward tagged packets based on the information carried by tags**

Tag Switching Devices



Functional Components

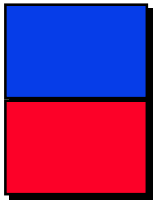
- **Forwarding component:**
  - **uses tag information carried in a packet and tag binding information maintained by a tag switch to forward the packet**
- **Control component:**
  - **responsible for maintaining correct tag binding information among tag switches**



Forwarding Component

- **Tag Information Base (TIB)**
  - **each entry consists of:**
    - **incoming tag**
    - **one or more sub-entries:**
      - **(outgoing tag, outgoing interface, outgoing MAC address)**
  - **TIB is indexed by incoming tag**
  - **TIB could be either per box, or per interface**

Cisco Systems

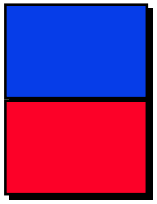


Forwarding Component (cont.)

- **Forwarding algorithm:**
  - **extract tag from a packet**
  - **find an entry in the TIB with the incoming tag equal to the tag in the packet**
  - **replace the tag in the packet with the outgoing tag(s) (from the found entry)**
  - **send the packet on the outgoing interface(s) (from the found entry)**

<u>Observation:</u> forwarding algorithm is Network Layer independent

CISCO SYSTEMS

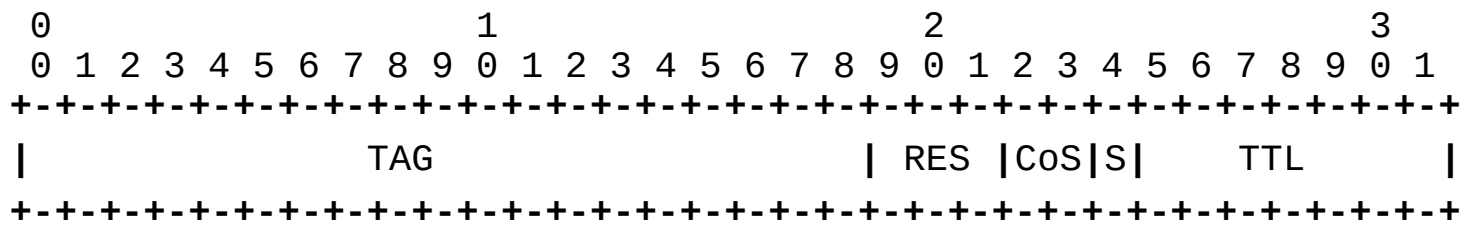


Forwarding Component (cont.)

- **Carrying tag information:**
  - **as part of the Network Layer header**
    - » **Flow Label field in IPv6**
      - **requires changes to the semantics of the Flow Label field**
  - **as part of the MAC header**
    - » **VCI/VPI in ATM**
    - » **DLCI in Frame Relay**
  - **via a “shim” between the MAC and the Network Layer header**

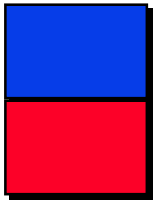
<u>Observation:</u> tag information could be carried over any media

“Shim” format



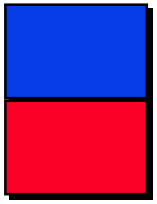
S = Bottom of stack
TTL = Time to live
CoS = Class of Service
RES = Reserved

- Can be used over Ethernet, 802.3, or PPP links
- Will require 2 new Ethertypes/PPP PIDs
  - one for unicast, one for multicast
- 4 octets per tag level



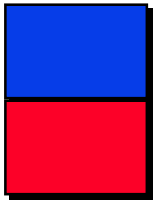
Forwarding Component (cont.)

- **Wide range of forwarding granularities:**
  - **tag is bound to a group of destinations (address prefix)**
    - **» tag is bound to a collection of address prefixes**
  - **hierarchy of tags**
    - **» reflects underlying routing hierarchy**
  - **tag is bound to a multicast tree**
  - **tag is bound to an application flow (e.g., RSVP flow)**
- **Enables scaleable routing**
- **Enables functionally rich routing**



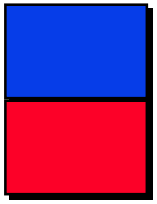
Forwarding Component - Summary

- Based on the exact match algorithm
- Wide range of forwarding granularities
- Could be implemented with any MAC/Link Layer technology
- Network Layer independent - multiprotocol



Control Component

- **Organized as a collection of modules:**
  - **each module is designed to support a particular routing function:**
    - **destination-based routing**
    - **hierarchy of routing knowledge**
    - **resource reservations**
    - **explicit routes**
    - **multicast**
  - **new modules could be added to support new routing functions without impacting the Forwarding Component**



Control Component (cont.)

