

New Internet and Networking Technologies for Grids and High-Performance Computing



Bangalore, India

December 22th, 2004

C. Pham

<http://www710.univ-lyon1.fr/~cpham>

University of Lyon, France

LIP (CNRS-INRIA-ENS-UCBL)

Computational Sciences

- Use of computers to solve complex problems
 - Modeling techniques
 - Simulation techniques
 - Analytic & Mathematic methods
 - ...
- Large problems require huge amount of processing power: supercomputers, high-performance clusters, etc.

Earth simulator: #3 TOP500

©JAMSTEC

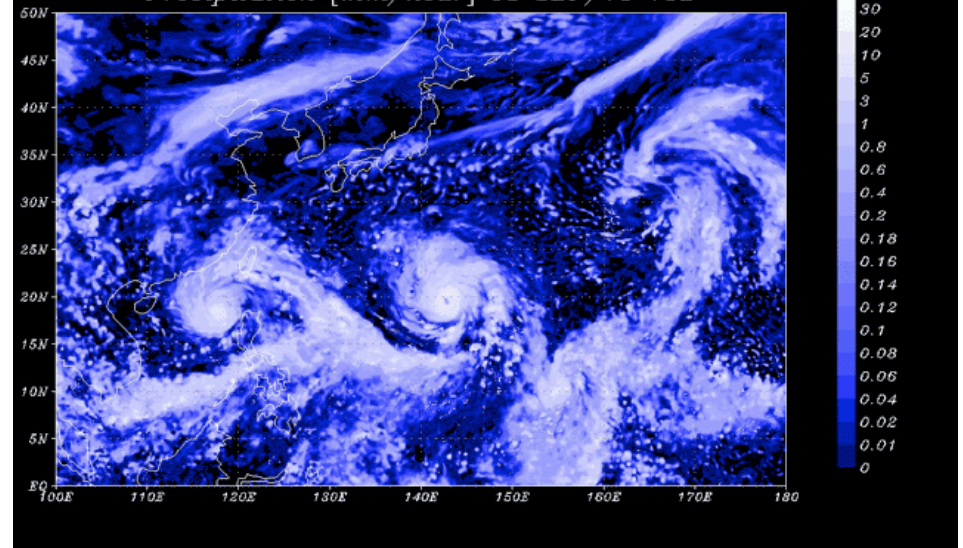
Previously #1 for a long time!

- ❑ Intensive numerical simulations
- ❑ Ex: Super High Resolution Global Atmospheric Simulation

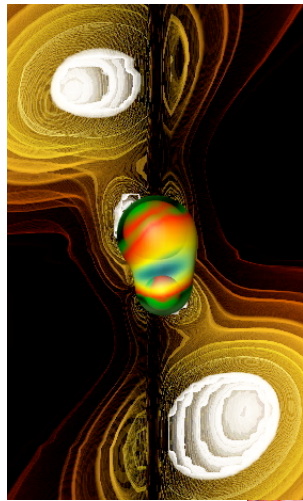


Super High Resolution Global Atmospheric Simulation

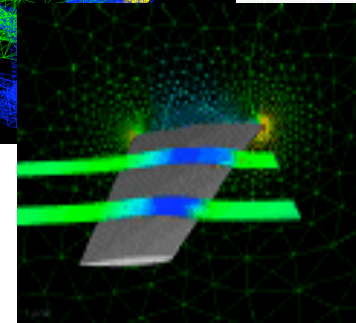
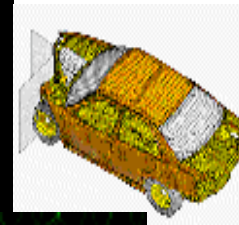
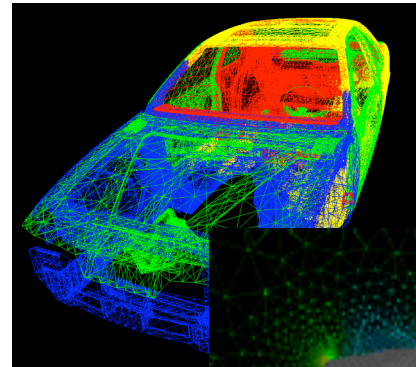
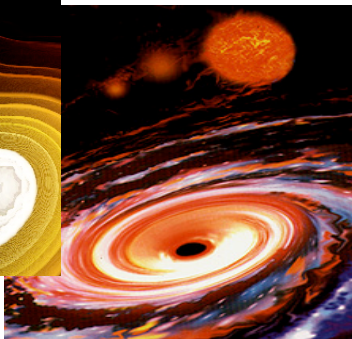
AFES T1279L96
Precipitation [mm/hour] 03 SEP/15 15Z



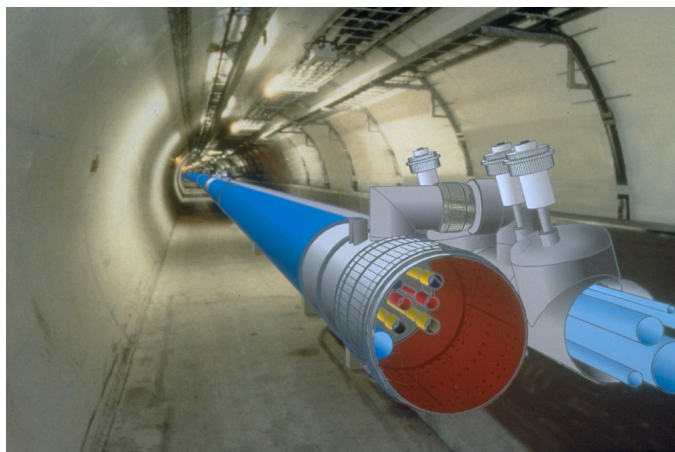
A large variety of applications



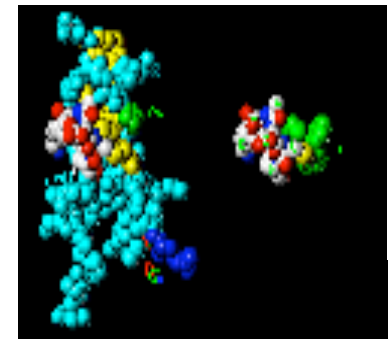
Astrophysics:
Black holes,
neutron stars,
Supernovae...



