Towards Service Differentiation on the Internet

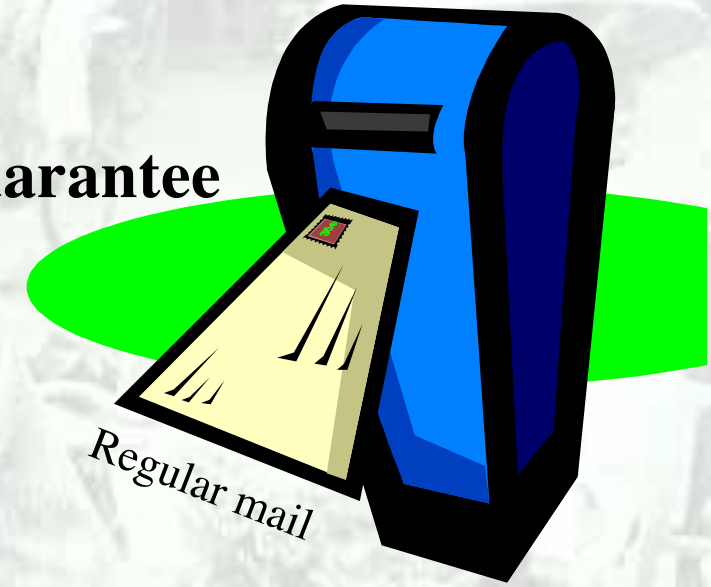
from

"New Internet and Networking Technologies for Grids and High-Performance Computing",
tutorial given at IEEE HOTI 2006, Stanford, California
August 25th, 2006

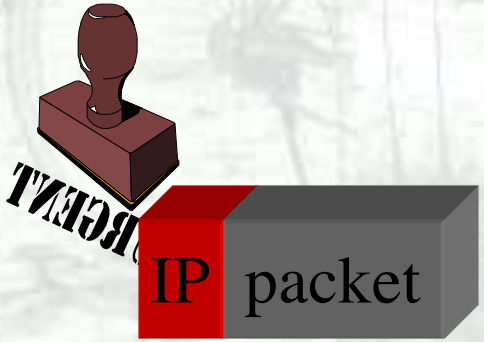
C. Pham
University of Pau, France
LIUPPA laboratory

Revisiting the *same service for all* paradigm

No delivery guarantee



Enhancing the best-effort service

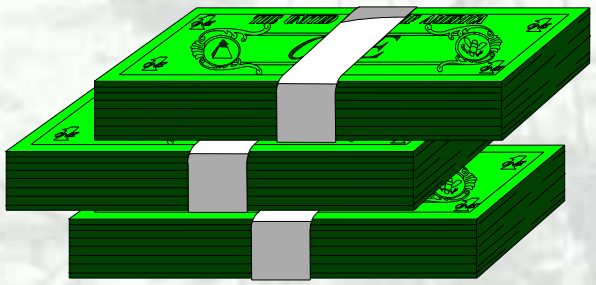


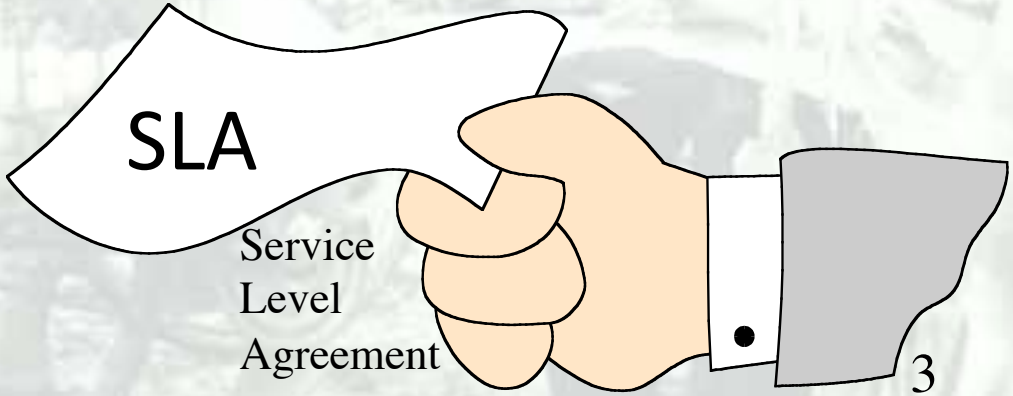
Introduce
Service Differentiation

Service Differentiation

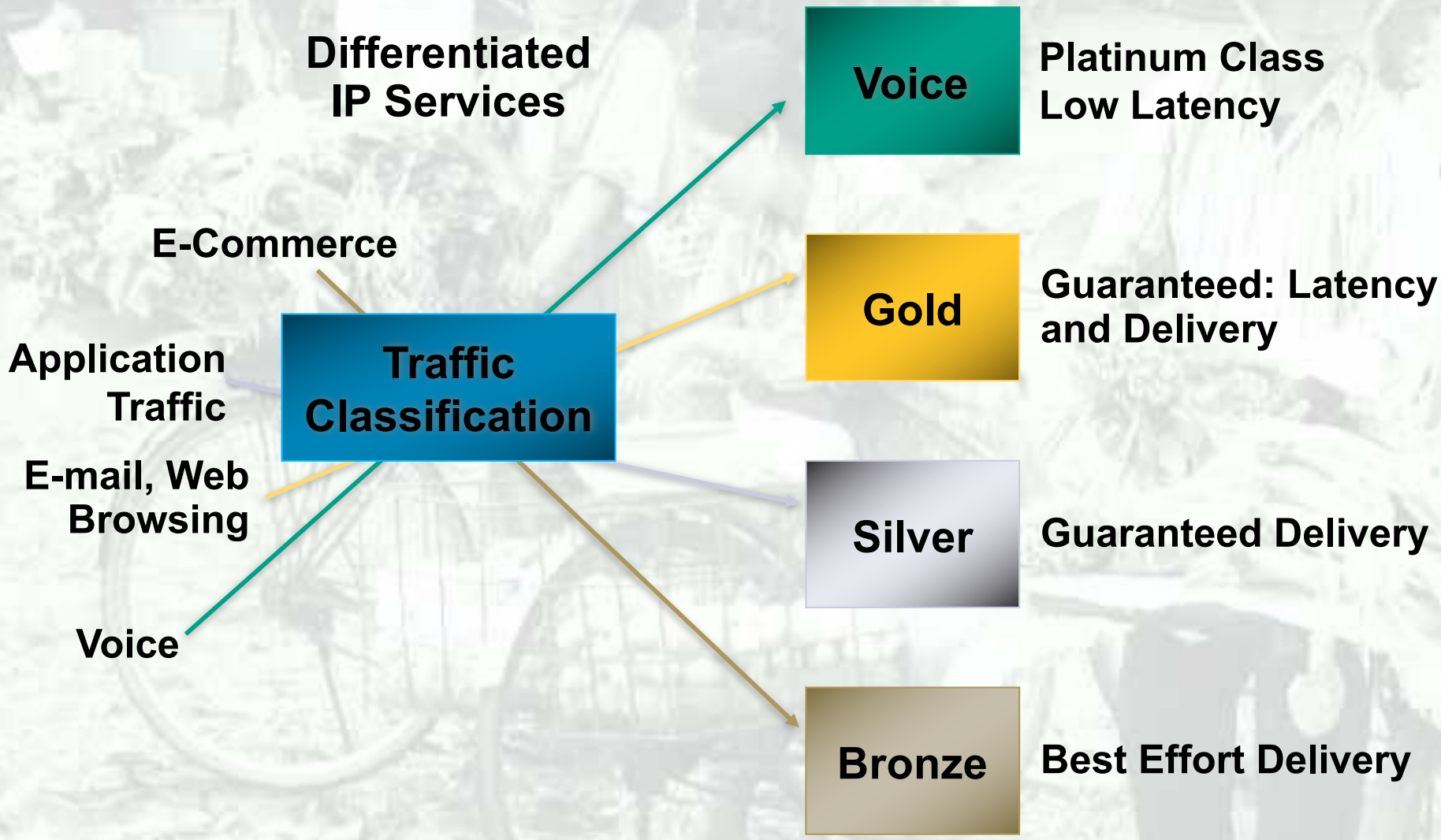
The real question is to choose which packets shall be dropped. The first definition of differential service is something like "not mine."
-- Christian Huitema

- Differentiated services provide a way to specify the relative priority of packets
- Some data is more important than other
- People who pay for better service get it!





Divide traffic into classes



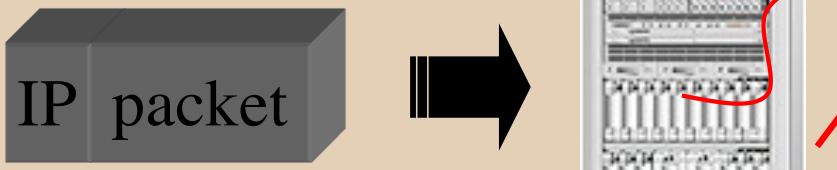
Design Goals/Challenges

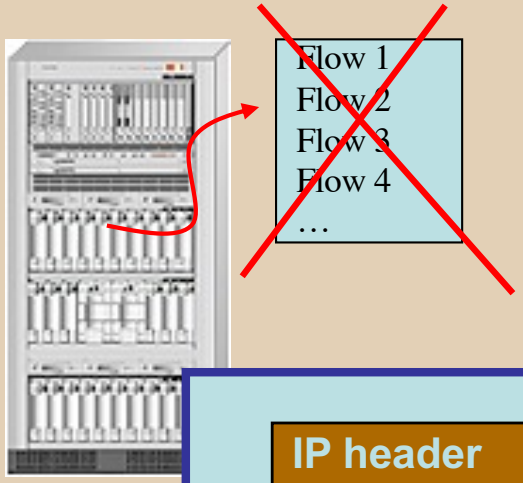
- Ability to charge differently for different services
- No per flow state or per flow signaling
- All policy decisions made at network boundaries
  - Boundary routers implement policy decisions by tagging packets with appropriate priority tag
- Traffic policing at network boundaries
- Deploy incrementally: build simple system at first, expand if needed in future