- **Responsible for binding between tags and routes**
  - **create tag binding**
    - **» allocate a tag**
    - **» bind a tag to a route**
  - **distribute tag binding information among tag switches:**
    - **» option 1: piggyback on existing protocols**
      - **BGP via the “tag” attribute**
      - **RSVP via the RSVP_TAG object**
      - **PIM**
    - **» option 2: use Tag Distribution Protocol (TDP)**

CISCO SYSTEMS

Creating Tag Binding

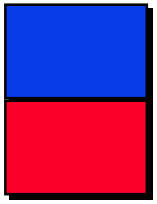
- Driven mostly by control traffic:
  - unicast routing updates
  - PIM Join/Prune
  - RSVP Path/Resv
- Advantages:
  - minimizes additional control traffic
  - independent of traffic pattern/profile
  - minimizes impact on forwarding performance
  - minimizes additional complexity
  - minimizes the amount of “guessing”

Distributing/Maintaining Tag Binding

- Consistent with the distribution of associated routing information:
  - tags for destination-based routing
    - » incremental updates with explicit acknowledgement
      - use reliable transport protocol
    - » no periodic refresh per tag
  - tags for multicast
    - » periodic updates/refresh per tag
    - » timeout

Destination-based Routing Module

- Three possible tag maintenance schemes for unicast:
  - downstream
    - incoming tag binding local, outgoing tag binding remote
  - downstream on demand
    - incoming tag binding local, outgoing tag binding remote
  - upstream
    - incoming tag binding remote, outgoing tag binding local



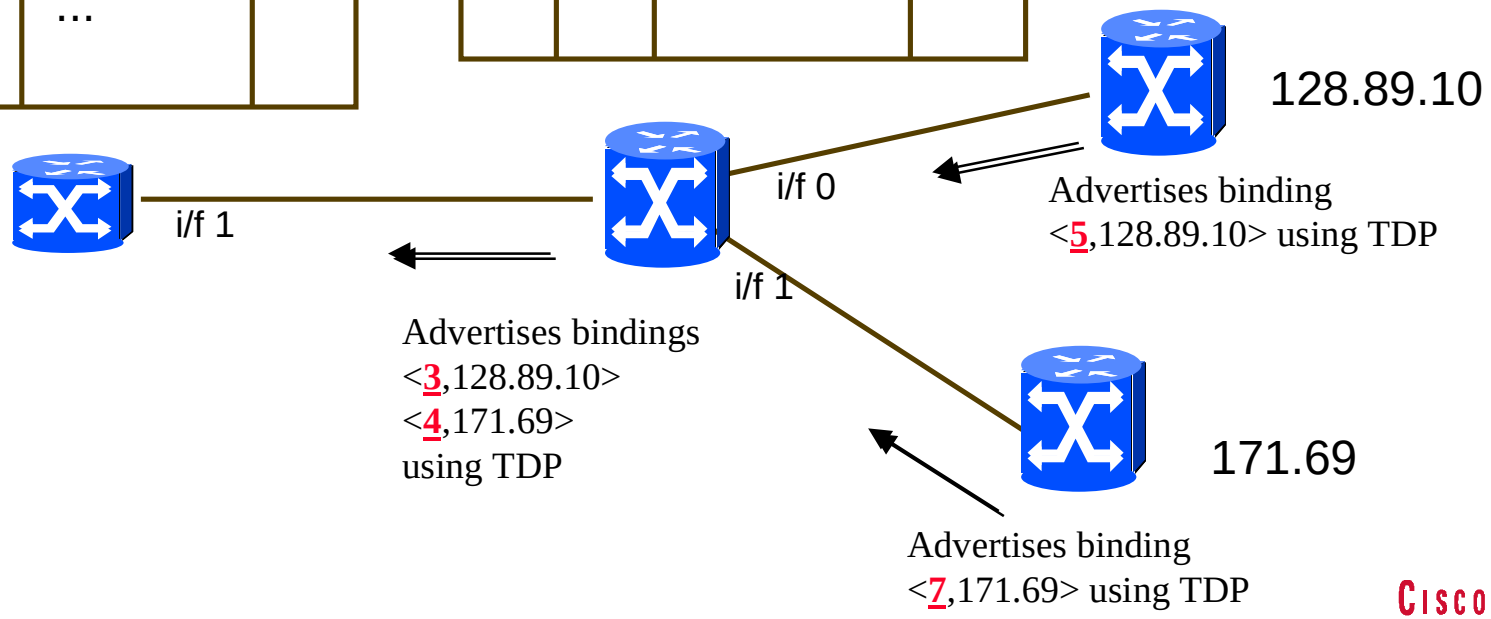
Downstream Scheme

- For each route in its Routing Information Base (RIB) a switch:
  - allocates a tag
  - creates an entry in its TIB with the incoming tag set to the allocated tag
  - advertises binding between the incoming tag and the route to all of the adjacent switches
- When a switch receives tag binding information for a route, if the information was received from the next hop for that route, the switch places the tag into the outgoing tag of the TIB entry associated with the route