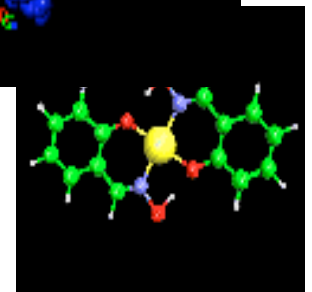
Mechanics:
Fluid dynamic,
CAD, simulation.



High-Energy Physics:
Fundamental particles of matter,
Mass studies...



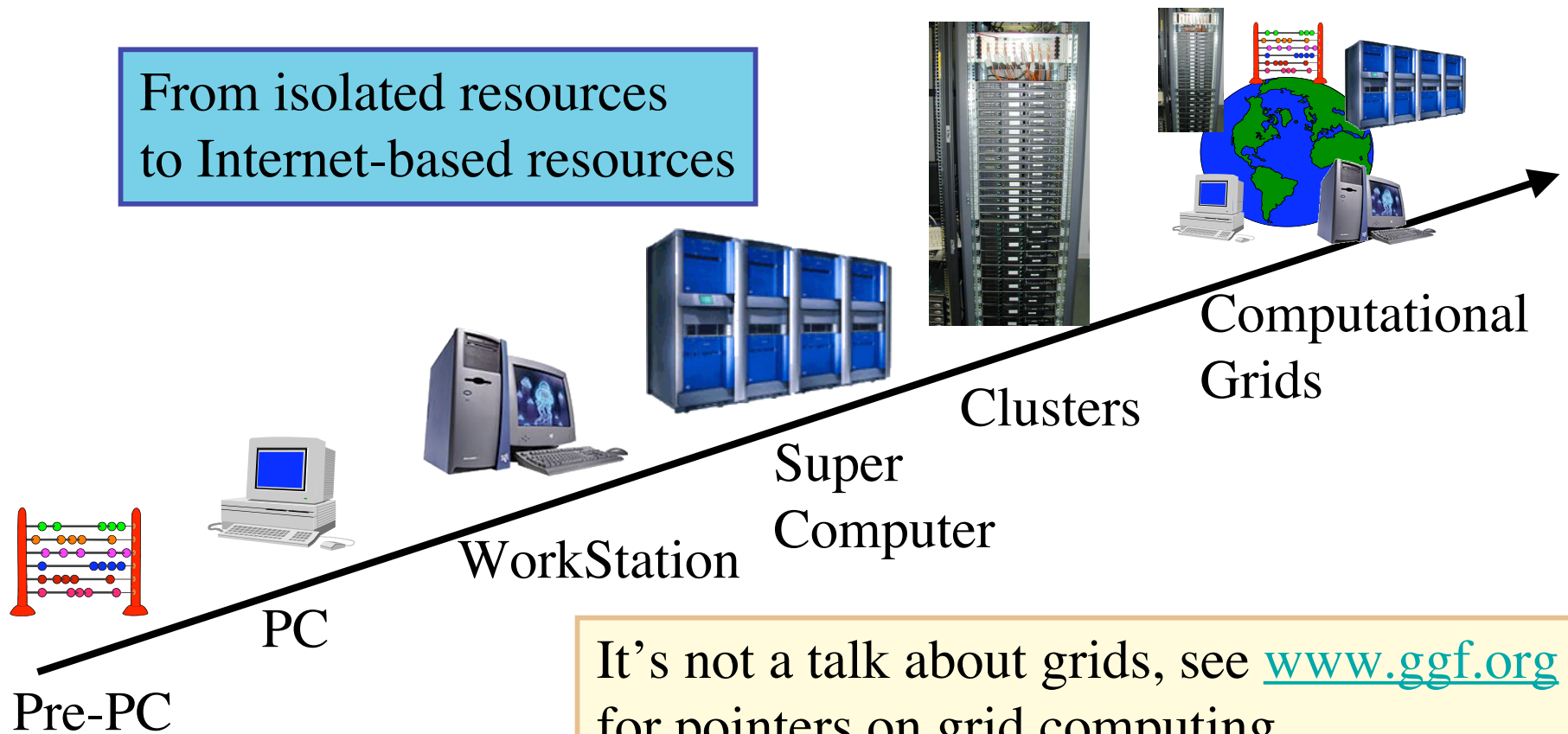
Chemistry&biology:
Molecular simulations,
Genomic simulations...



This talk is about...

- How the Internet revolution could be beneficial to computational sciences

From isolated resources
to Internet-based resources



It's not a talk about grids, see www.ggf.org
for pointers on grid computing

Purpose of this tutorial

□ Audience

- Scientists/students from parallel, distributed, computer or grid and computational sciences

□ Purpose

- Provides a comprehensive survey of advanced networking technologies

□ Expected results

- Understanding of why the network is important in a grid infrastructure
- Knowledge of current advanced technologies for decision making processes

Outline

- ❑ Introduction: new technologies, new challenges
- ❑ Service differentiation
- ❑ MPLS and bandwidth provisioning
- ❑ TCP and beyond
- ❑ Multicast communication models

