N E W C H A P T E R

Revisiting the *same service for all* paradigm

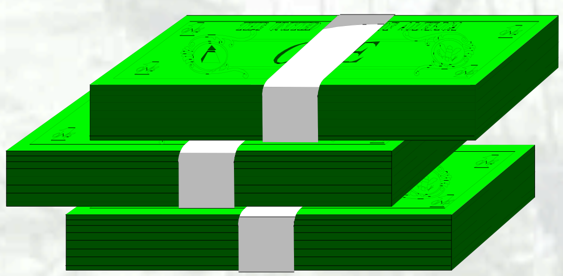
Enhancing the best-effort service

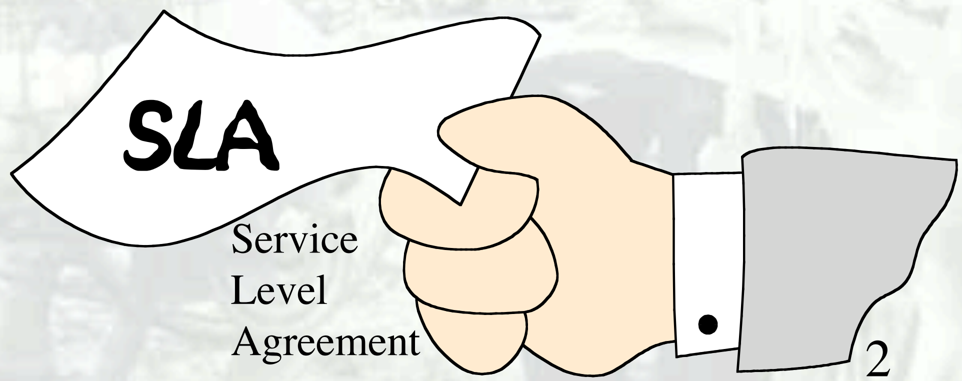
Introduce Service Differentiation

Service Differentiation

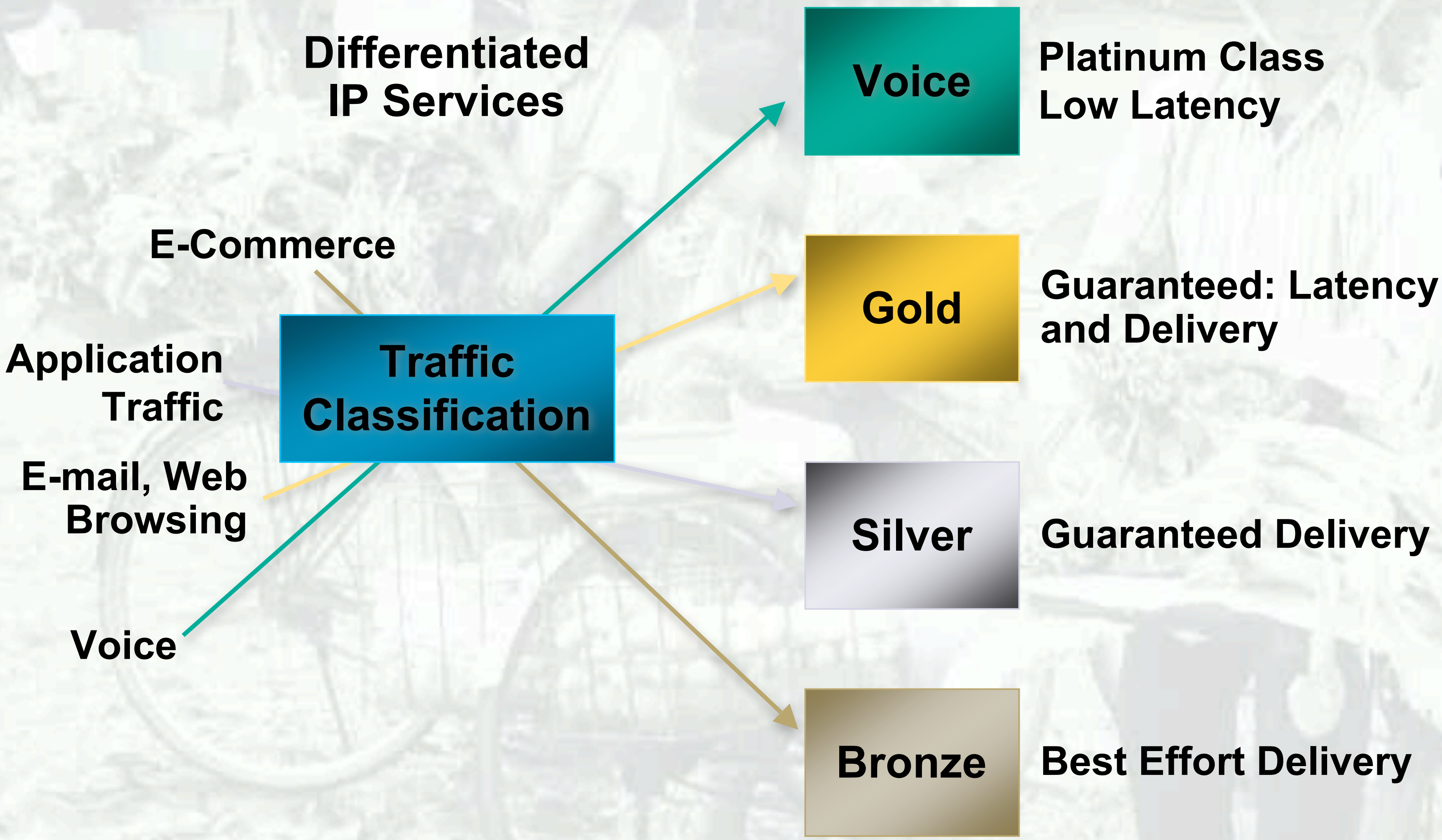
The real question is to choose which packets shall be dropped. The first definition of differential service is something like "not mine."
-- Christian Huitema

- 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



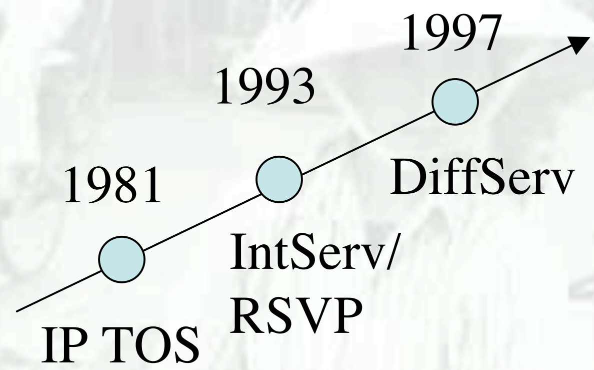
Borrowed from Cisco

Design Goals/Challenges

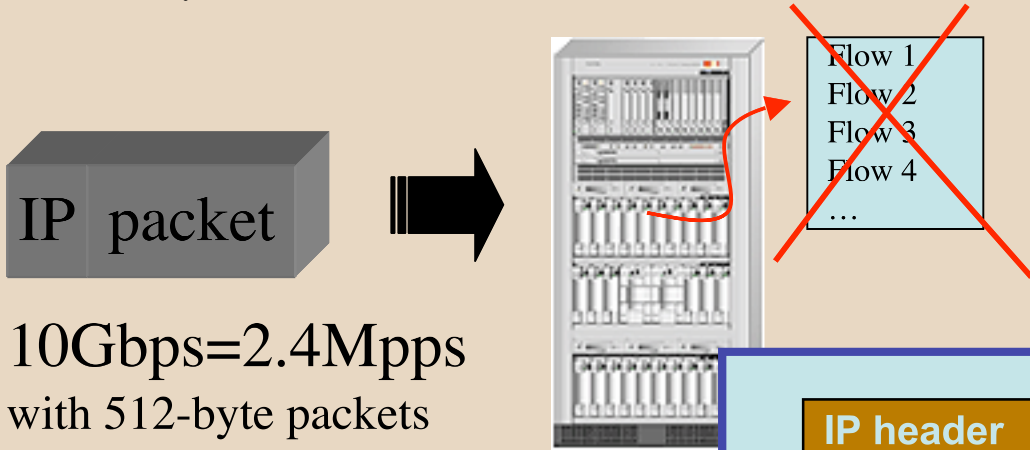
- Ability to charge differently for different services
- No per flow state or per flow signaling
- All policy decisions made at network boundaries
  - Boundary routers implement policy decisions by tagging packets with appropriate priority tag
- Traffic policing at network boundaries
- Deploy incrementally: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

IP implementation: DiffServ

RFC 2475

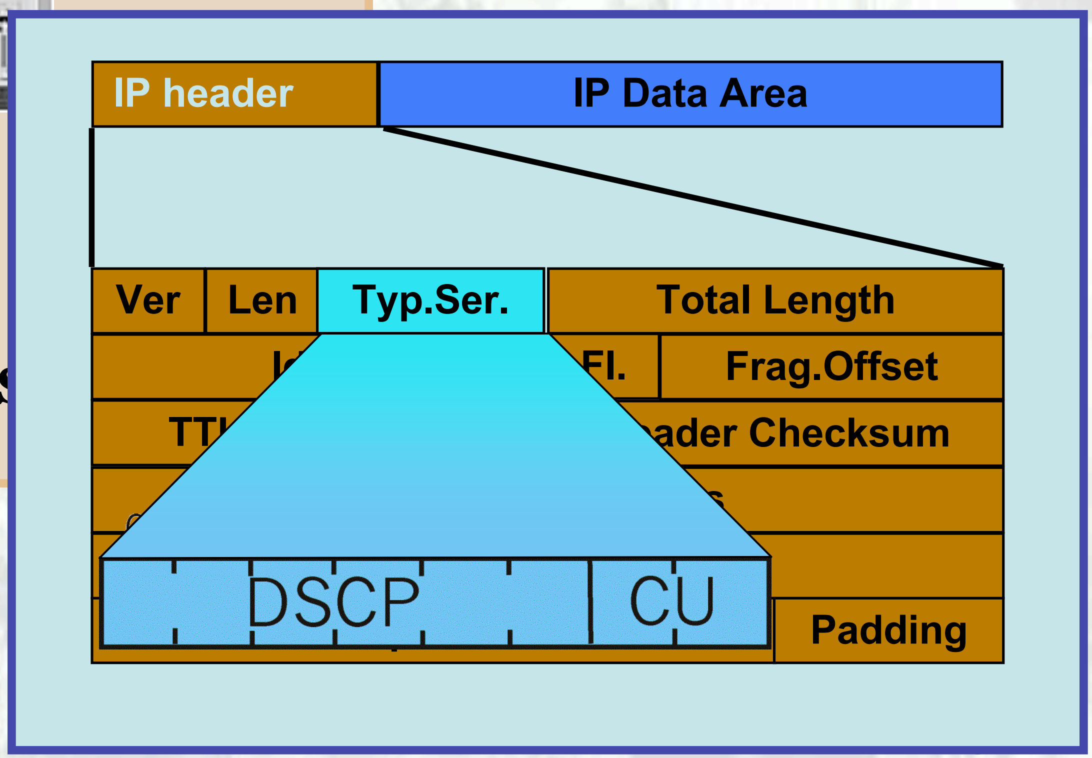


No per flow state in the core



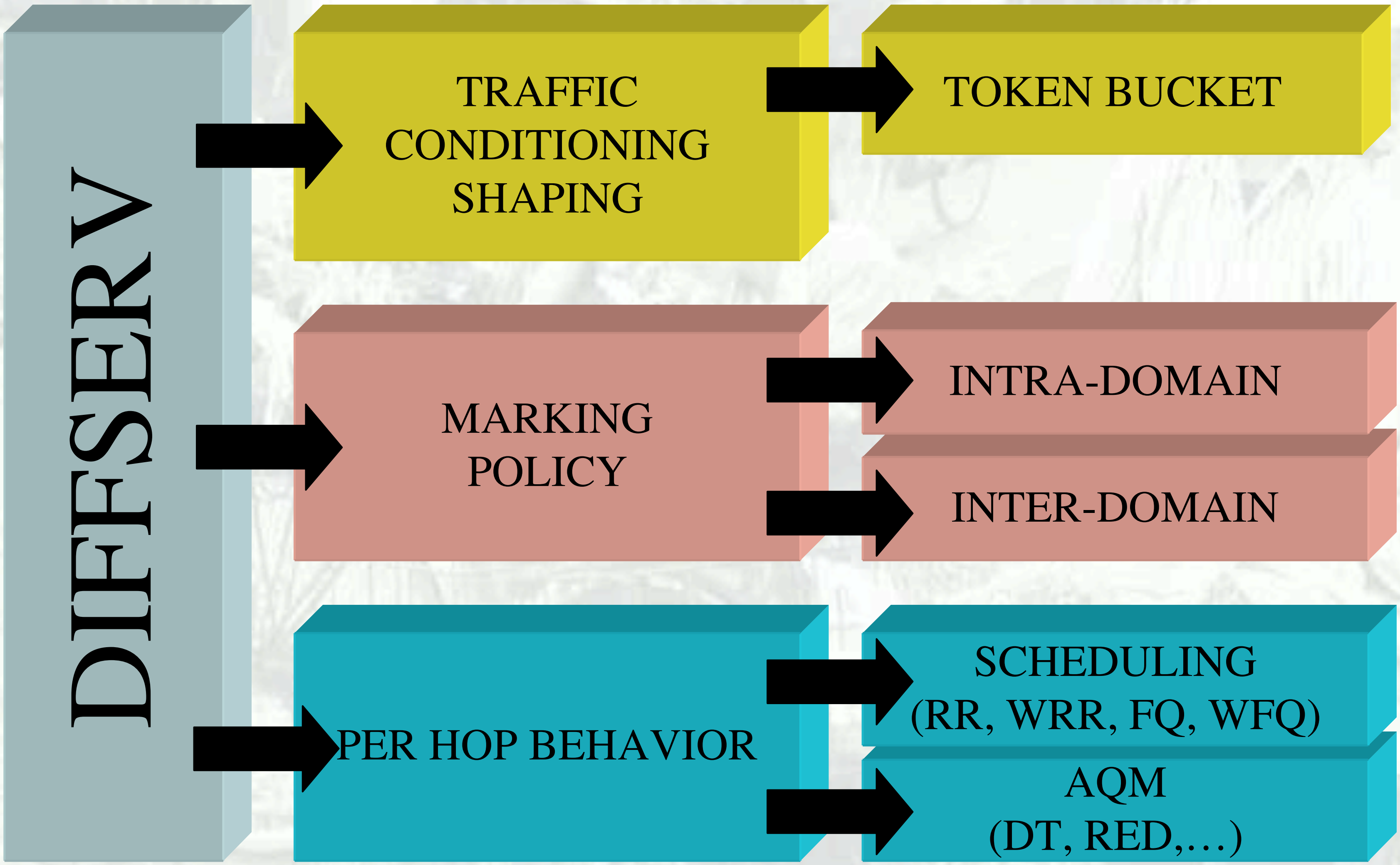
10Gbps=2.4Mpps
with 512-byte packets

Stateful approaches
scalable
at gigabit rates



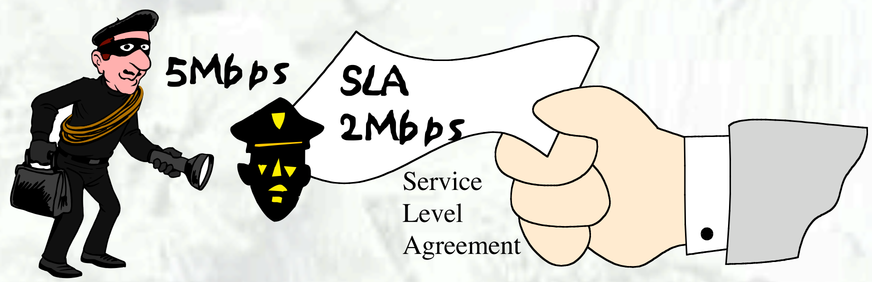
6 bits used for Differentiated Service
Code Point (DSCP) and determine PHB
that the packet will receive

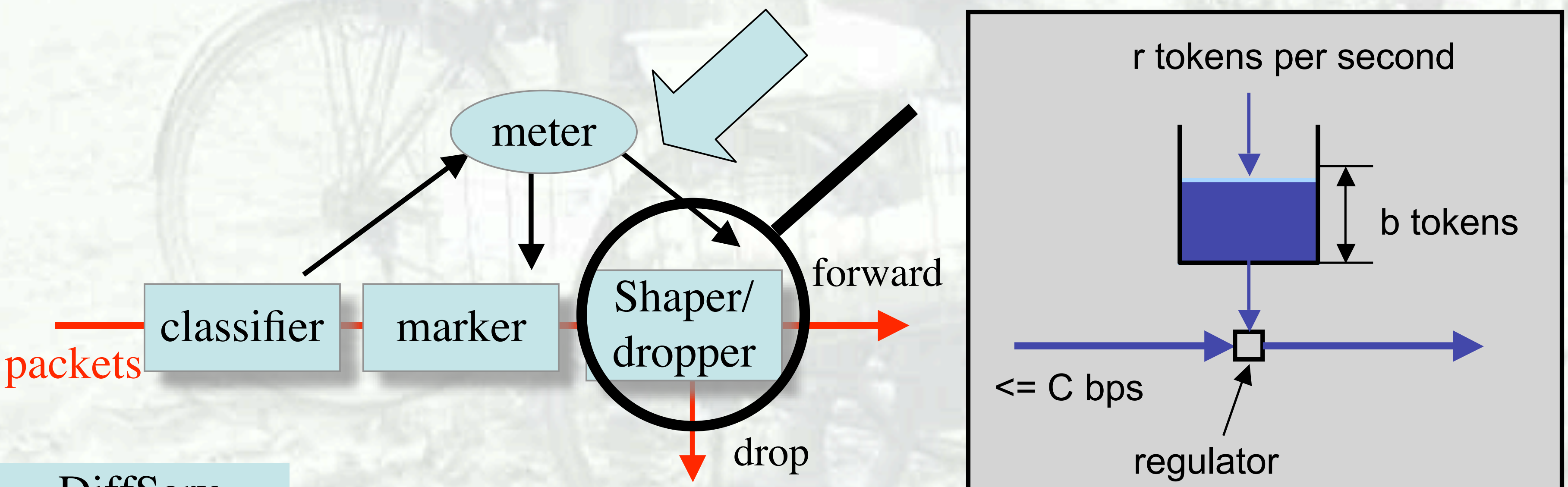
DiffServ building blocks



Traffic Conditioning

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

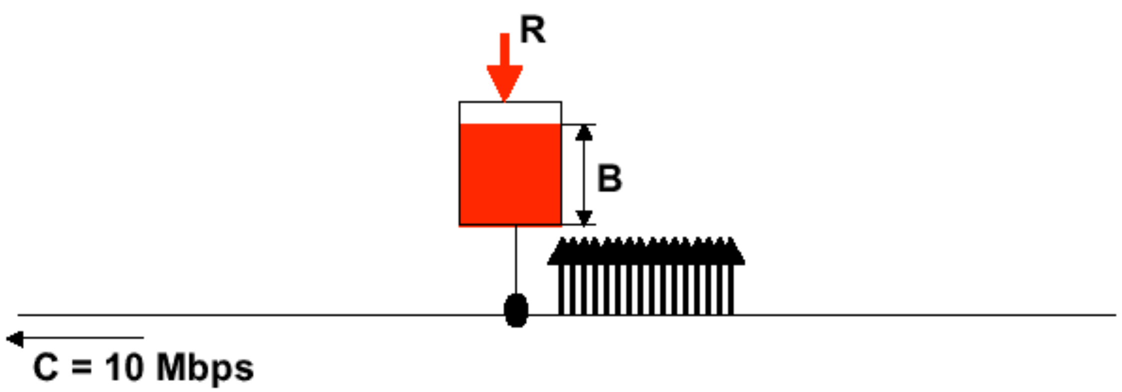




Token Bucket (1)

Example

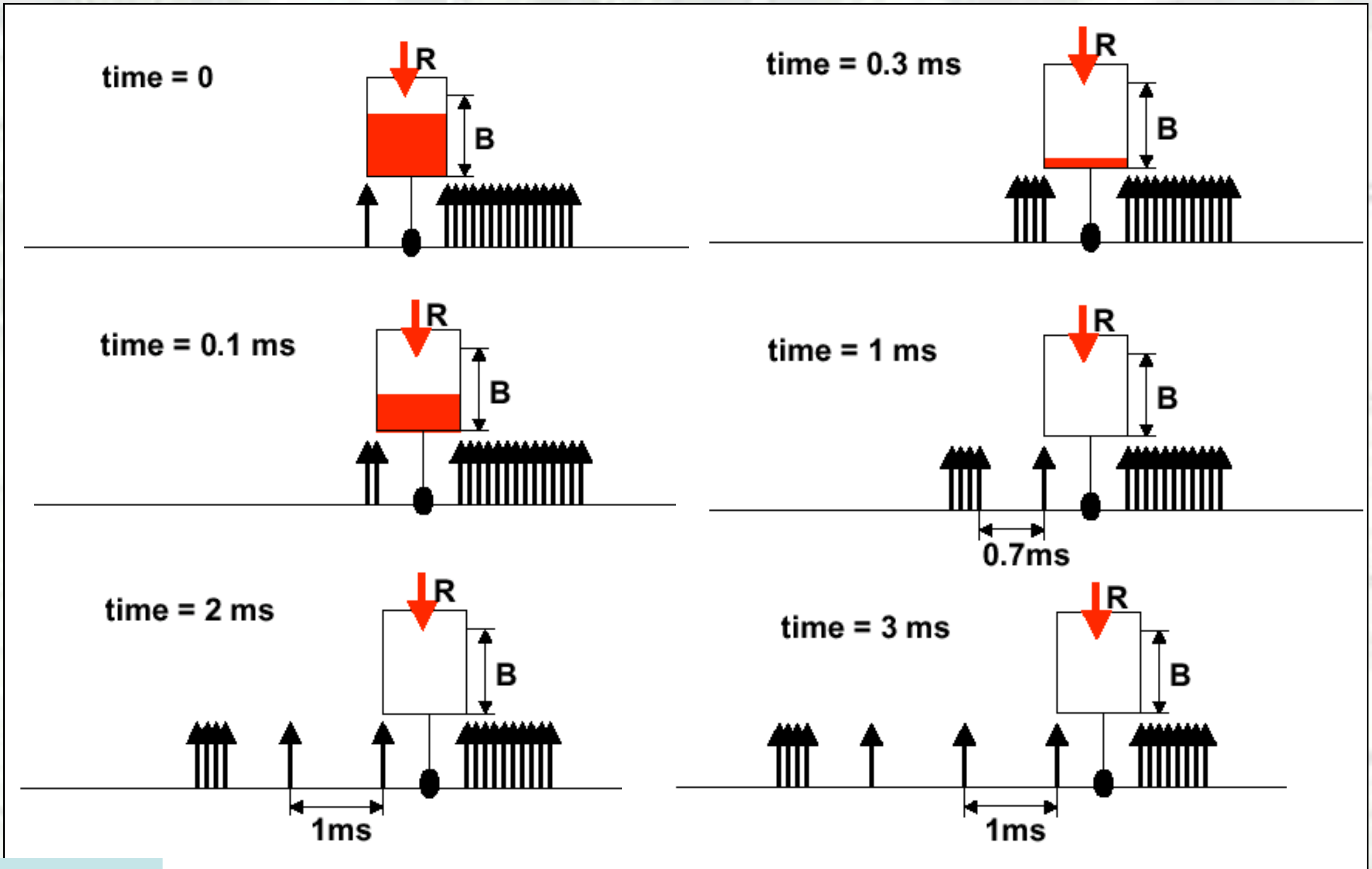
- B = 4000 bits, R = 1 Mbps, C = 10 Mbps
- Packet length = 1000 bits
- Assume the bucket is initially full and a "large" burst of packets arrives



istoica@cs.cmu.edu

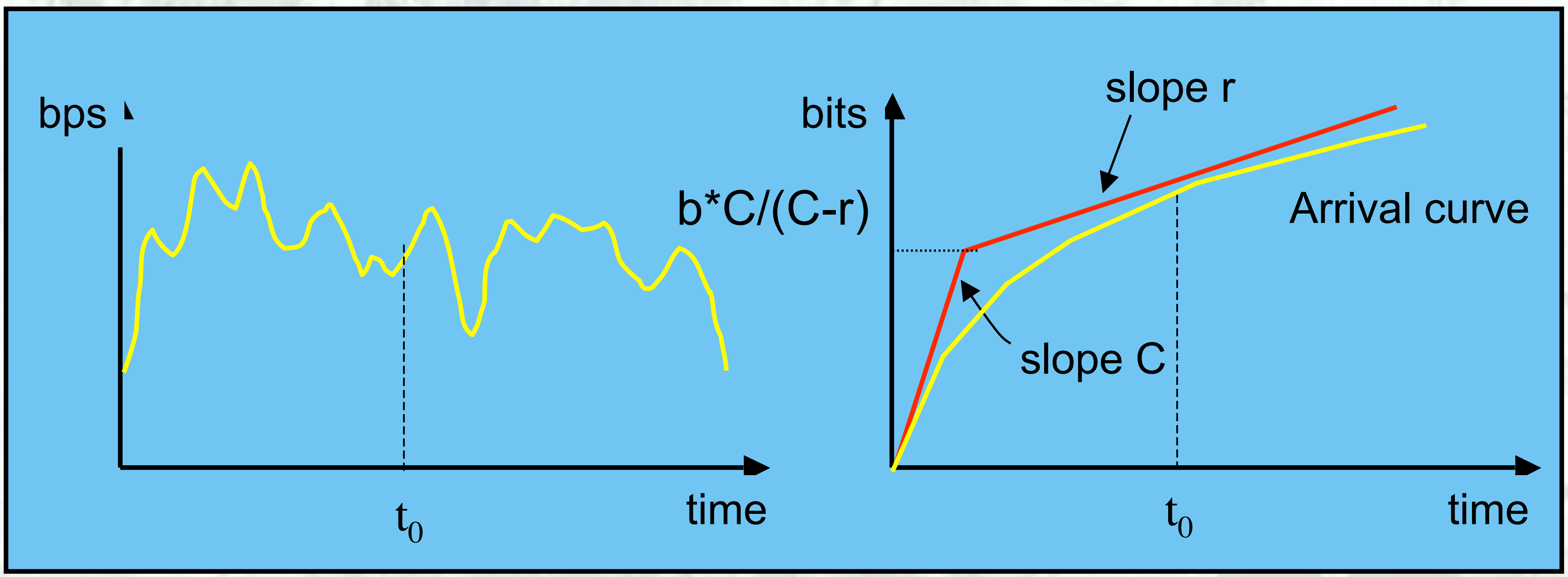
Token Bucket (2)