Downstream scheme - example

inc tag	out tag	address prefix	interface
x	3	128.89.10	1
x	4	171.69	1
		…	

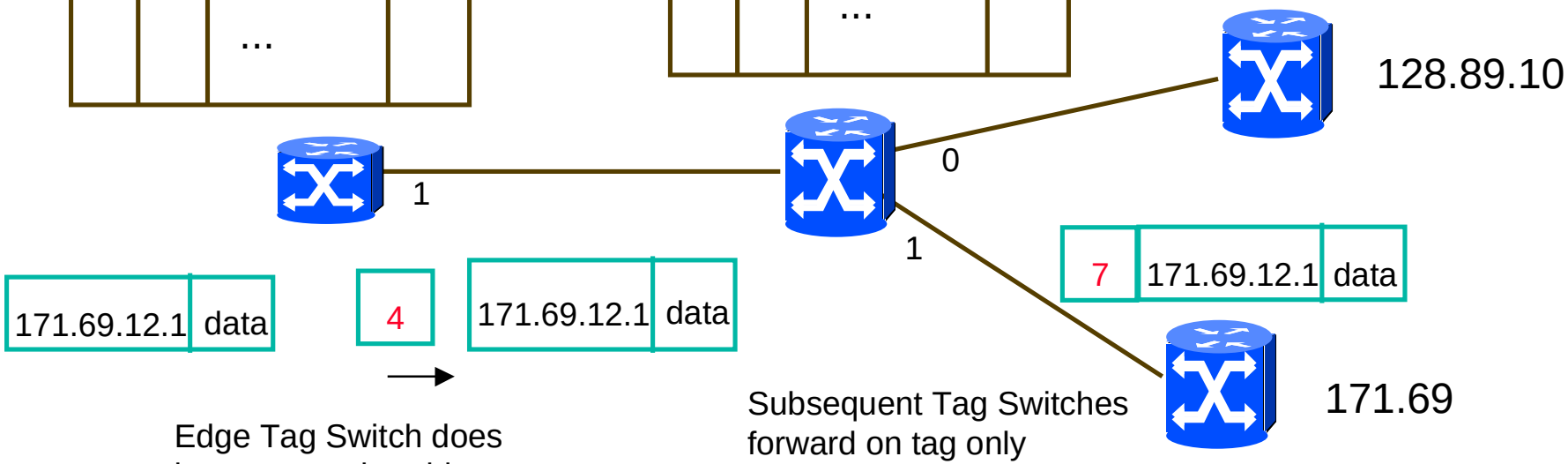
inc tag	out tag	address prefix	interface
3	5	128.89.10	0
4	7	171.69	1
		…	



Downstream scheme - example (cont.)

inc tag	out tag	address prefix	interface
x	3	128.89.10	1
x	4	171.69	1
		…	

inc tag	out tag	address prefix	interface
3	5	128.89.10	0
4	7	171.69	1
		…	

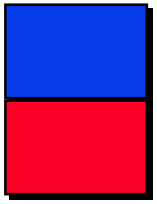


Downstream on Demand Scheme

- For each route in its Routing Information Base (RIB) a switch requests (via TDP) the next hop (associated with the route) to provide the switch with the tag binding information
- When the next hop receives the request, it:
  - allocates a tag
  - creates an entry in its TIB with the incoming tag set to the allocated tag
  - returns the binding to the requester
- When the requester receives the tag binding information for a route from the next hop for that route, the requester places the tag into the outgoing tag of the TIB entry associated with the route

Destination-based Routing Module (cont.)