Layout explanation

N
E
W

Body text

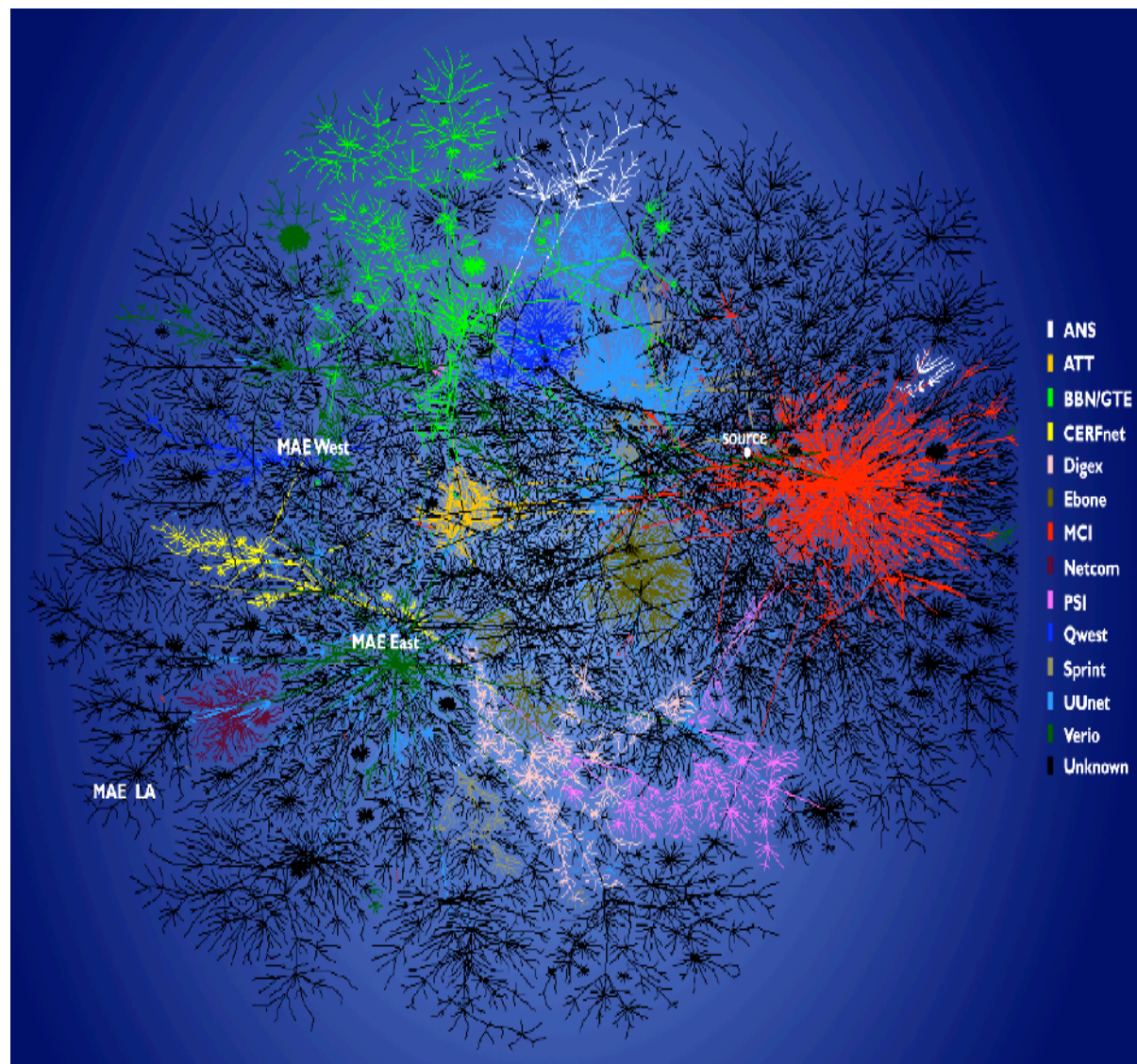
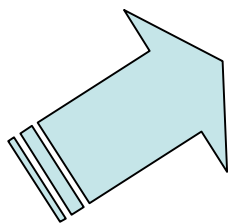
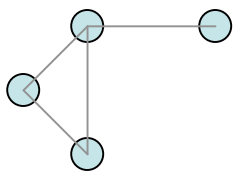
C
H
A
P
T
E
R