IP implementation: DiffServ

RFC 2475

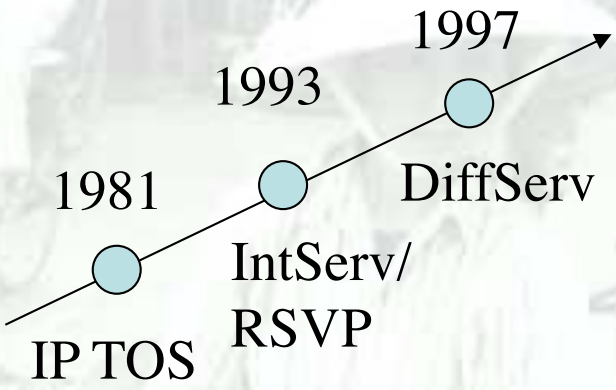
No per flow state in the core

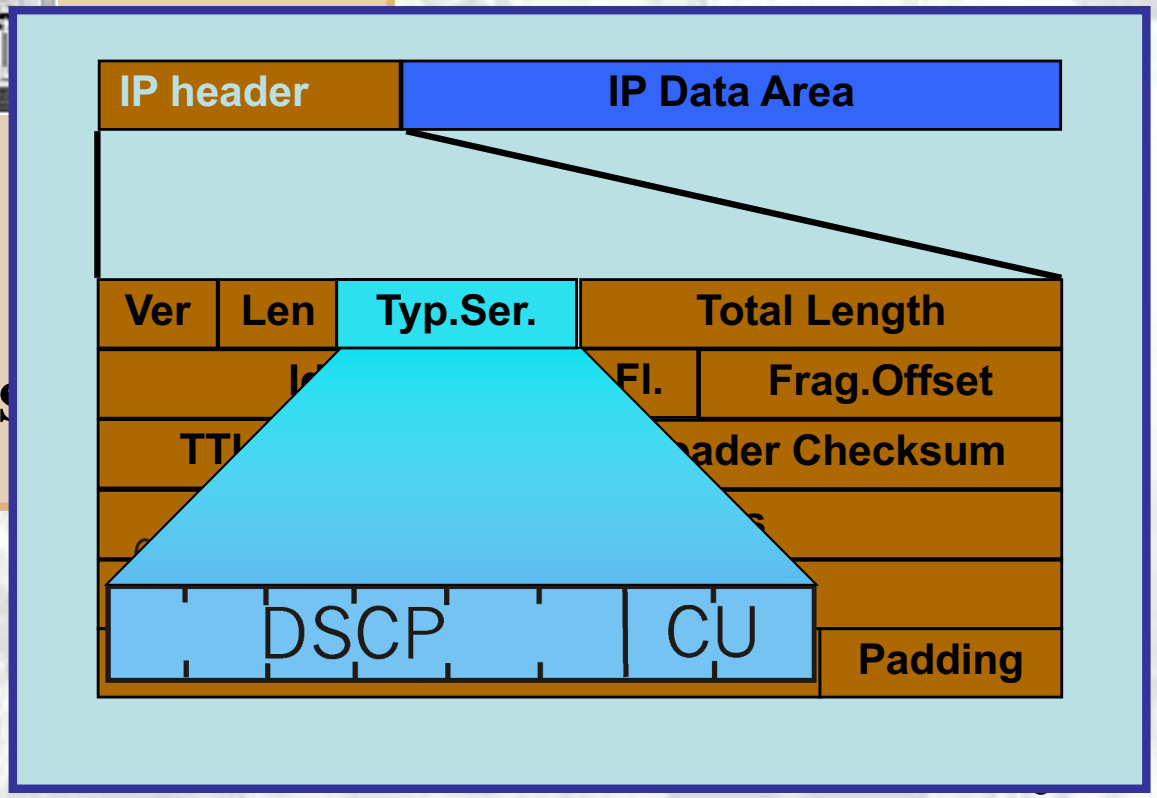




10Gbps=2.4Mpps
with 512-byte packets

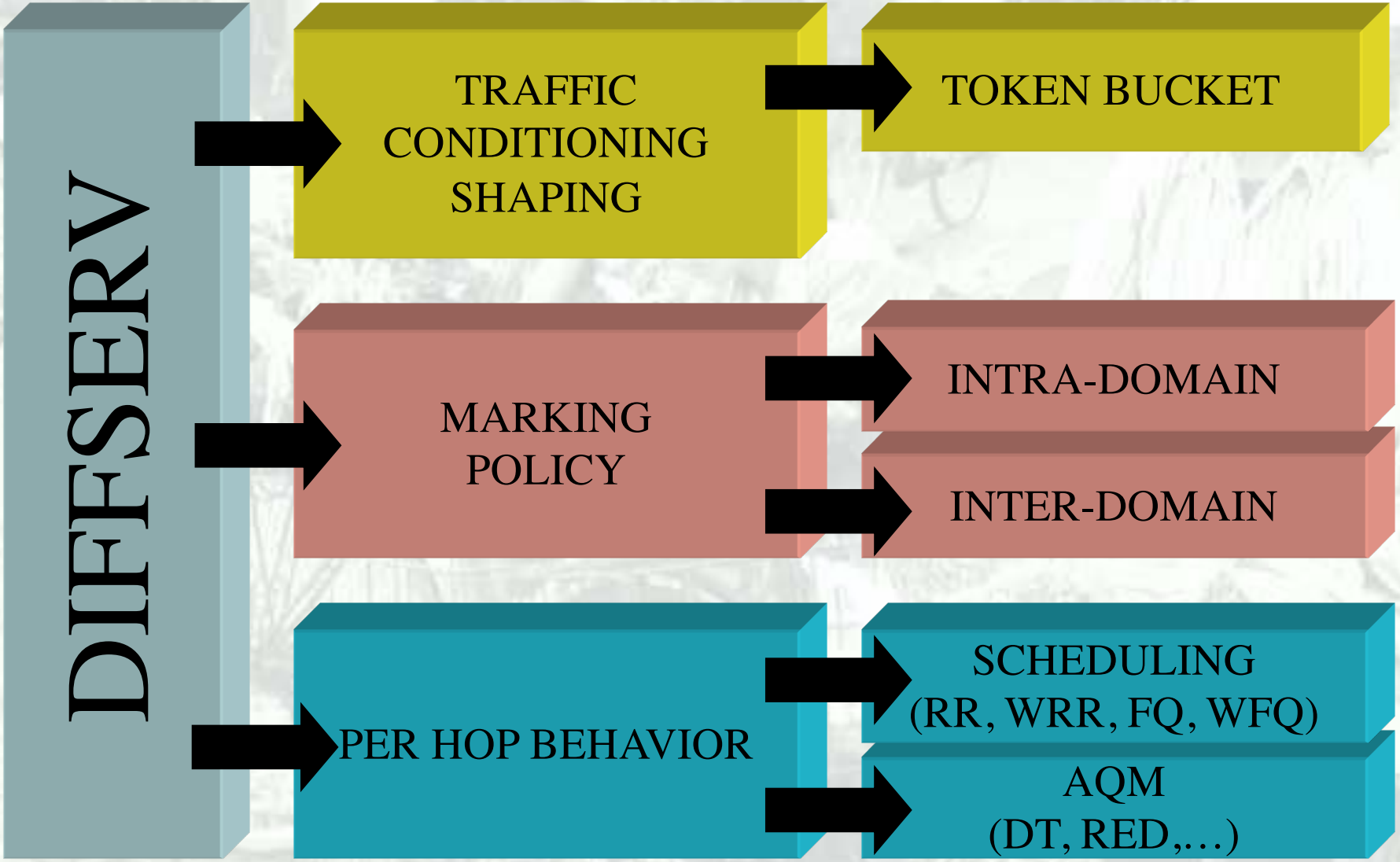
Stateful approaches
scalable
at gigabit rates





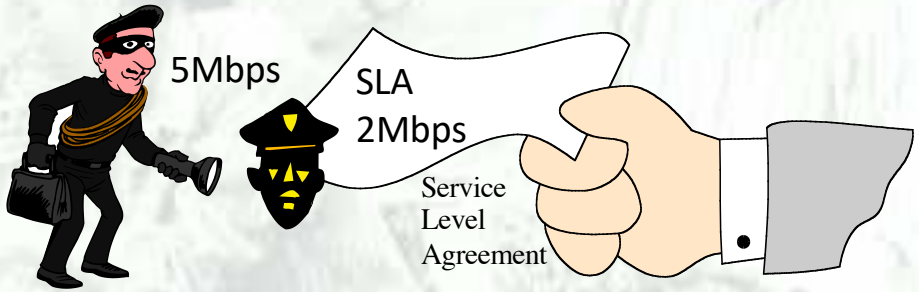
6 bits used for Differentiated Service Code Point (DSCP) and determine PHB that the packet will receive

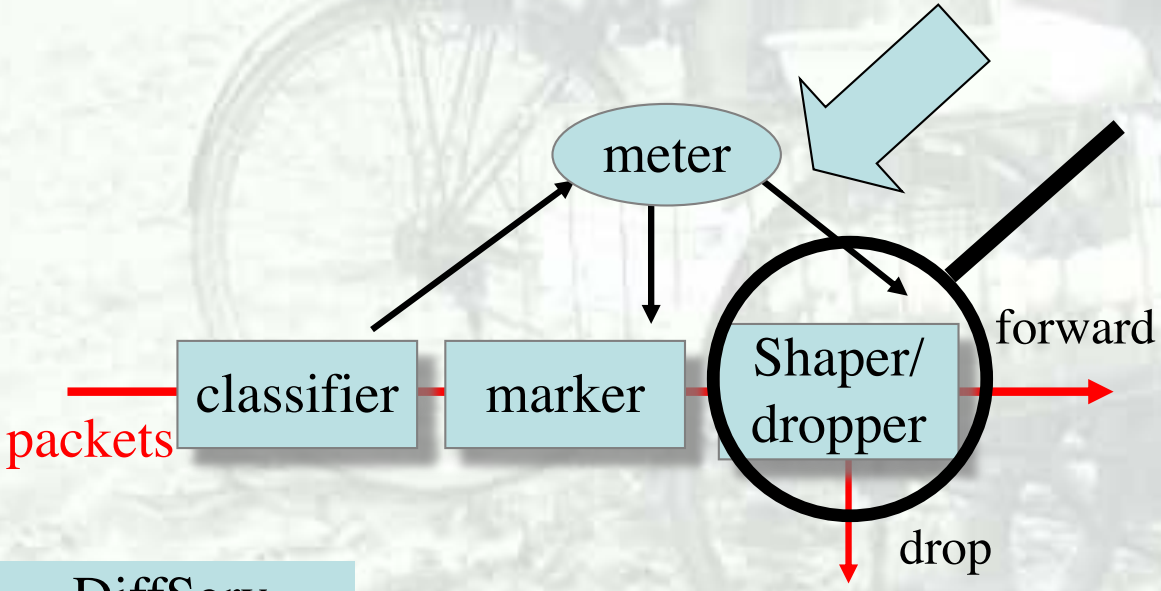
DiffServ building blocks

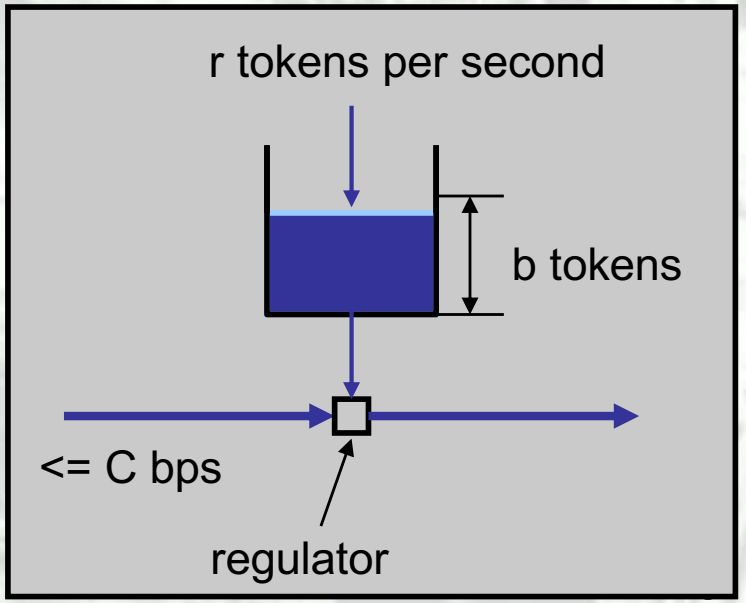


Traffic Conditioning

- User declares traffic profile (eg, rate and burst size); traffic is metered and shaped if non-conforming

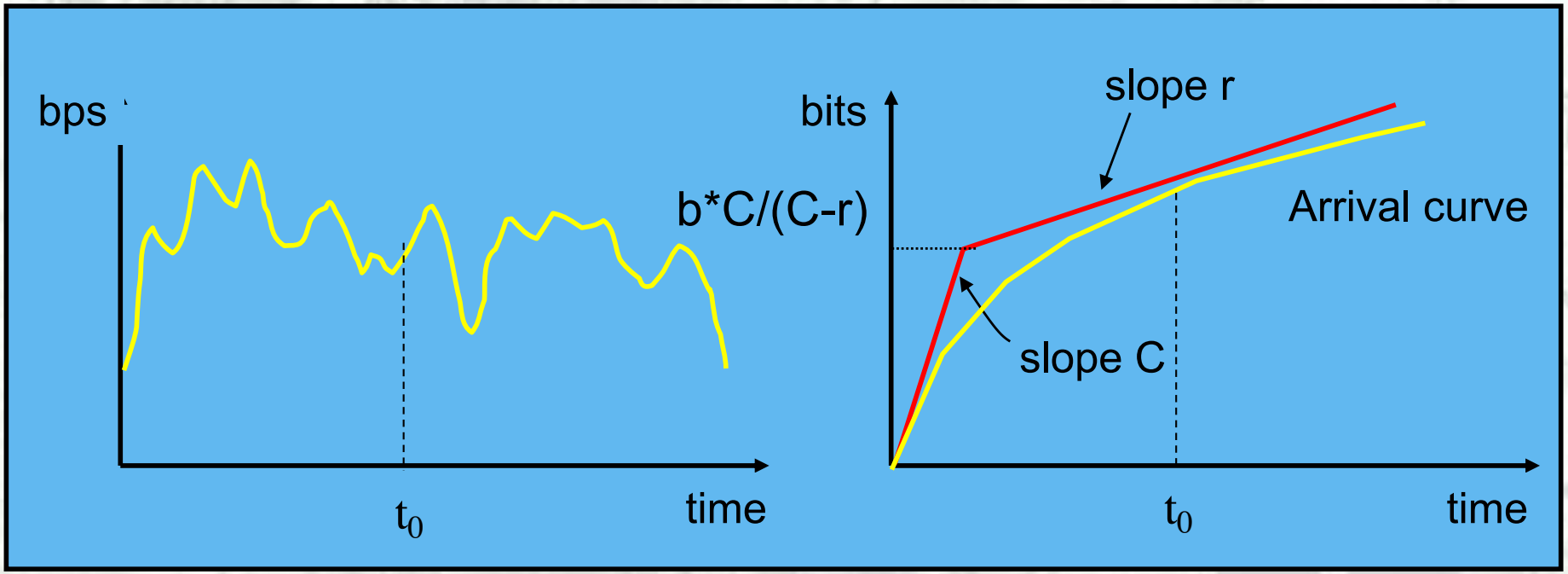




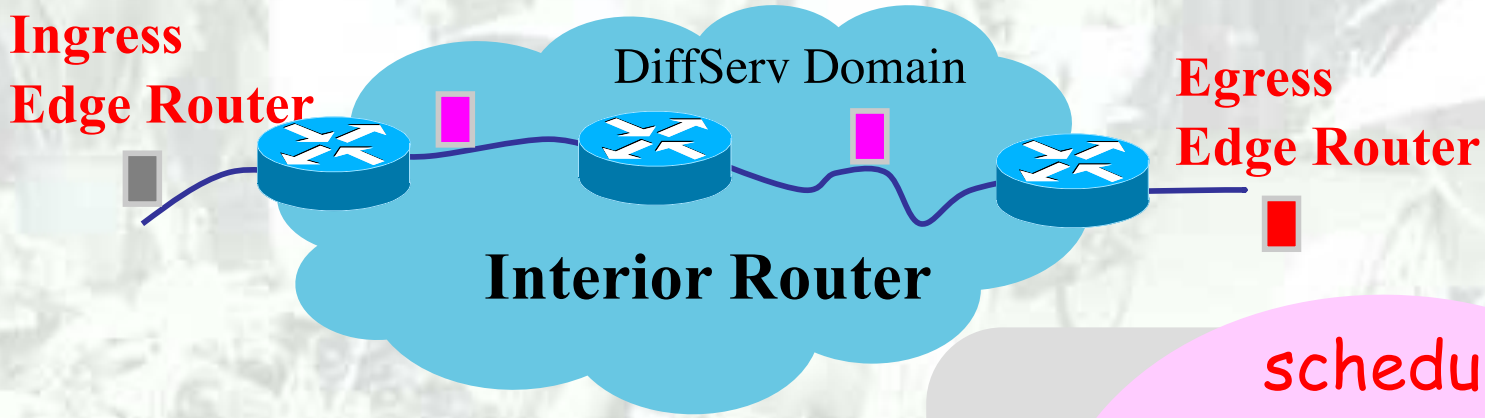


Token Bucket for traffic characterization

- Given b=bucket size, C=link capacity and r=token generation rate



Differentiated Architecture



Marking:

per-flow traffic management

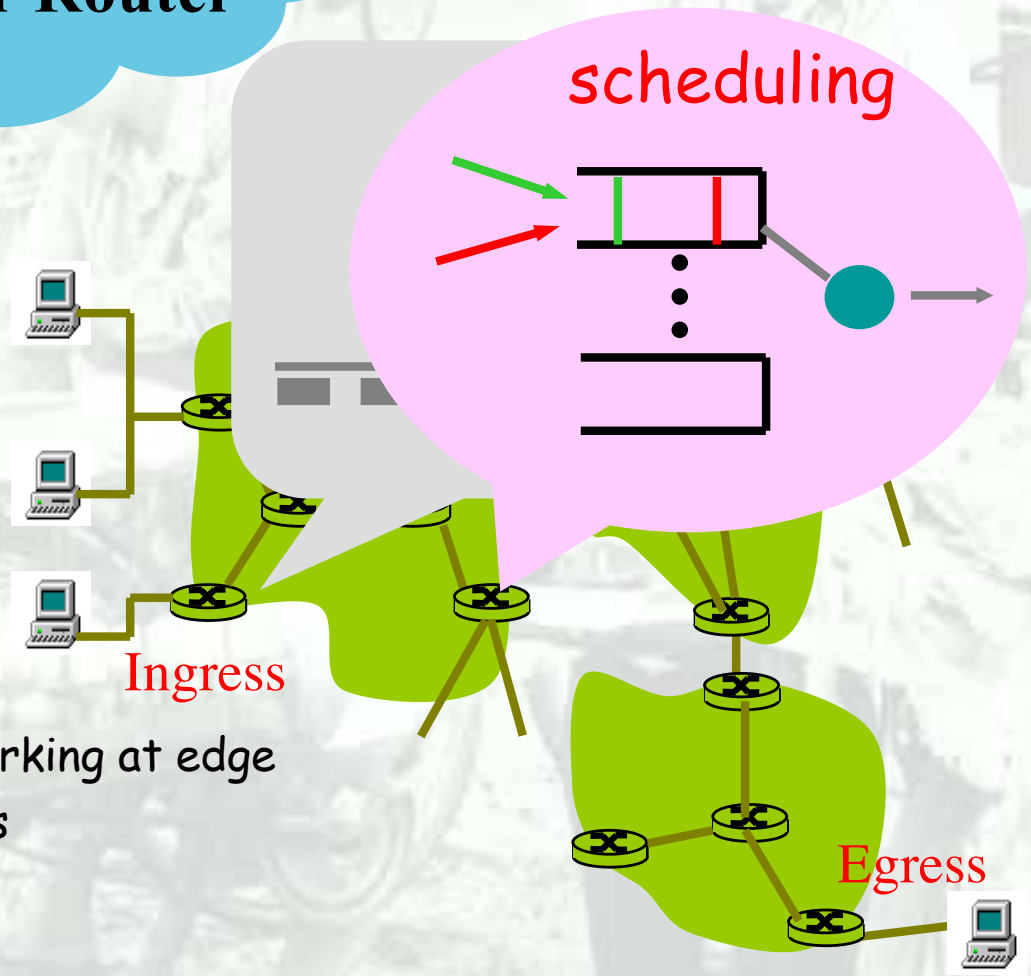
marks packets as in-profile and out-profile