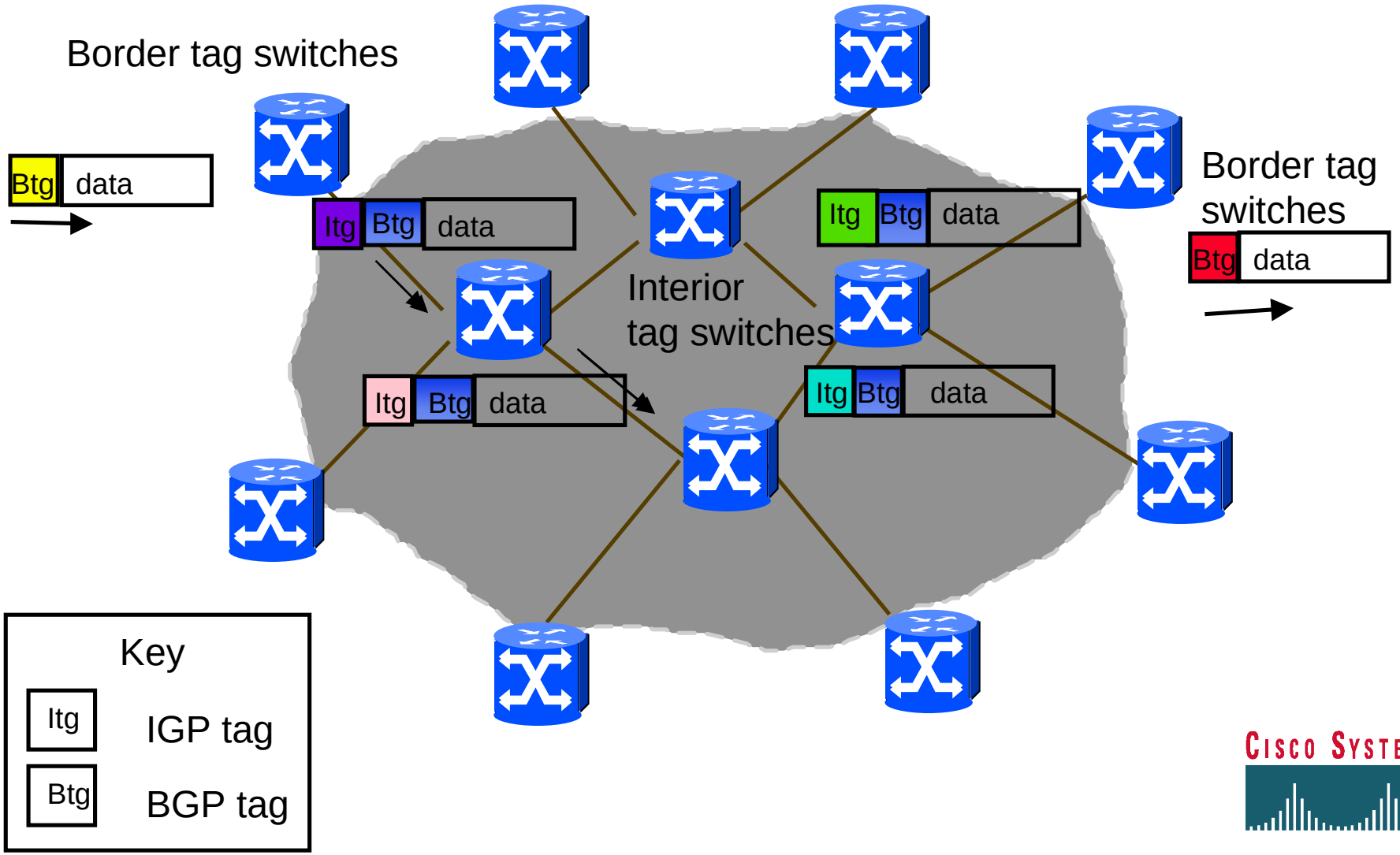
- **Scaling properties:**
  - **total number of incoming tags is no greater than the number of routes in the RIB**
    - **could be less if a tag is associated with a group of routes**



Hierarchy of Routing Knowledge Module

- How to keep interior routers away from maintaining exterior routing information ?
  - allow a packet to carry a “stack” of tags
    - » between domains use tags associated with exterior routes
      - BGP tag
    - » within a domain use tags associated with interior routes to BGP border routers (BGP NEXT_HOP) of the domain
      - IGP tag + BGP tag
    - » IGP tag could be associated with a group of BGP routes (address prefixes)
      - all reachable through the same BGP border router (same BGP NEXT_HOP)
      - tag per BGP border router (per BGP NEXT_HOP)
    - » IGP tags could be used without BGP tags

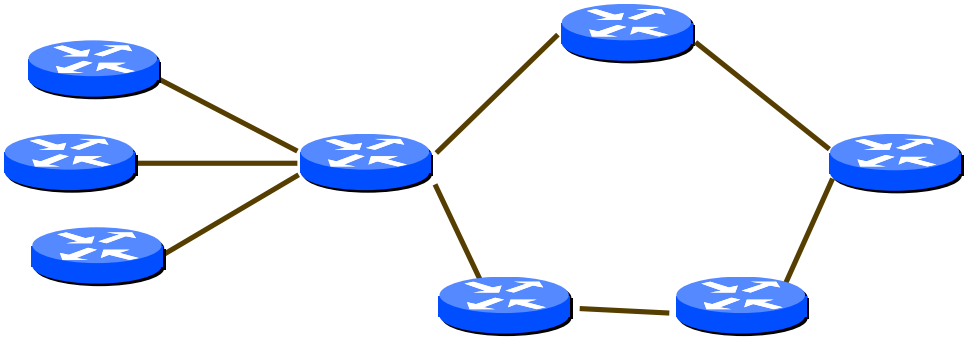
Hierarchy of Routing Knowledge (example)