Indicates a new chapter

Indicates which chapter the
slide belongs to

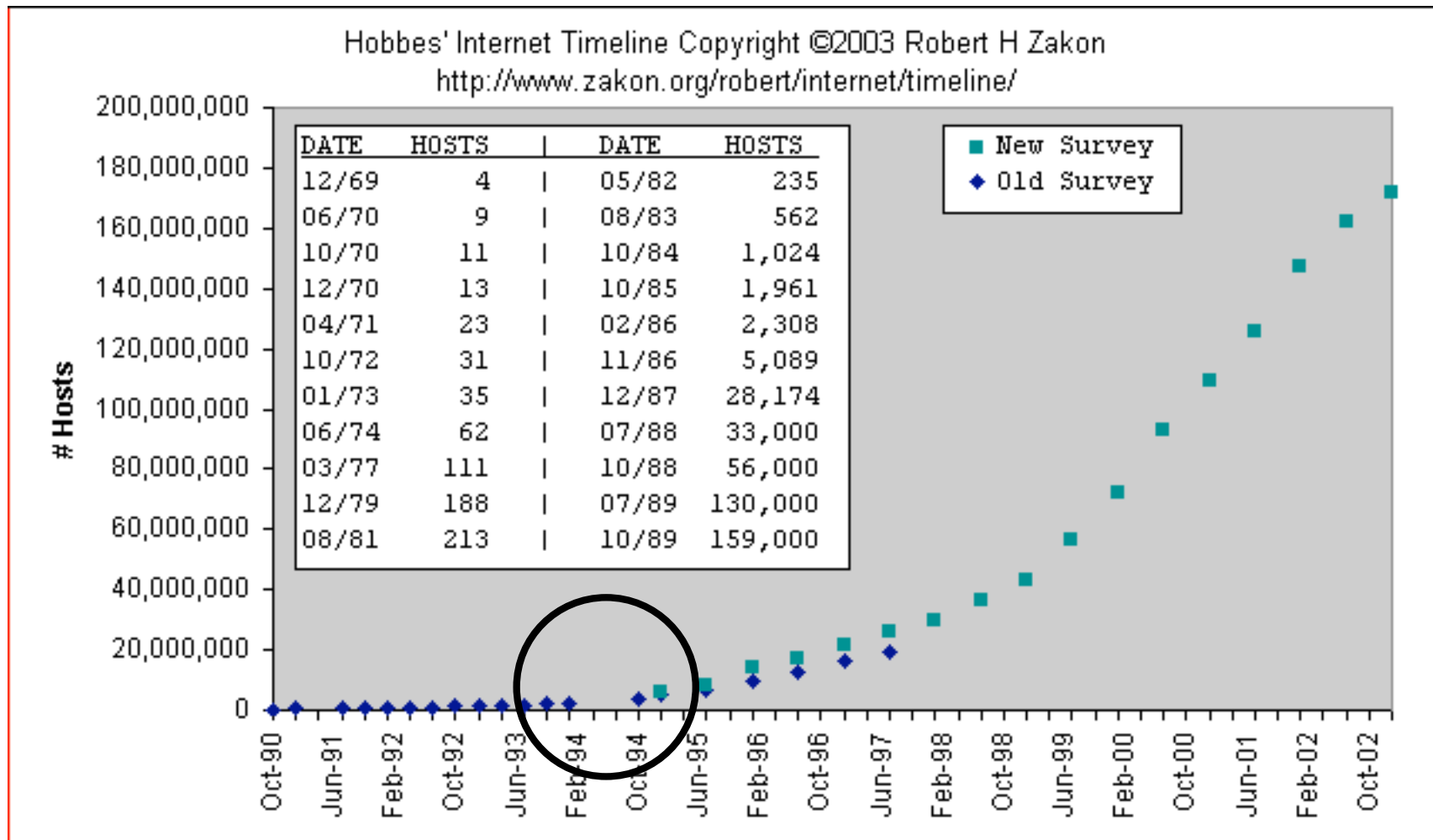
Introduction

The big-bang of the Internet

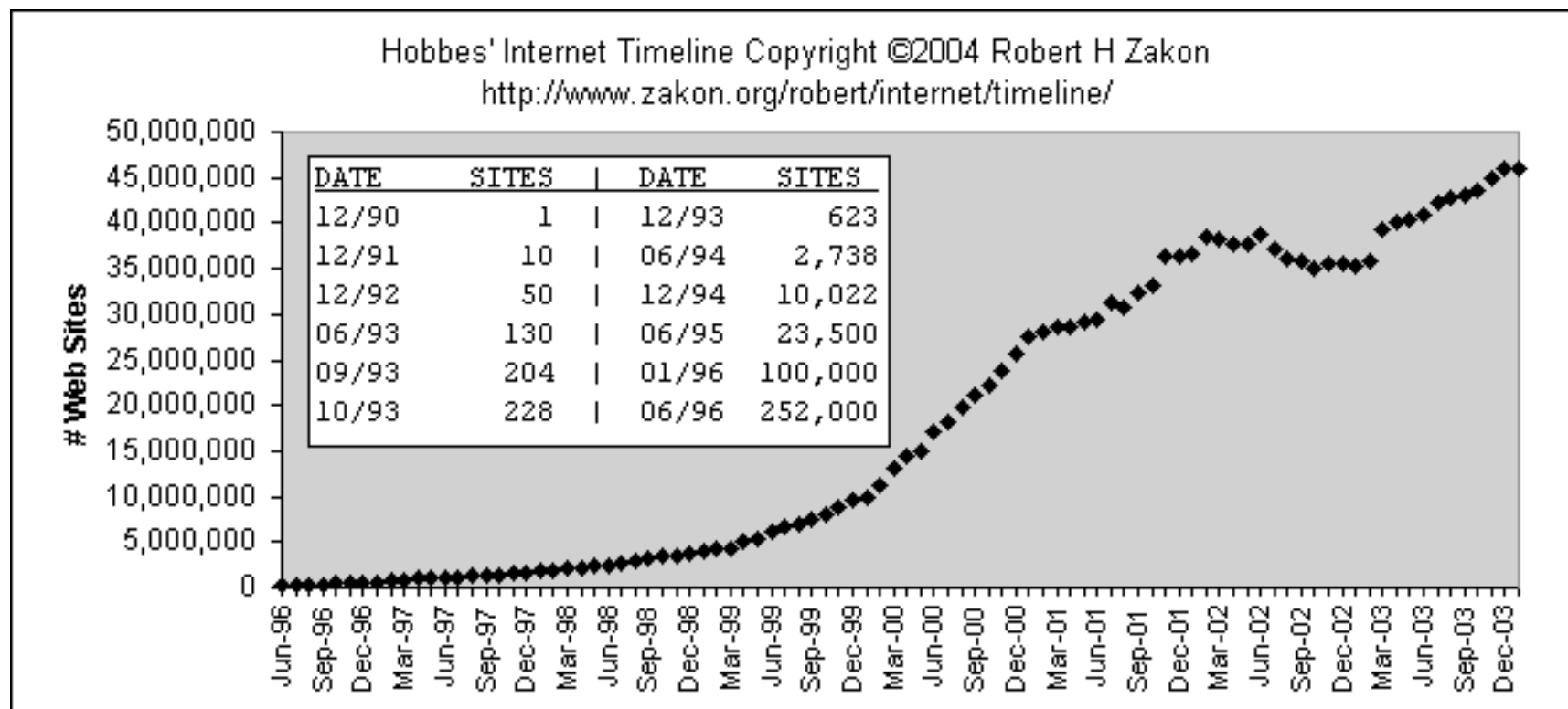


Introduction

Internet host

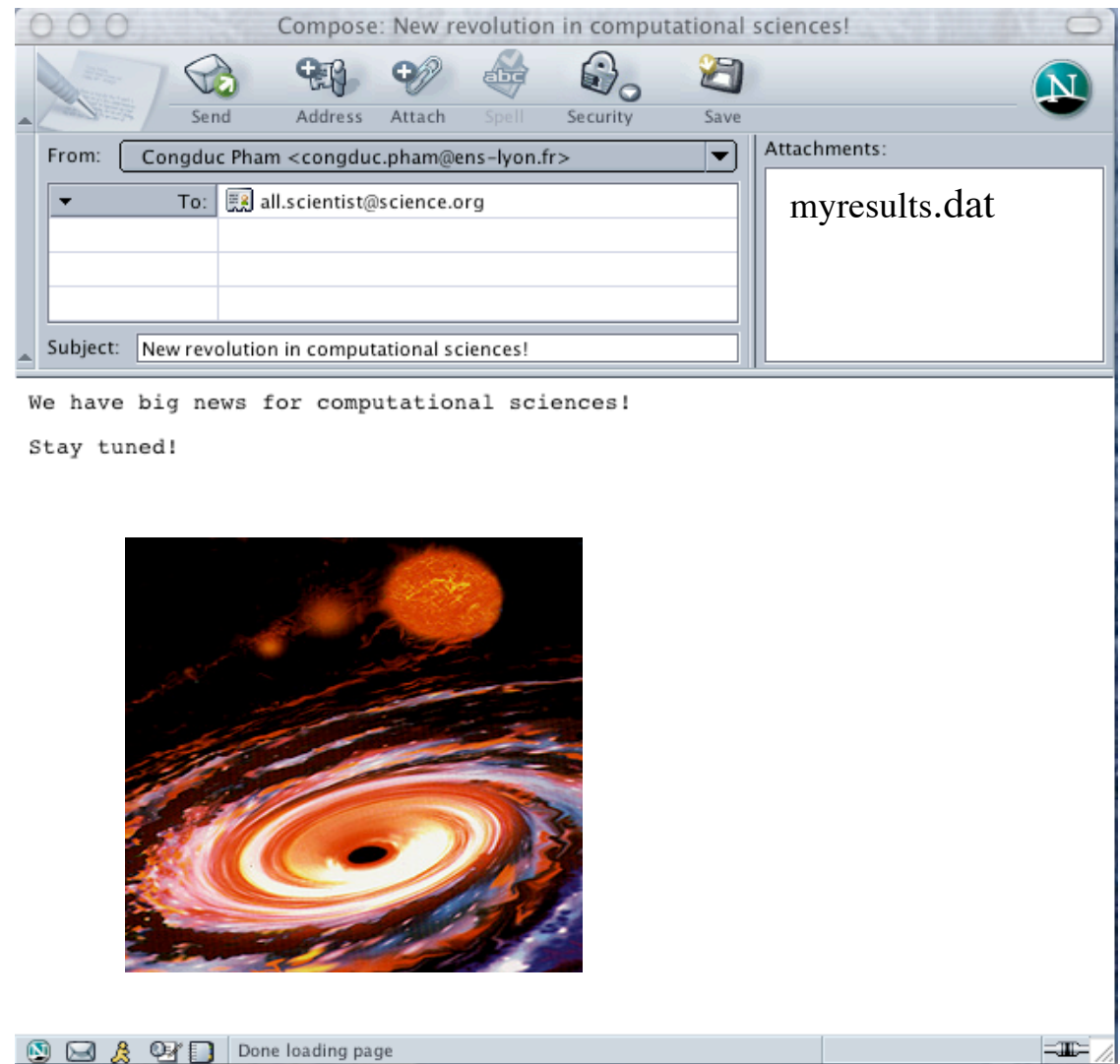


www.web-the-big-bang.org