B=4000 bits, R=1Mbps, C=10Mbps

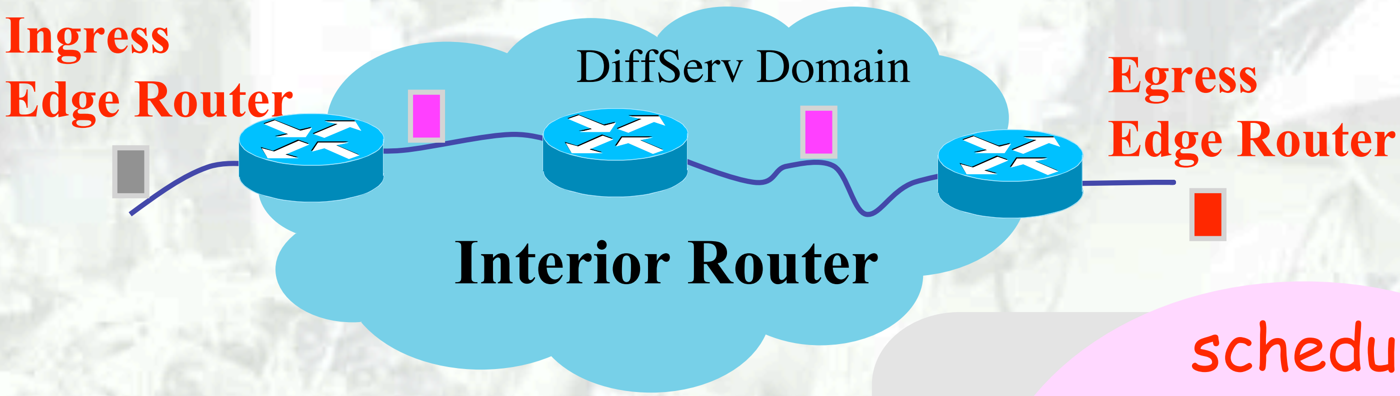


Token Bucket for traffic characterization

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



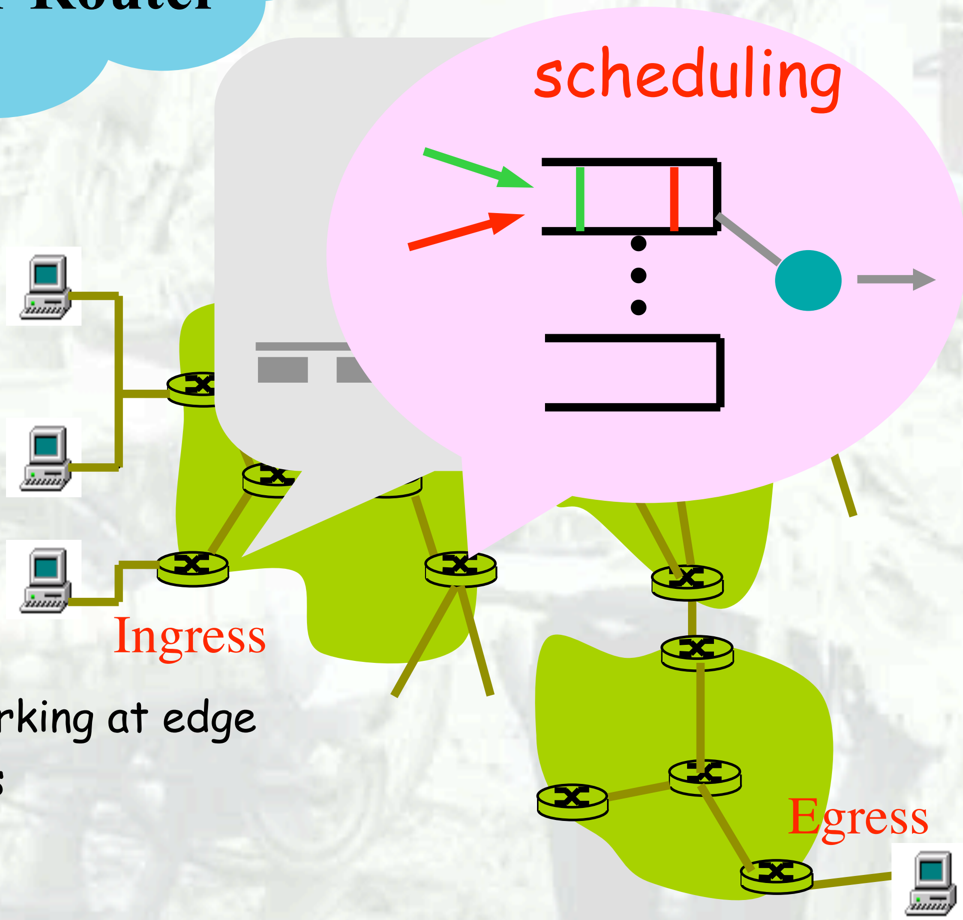
Differentiated Architecture



Marking:

per-flow traffic management

marks packets as in-profile and out-profile



Per-Hop-Behavior (PHB):

per class traffic management

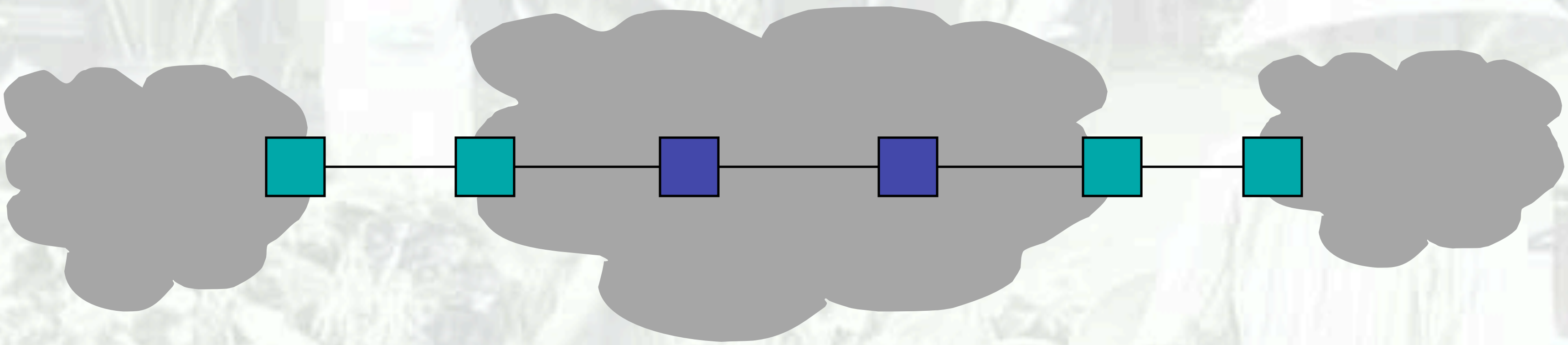
buffering and scheduling based on marking at edge

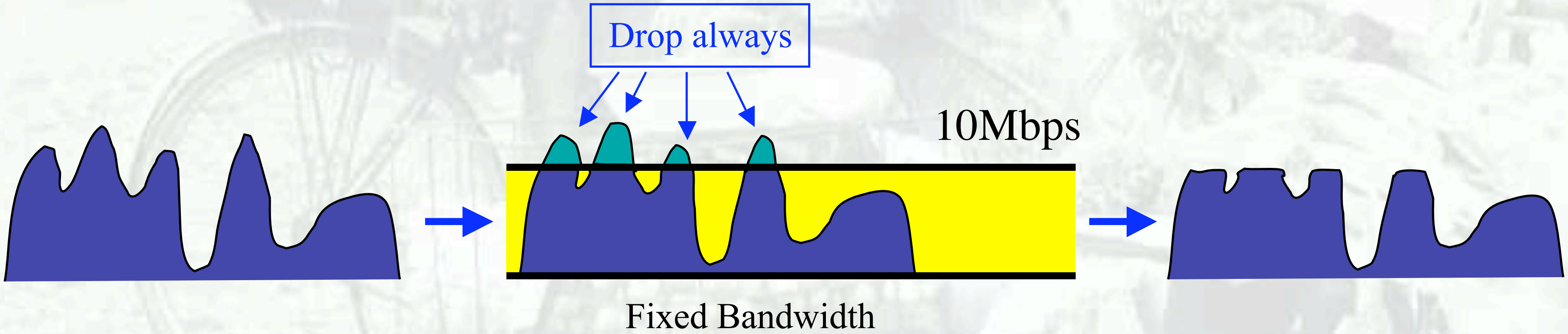
preference given to in-profile packets

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

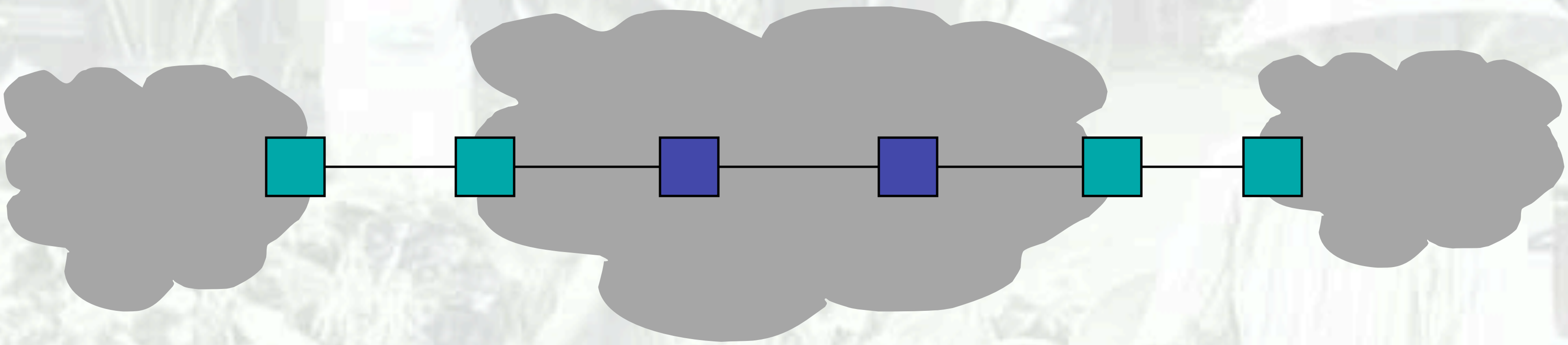
Premium Service Example

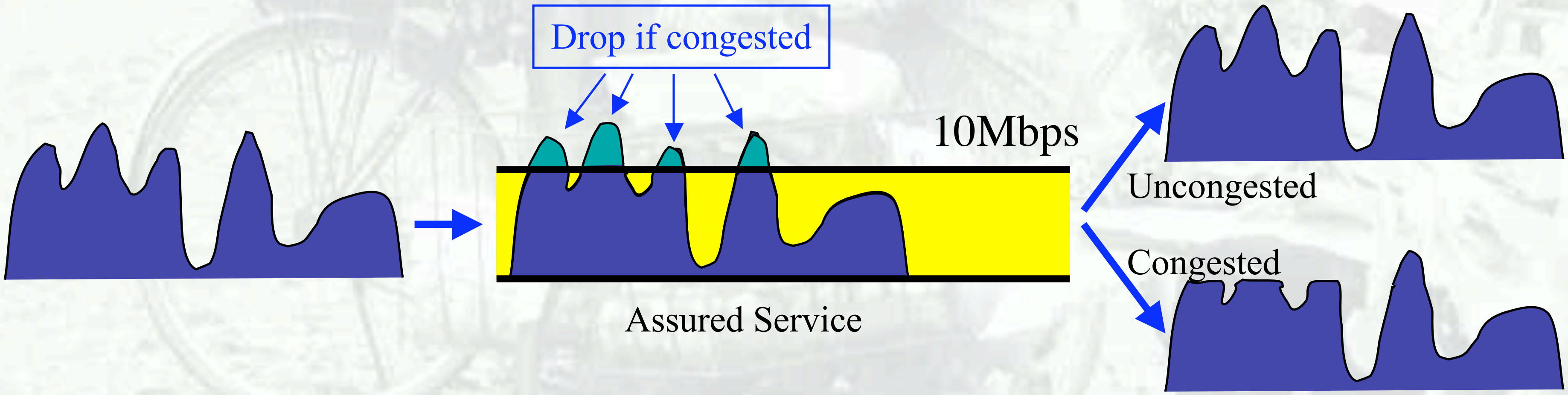




source Gordon Schaffee

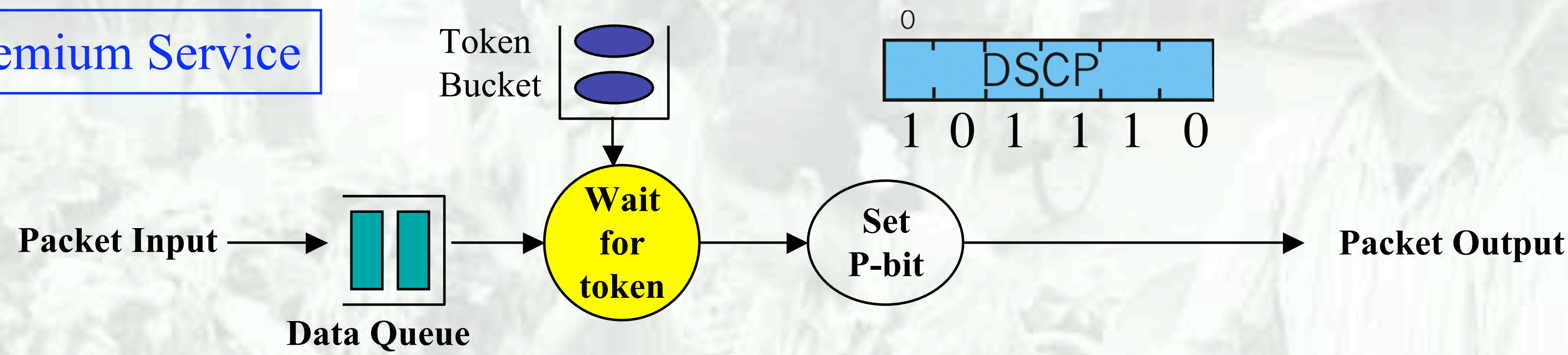
Assured Service Example





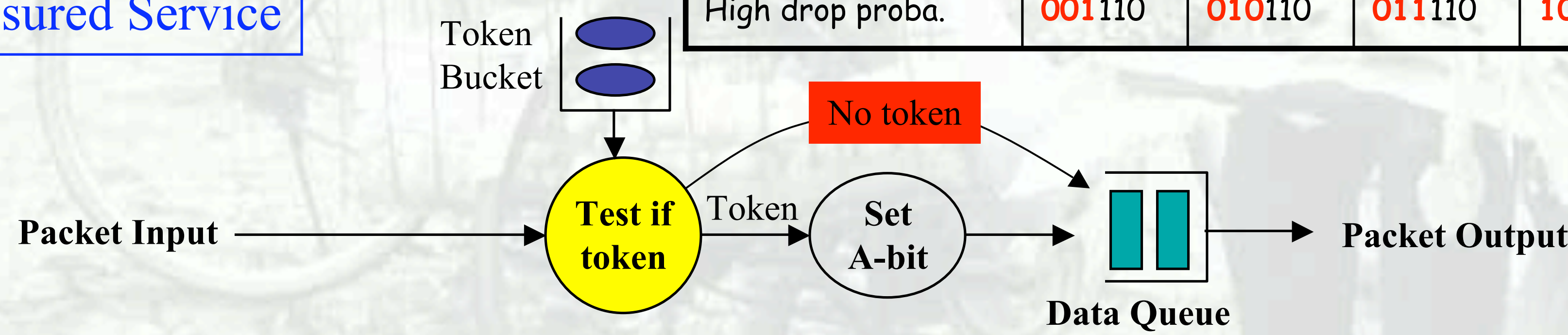
Border Router Functionality

Premium Service



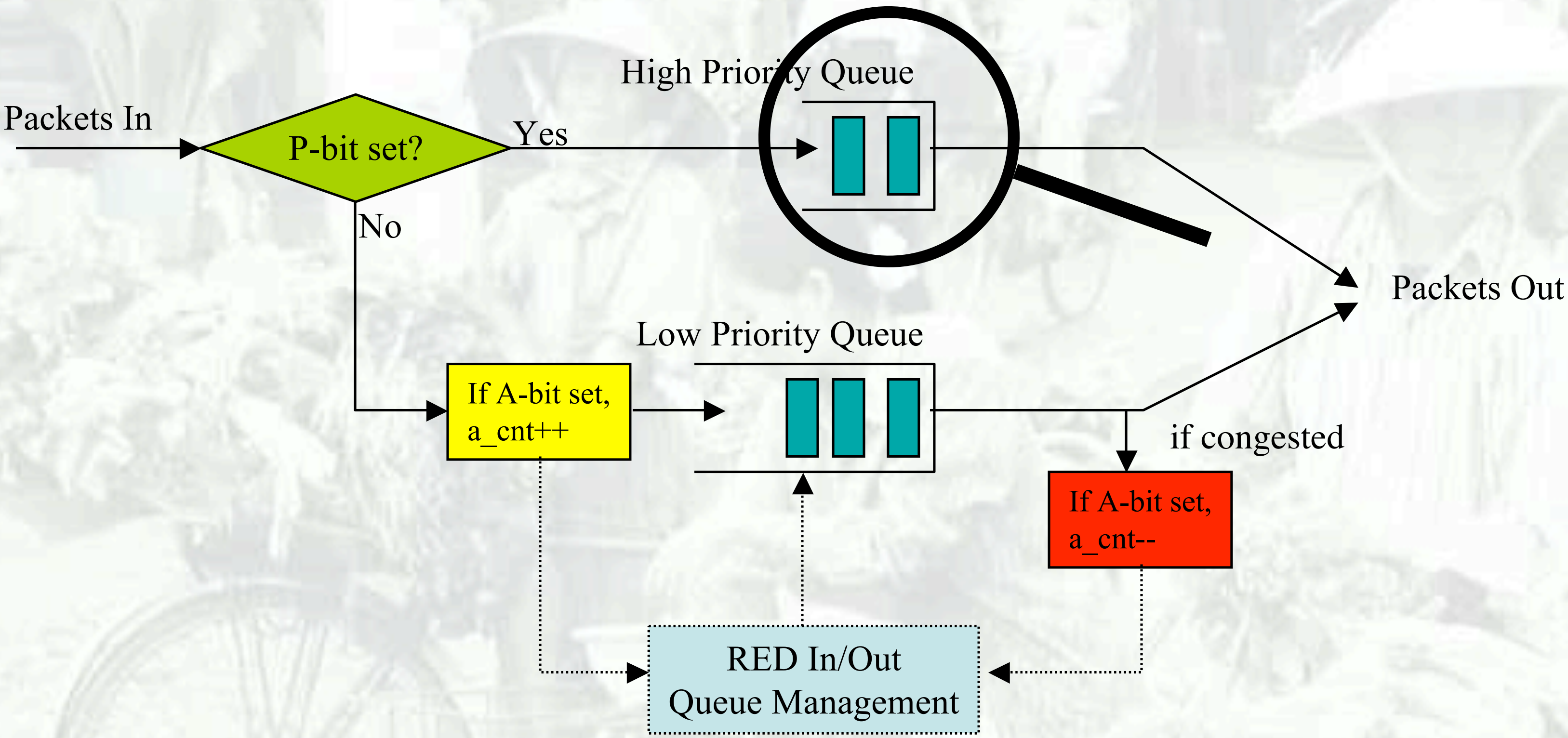
DSCP	Class 1	Class 2	Class 3	Class 4
Low drop probability	001010	010010	011010	100010
Medium drop proba.	001100	010100	011100	100100
High drop proba.	001110	010110	011110	100110