Per-Hop-Behavior (PHB):

per class traffic management

buffering and scheduling based on marking at edge

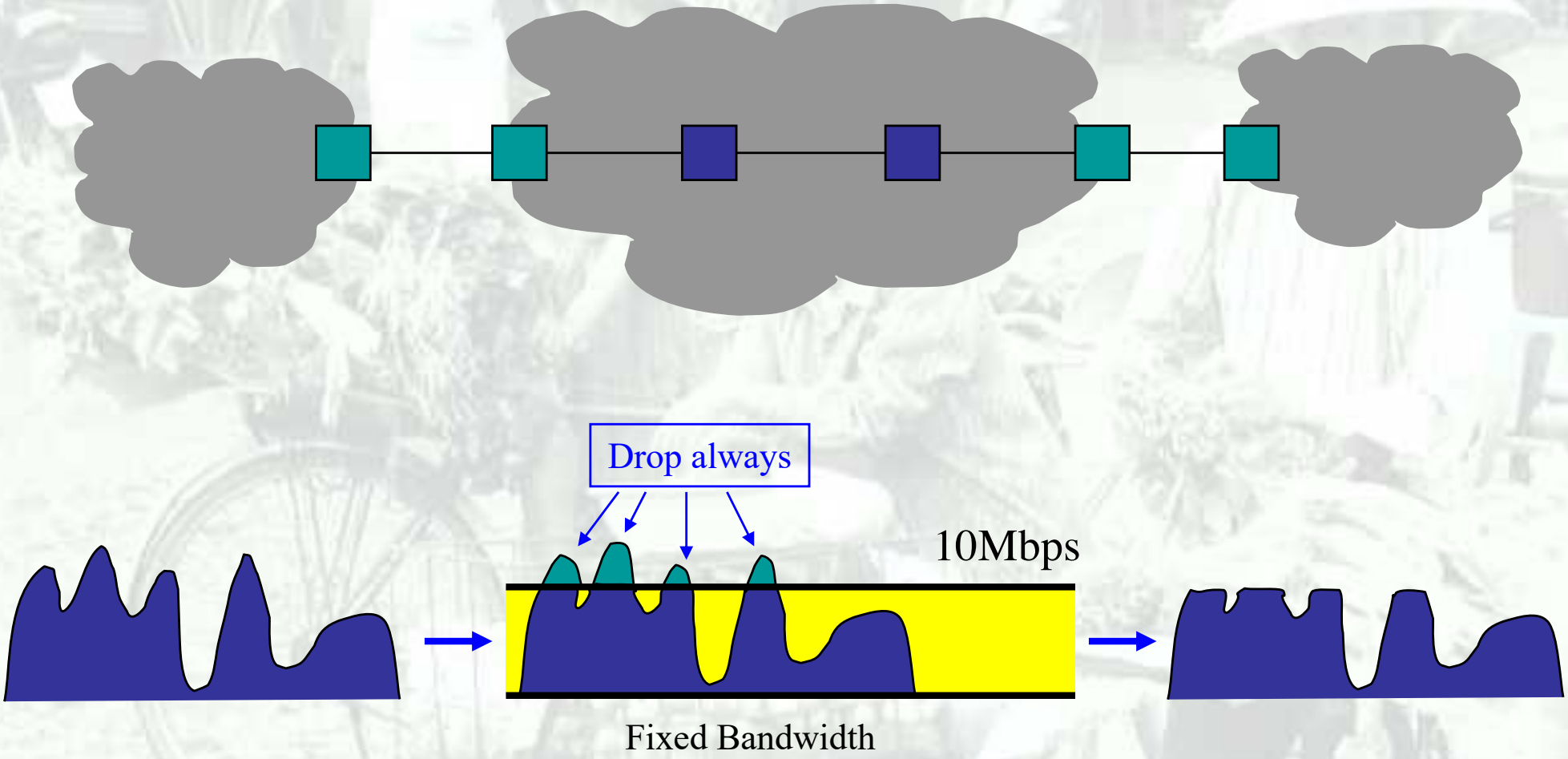
preference given to in-profile packets



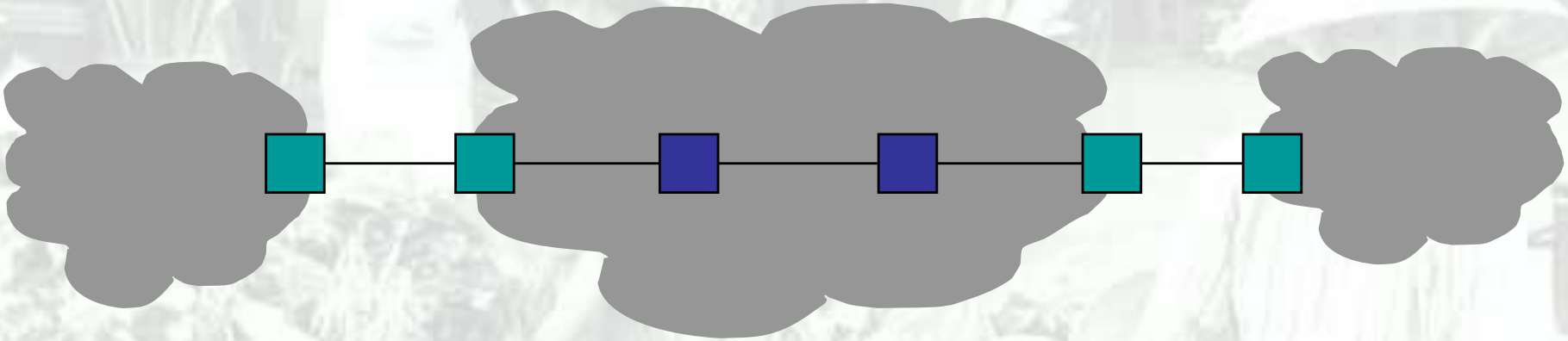
Pre-defined PHB

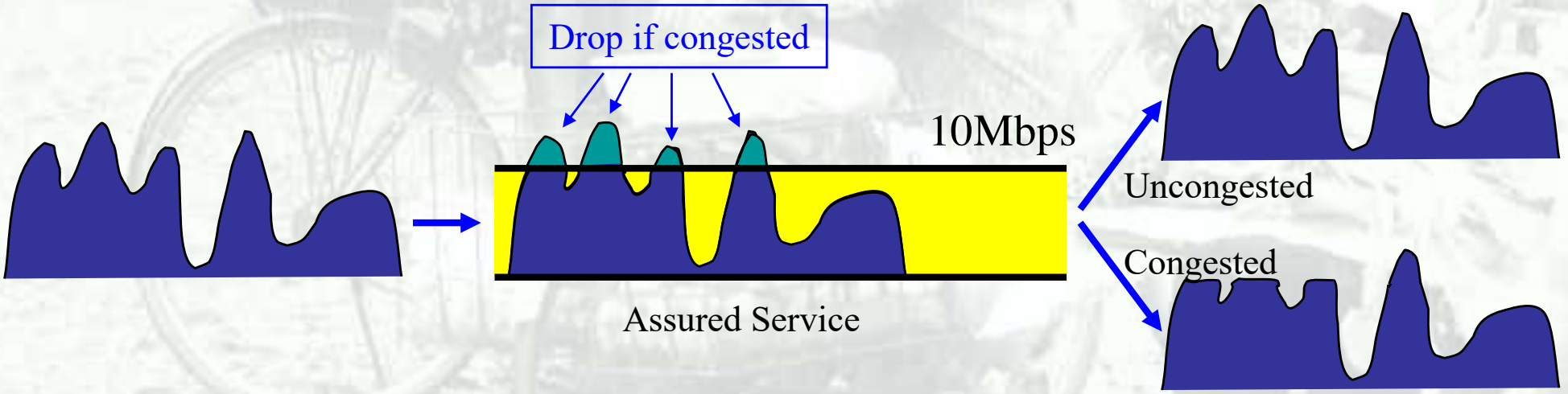
- Expedited Forwarding (EF, premium):
  - departure rate of packets from a class equals or exceeds a specified rate (logical link with a minimum guaranteed rate)
  - Emulates leased-line behavior
- Assured Forwarding (AF):
  - 4 classes, each guaranteed a minimum amount of bandwidth and buffering; each with three drop preference partitions
  - Emulates frame-relay behavior