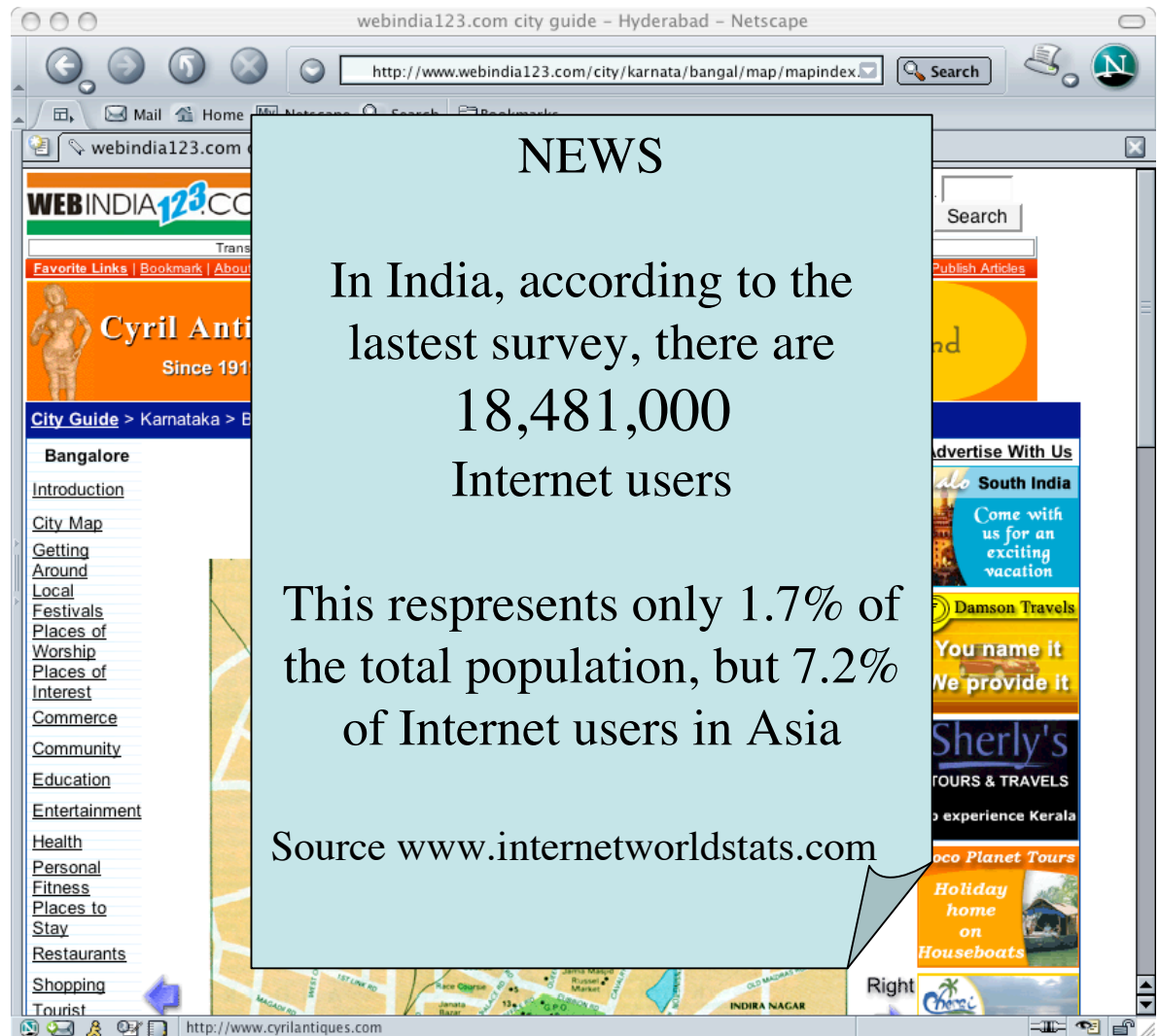
Internet usage: e-mail...

- ❑ Convenient way to communicate in an informal manner
- ❑ Attachments as a easy way to exchange data files, images...



...and surfing the web

- ❑ A true revolution for rapid access to information
- ❑ Increasing number of apps:
 - ❑ e-science,