Assured Service



source Gordon Schaffee, modified by C. Pham

Internal Router Functionality

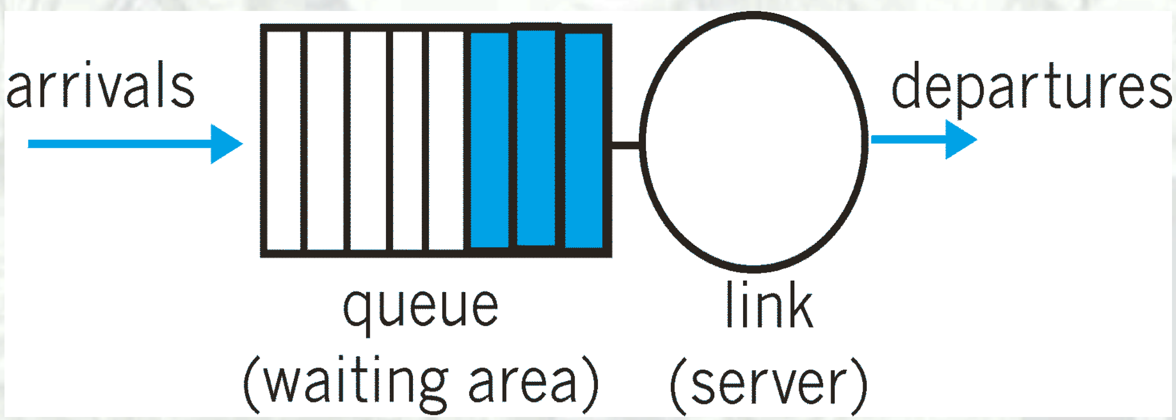


A DSCP codes aggregates, not individual flows
No state in the core
Should scale to millions of flows

source Gordon Schaffee, modified by C. Pham

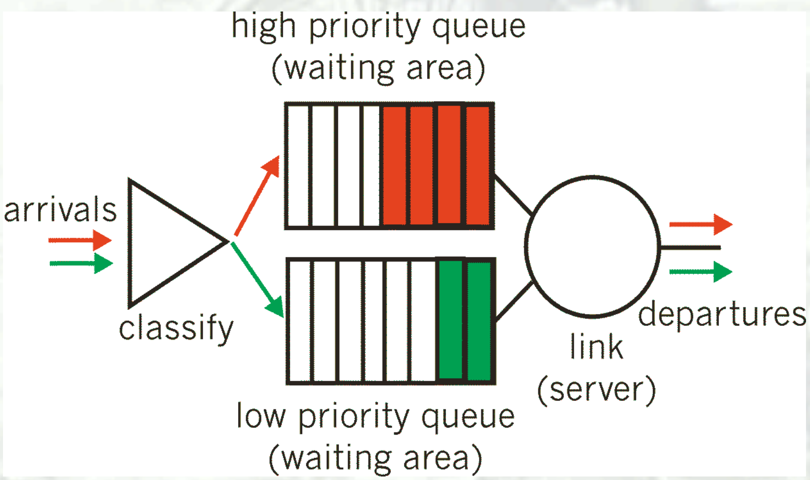
Scheduling

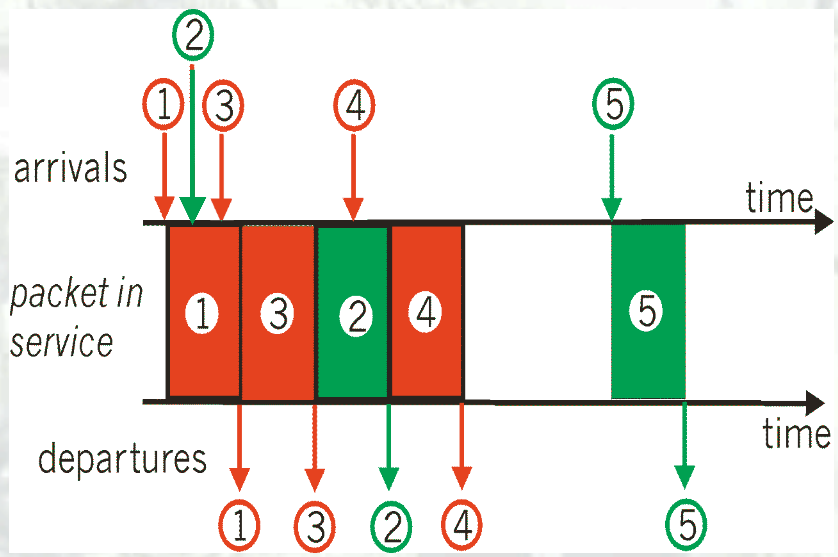
- DiffServ PHB relies mainly on scheduling
  - choose the next packet for transmission
  - FIFO: in order of arrival to the queue; packets that arrive to a full buffer are either discarded, or a discard policy is defined.
  - More complex policies: FCFS, PRIORITY, EDD…



Priority Queueing

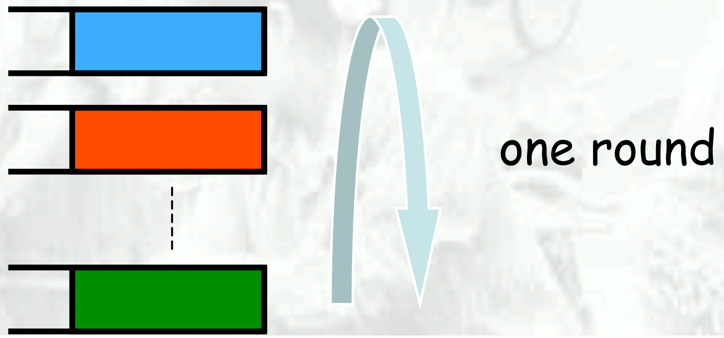
- Priority Queuing: classes have different priorities;
- Transmit a packet from the highest priority class with a non-empty queue
- Preemptive and non-preemptive versions

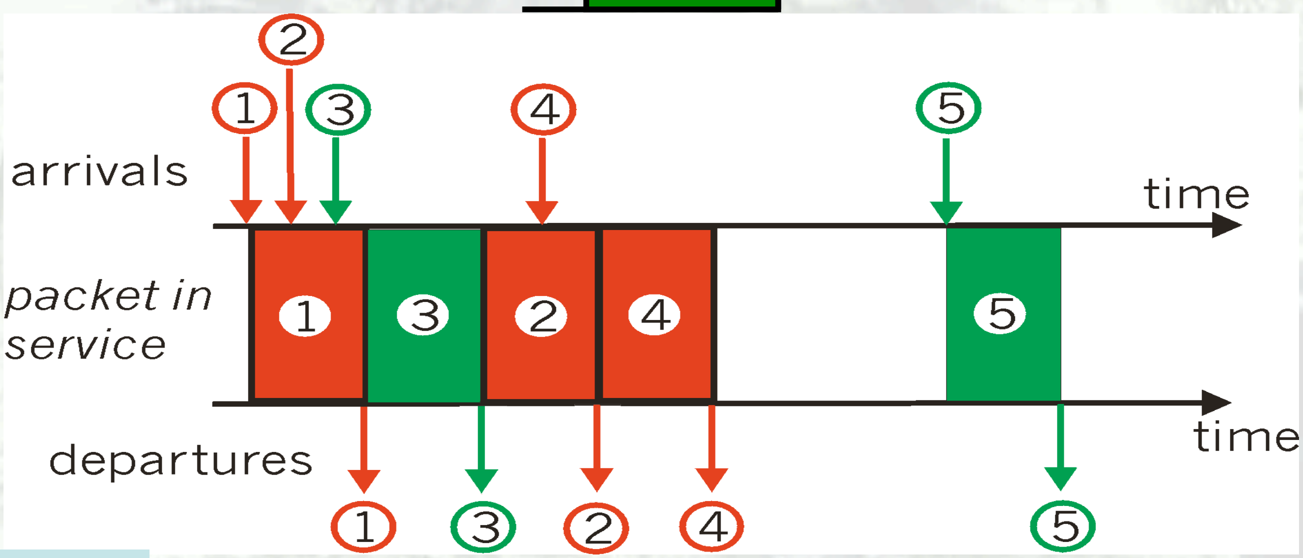




Round Robin (RR)

- Round Robin: scan class queues serving one from each class that has a non-empty queue



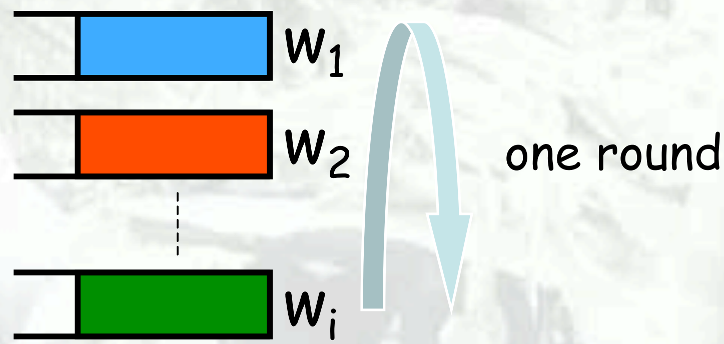


Weighted Round Robin, WRR

- Assign a weight to each connection and serve a connection in proportion to its weight

Connection A, B and C with same packet size and weight 0.5, 0.75 and 1. How many packets from each connection should a round-robin server serve in each round?

A: Normalize each weight so that they are all integers: we get 2, 3 and 4. Then in each round of service, the server serves 2 packets from A, 3 from B and 4 from C.

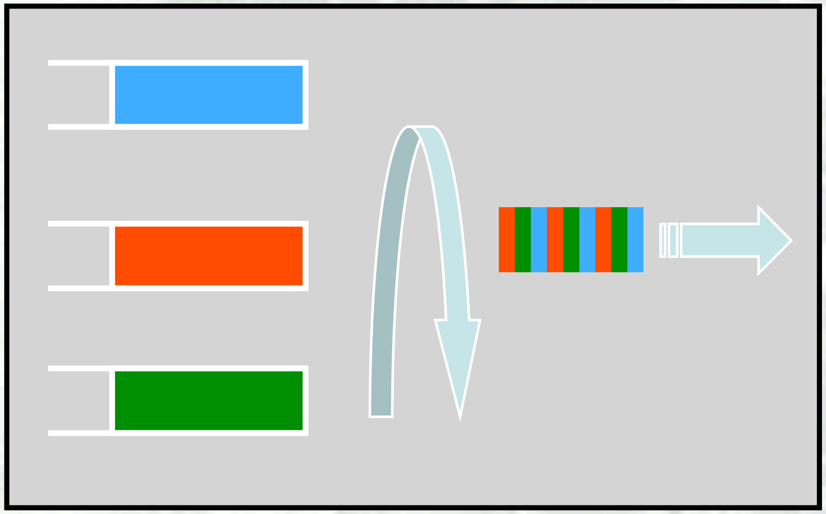


(Weighted) Round-Robin Discussion

- Advantages: protection among flows
  - Misbehaving flows will not affect the performance of well-behaving flows
  - FIFO does not have such a property
- Disadvantages:
  - More complex than FIFO: per flow queue/state
  - Biased toward large packets: a flow receives service proportional to the number of packets
- If packet size are different, we normalize the weight by the packet size
  - ex: 50, 500 & 1500 bytes with weight 0.5, 0.75 & 1.0

Generalized Processor Sharing (GPS)

- Assume a fluid model of traffic
  - Visit each non-empty queue in turn (like RR)
  - Serve infinitesimal from each
  - Leads to “max-min” fairness
- GPS is un-implementable!
  - We cannot serve infinitesimals, only packets



max-min fairness

Let n sources requiring resources $x_1,..,x_n$ with $x_1<x_2..<x_n$ for instance. Server has a capacity of C.

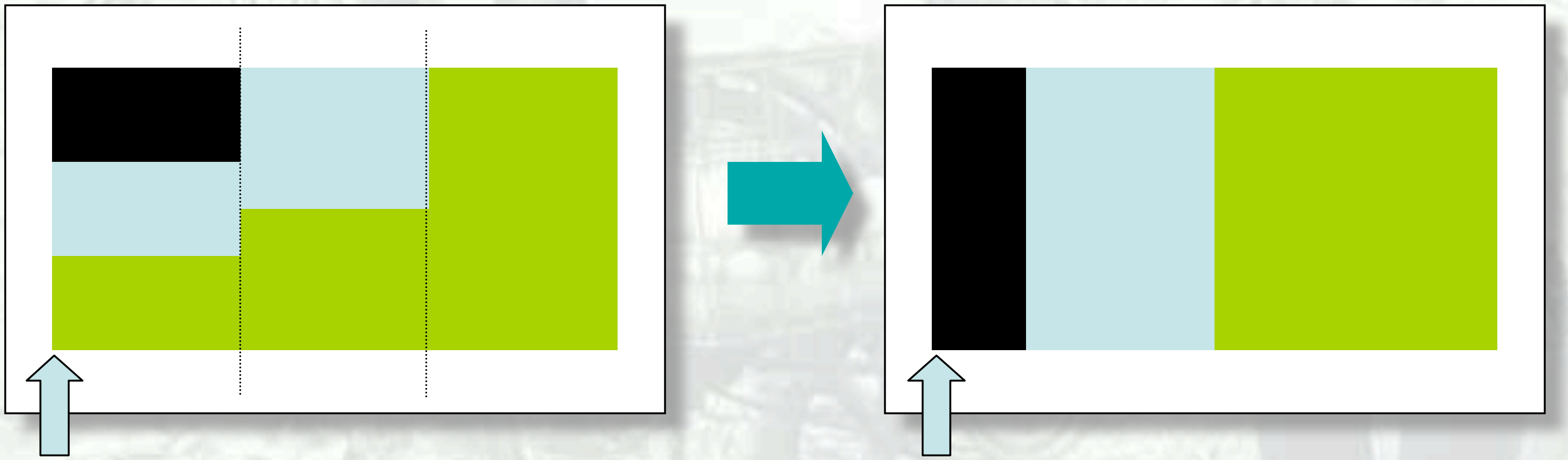
We assign C/n to source 1. If $C/n>x_1$, give $C/n+(C/n-x_1)/(n-1)$ to the (n-1) remaining sources. If this amount is greater than x_2, process again.

Packet Approximation of Fluid System

- GPS un-implementable
- Standard techniques of approximating fluid GPS
  - Select packet that finishes first in GPS assuming that there are no future arrivals (emulate GPS on the side)
- Important properties of GPS
  - Finishing order of packets currently in system independent of future arrivals
- Implementation based on virtual time
  - Assign virtual finish time to each packet upon arrival
  - Packets served in increasing order of virtual times

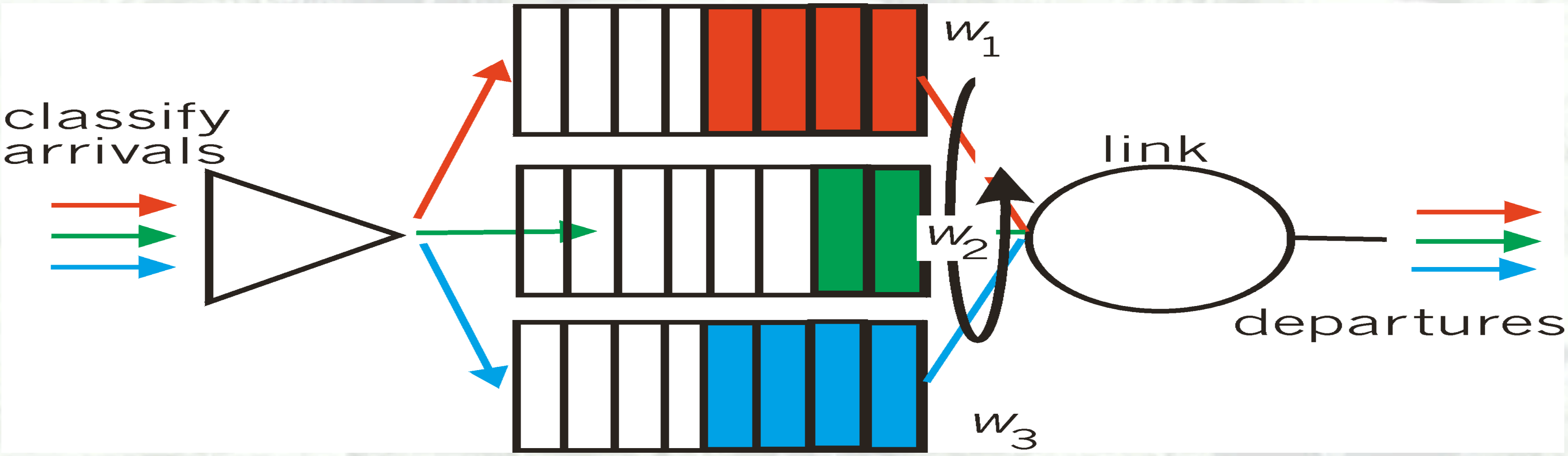
Fair Queuing (FQ)

- Idea: serve packets in the order in which they would have finished transmission in the fluid flow system
- Mapping bit-by-bit schedule onto packet transmission schedule
- Transmit packet with the lowest finish time at any given time



Weighted Fair Queueing

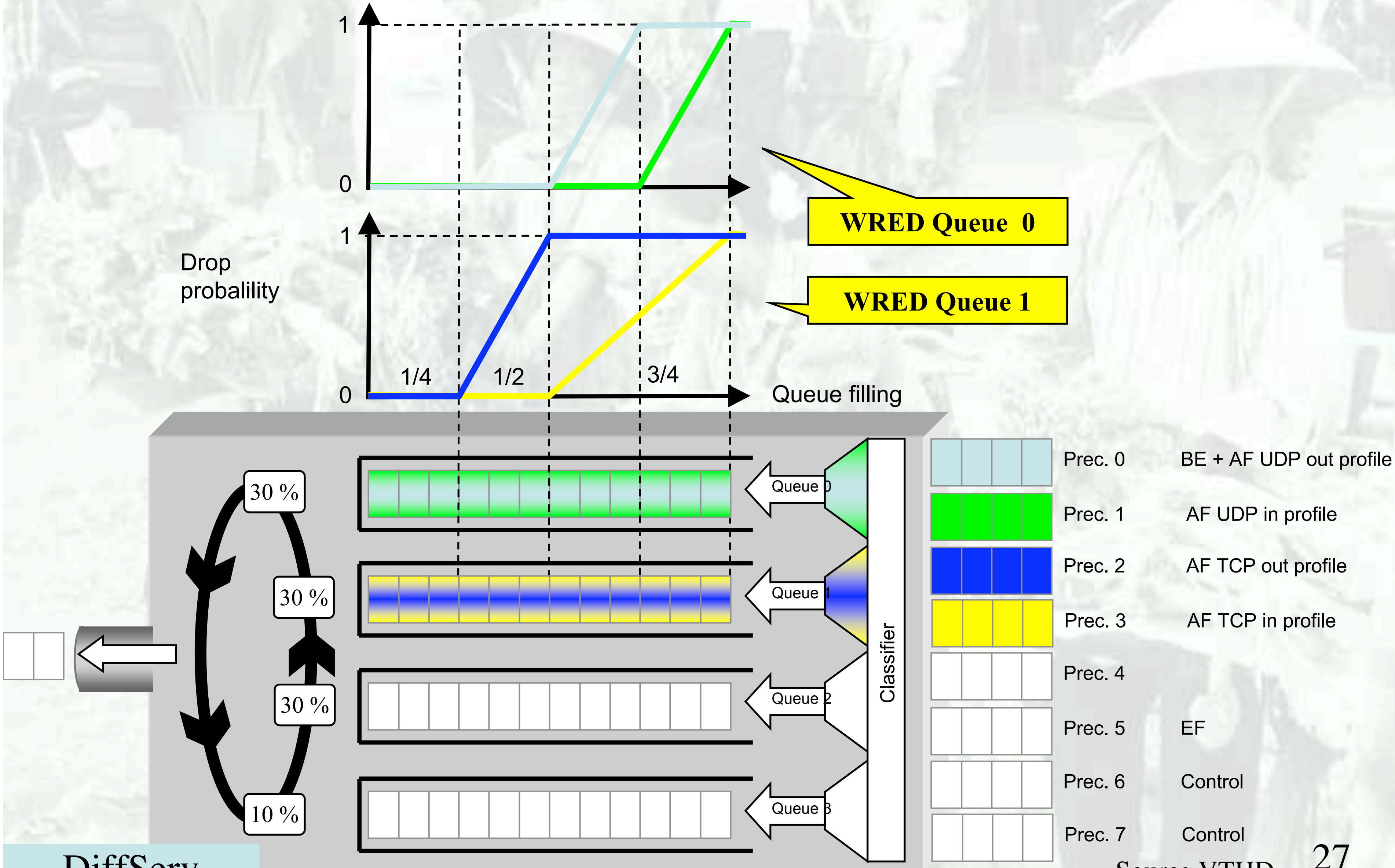
- Variation of FQ: Weighted Fair Queuing (WFQ)
- Weighted Fair Queuing: is a generalized Round Robin in which an attempt is made to provide a class with a differentiated amount of service over a given period of time



Implementing WFQ

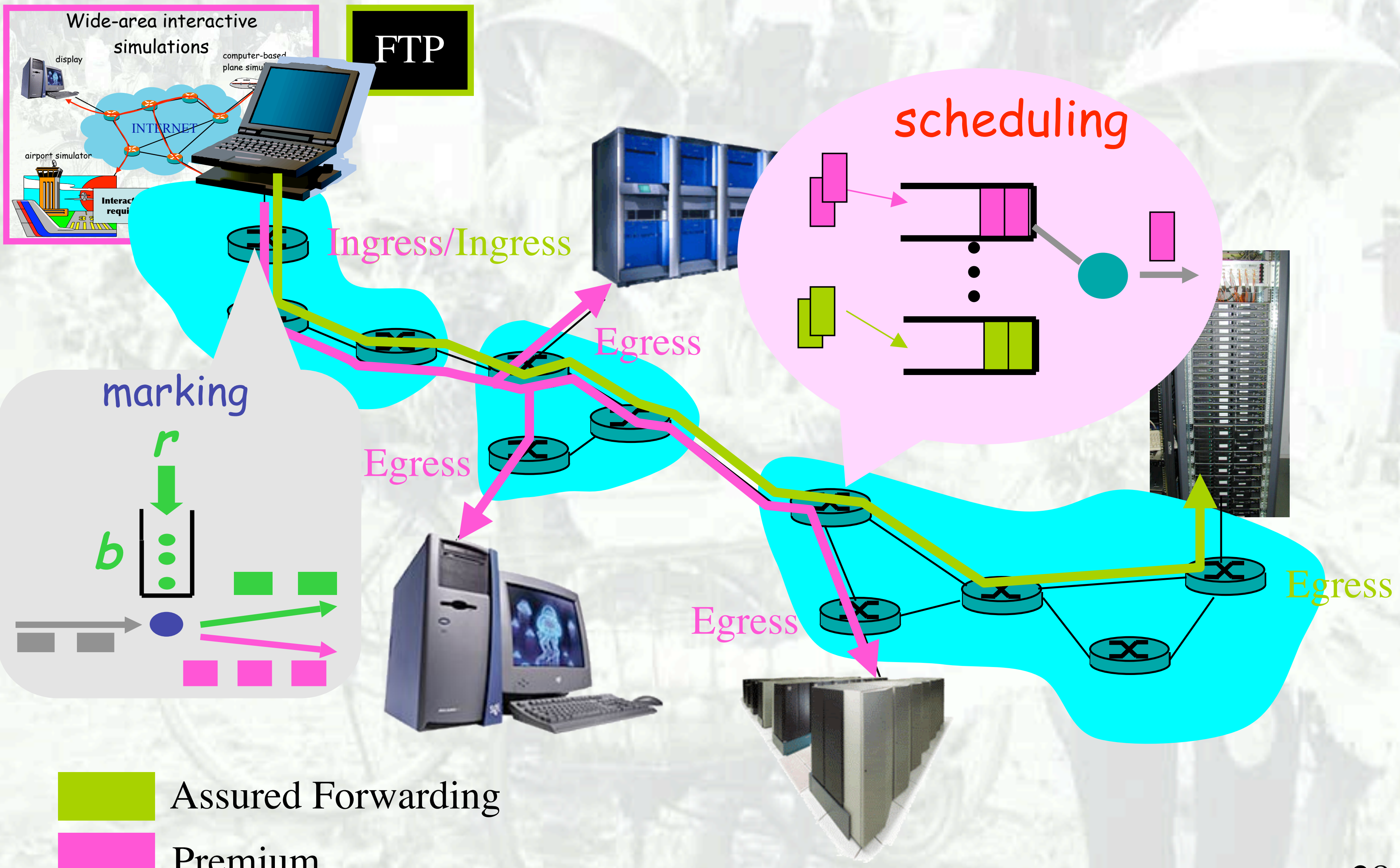
- WFQ needs per-connection (or per-aggregate) scheduler state→implementation complexity.
  - complex iterated deletion algorithm
  - complex sorting at the output queue on the service tag
- WFQ needs to know the weight assigned for each queue →manual configuration, signalling.
- WFQ is not perfect…
- Router manufacturers have implemented as early as 1996 WFQ in their products
  - from CISCO 1600 series
  - Fore System ATM switches

Putting it together!