Premium Service Example

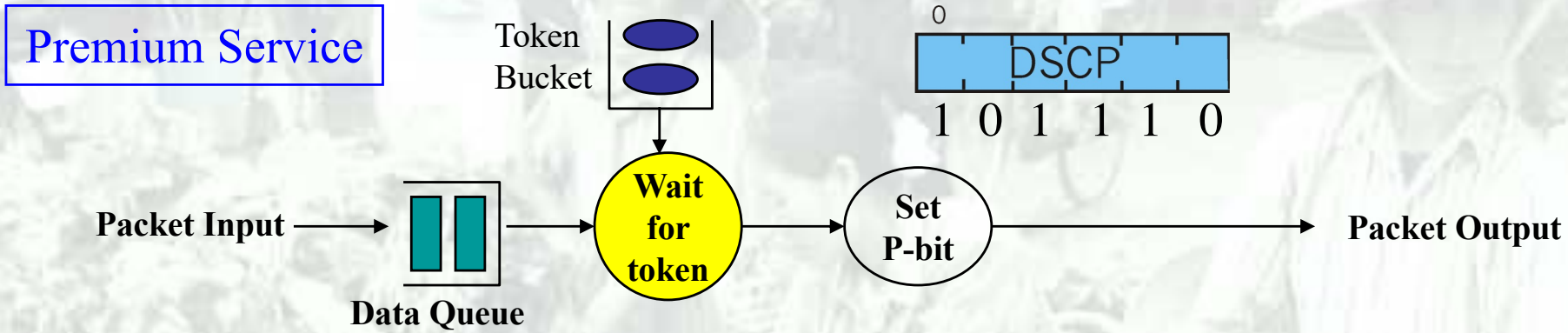


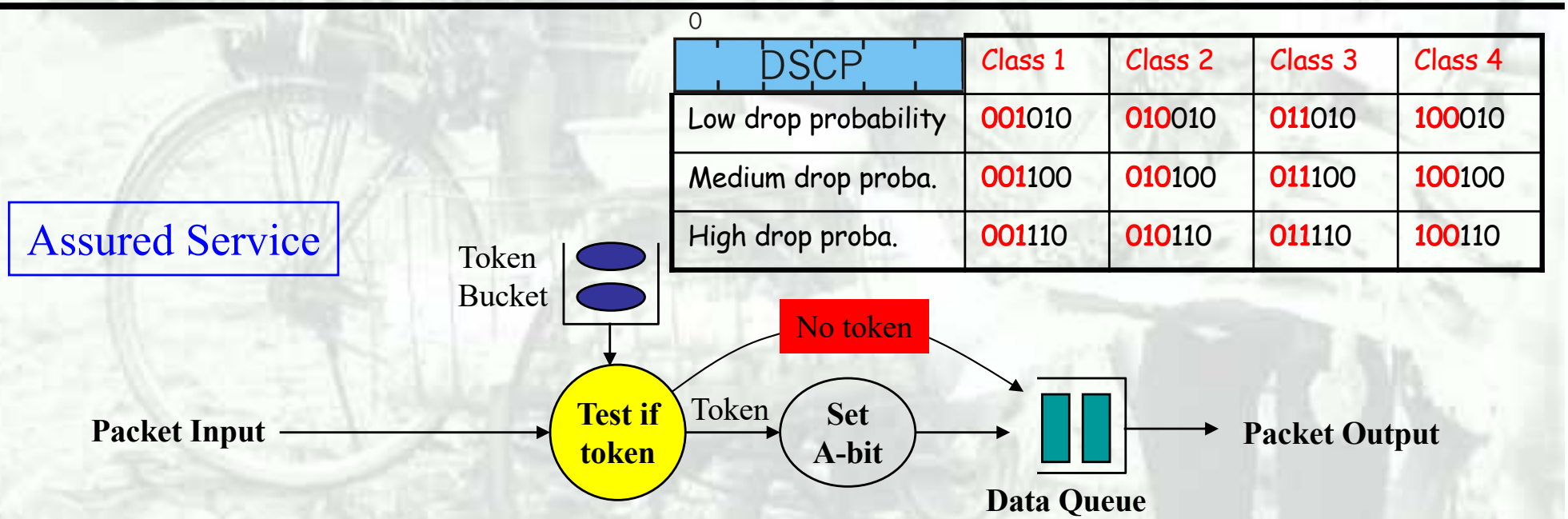
Assured Service Example



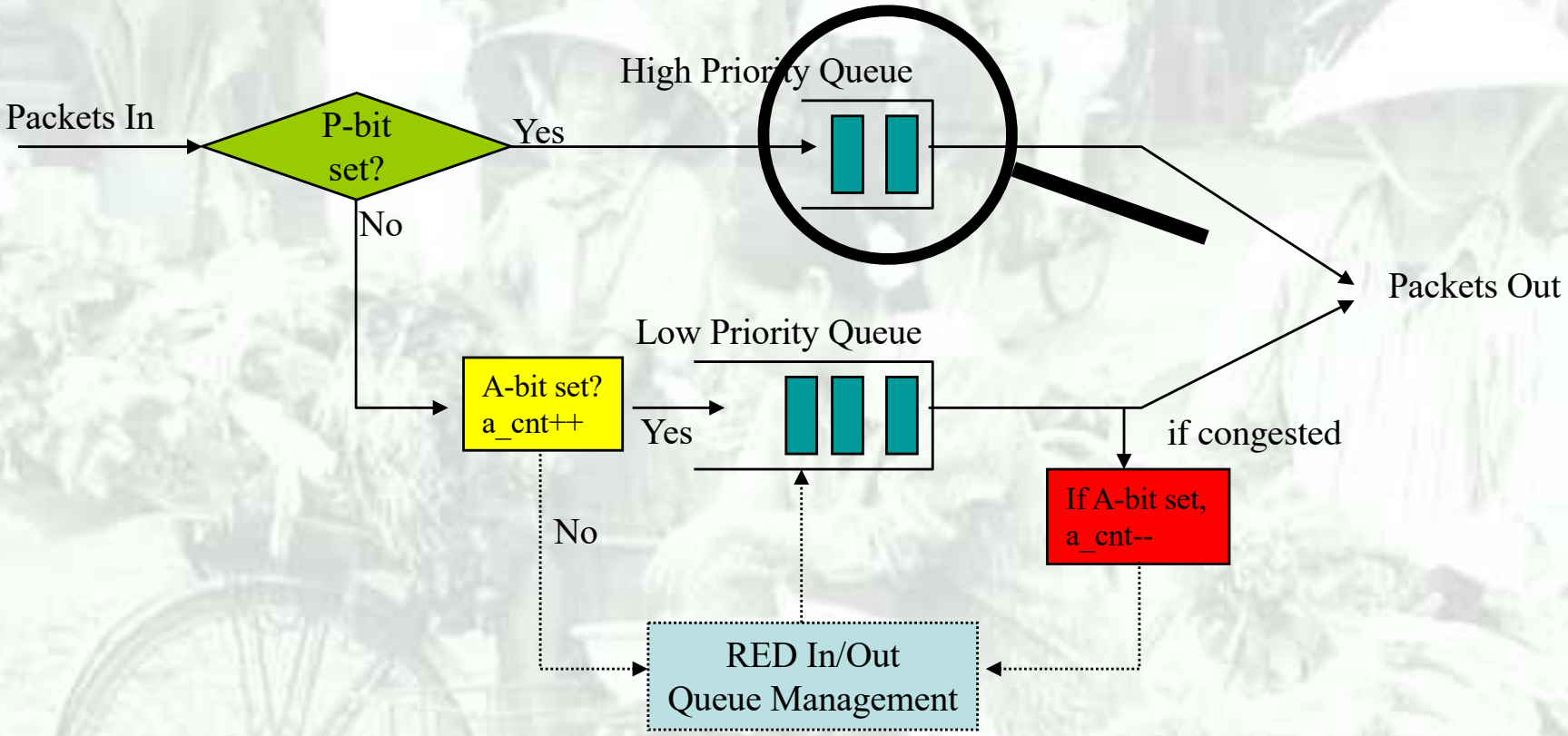


Border Router Functionality



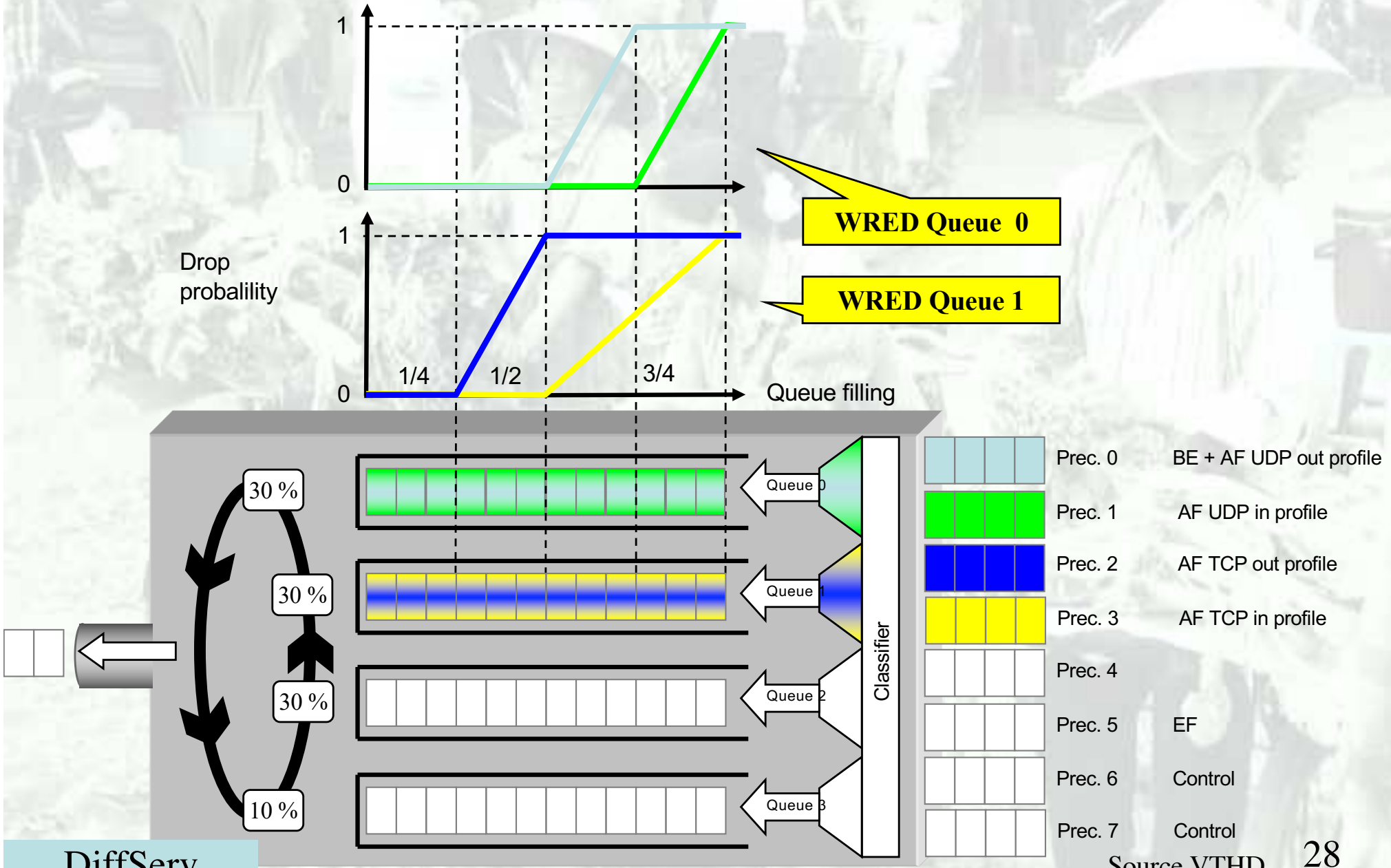


Internal Router Functionality

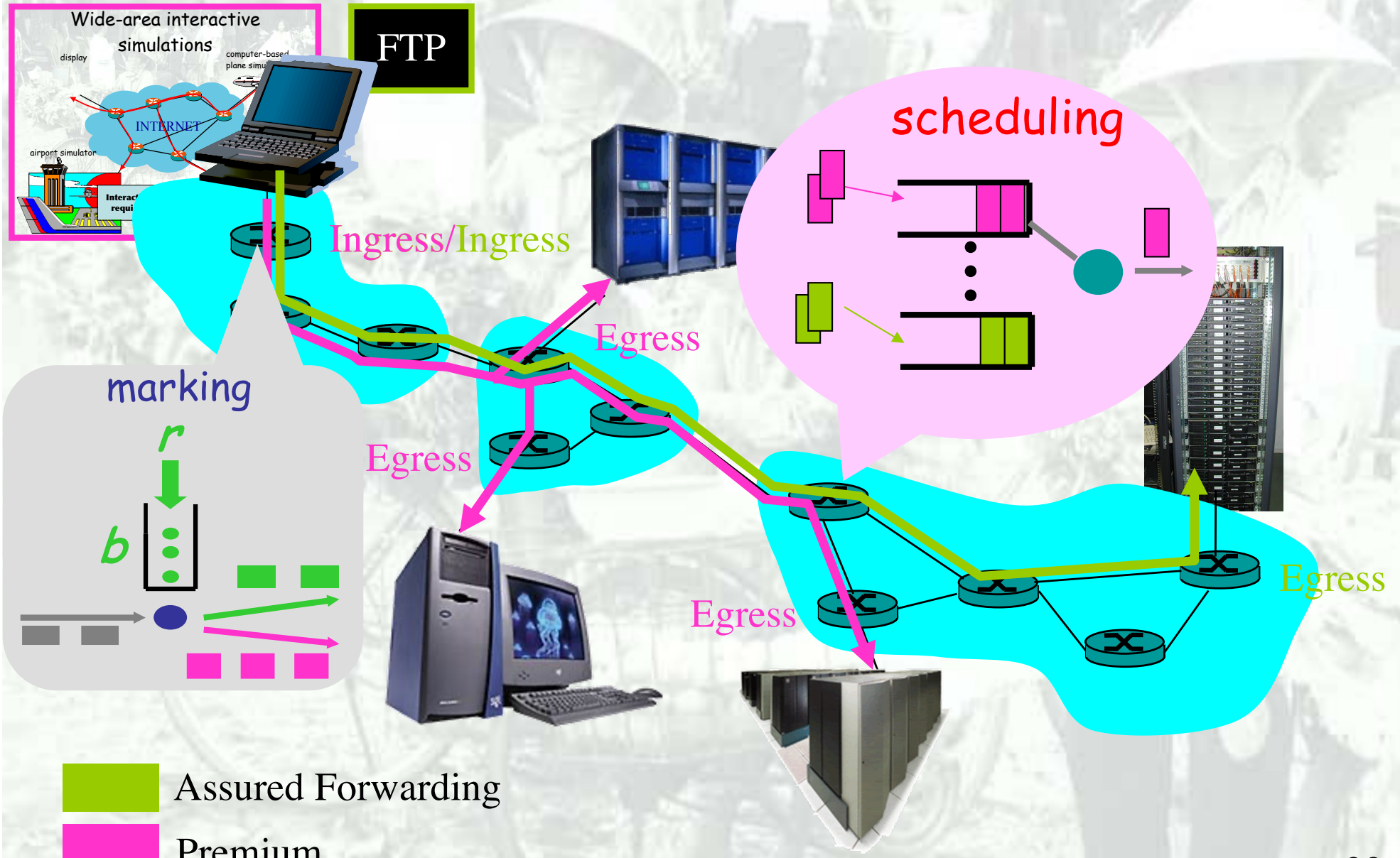


A DSCP codes aggregates, not individual flows
No state in the core
Should scale to millions of flows

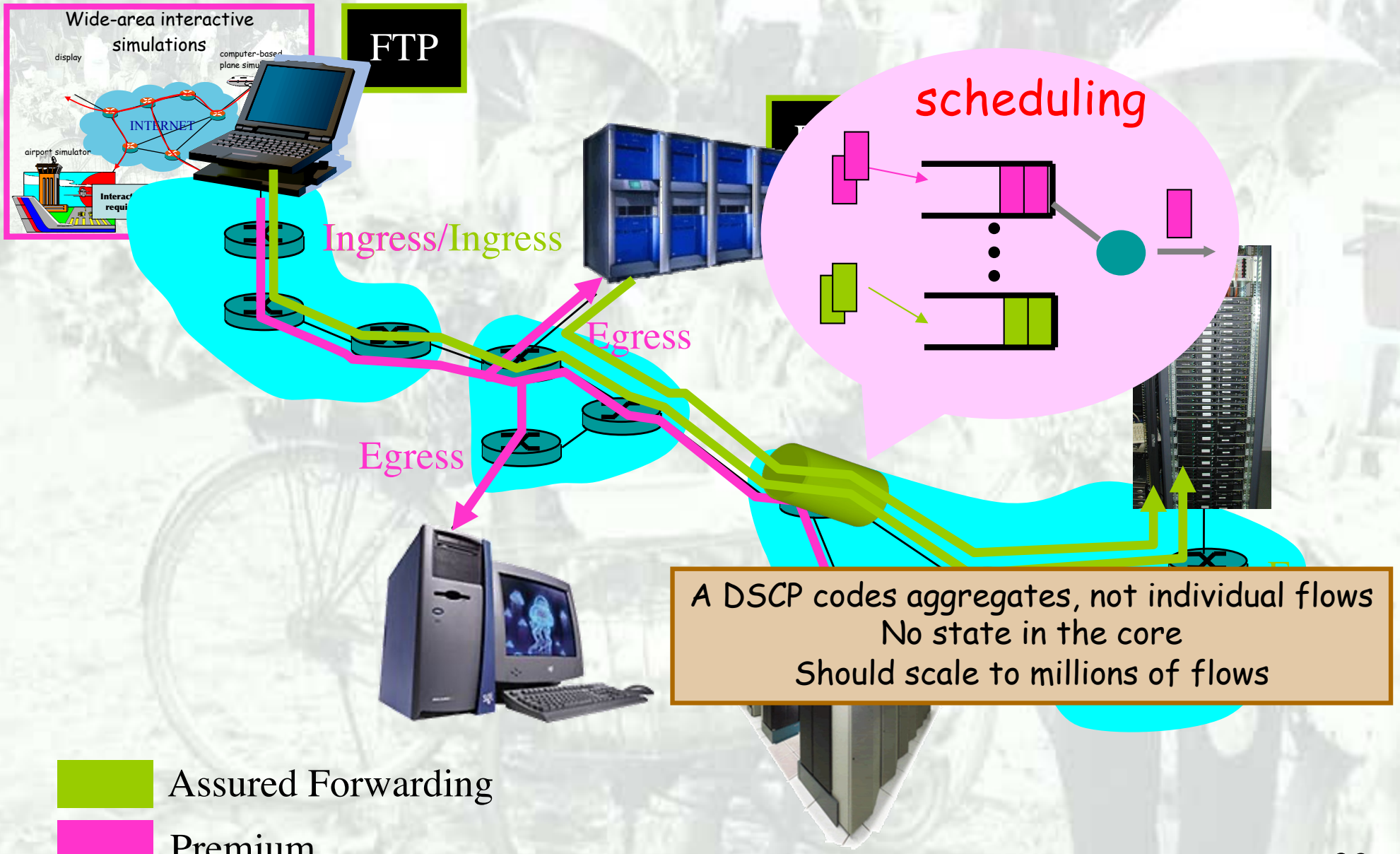
Putting it together!