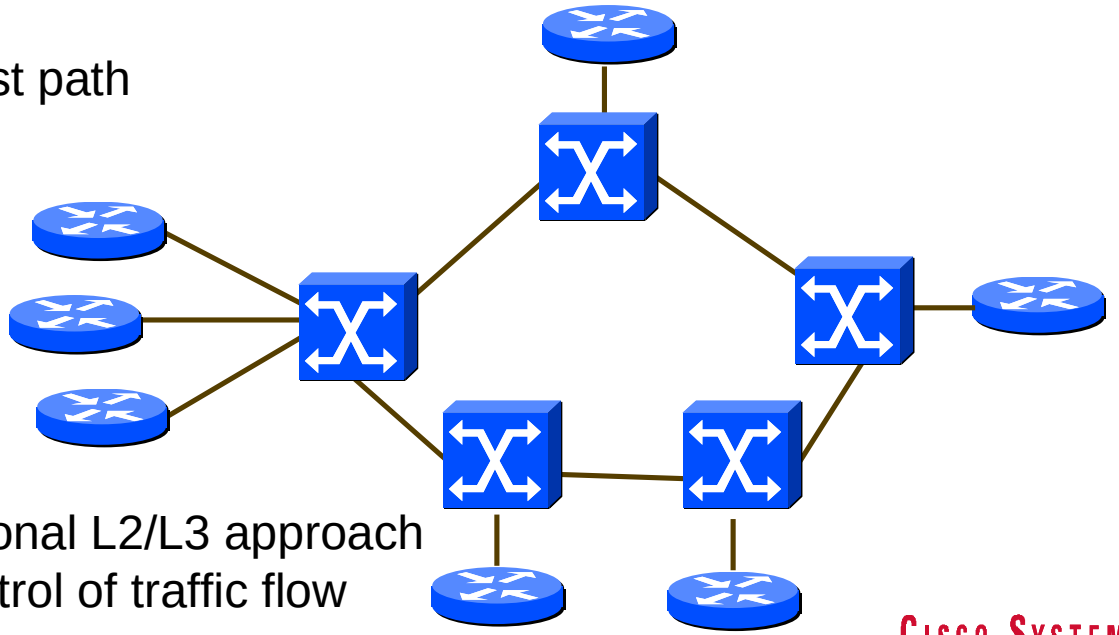
Explicit Route Module

- Overrides the destination-based routing paths
- Requires the ability to install tag bindings that are independent from the tags installed via the destination-based routing
  - may be coupled with resource reservations
- Possible applications:
  - allows finer control over traffic distribution over multiple links (traffic engineering)
  - support for forwarding with QoS-based routing