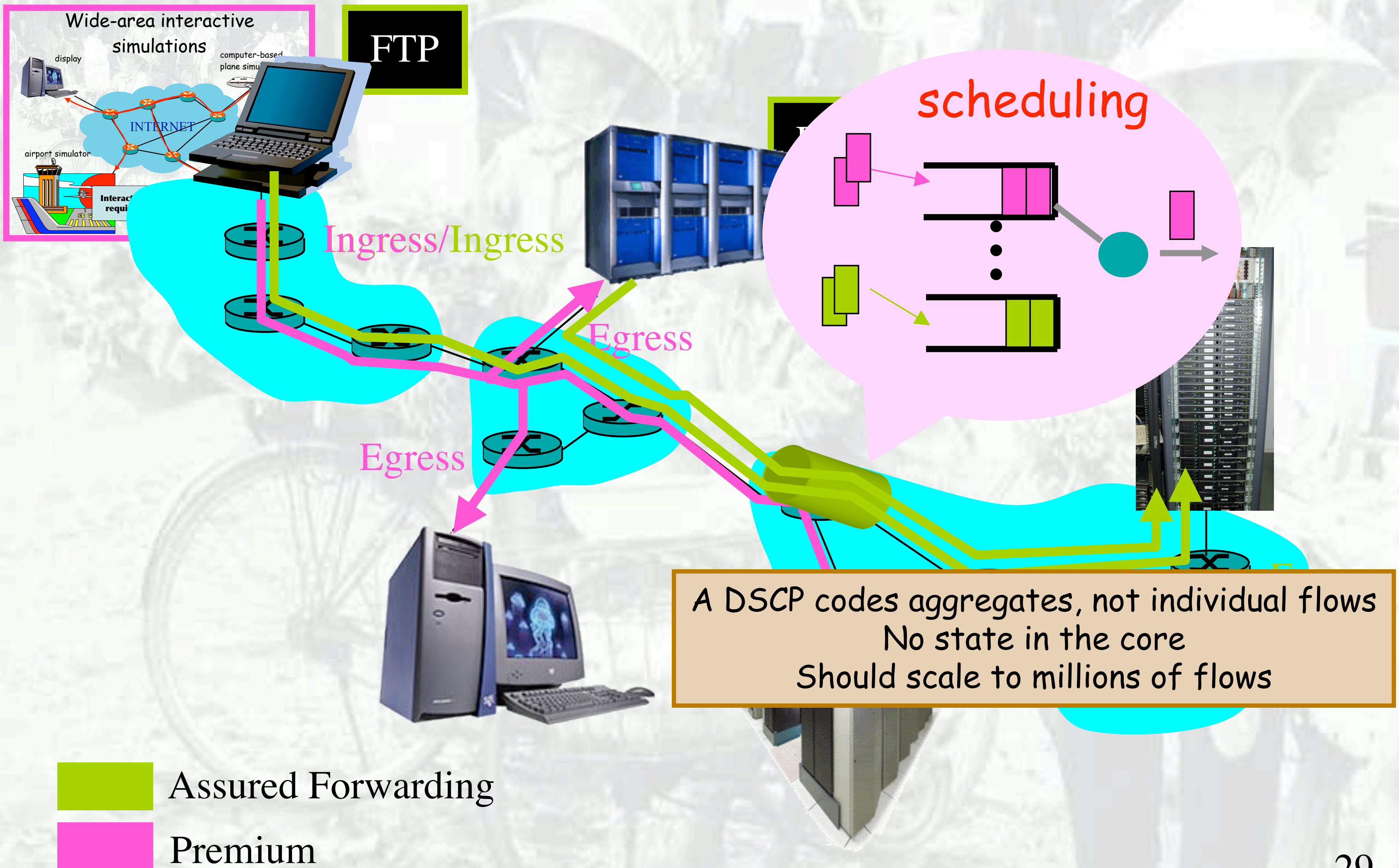


Source VTHD

DiffServ for grids

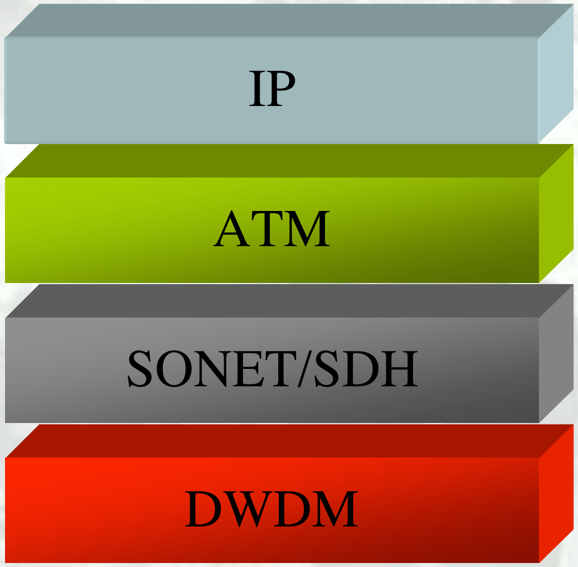


DiffServ for grids (con't)



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

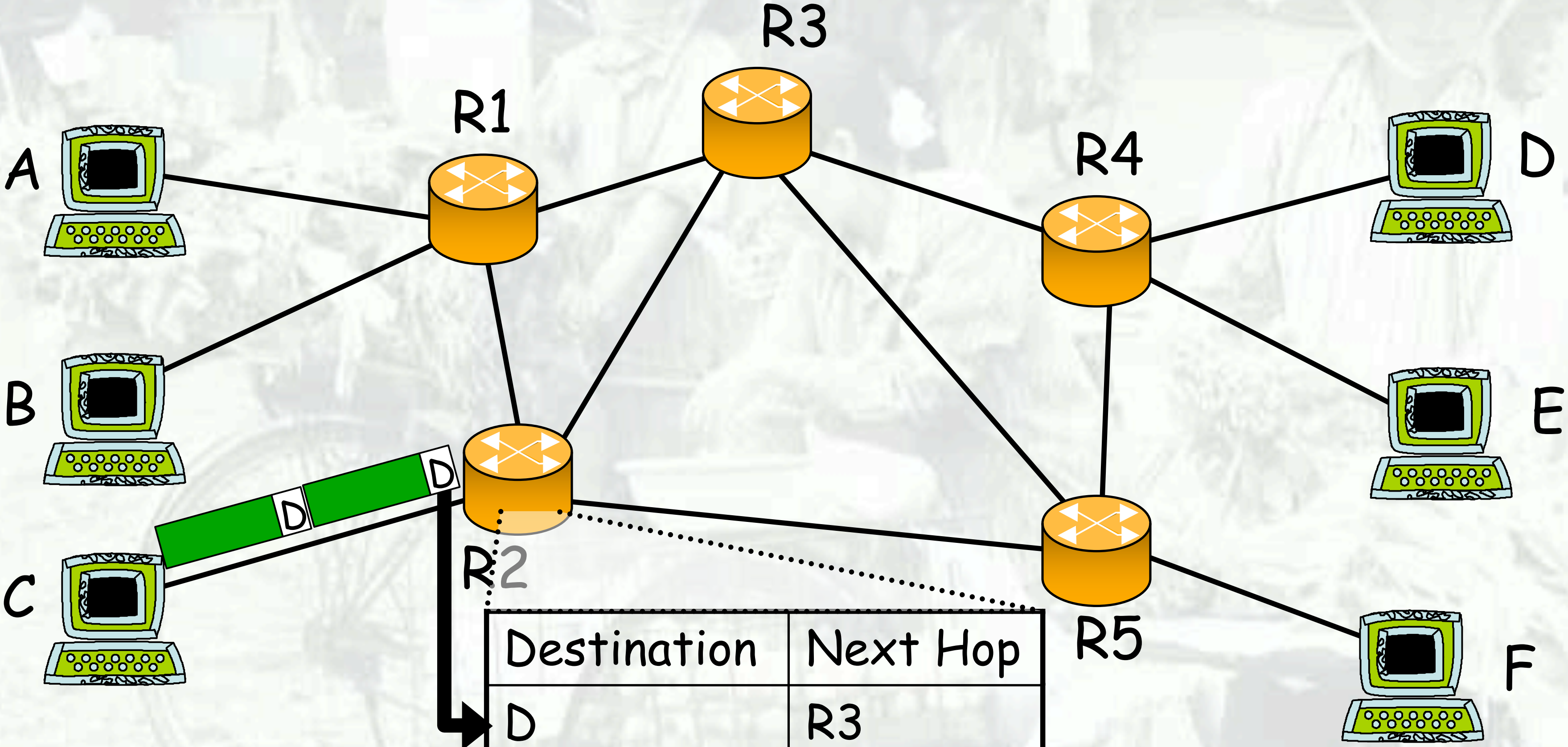


NEW CHAPTER

Provider's view

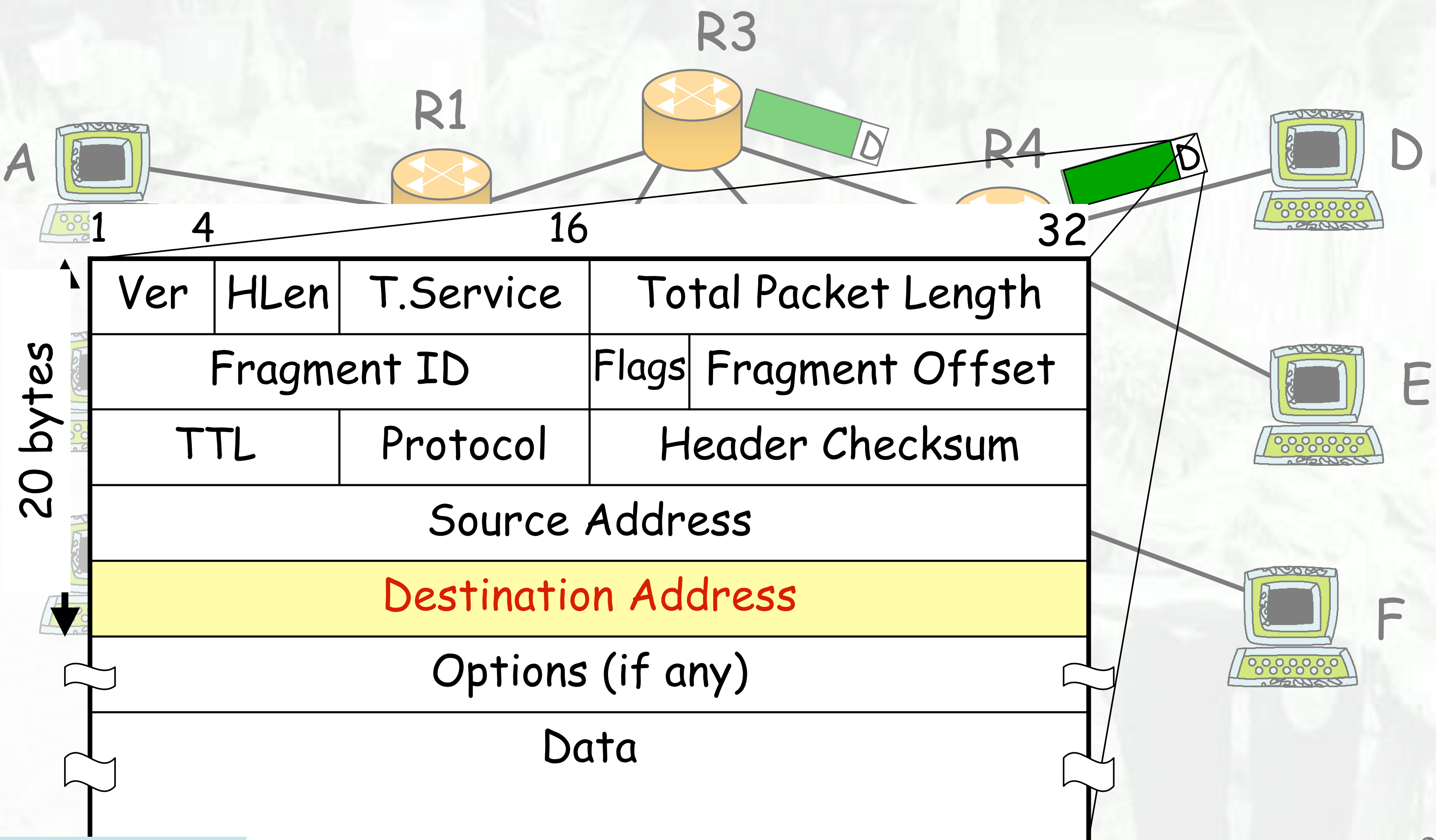
Today's setting time is several weeks/months! We want to set dynamic links within hours

Review of IP routing



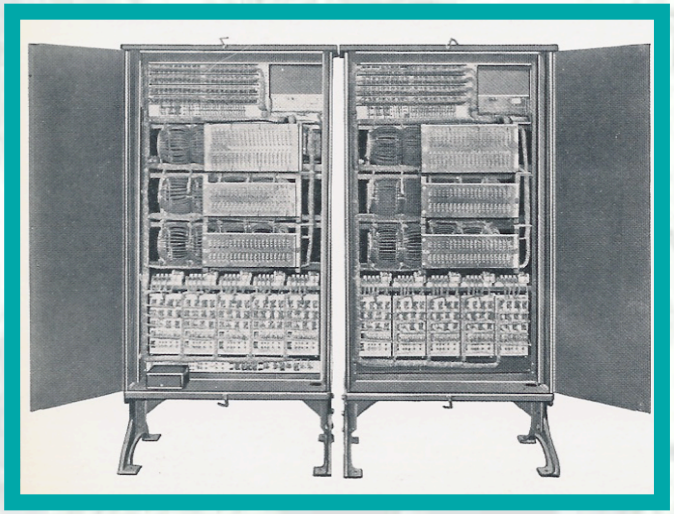
Destination	Next Hop
D	R3
E	R3
F	R5

Review of IP routing



Review of telephone network

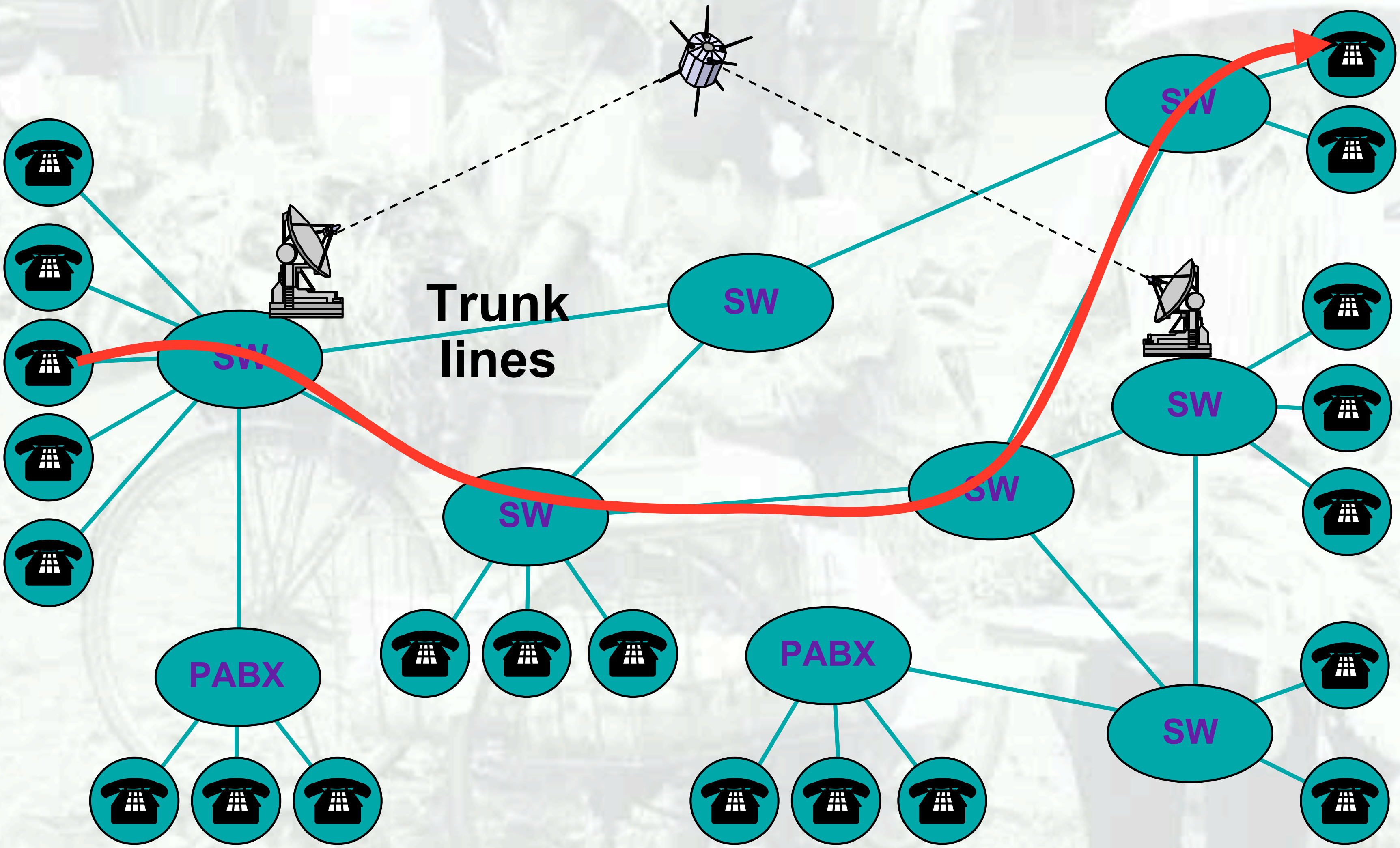
First automatic Branch Exchange Almond B. Strowger, 1891…



Signaling replaces the operator

Source J. Tiberghien, VUB

The telephone circuit view



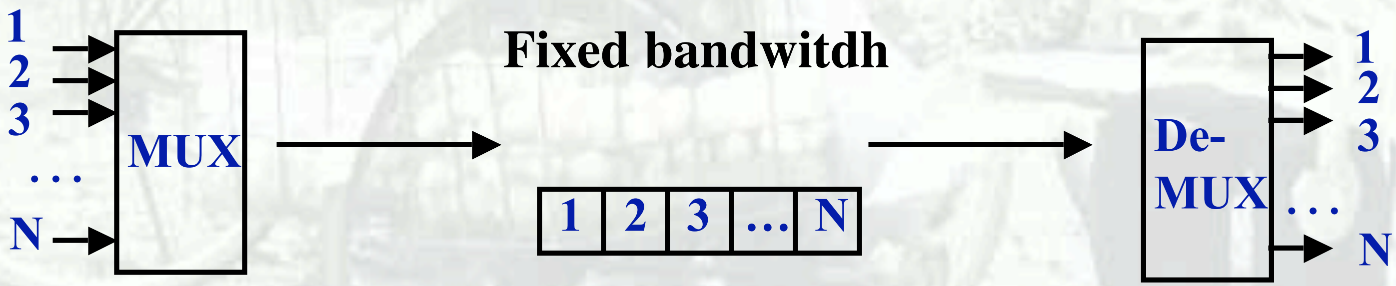
Advantages of circuits

- 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

Time Division Circuits

- Most trunks time division multiplex voice samples
- At a central office, trunk is demultiplexed and distributed to active circuits
- Synchronous multiplexor
  - N input lines
  - Output runs N times as fast as input

Simple, efficient, but low flexibility and wastes resources

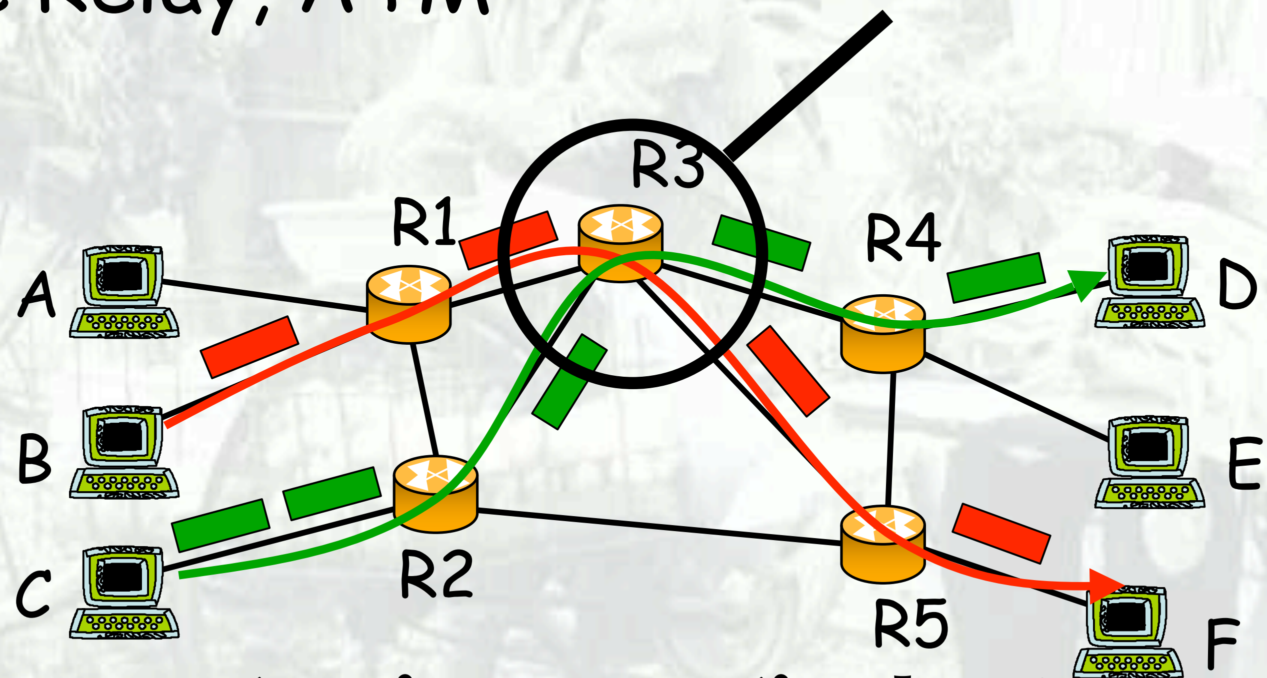


1 sample every 125us gives a 64Kbits/s channel

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

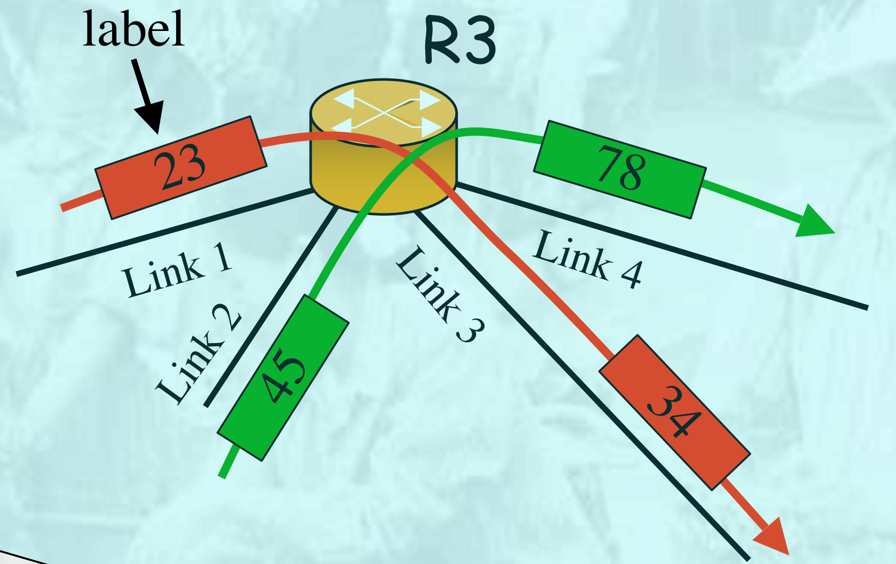


But IP is connectionless!