DiffServ for grids

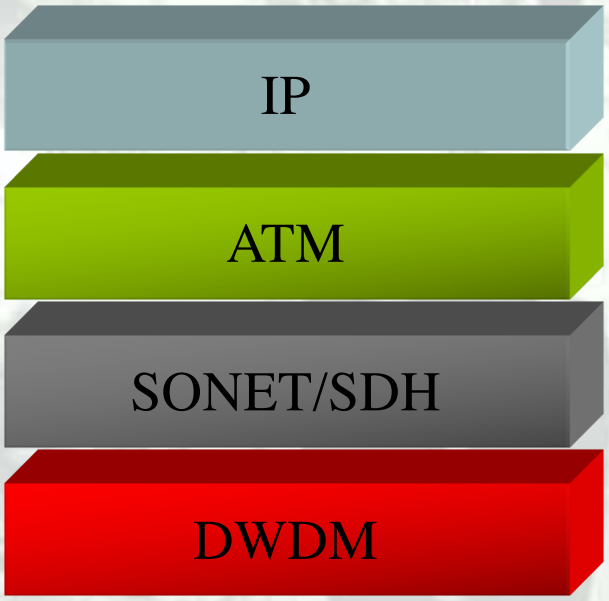


DiffServ for grids (con't)

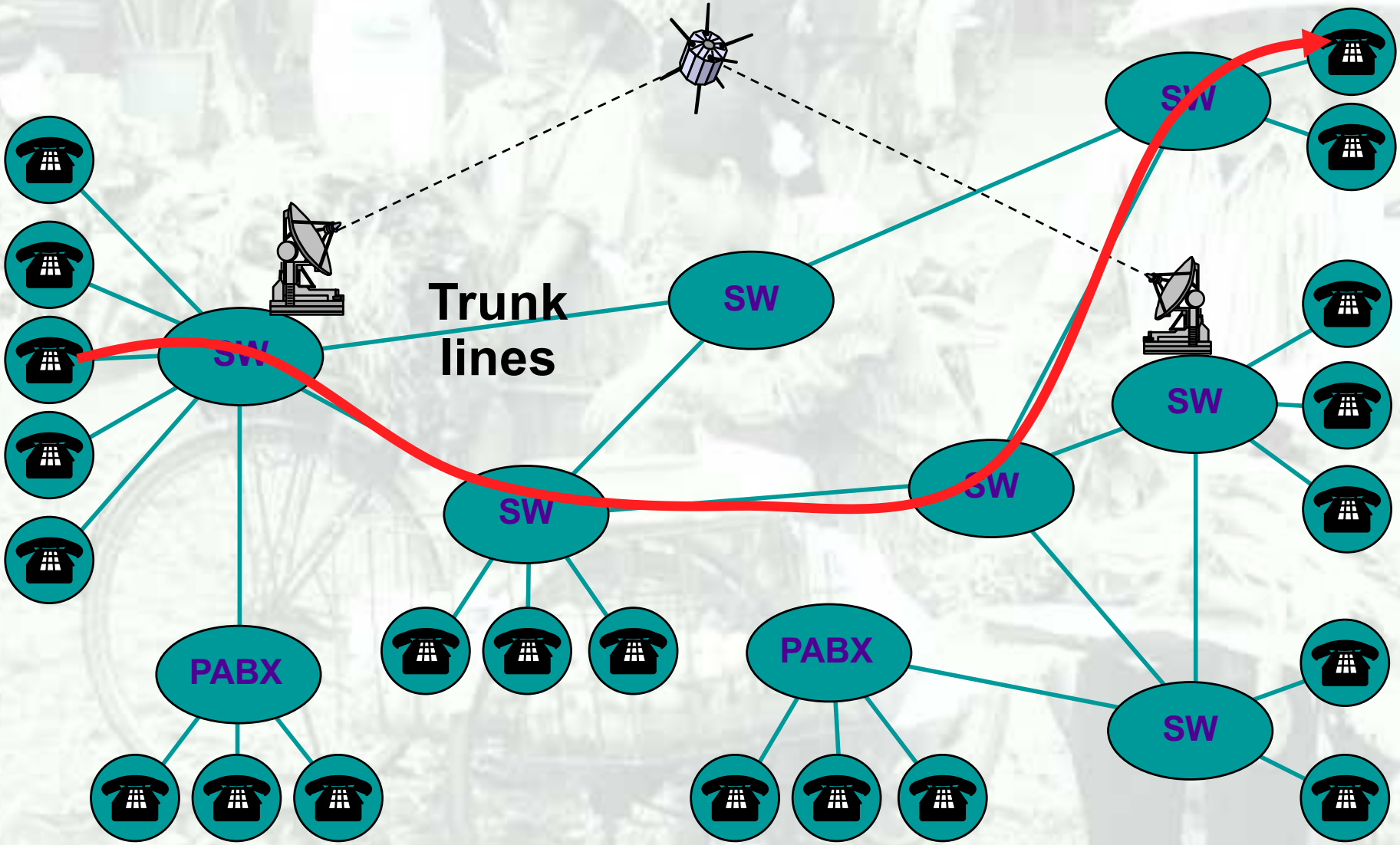


Bandwidth provisioning

- DWDM-based optical fibers have made bandwidth very cheap in the backbone
- On the other hand, dynamic provisioning is difficult because of the complexity of the network control plane:
  - Distinct technologies
  - Many protocols layers
  - Many control software



The telephone circuit view

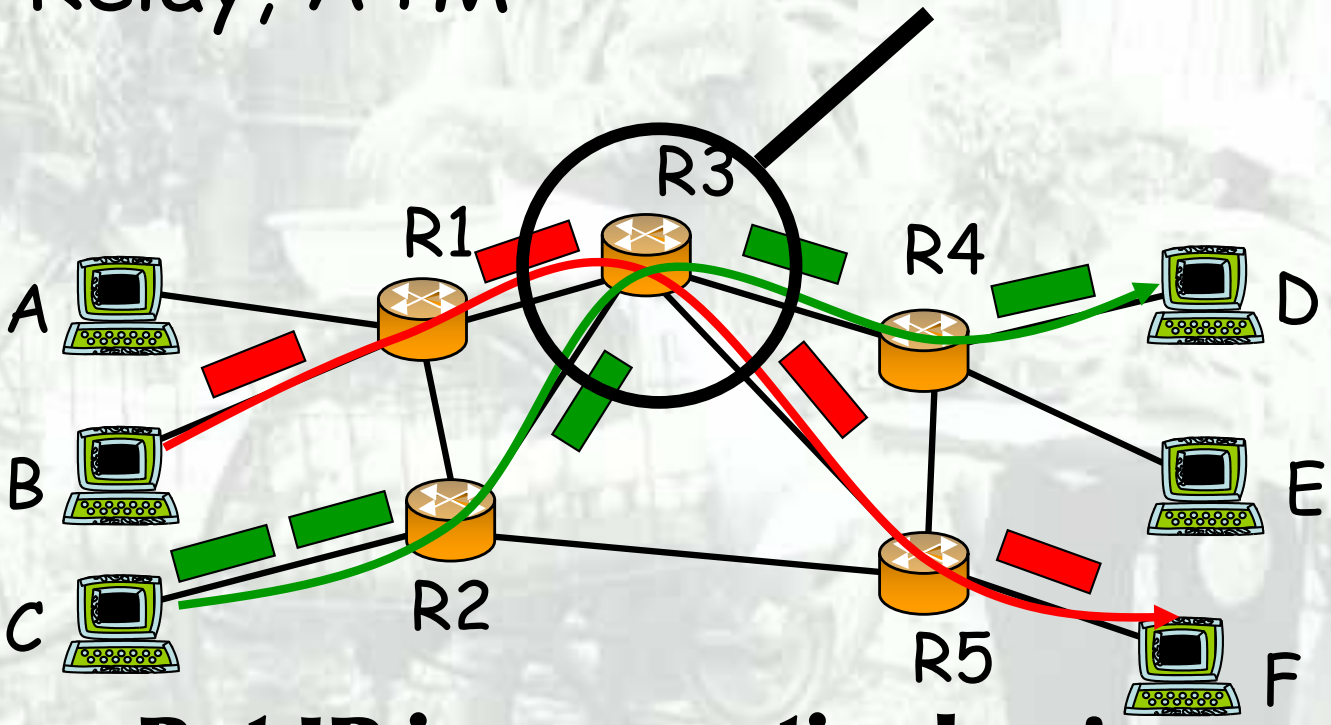


Advantages of circuits

- Provides the same path for information of the same connection: less out-of-order delivery
- Easier provisioning/reservation of network's resources: planning and management features

Back to virtual circuits

- Virtual circuit refers to a connection oriented network/link layer: e.g. X.25, Frame Relay, ATM

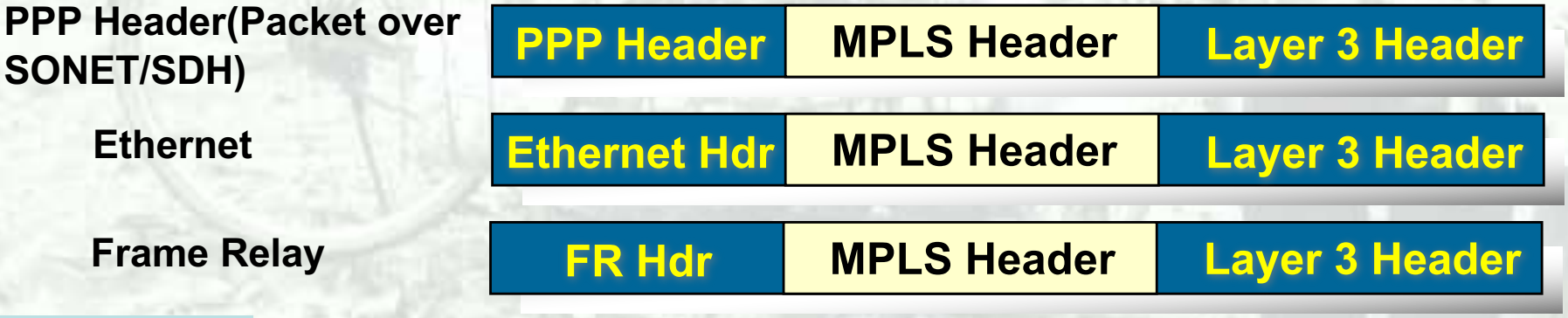


Virtual Circuit Switching: a path is defined for each connection

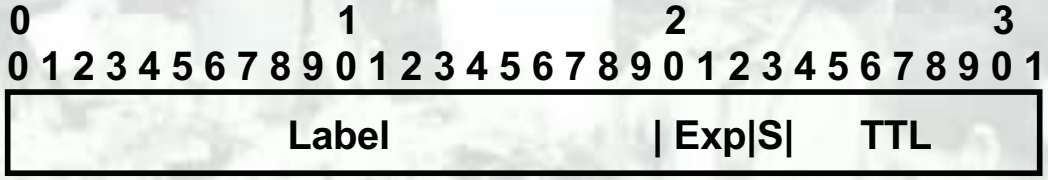
But IP is connectionless!

Virtual circuits in IP networks

- Multi-Protocol Label Switching
  - Fast: use label switching➔ LSR
  - Multi-Protocol: above link layer, below network layer
  - Facilitate traffic engineering