 - ❑ e-commerce, B2B, B2C,
 - ❑ e-training, e-learning,
 - ❑ e-tourism
 - ❑ ...



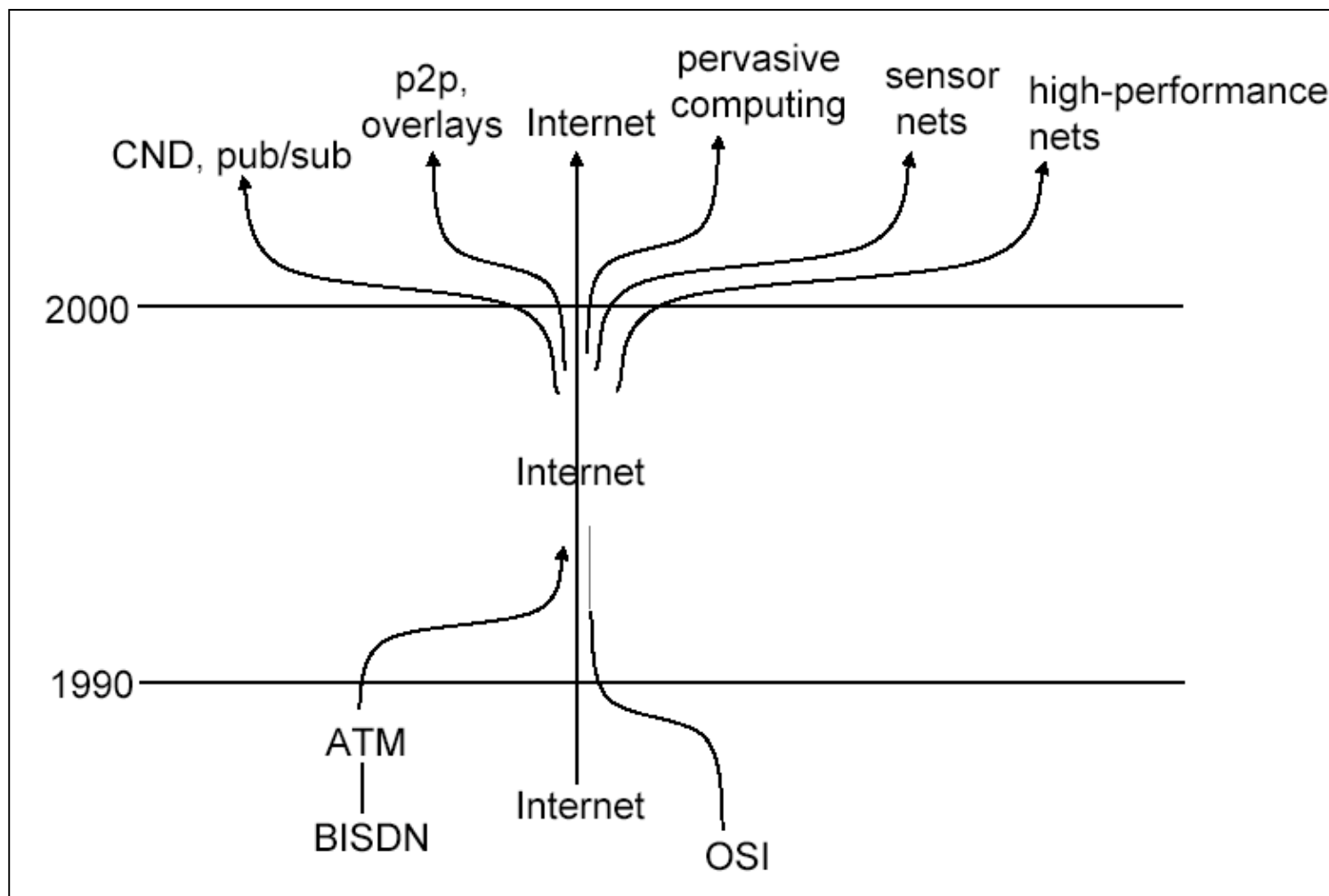
NEWS

In India, according to the latest survey, there are 18,481,000 Internet users

This represents only 1.7% of the total population, but 7.2% of Internet users in Asia