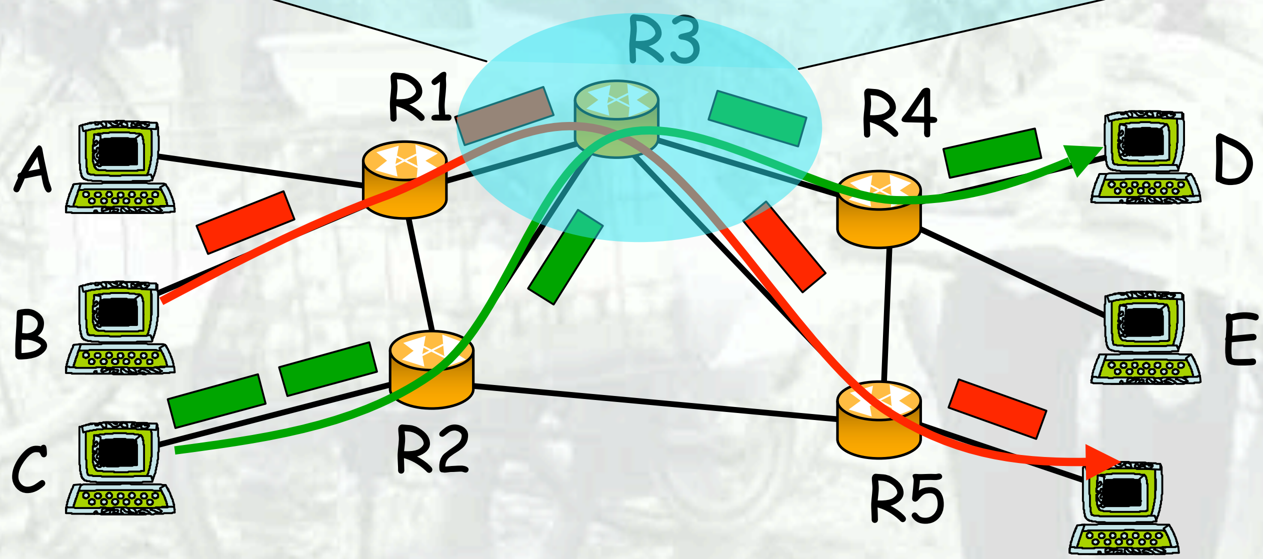
Virtual circuit principles

Connections &
Virtual circuits table

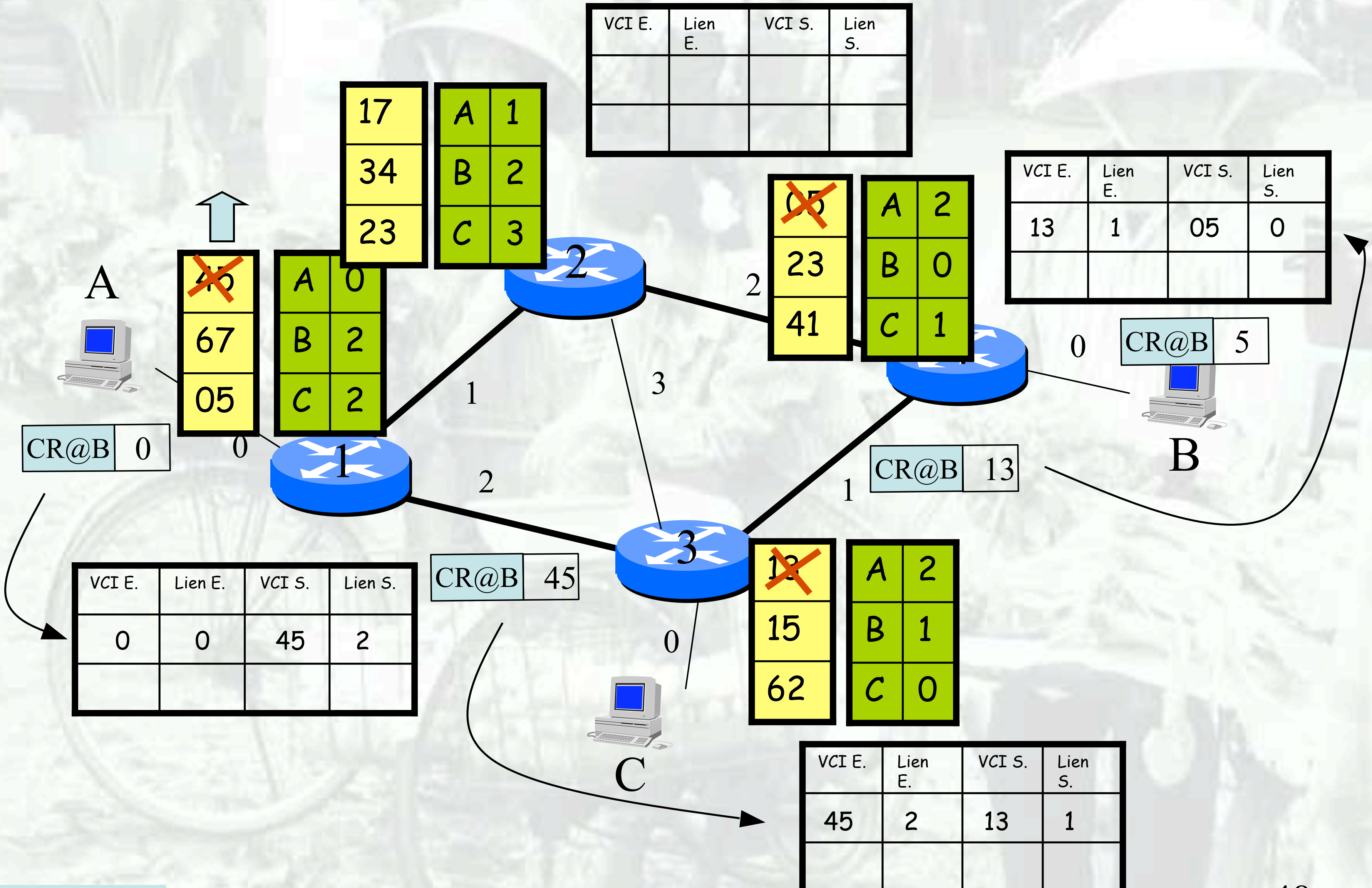
Label IN	Link IN	Label OUT	Link OUT
23	1	34	3
45	2	78	4



Virtual
Circuit
Switching



End-to-end operation (1)



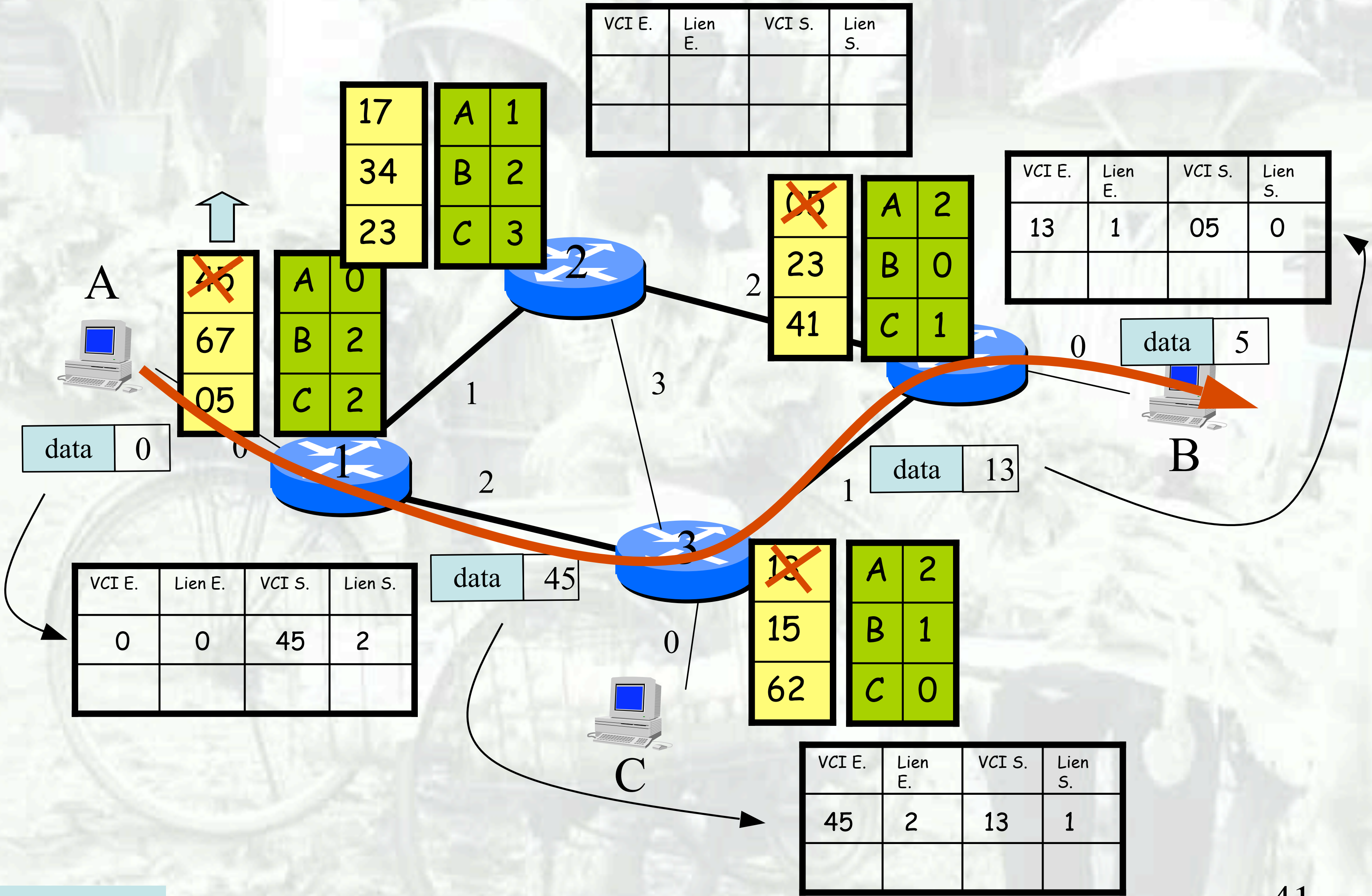
VCI E.	Lien E.	VCI S.	Lien S.
0	0	45	2

VCI E.	Lien E.	VCI S.	Lien S.

VCI E.	Lien E.	VCI S.	Lien S.
45	2	13	1

VCI E.	Lien E.	VCI S.	Lien S.
13	1	05	0

End-to-end operation (2)



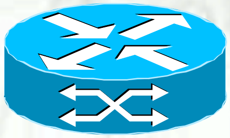
Why virtual circuit?

- Initially to speed up router's forwarding tasks: X.25, Frame Relay, ATM.

Now: Virtual circuits for traffic engineering!

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



PPP Header(Packet over SONET/SDH)	PPP Header	MPLS Header	Layer 3 Header
Ethernet	Ethernet Hdr	MPLS Header	Layer 3 Header
Frame Relay	FR Hdr	MPLS Header	Layer 3 Header

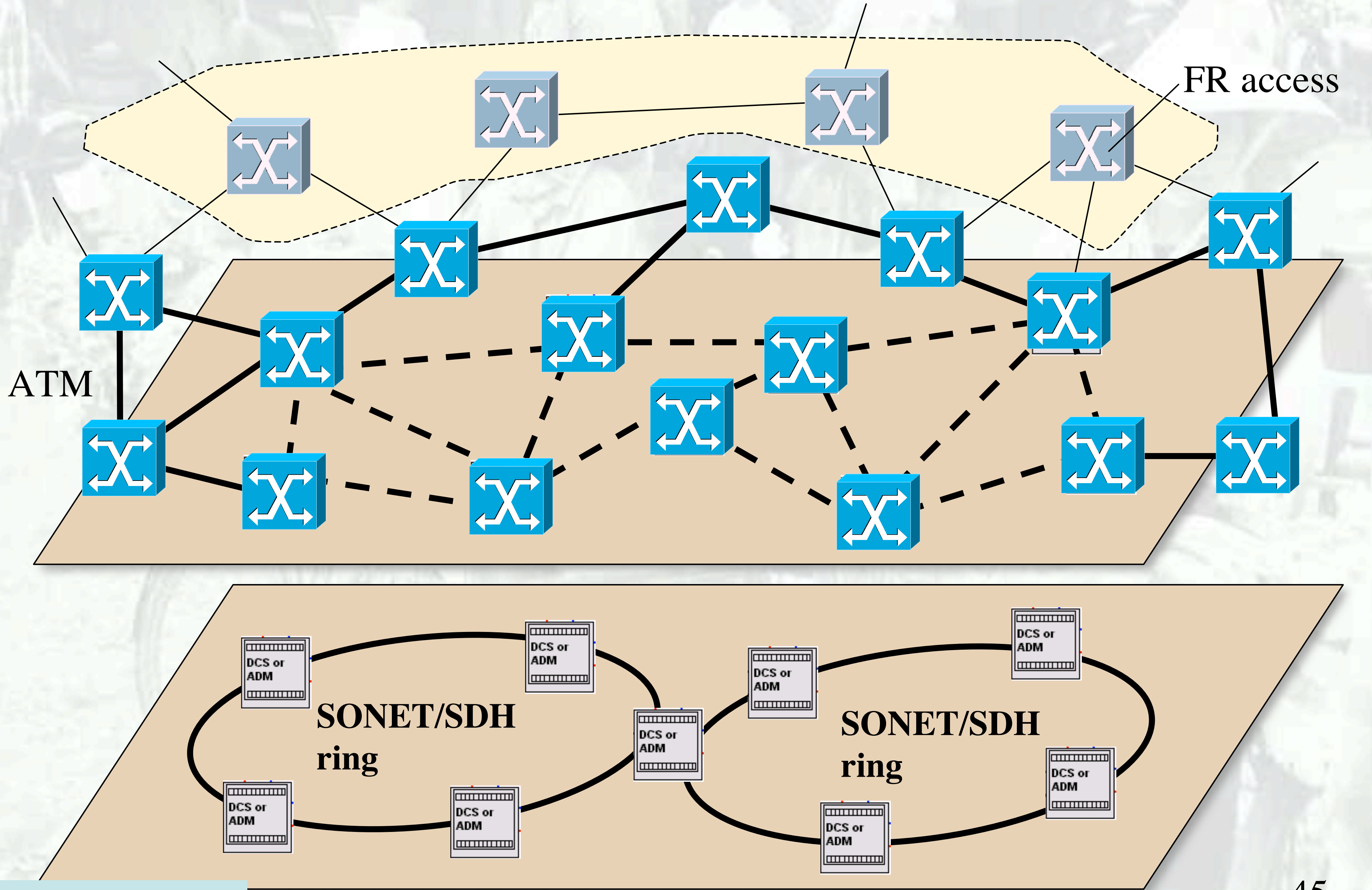
Label structure

```
0                   1                   2                   3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+---------------------------------------------------------------+
|               Label                   | Exp|S|      TTL       |
+---------------------------------------------------------------+
```

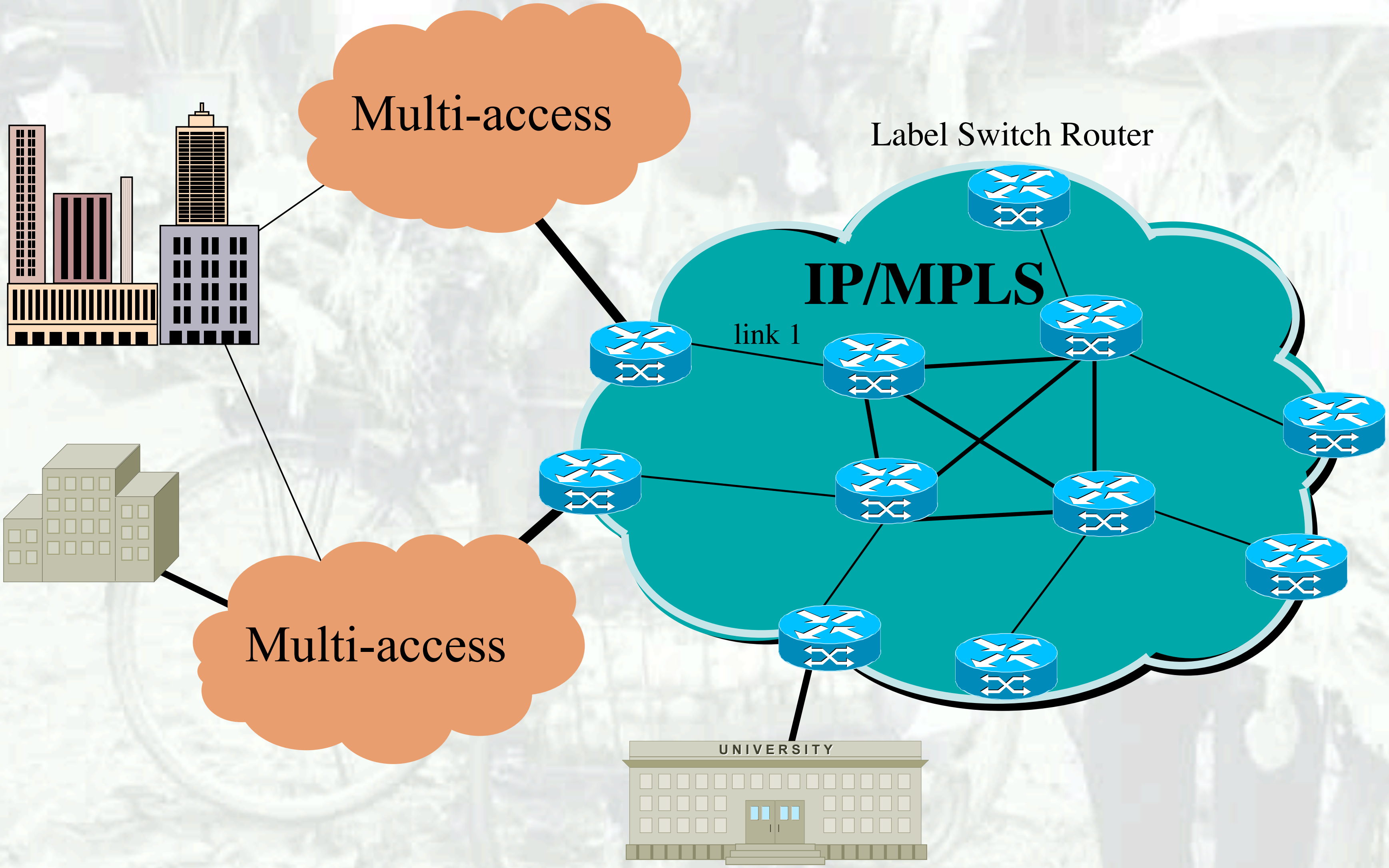
Label = 20 bits
Exp = Experimental, 3 bits
S = Bottom of stack, 1bit
TTL = Time to live, 8 bits