Traffic engineering example

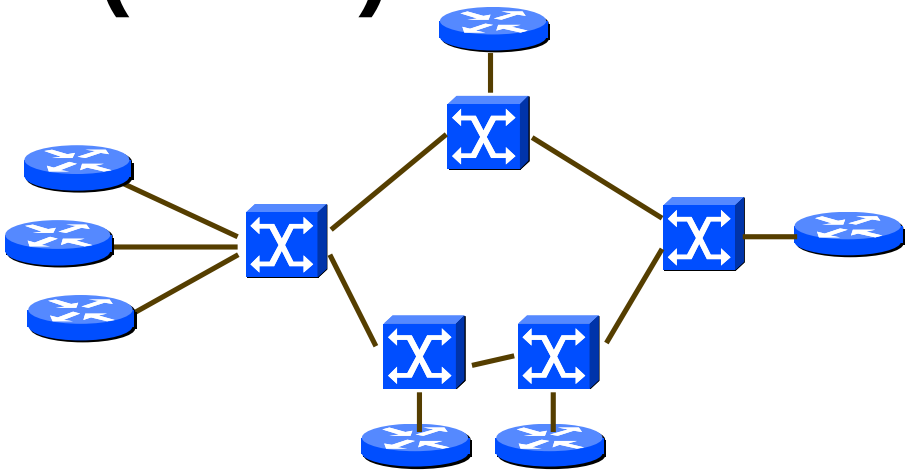


Pure routed network
All traffic follows L3 shortest path

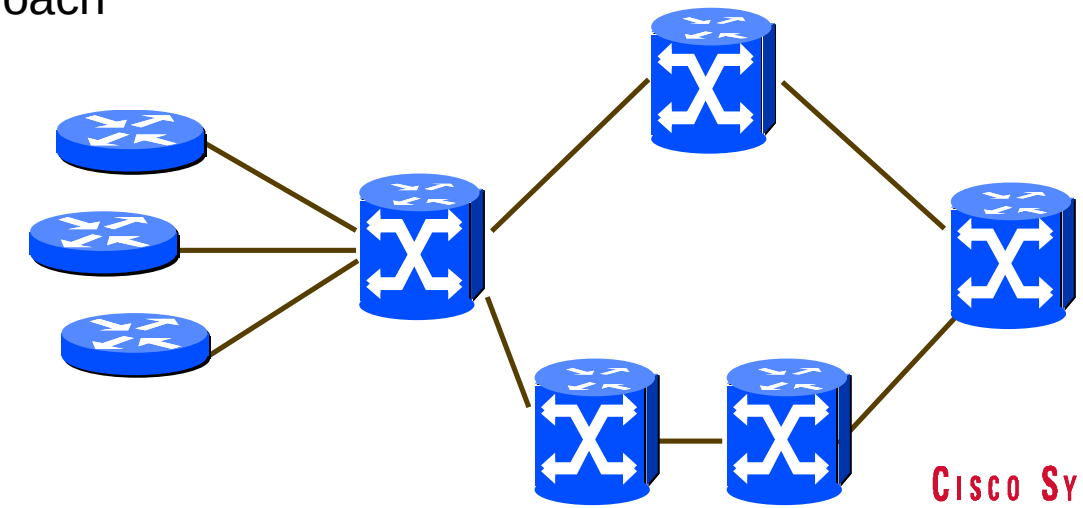


Conventional L2/L3 approach
Finer control of traffic flow

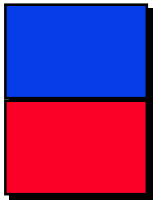
Traffic engineering example (cont.)



Conventional L2/L3 approach



Tag switching approach



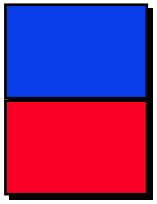
Multicast Module

- Multicast uses spanning trees for distribution of multicast data
- Binding a tag to a multicast tree
  - when a tag switch receives a packet the tag must identify both a particular multicast group and the previous (upstream) tag switch that sent a packet
- Utilizing Data Link layer multicast capabilities
  - an upstream tag switch should use the same tag when forwarding to all the downstream tag switches on a common LAN

Cisco Systems

Multicast Module (cont.)

- **Design requirements:**
  - **tag space used for multicast is partitioned into non-overlapping regions among all tag switches on a common Data Link subnetwork**
  - **multicast tags are associated with interfaces**
  - **tag switches that belong to a common multicast tree and are on a common Data Link subnetwork agree on the tag switch that is responsible for allocating and binding a tag to a tree**
  - **the tag switch that allocates and binds the tag is responsible for distributing tag binding information to other tag switches on a common Data Link subnetwork**



Multicast Tag Space Partition

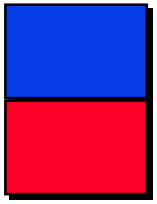
- Each tag switch claims a region of tag space, and announces the region to other tag switches on a common subnetwork
- Use IP addresses of the contending tag switches for conflict resolution
- Once the tag space is partitioned, the switched may create bindings between tags and multicast trees

Multicast - Upstream Binding

- Upstream tag switch creates tag binding and advertises it downstream
- Advertisement of binding:
  - piggyback on data traffic
    - » downstream routers would use tag fault
      - merging data and control functions - not a good idea
  - use control messages
    - » creates race conditions - routing updates and distribution of tag binding information flow in opposite directions
- Uneven distribution of binded tags
- Upstream neighbor change requires tag rebinding