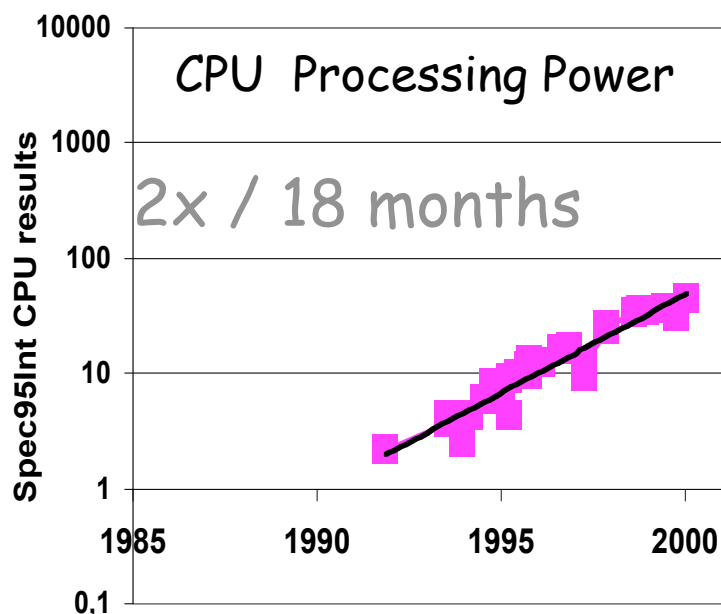
Source www.internetworldstats.com

Towards all IP

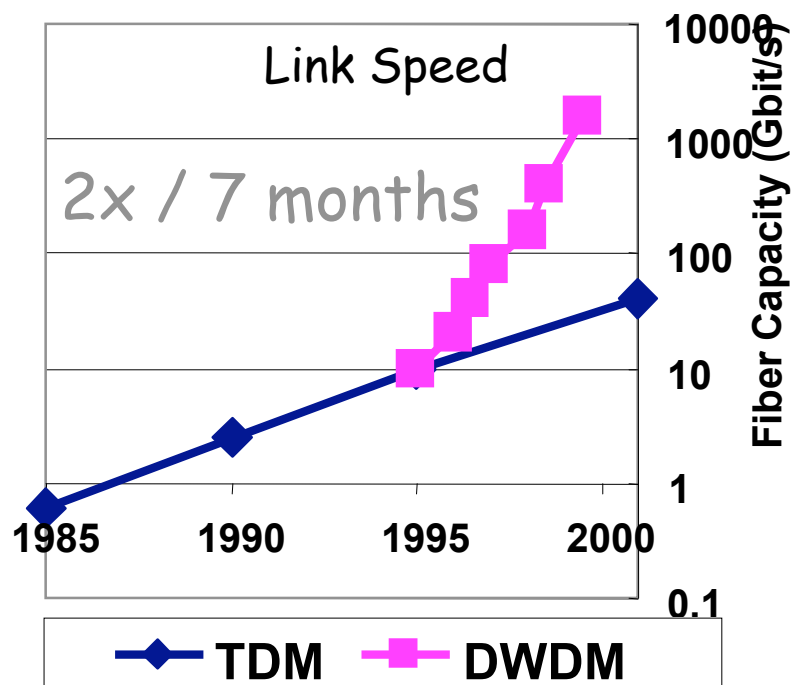


From Jim Kurose

The optical revolution



From McKeown