- 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

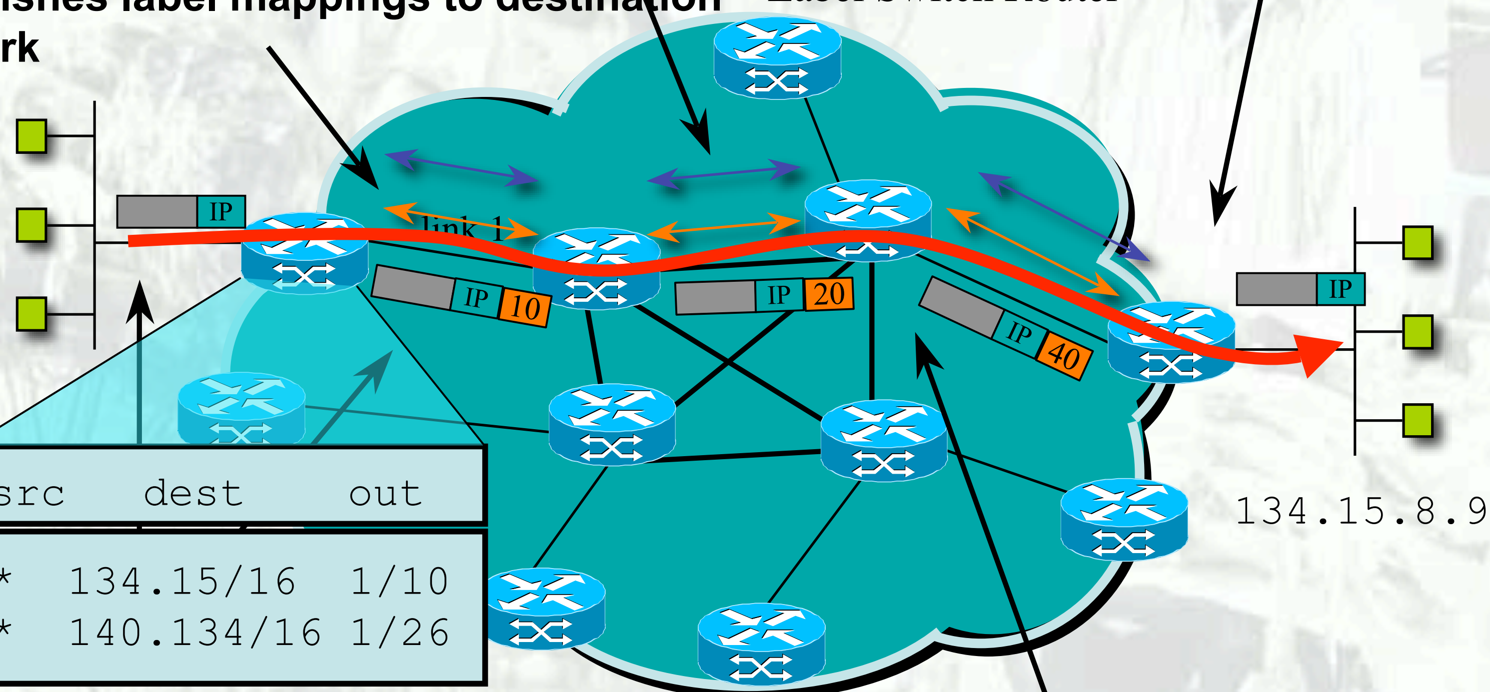


MPLS operation

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



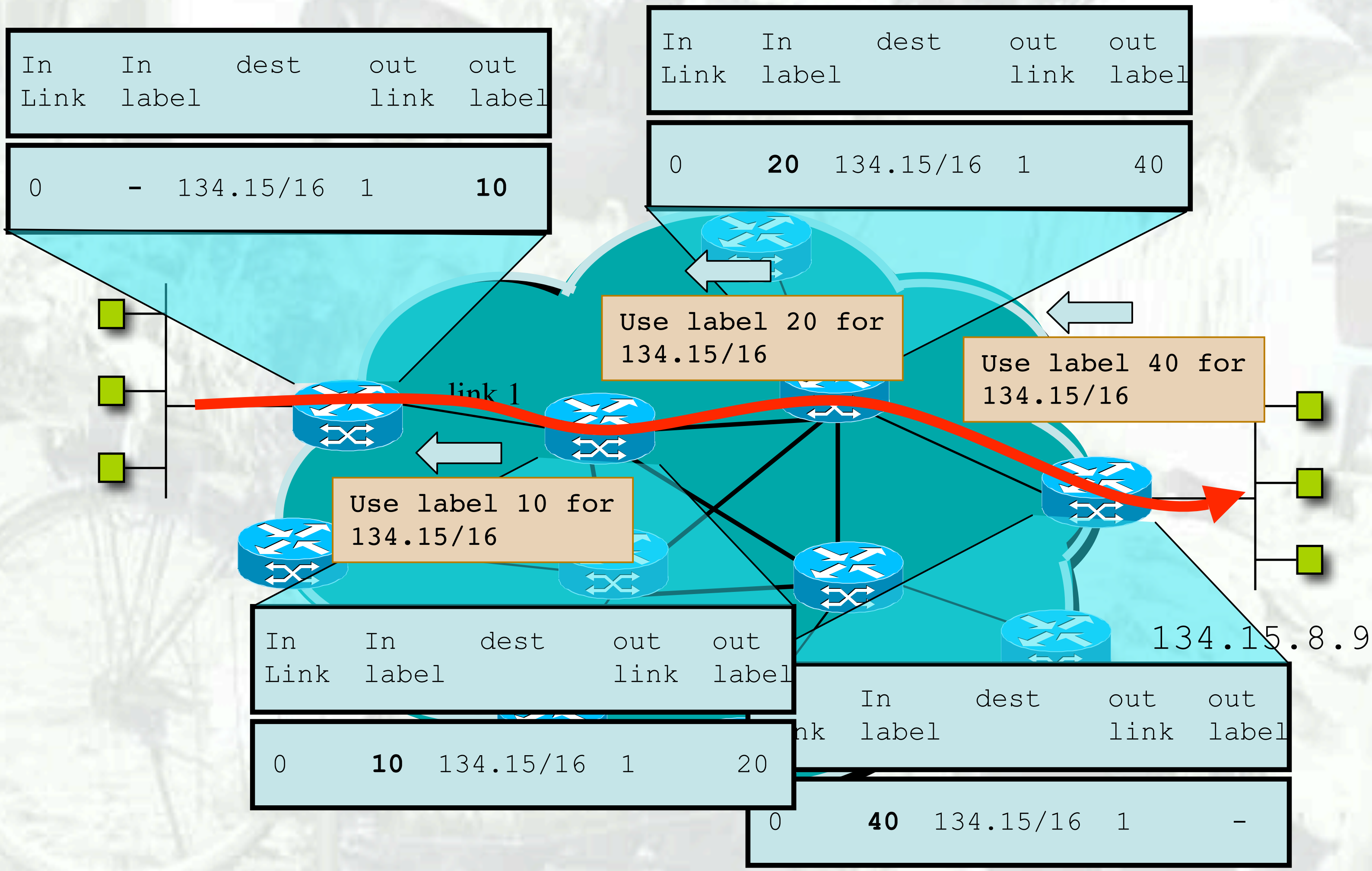
src	dest	out
*	134.15/16	1/10
*	140.134/16	1/26

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

Source Yi Lin, modified C. Pham

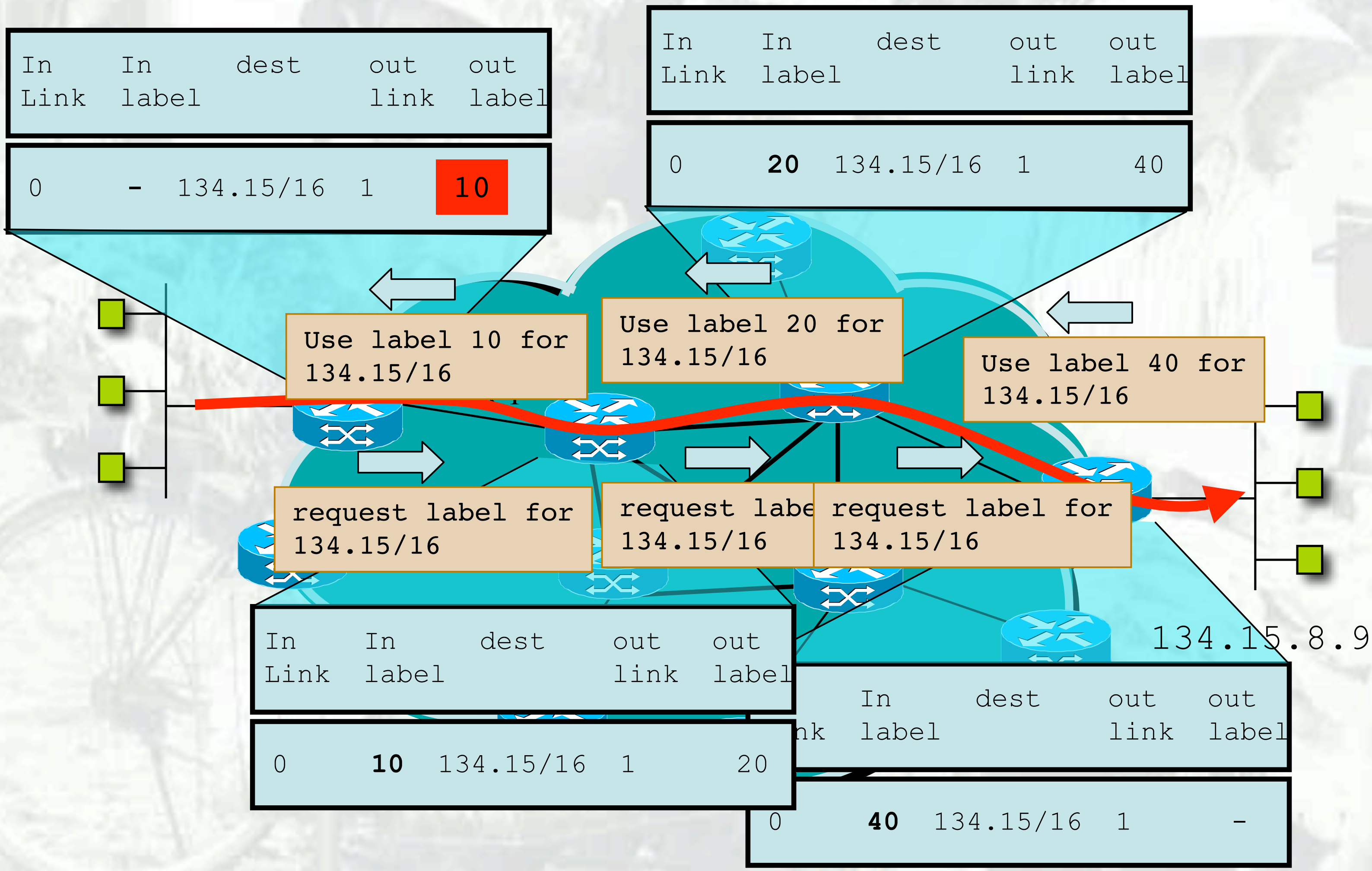
3. LSR forwards packets using label switching

Label Distribution



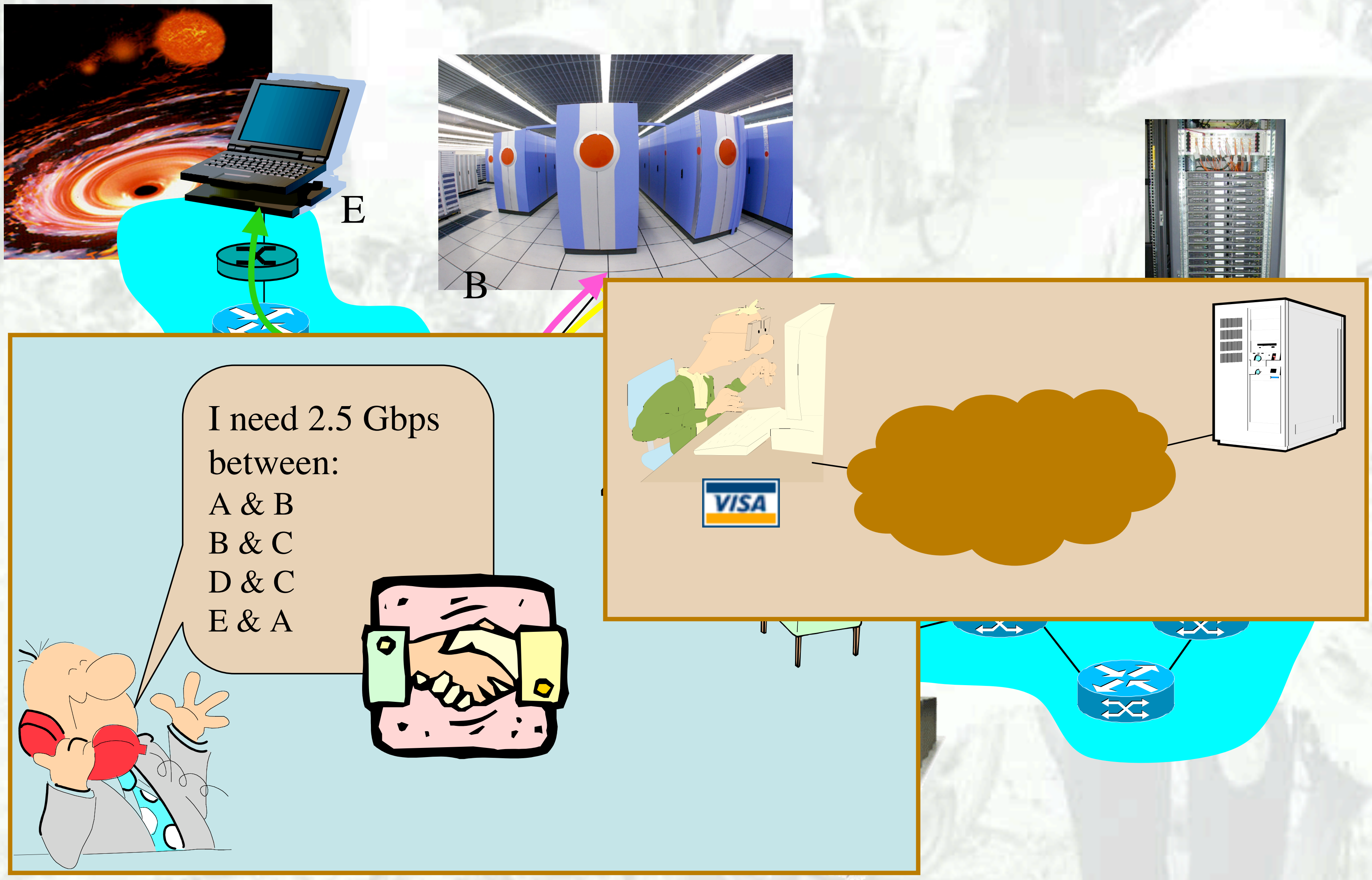
Unsolicited downstream label distribution

Label Distribution (con't)

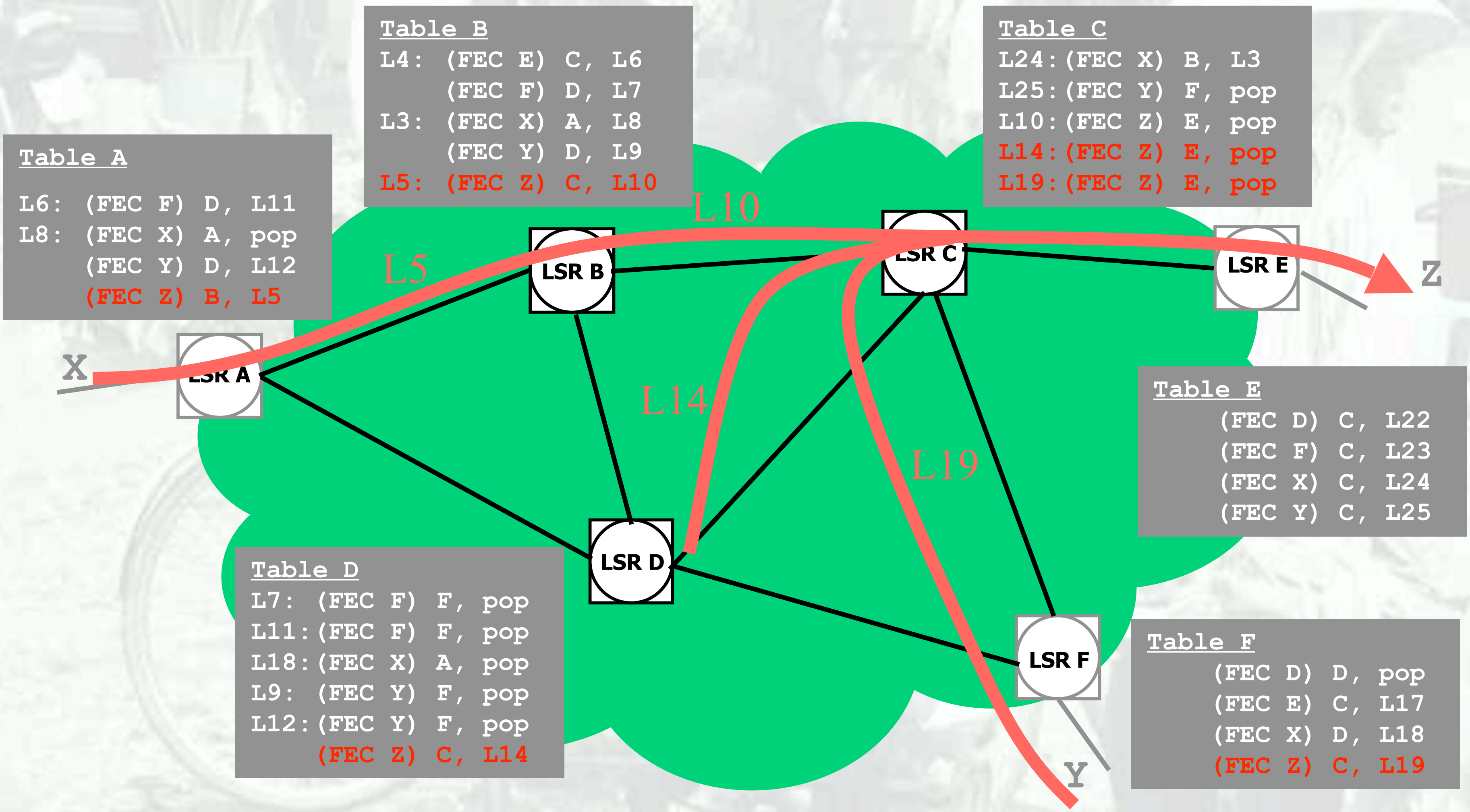


On-demand downstream label distribution

Dynamic circuits for grids

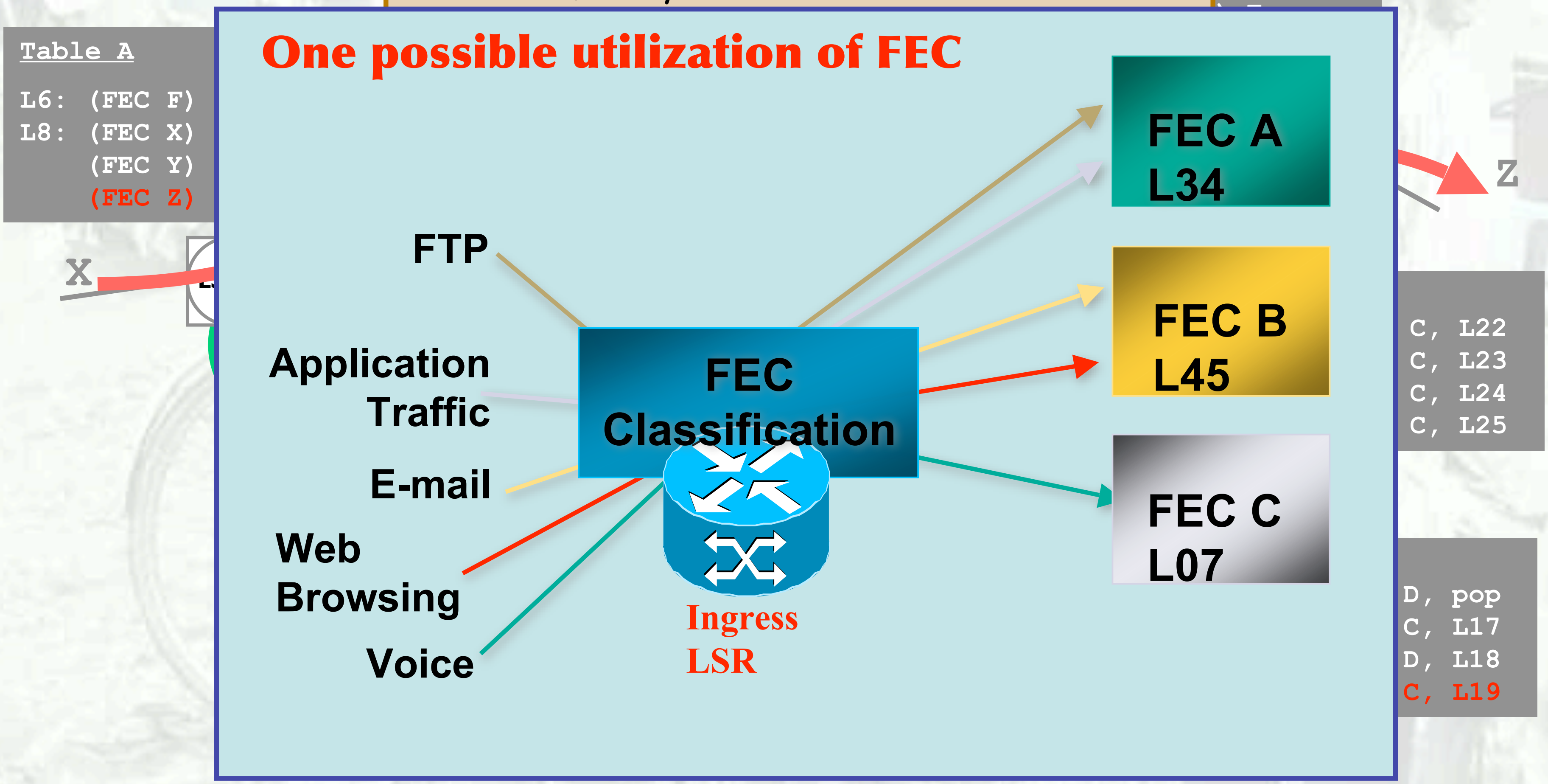


Forwarding Equivalent Class: high-level forwarding criteria

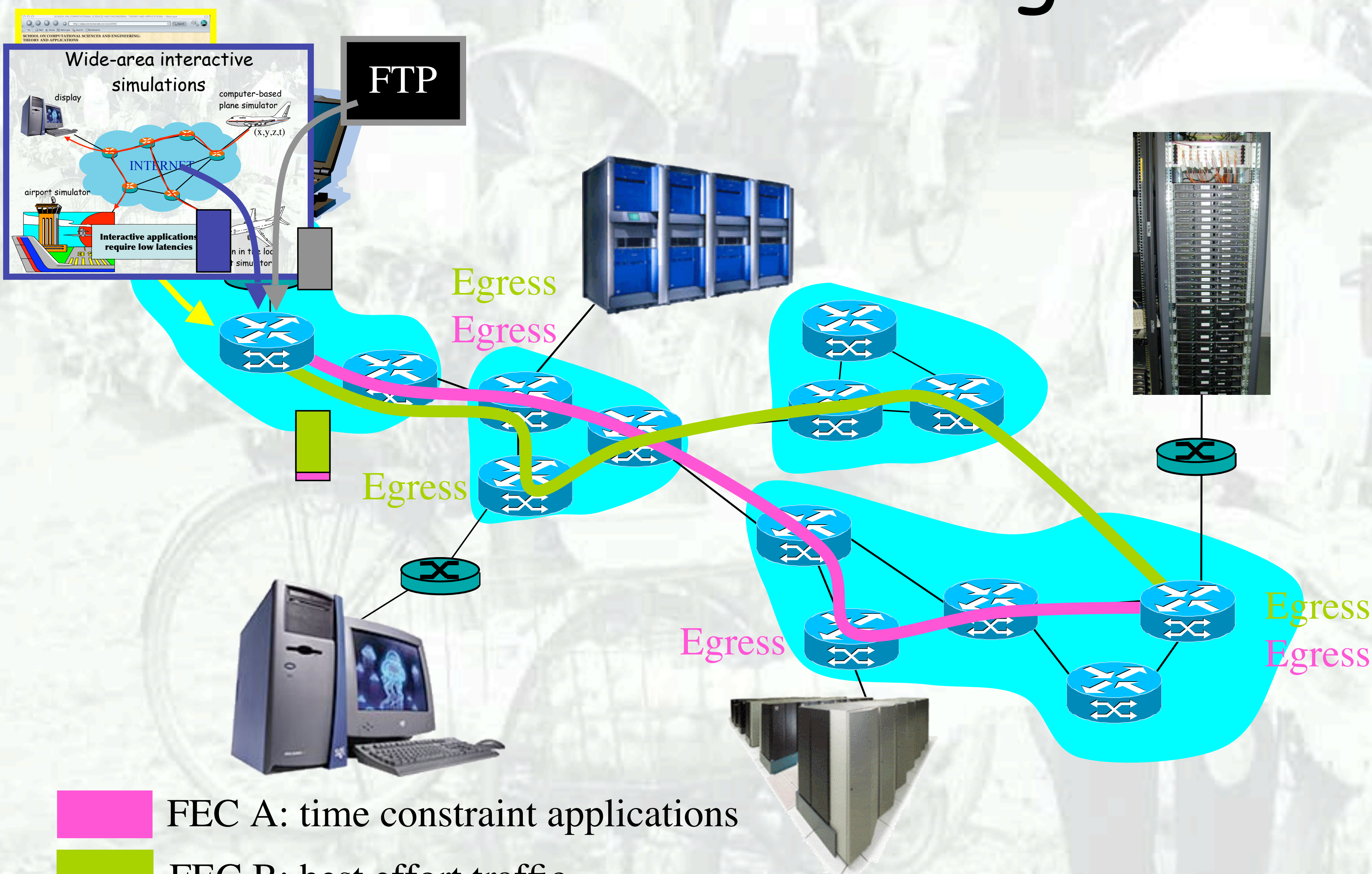


Forwarding Equivalent Class

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



FEC A: time constraint applications

FEC B: best effort traffic

Label & FEC

- Independent LSP control
  - An LSR binds a label to a FEC, whether or not the LSR has received a label from the next-hop for the FEC
  - The LSR then advertises the label to its neighbor