Multicast - Downstream Binding

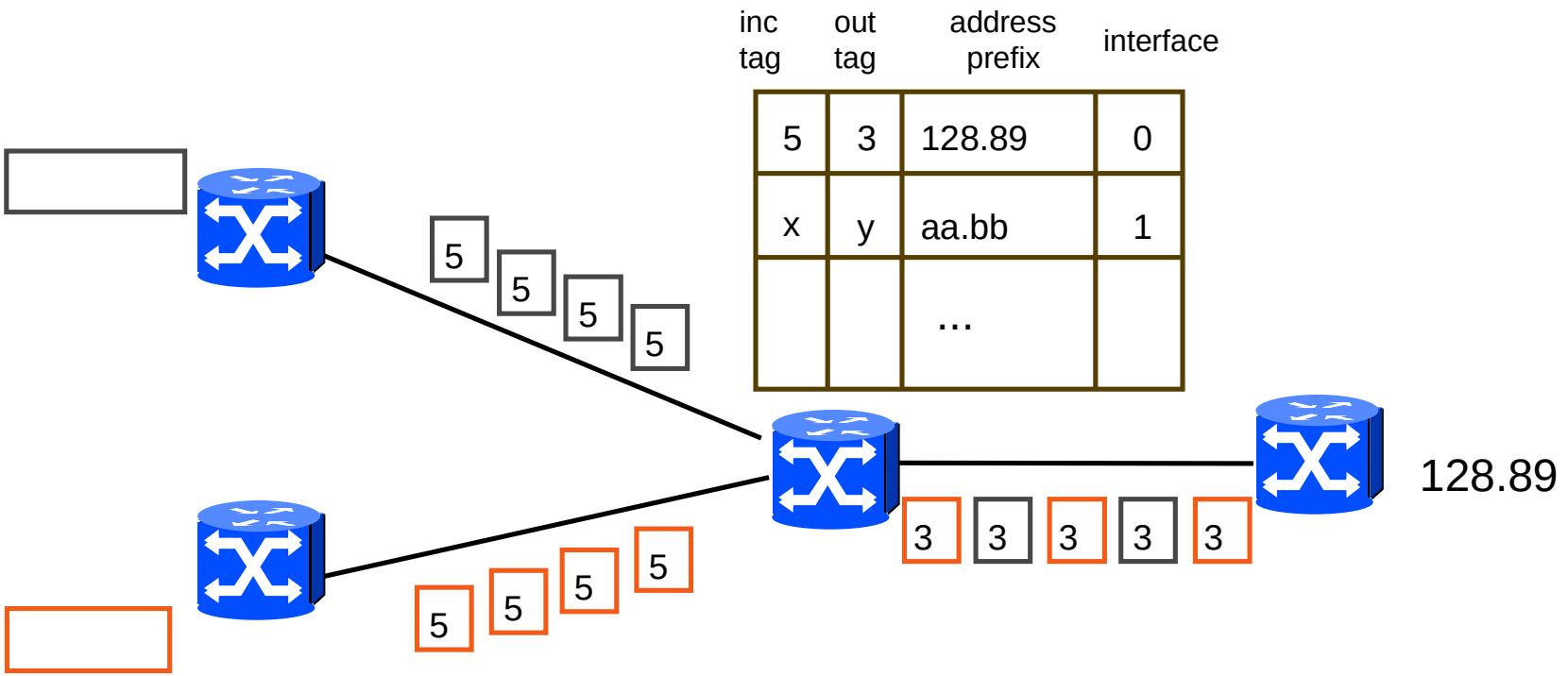
- One of the downstream tag switches creates tag binding and advertises it to other tag switches
  - requres choosing the tag switch from among the downstream tag switches on a subnetwork
- Consistent with the distribution of multicast routing information
  - enables to piggyback tag binding information on top of existing multicast routing protocols (PIM)
- Randomizes tag binding
- Upstream neighbor change does not require tag rebinding
- Better choice than the upstream binding



Tag Switching with ATM

- Common forwarding paradigm - label swapping
- Use ATM user plane
  - use VCI for tags
    - » use of VPI is possible for two levels of tags
- Replace ATM control plane defined by the ATM Forum with the tag switching control component:
  - Network Layer routing protocols (e.g., OSPF, BGP, PIM) + Tag Distribution Protocol (TDP)