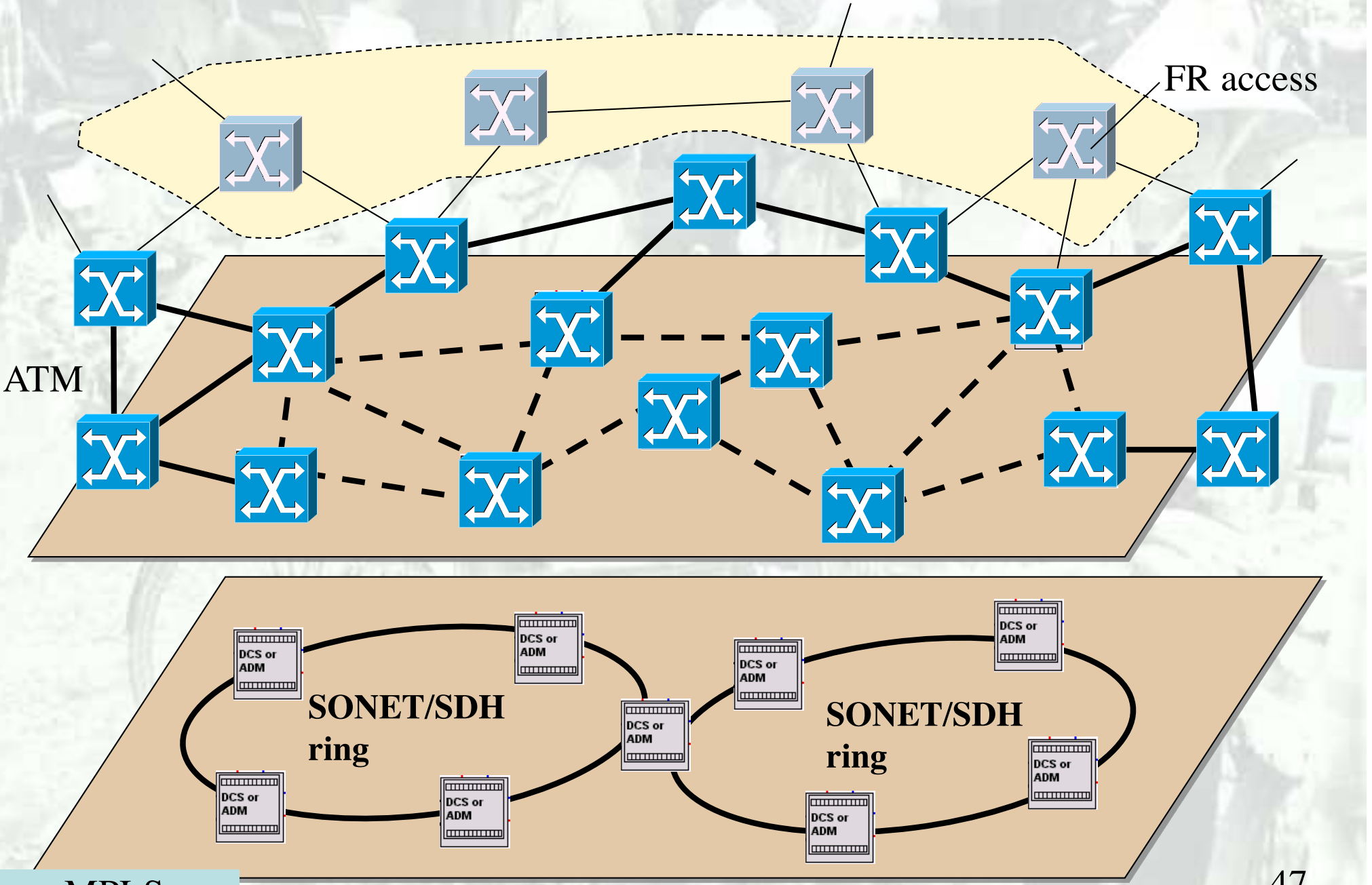
Label structure



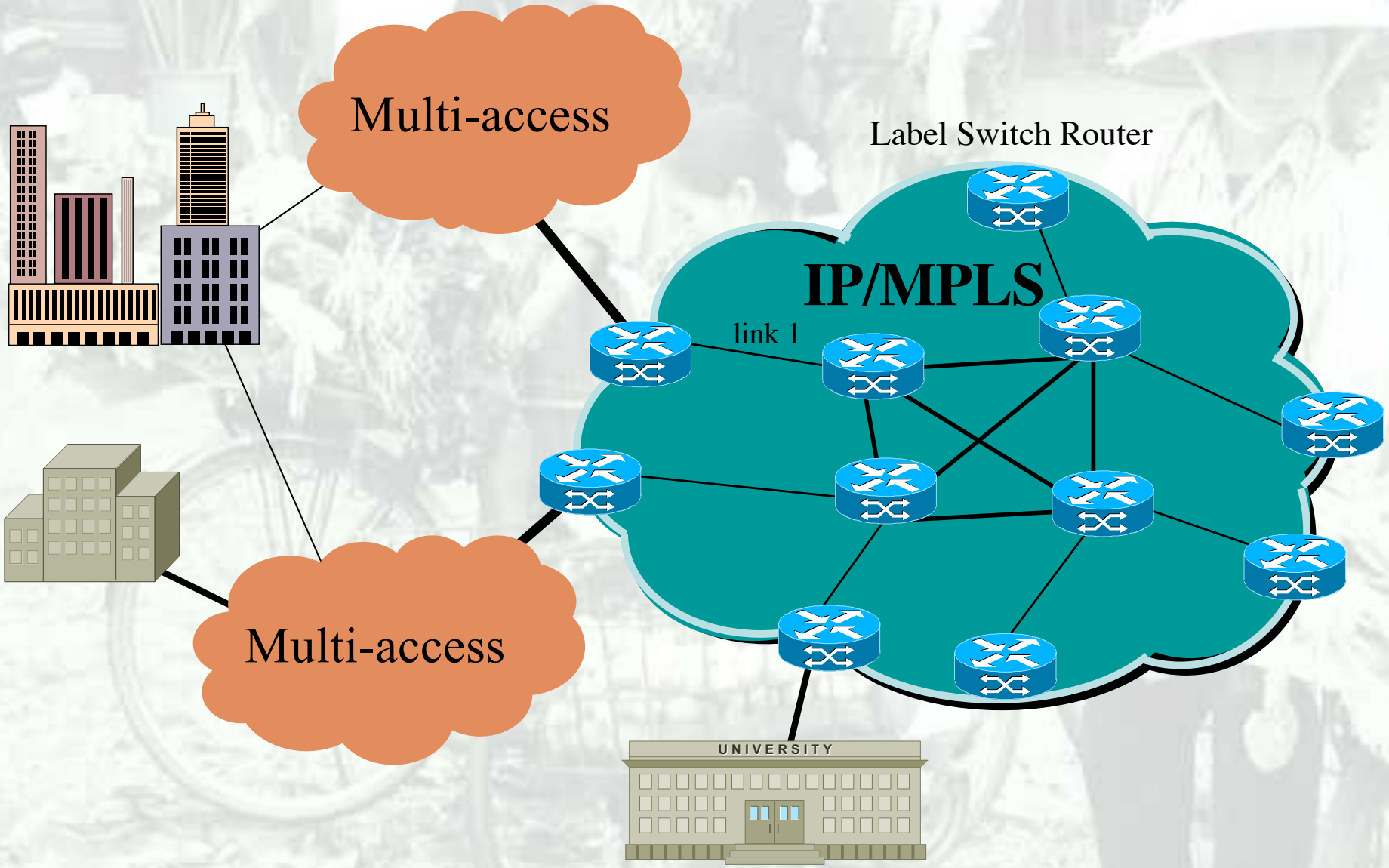
Label = 20 bits
Exp = Experimental, 3 bits
S = Bottom of stack, 1bit
TTL = Time to live, 8 bits

- More than one label is allowed -> Label Stack
- MPLS LSRs always forward packets based on the value of the label at the top of the stack

From multilayer networks…



...to IP/MPLS networks

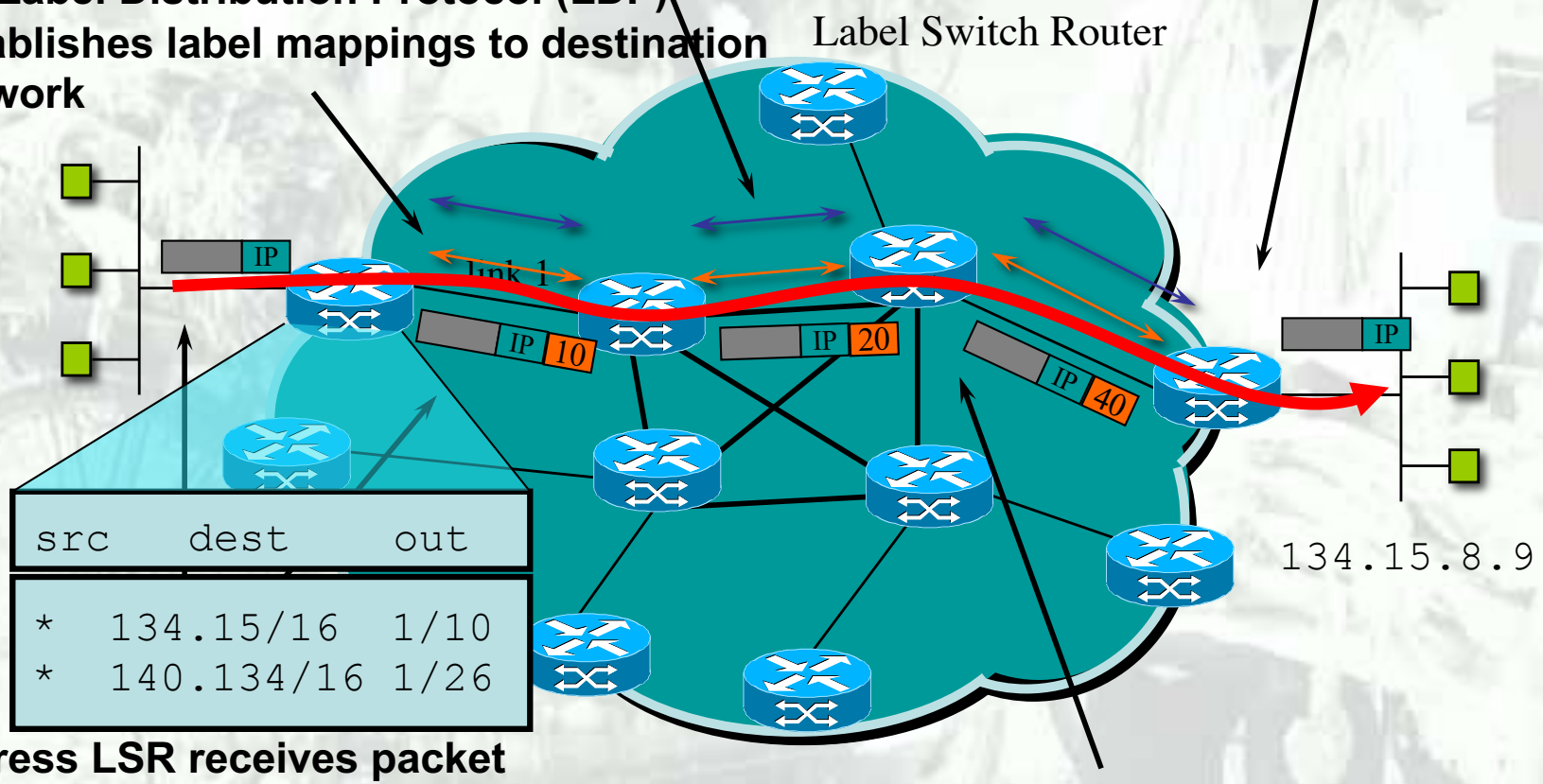


MPLS operation

1a. Routing protocols (e.g. OSPF-TE, IS-IS-TE) exchange reachability to destination networks

1b. Label Distribution Protocol (LDP) establishes label mappings to destination network