- Ordered LSP control
  - An LSR only binds and advertises a label for a particular FEC if:
    - it is the egress LSR for that FEC or
    - it has already received a label binding from its next-hop

Label Distribution Protocols

- LDP
  - Maps unicast IP destinations into labels
- RSVP-TE, CR-LDP
  - Used in traffic engineering
- BGP
  - External labels (VPN)
- PIM
  - For multicast states label mapping

MPLS for resiliency

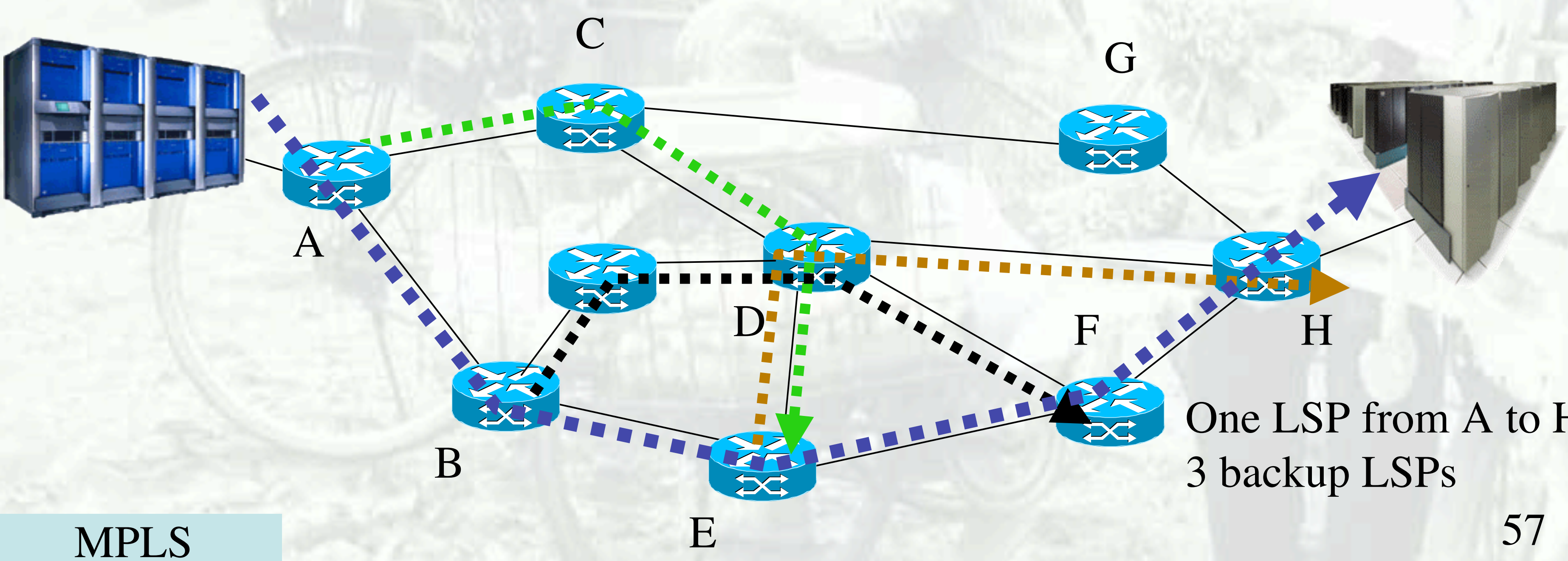
MPLS FastReroute

- 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

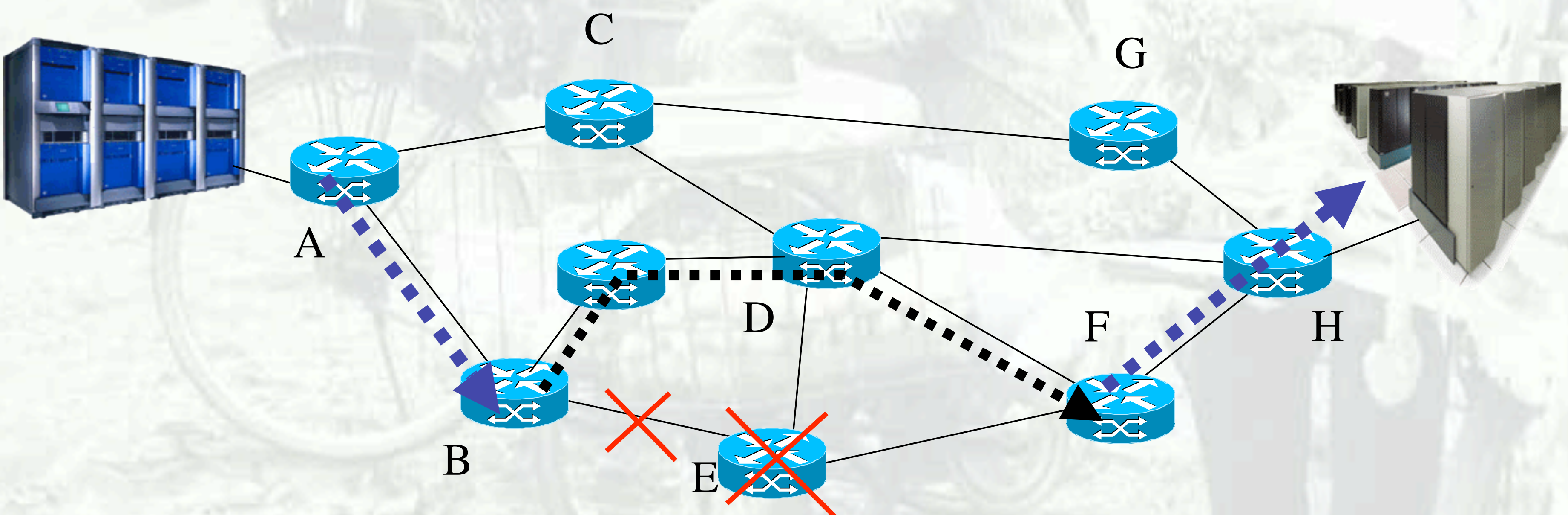


One LSP from A to H
3 backup LSPs

MPLS for resiliency, con't

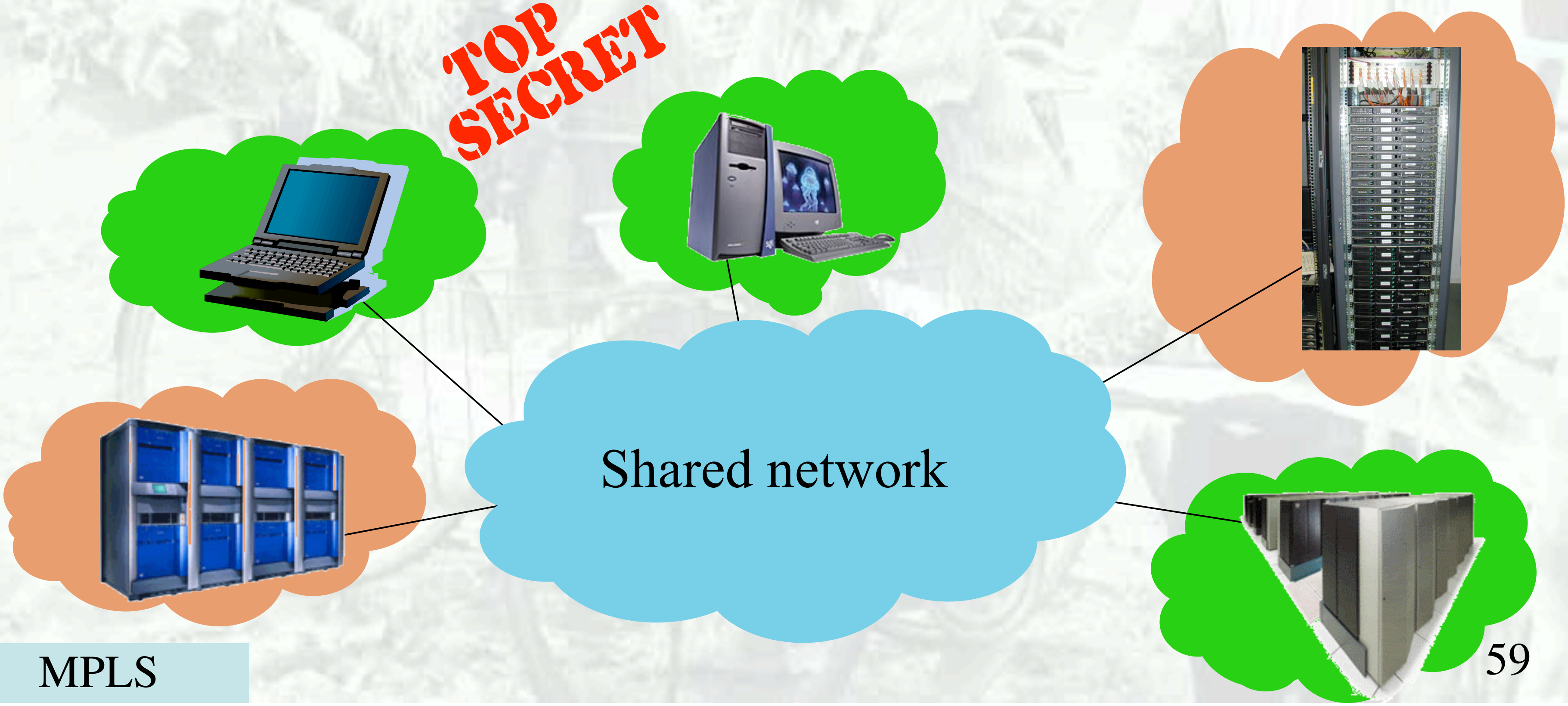
Recovery on failures

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



MPLS for VPN
(Virtual Private Networks)

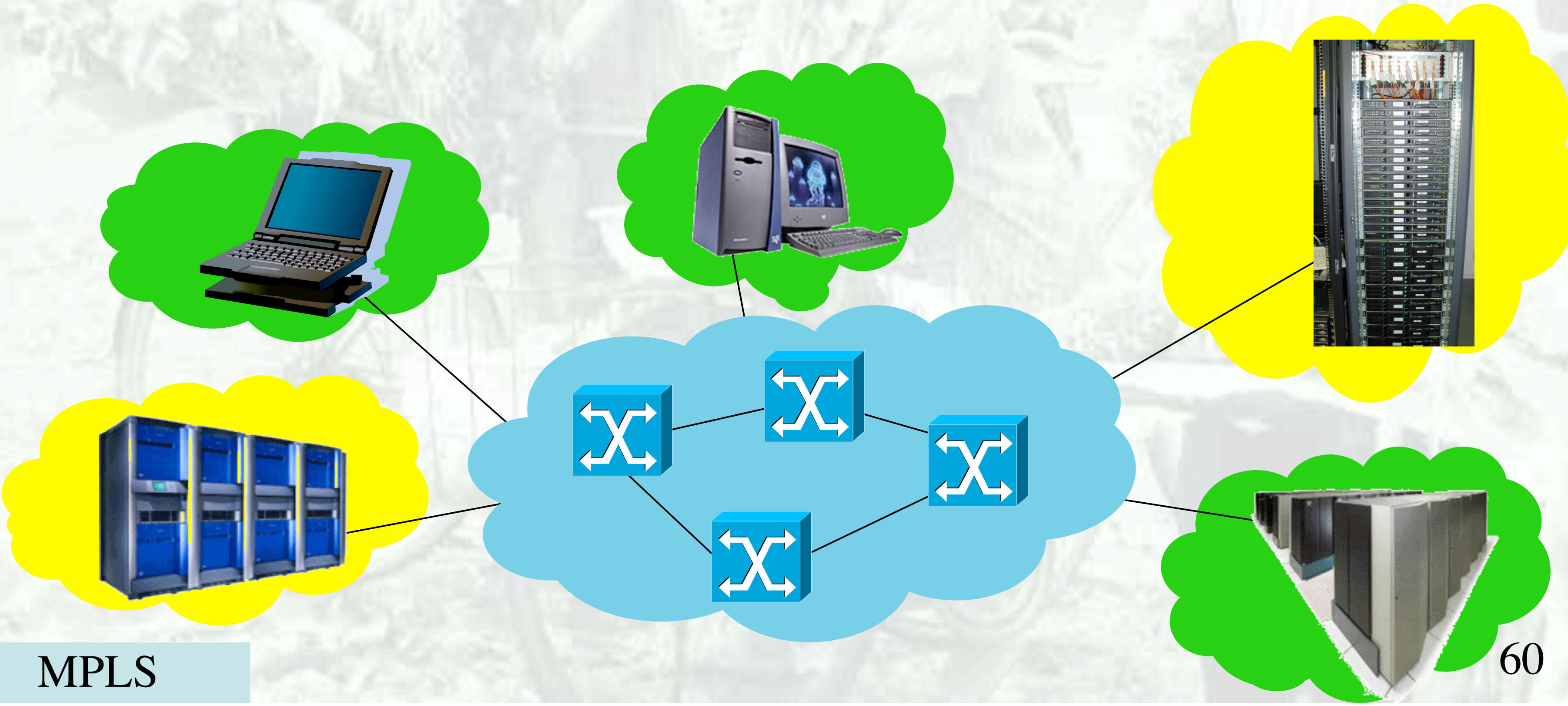
- Virtual Private Networks: build a secure, confidential communication on a public network infrastructure using routing, encryption technologies and controlled accesses



MPLS for VPN, con't

The traditional way of VPN

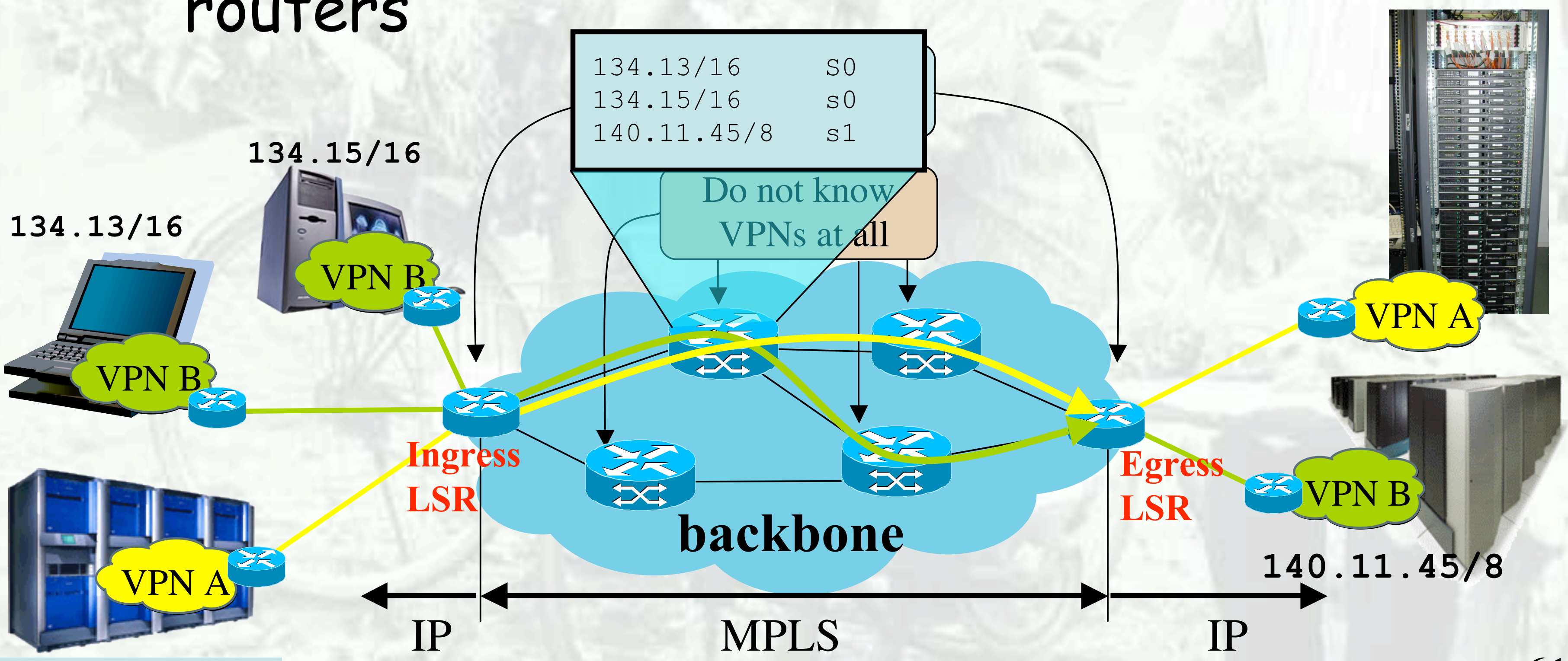
- Uses leases lines, Frame Relay/ATM infrastructures…



MPLS for VPN, con't

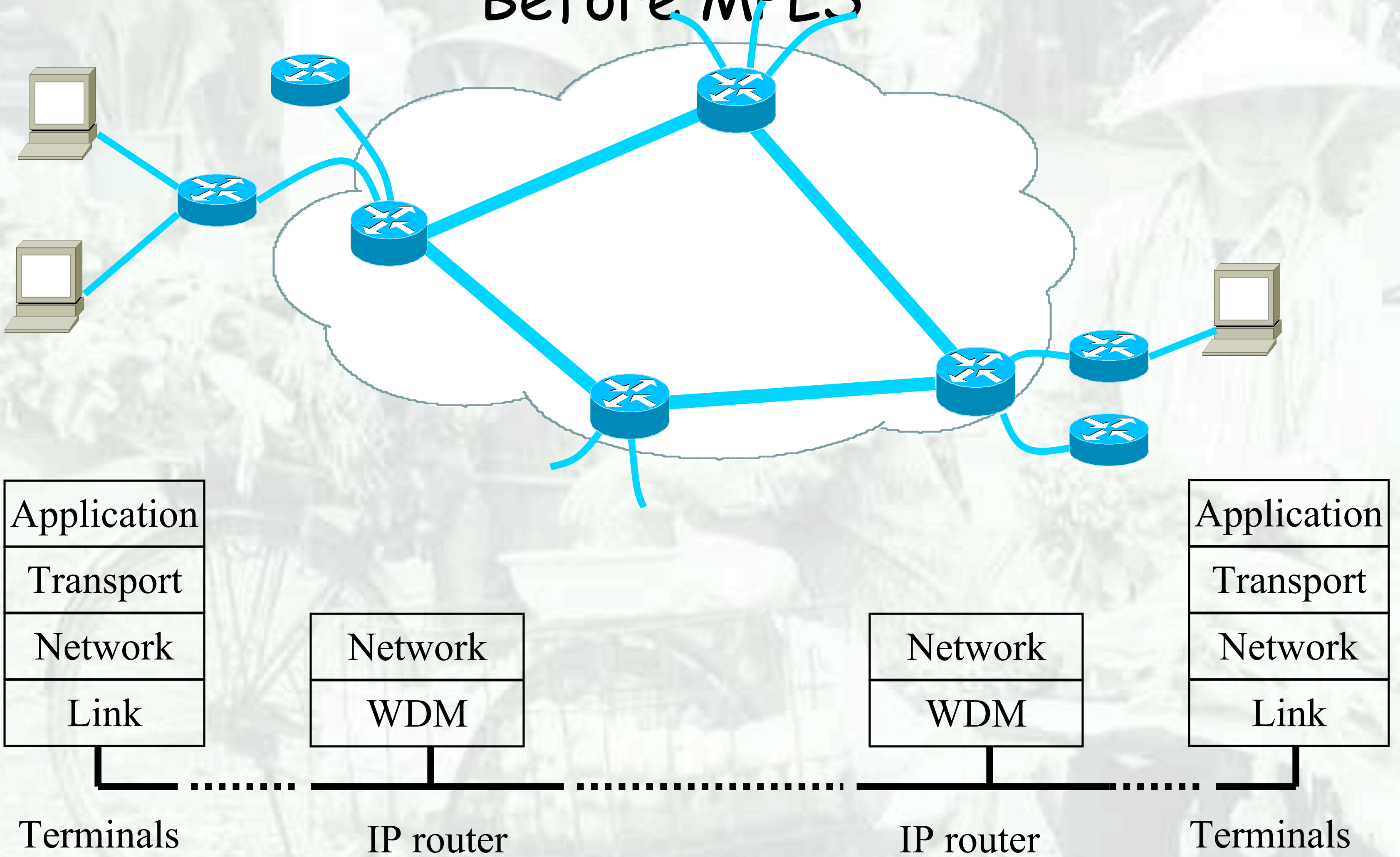
VPN over IP/MPLS

- IP/MPLS replace dedicated networks
- MPLS reduces VPN complexity by reducing routing information needed at provider's routers



MPLS for optical networks

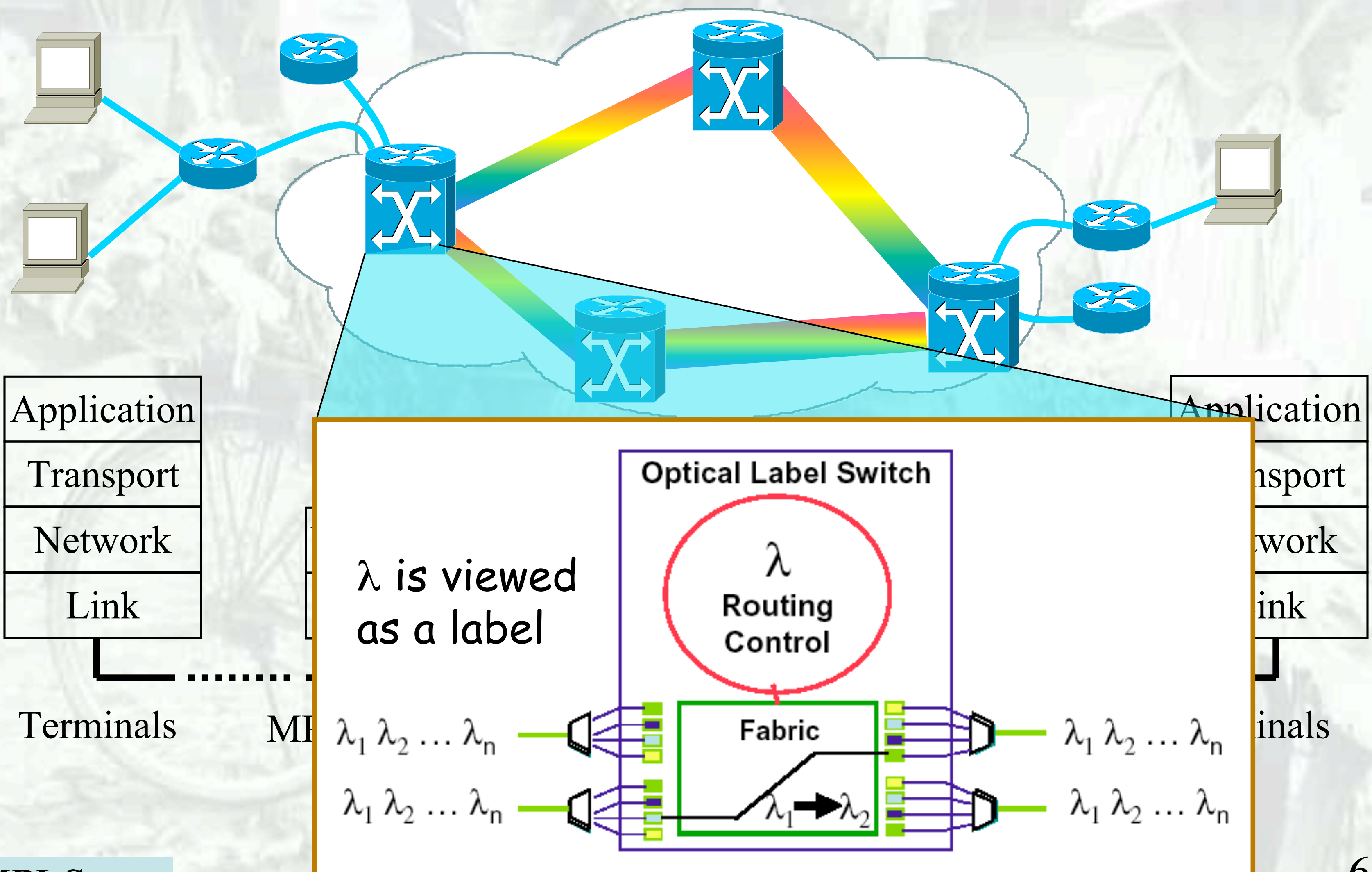
Before MPLS



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

MPLS for ON, con't

MPλS=MPLS+λ lightpath



MPLS for ON, con't

GMPLS

- GMPLS stands for "Generalized Multi-Protocol Label Switching"
- Extends the concept of MPLS beyond data networks to address legacy transport networks
- Reduce OPEX cost for operators
- A suite of protocols that provides a common set of control functions for disparate transport technologies (IP, ATM, SONET/SDH, DWDM)
- Hot issue at IETF!