Cell interleave issue - example

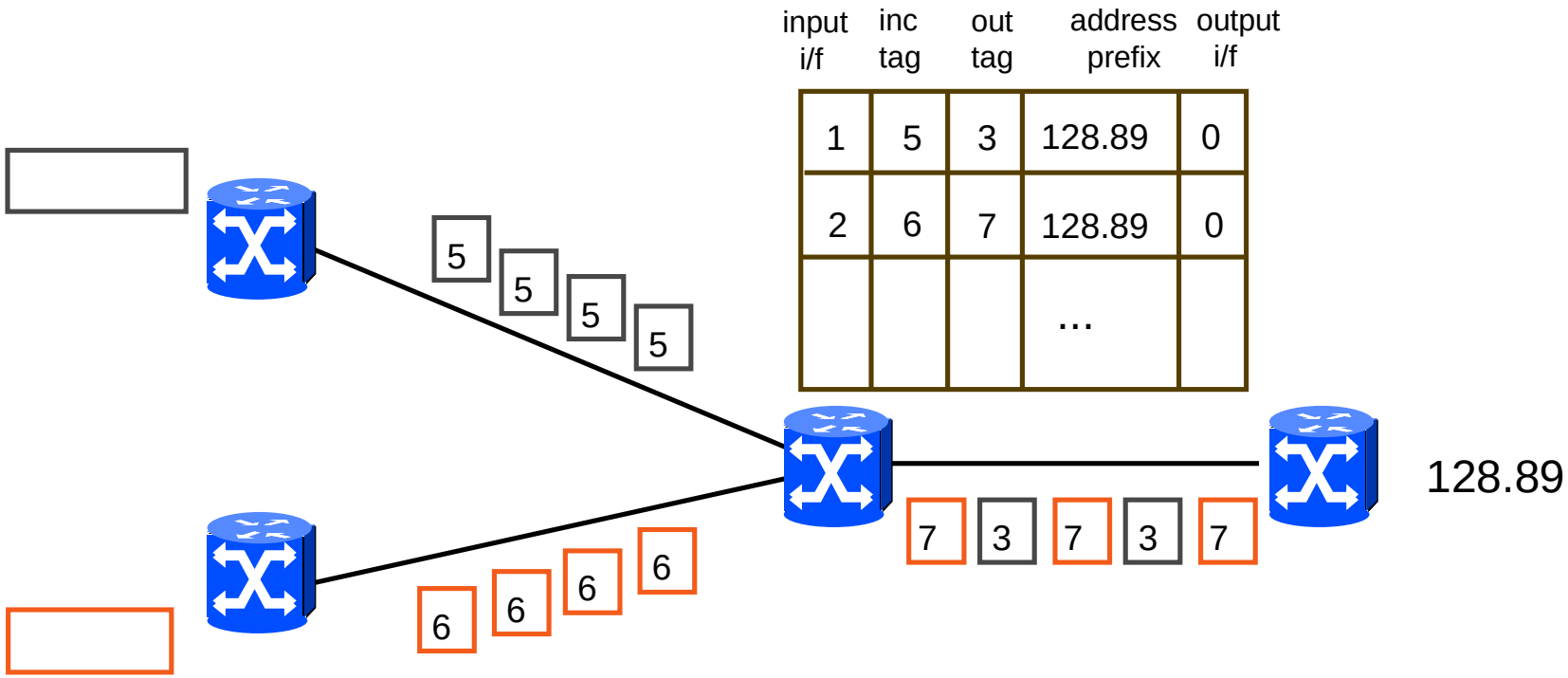


ATM switch interleaves cells of different packets onto same tag

Handling cell interleave

- **Option 1: maintain multiple tags per route**
  - **use IGP tags to improve scaling properties**
  - **use tags to egress Tag Edge Routers**
- **Option 2: use VPI for tags, VCI for demultiplexing among multiple sources**
  - **one tag per route**
  - **scalability limited by the size of the VPI space**
- **Option 3: use VC-merge capabilities**
  - **one tag per route**
  - **scalability limited by the size of the VCI space**

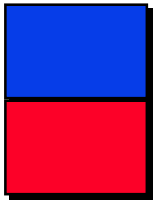
Handling cell interleave with multiple tags - example



- Multiple tags per prefix must be assigned
- One tag per (ingress, egress) router pair
- Can be reduced further with `VC-merge'

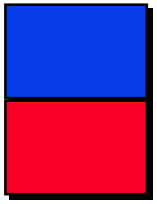
Tag Switching with ATM (cont.)

- **Simplifies integration of ATM switches and routers**
  - **ATM switch with tag switching capabilities appears as a router to an adjacent router**
    - **common routing and addressing plan for routers and ATM switches**
- **Enables better routing**
  - **exposes physical topology to the Network Layer routing**
- **Doesn't preclude the ability to support the ATM Forum defined control plane on the same switch**
  - **use "ships in the night" approach**



Tag Switching - Summary

- Provides functionally rich routing system
- Provides scaleable routing system
- Provides high forwarding performance
- Leverages widely deployed technology
- Multiprotocol solution
- Enables graceful evolution of routing system to meet new and emerging requirements



Suggested reading

- draft-rekhter-tagsw-arch-00.txt
- draft-doolan-tdp-spec-00.txt
- draft-davie-tag-switching-atm-01.txt
- draft-rosen-tag-stack-01.txt
- draft-baker-flow-label-00.txt
- draft-farinacci-multicast-tagswitching-00.txt
- draft-farinacci-multicast-tag-partition-00.txt
- draft-baker-tags-rsvp-00.txt
- more Internet Drafts are coming...
- mailing list: mpls@external.cisco.com