4. LSR at egress removes label and delivers packet

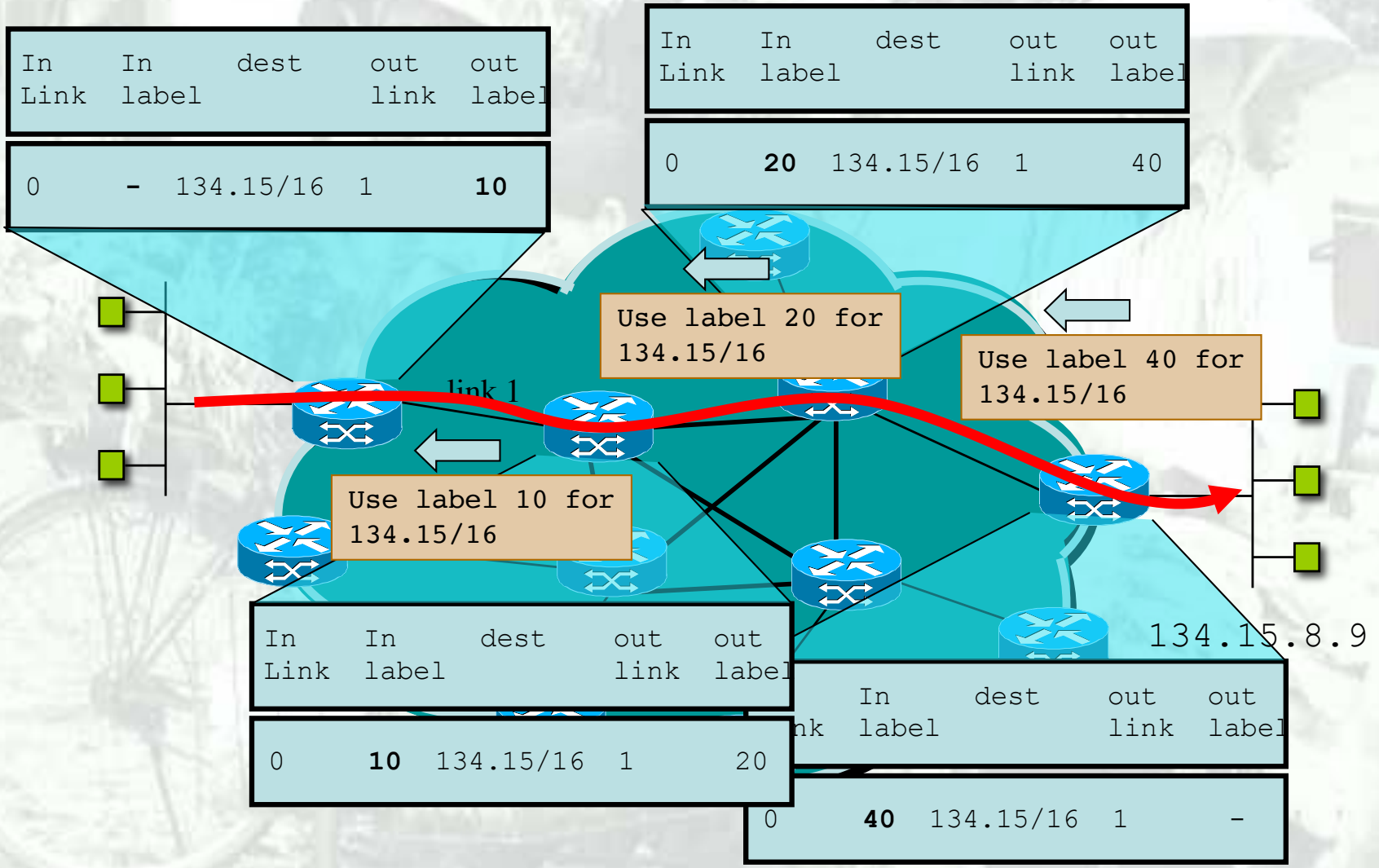


2. Ingress LSR receives packet and "label"s packets

3. LSR forwards packets using label switching

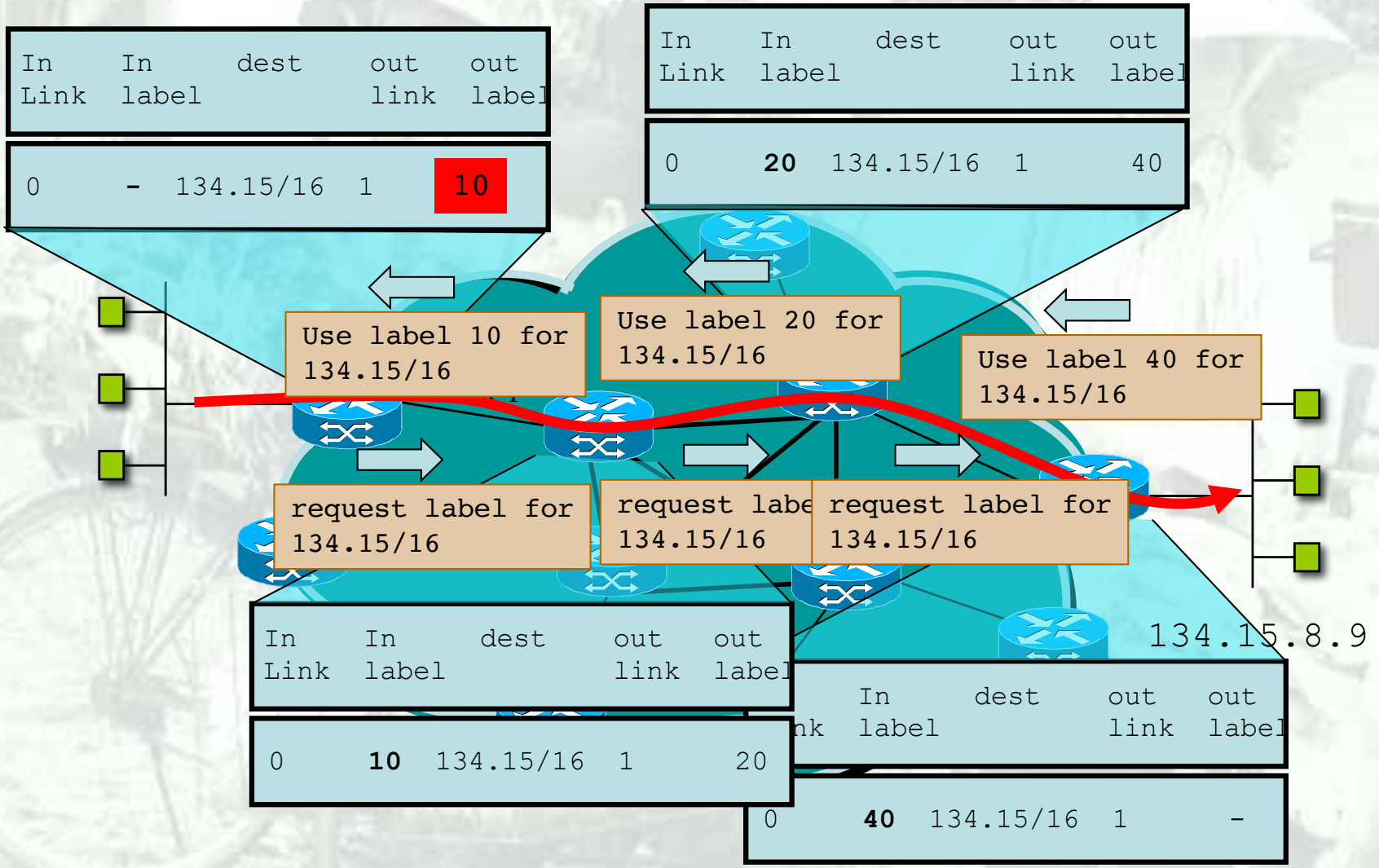
Source Yi Lin, modified C. Pham

Label Distribution



Unsolicited downstream label distribution

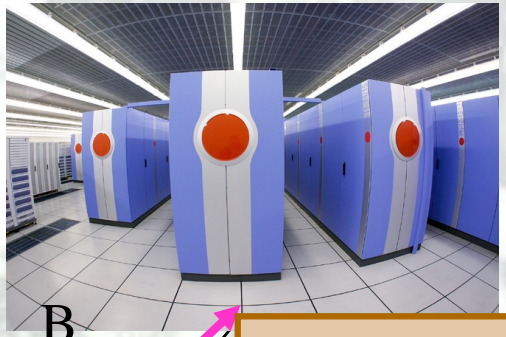
Label Distribution (con't)



On-demand downstream label distribution

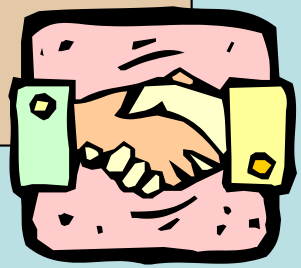
Dynamic circuits for grids

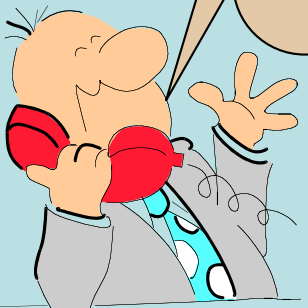
E



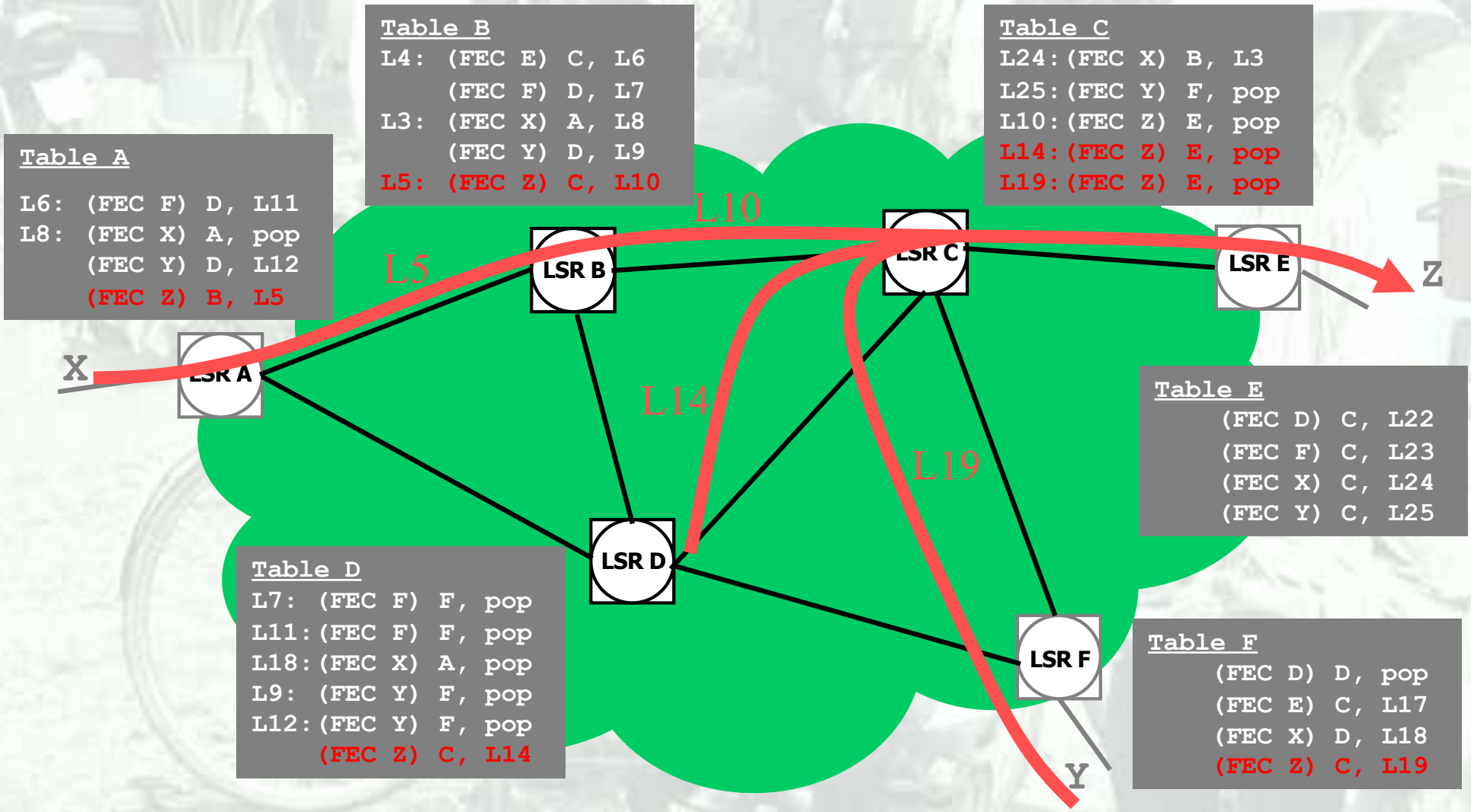
B

I need 2.5 Gbps between:
A & B
B & C
D & C
E & A



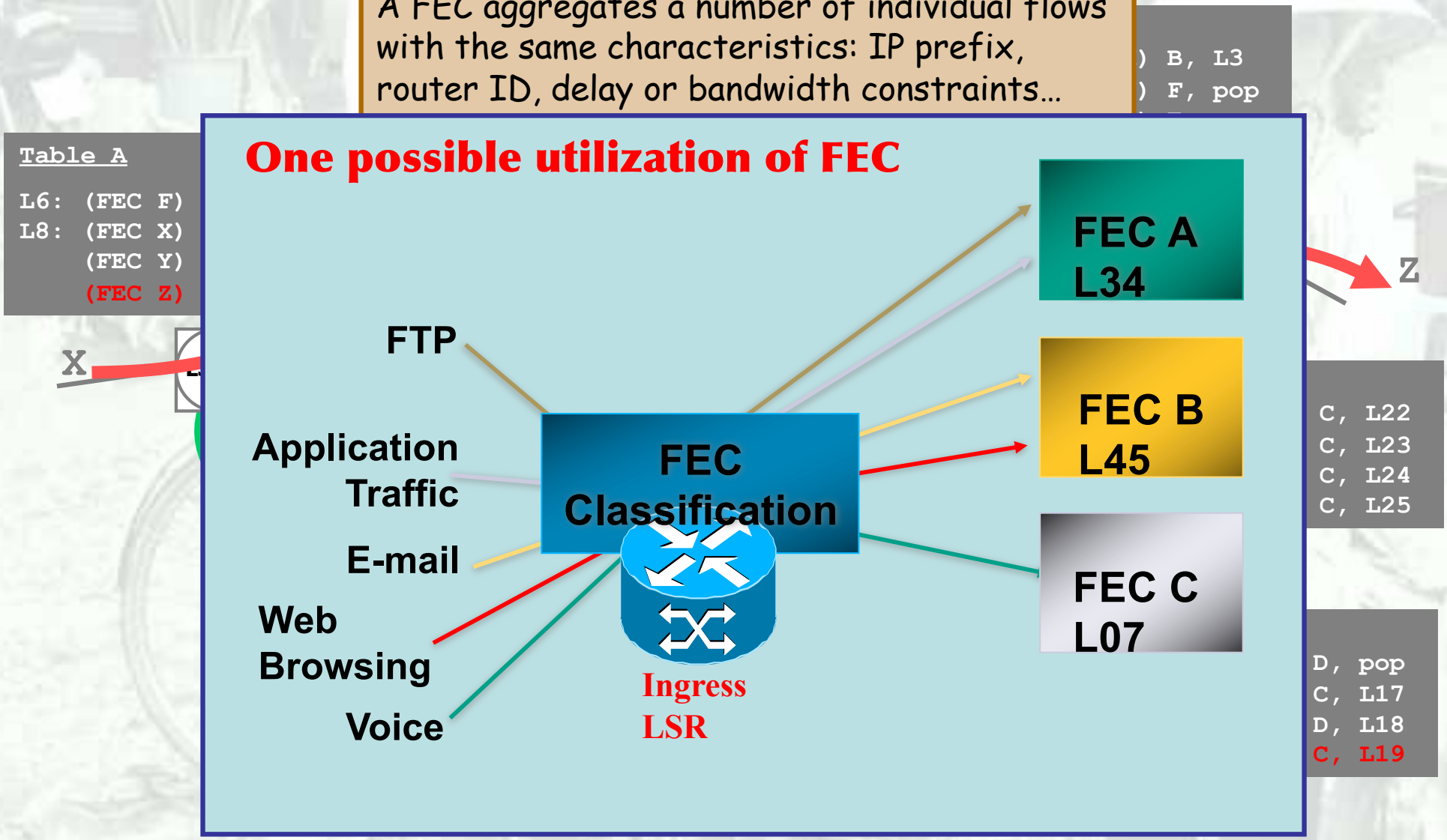


Forwarding Equivalent Class: high-level forwarding criteria

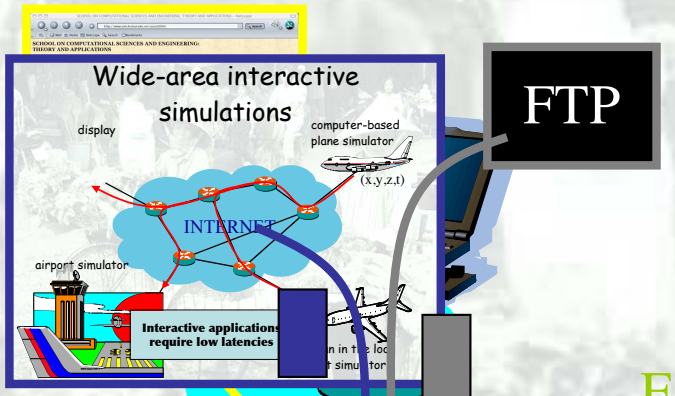


Forwarding Equivalent Class

A FEC aggregates a number of individual flows with the same characteristics: IP prefix, router ID, delay or bandwidth constraints…



MPLS FEC for the grid



FTP

Egress
Egress
Egress
Egress
Egress
Egress

FEC A: time constraint applications

FEC B: best effort traffic

MPLS for resiliency

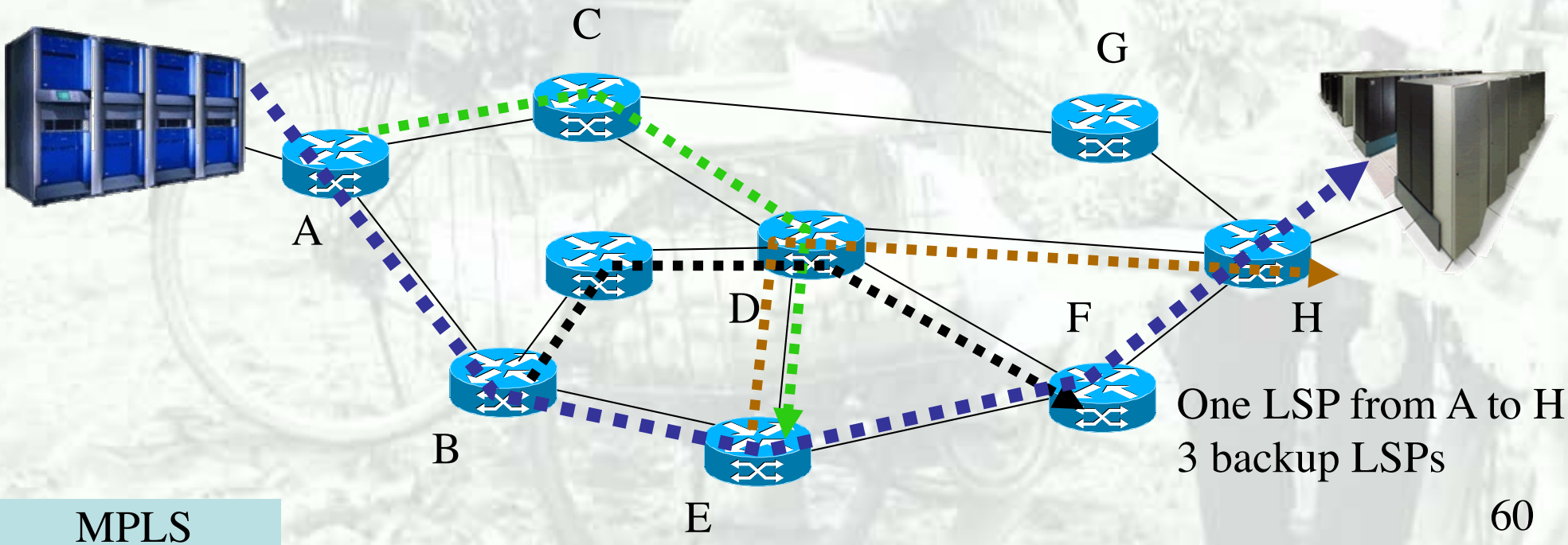
MPLS FastReroute

- Intended to provide SONET/SDH-like healing capabilities
- Selects an alternate route in tenth of ms, provides path protection
- Traditional routing protocols need minutes to converge!
- FastReroute is performed by maintaining backup LSPs