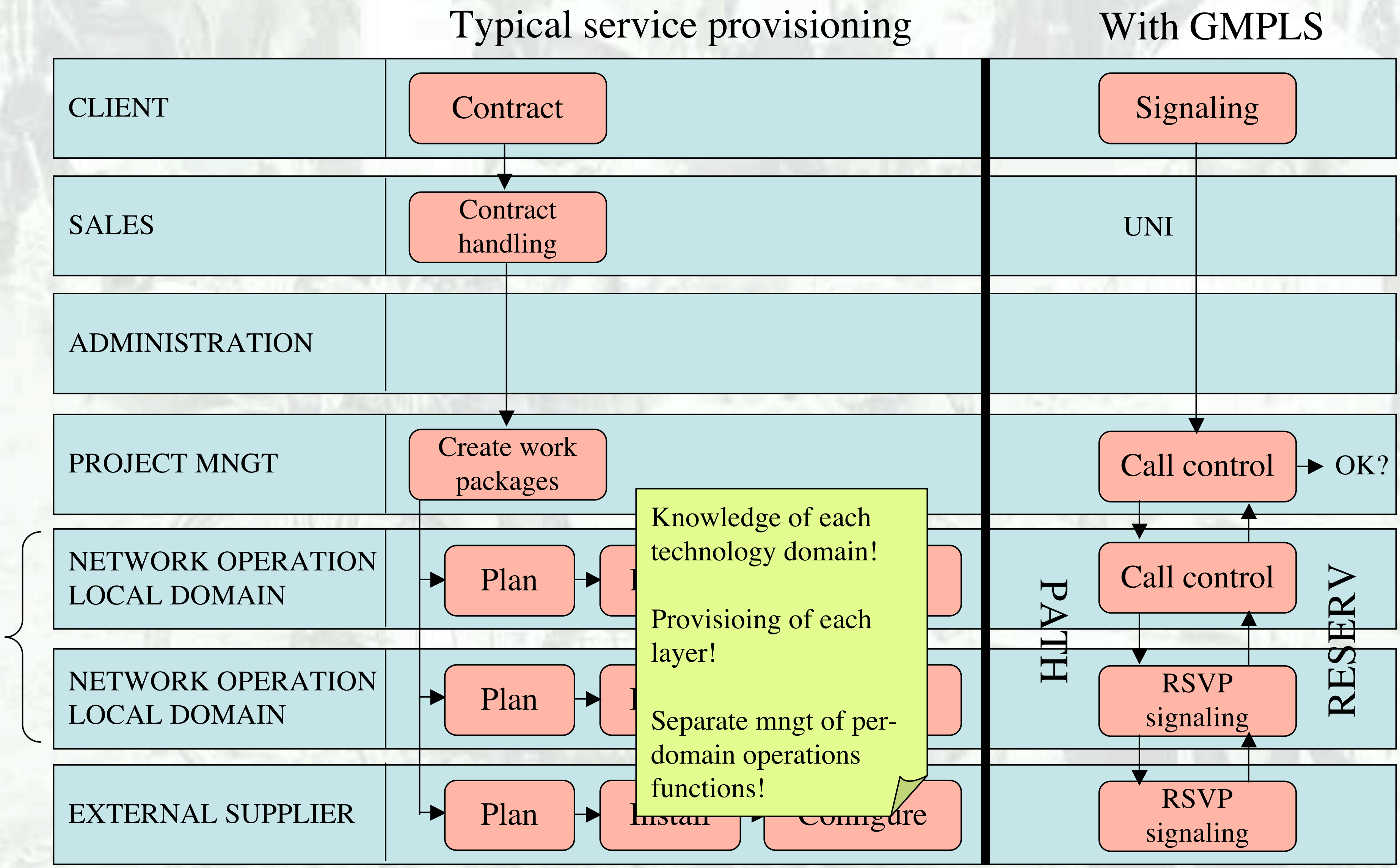
MPLS for ON, con't

GMPLS control plane

LINK MANAGEMENT: Link Management Protocol (LMP)	-Neighbor discovery -Maintain control channel connectivity -Verify data link connectivity -Correlate link property information -Suppress downstream alarms -Localize link failures
ROUTING: Open Shortest Path First-Traffic Engineering (OSPF-TE)	-Distribute TE link information -Advertise nodes in the network and create topology -Calculate constrained shorted path (CSPF) -Routing information for control and data plane
SIGNALING: Resource ReserVation Protocol-Traffic Engineering (RSVP-TE)	-Signals setup/teardown/refresh of paths with QoS requirements (e.g., circuit size) -Uses control channel to setup an optical LSP -Supports refresh reduction -Supports Explicit Route Object (ERO) and Record Route Object (RRO)

Source S. Kinoshita, R. Rabbat, APNOMS 2005

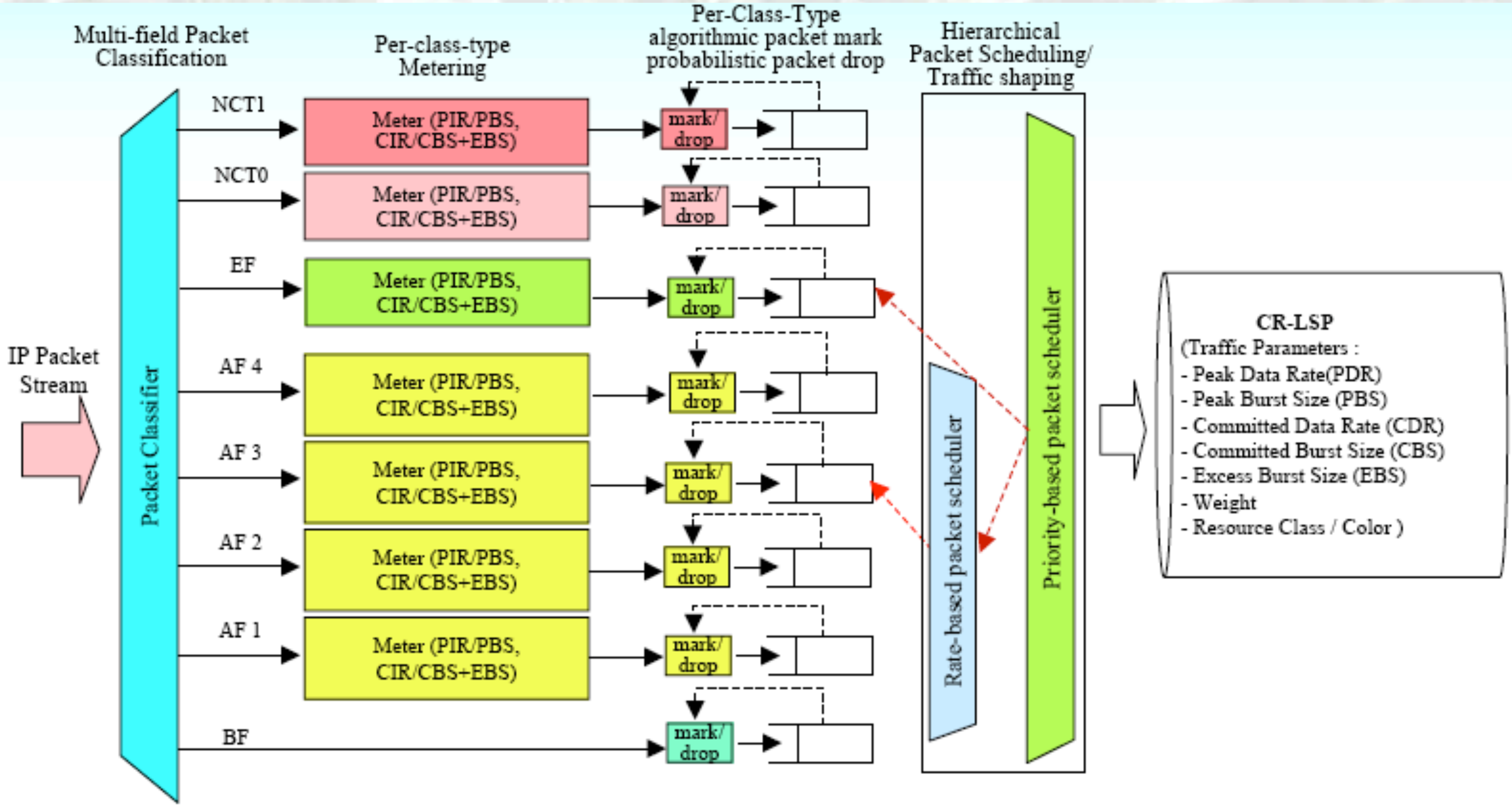
Ex: Service Provisioning



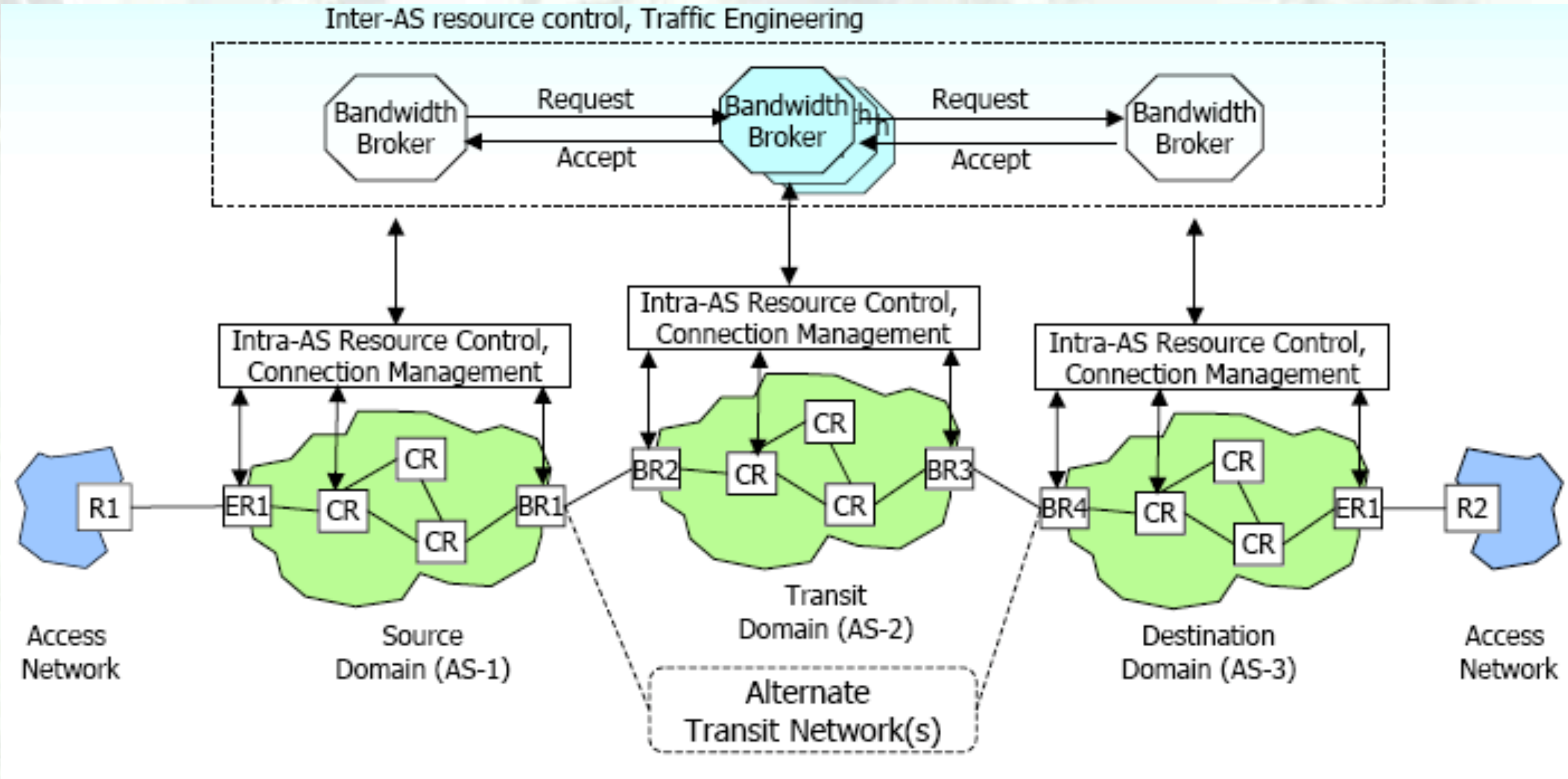
From Pascalini et al., IEEE Comm. Mag. July 2005

DiffServ over (G)MPLS

map DiffServ class on MPLS FEC



Some words on inter-domain

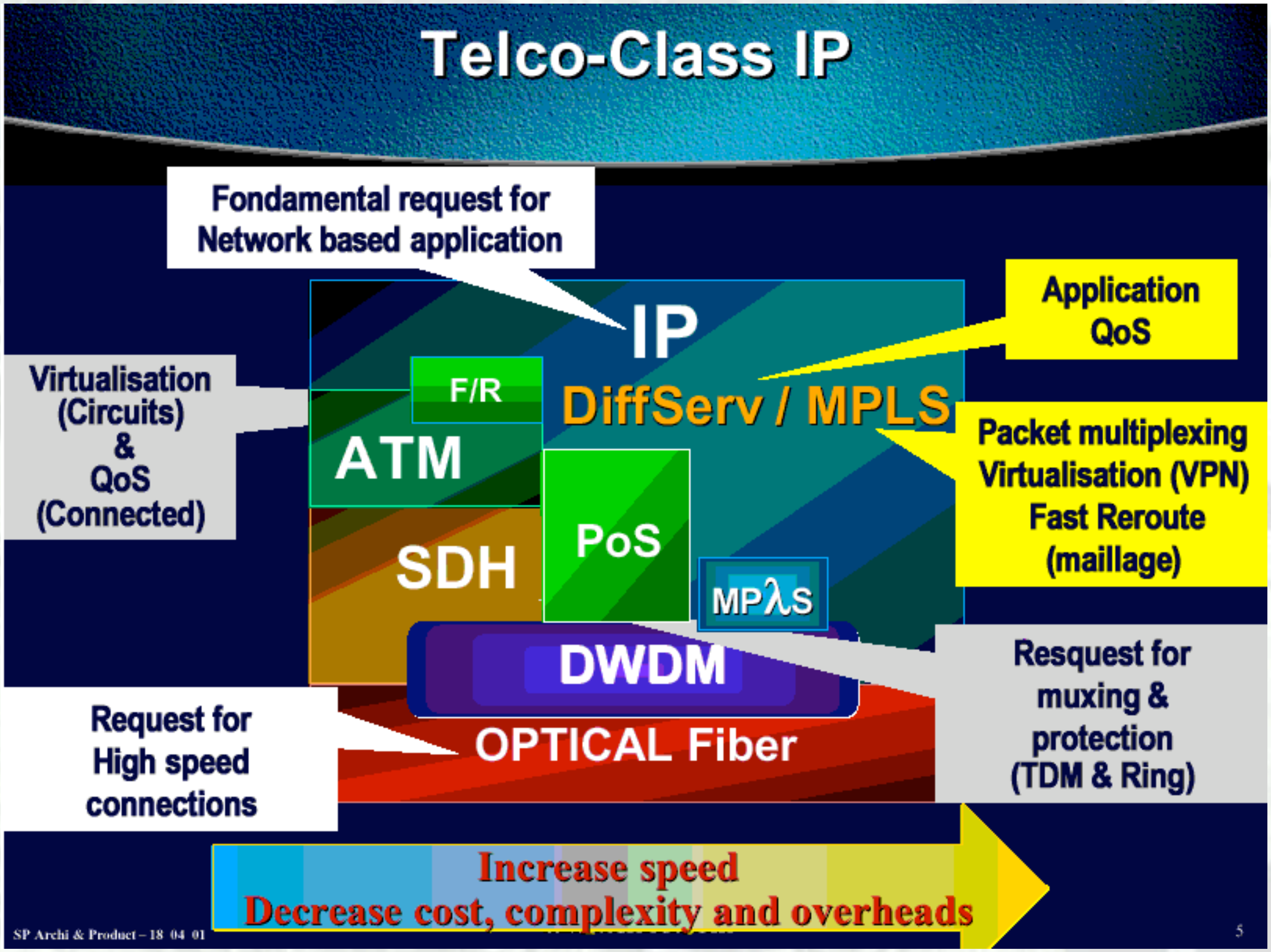


Source Yong-Tak Kim

Summary

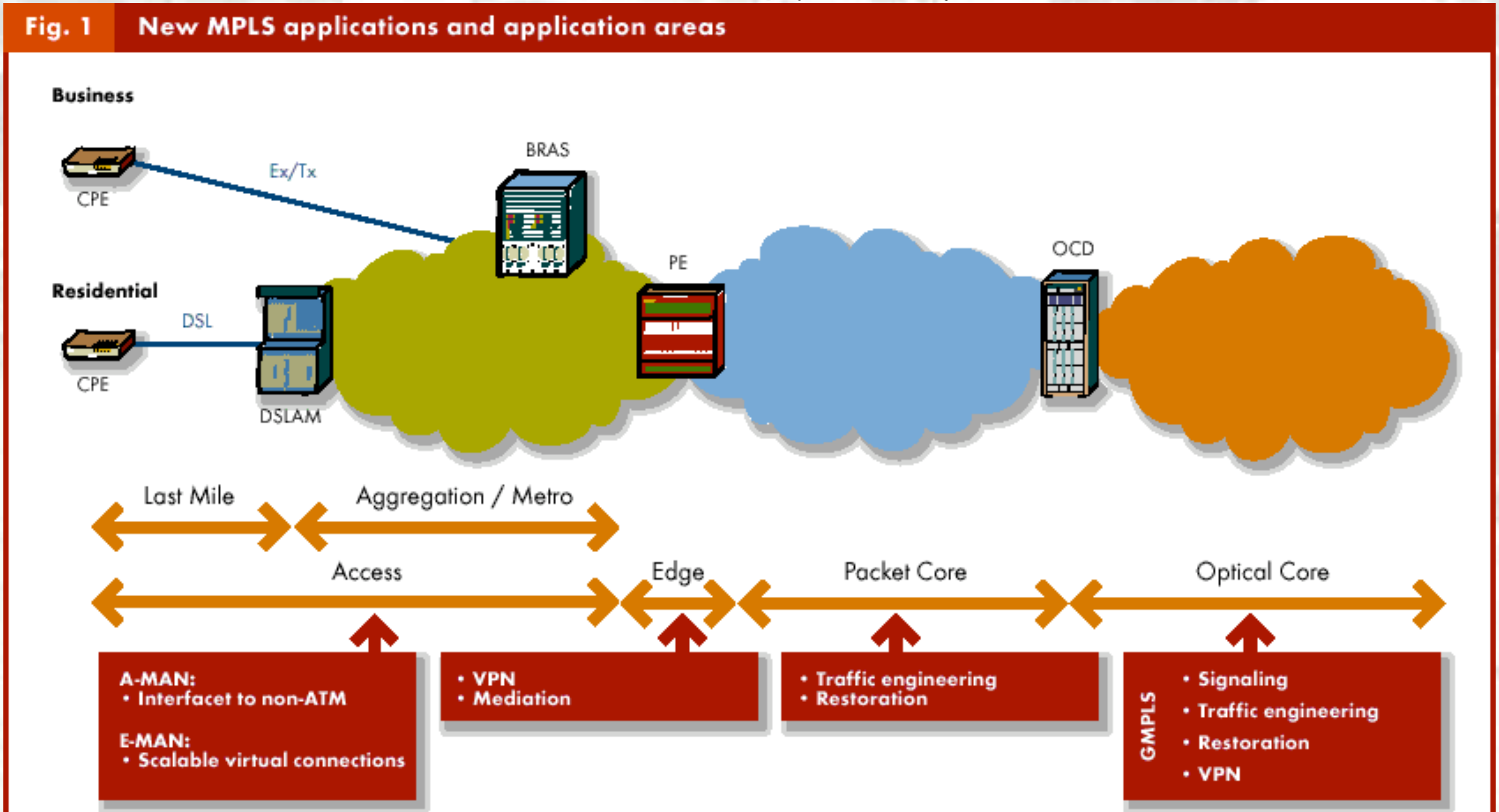
Towards IP/(G)MPLS/DWDM

From cisco



Summary
Technology scope

Source Alcatel



A-MAN: ATM Metropolitan Area Network
BRAS: Broadband Remote Access Server
CPE: Customer Premises Equipment
DSLAM: Digital Subscriber Line Access Multiplexer
Ex/Tx: E1/T1 or E3/T3
OCD: Optical Core Device
PE: Provider Edge

Want to know more?

- GMPLS: IEEE Comm. Mag., Vol. 43(7), July 2005
- Optical Control Plane for the Grid Community: IEEE Comm. Mag., Vol. 44(3), March 2006.
- "Optical Transport Systems/Networks" by S. Kinoshita & R. Rabbat, APNOMS 2005. http://www.apnoms.org/2005/tutorial/Tutorial%202.pdf
- « Inter-domain Traffic Engineering for QoS-guaranteed DiffServ Provisioning », Young-Tak Kim, APNOMS 2005. http://www.apnoms.org/2005/tutorial/Tutorial%203.pdf
- See Tutorial IV of HOTI 2006: Dynamic Optimal Networks for Grid Computing

End of part 1, go to part 2