MPLS for resiliency, con't

Backup LSPs

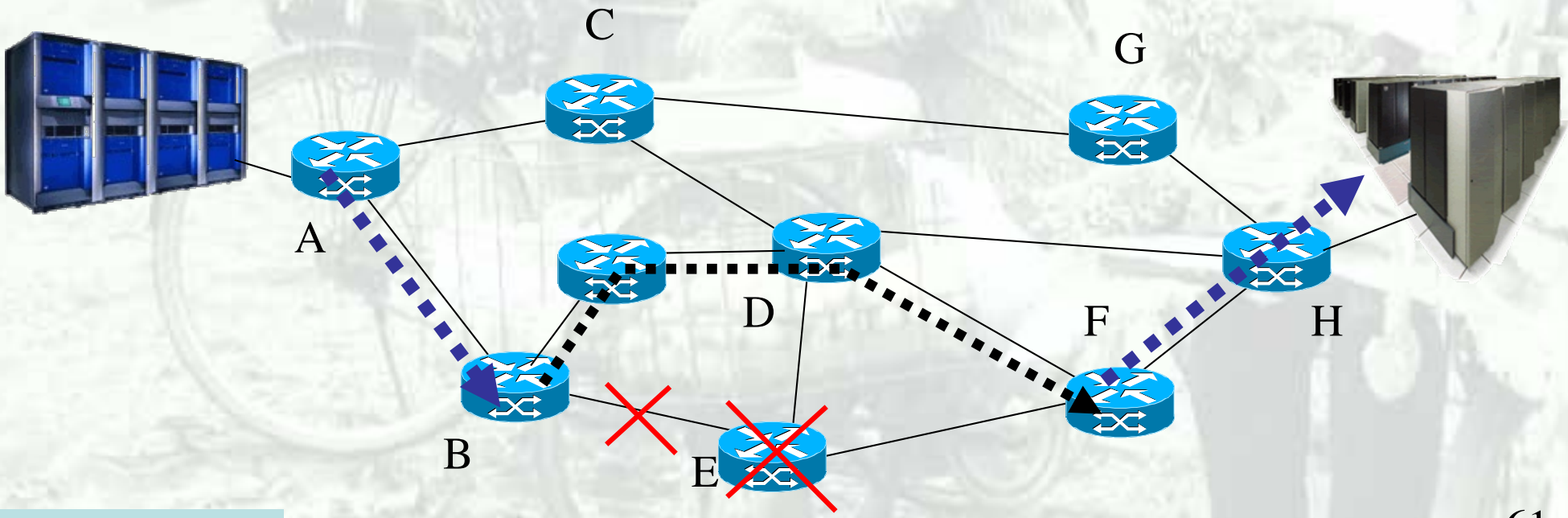
- One-to-one
- Many-to-one: more efficient but needs more configurations



MPLS for resiliency, con't

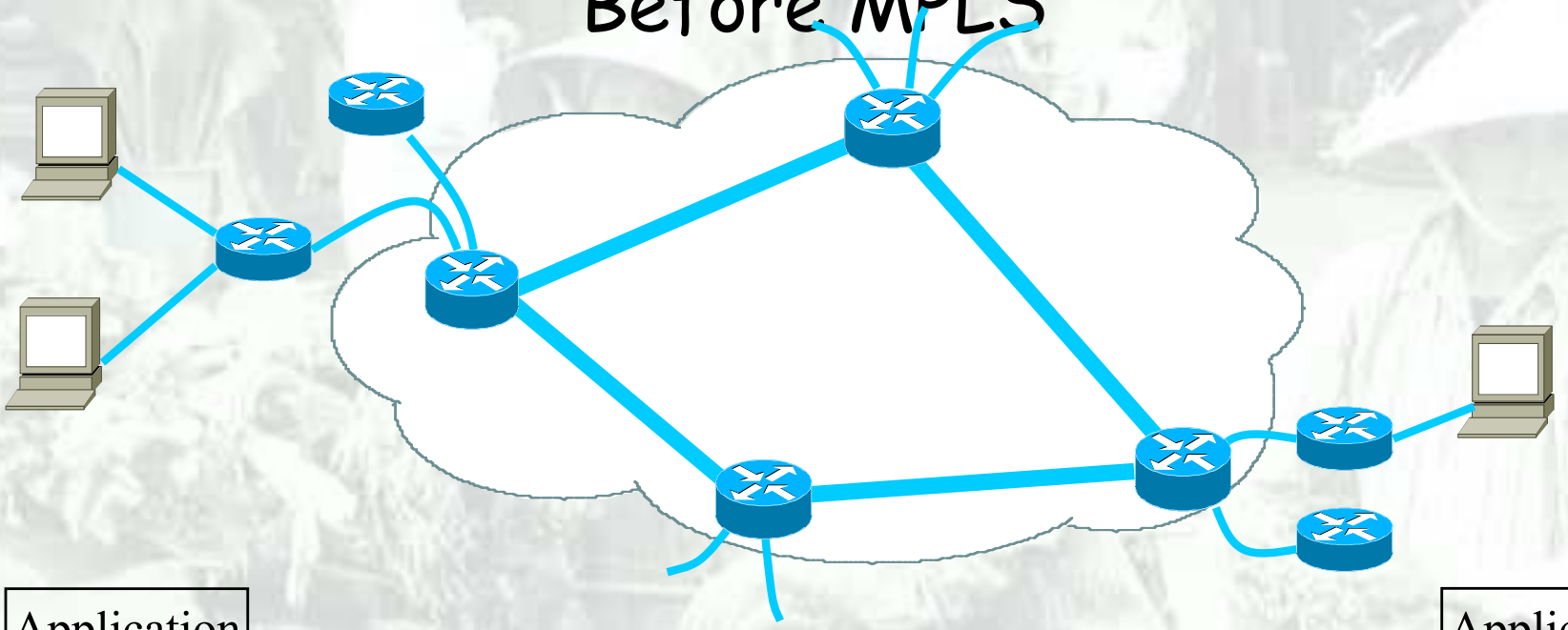
Recovery on failures

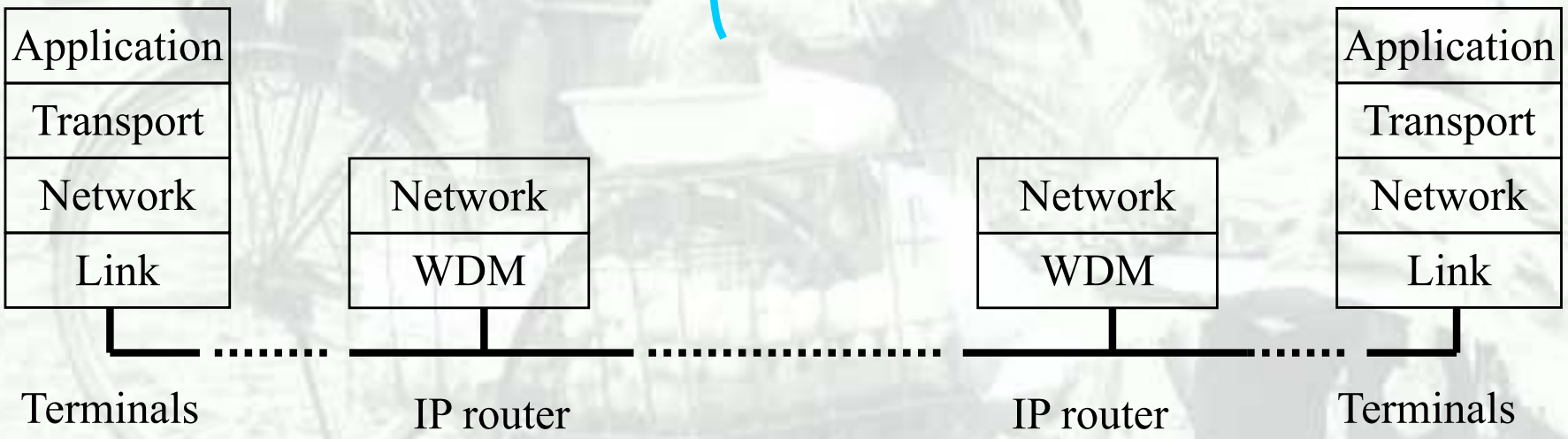
- Suppose E or link B-E is down…
- B uses detour around E with backup LSP



MPLS for optical networks

Before MPLS

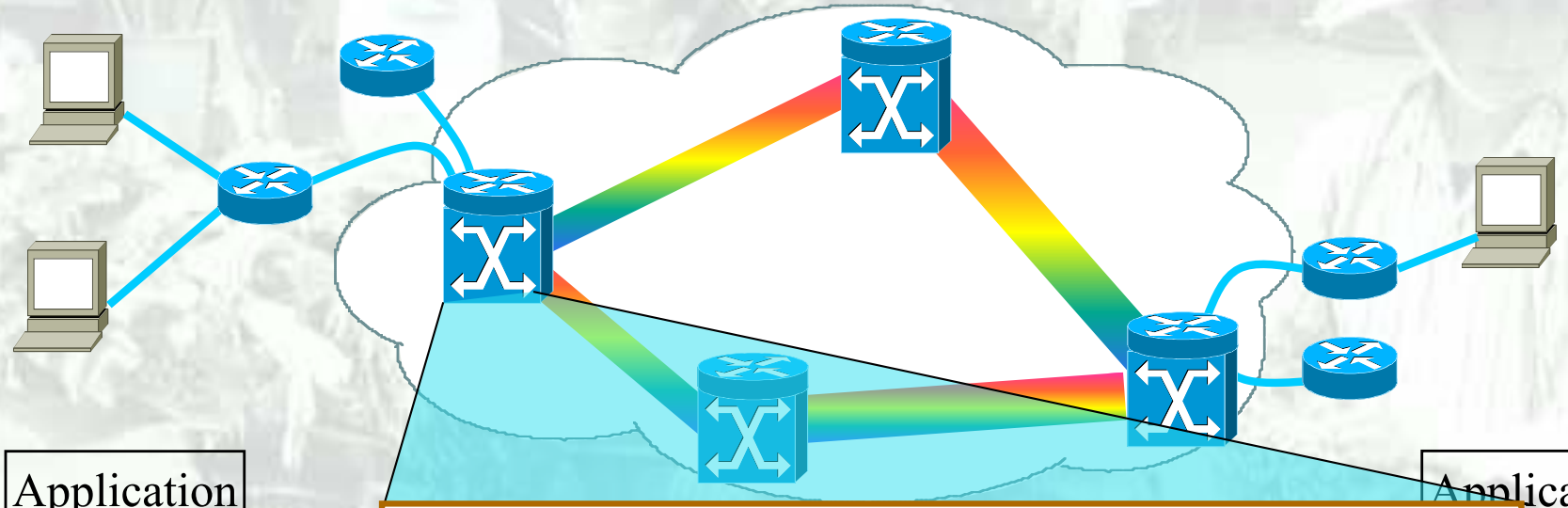


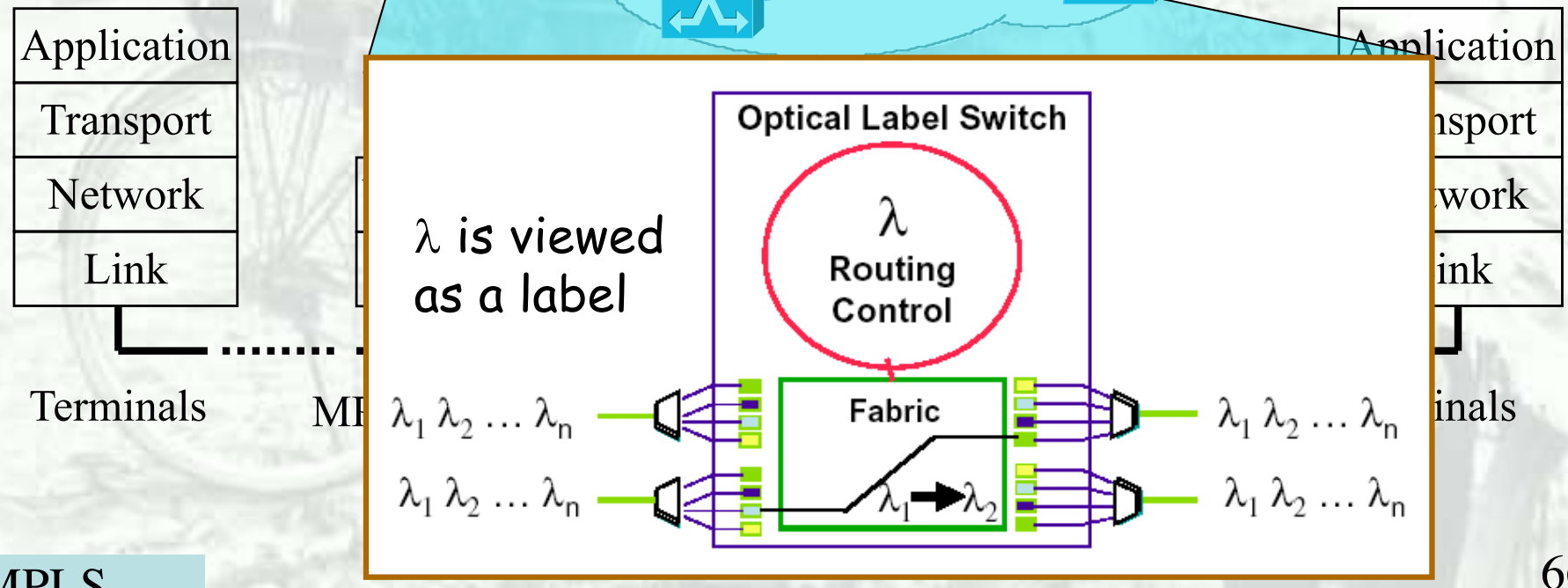


Source J. Wang, B. Mukherjee, B. Yoo

MPLS for ON, con't

MPλS=MPLS+λ lightpath





Summary
Towards IP/(G)MPLS/DWDM

From cisco