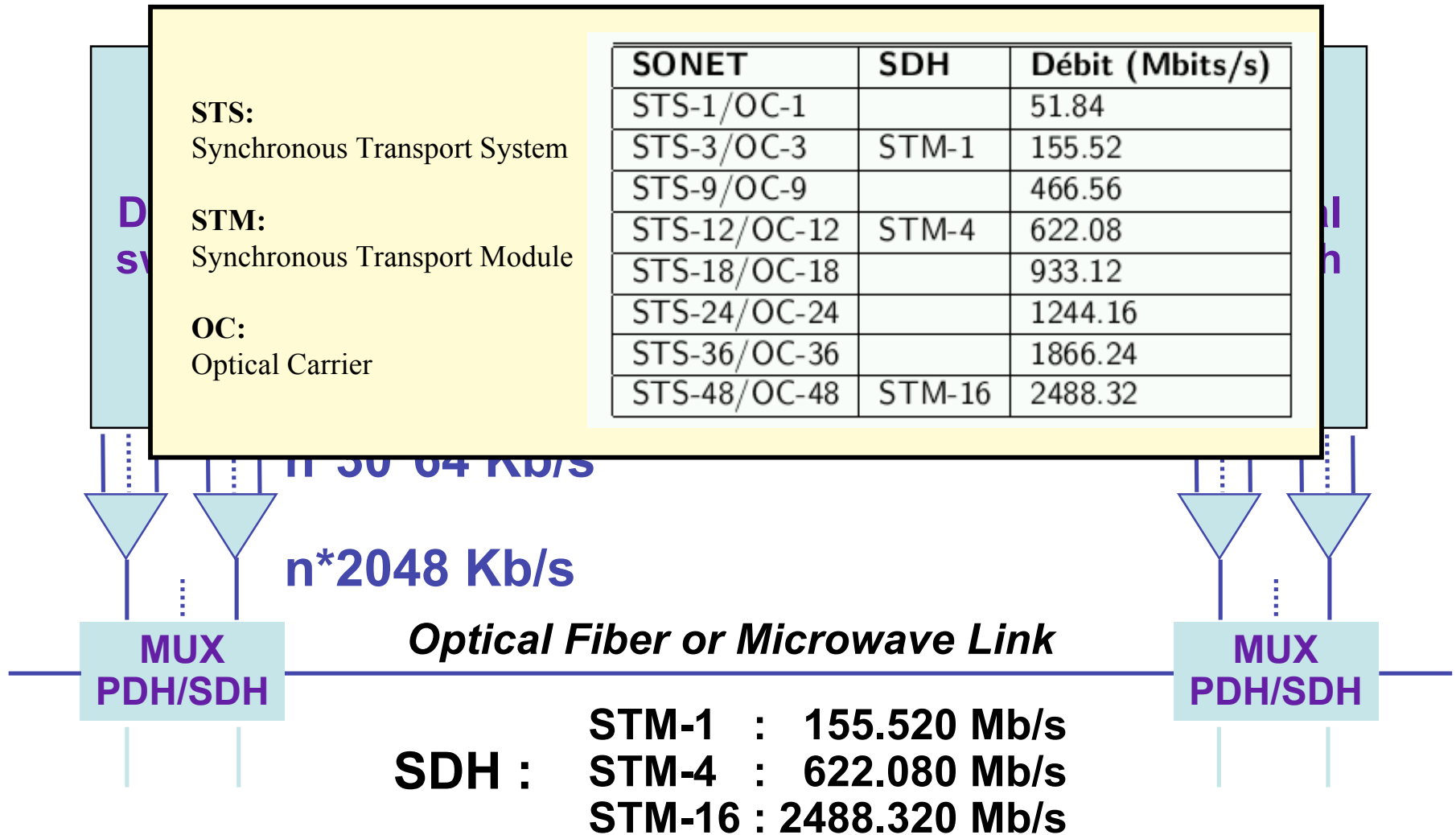


Demand: about 111 million km of cabled optical fiber / year



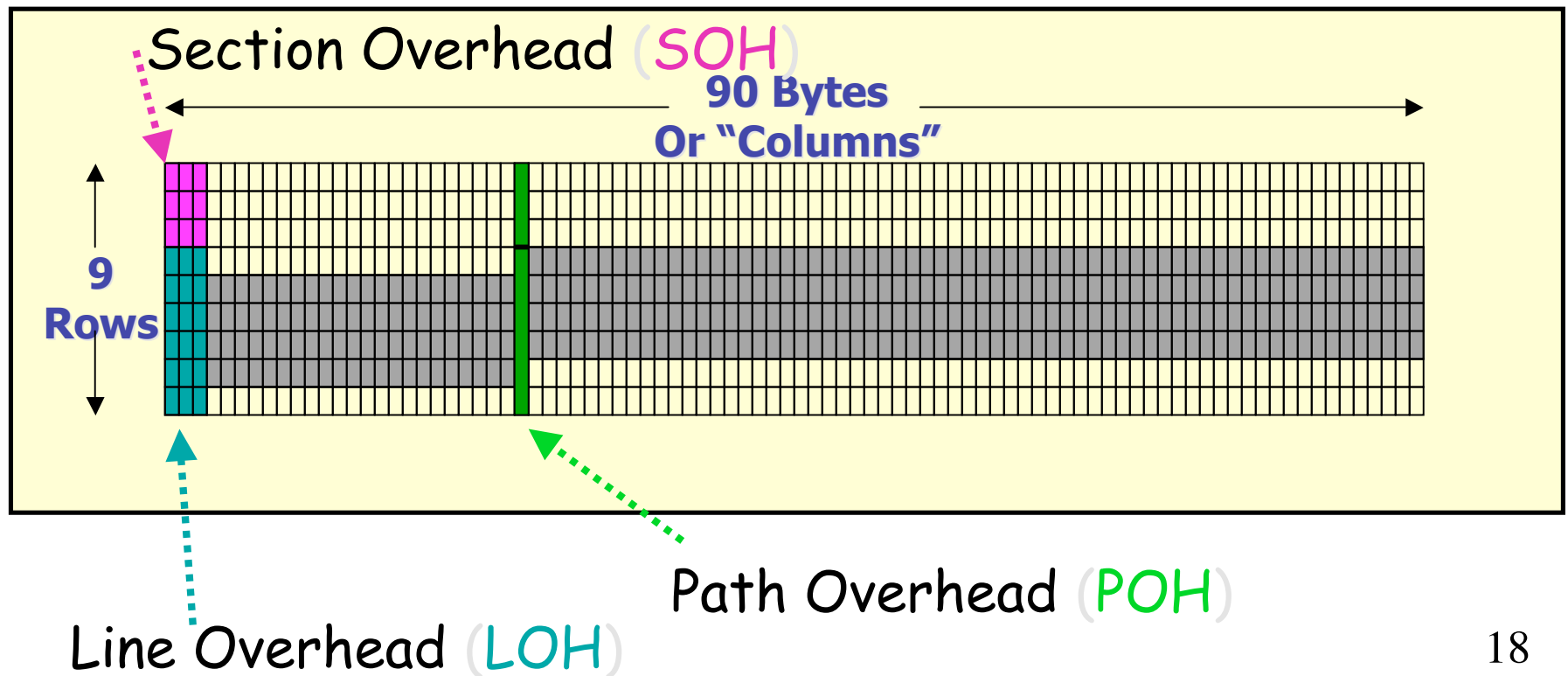
SONET/SDH in the core

95% of exploited OF use SONET/SDH

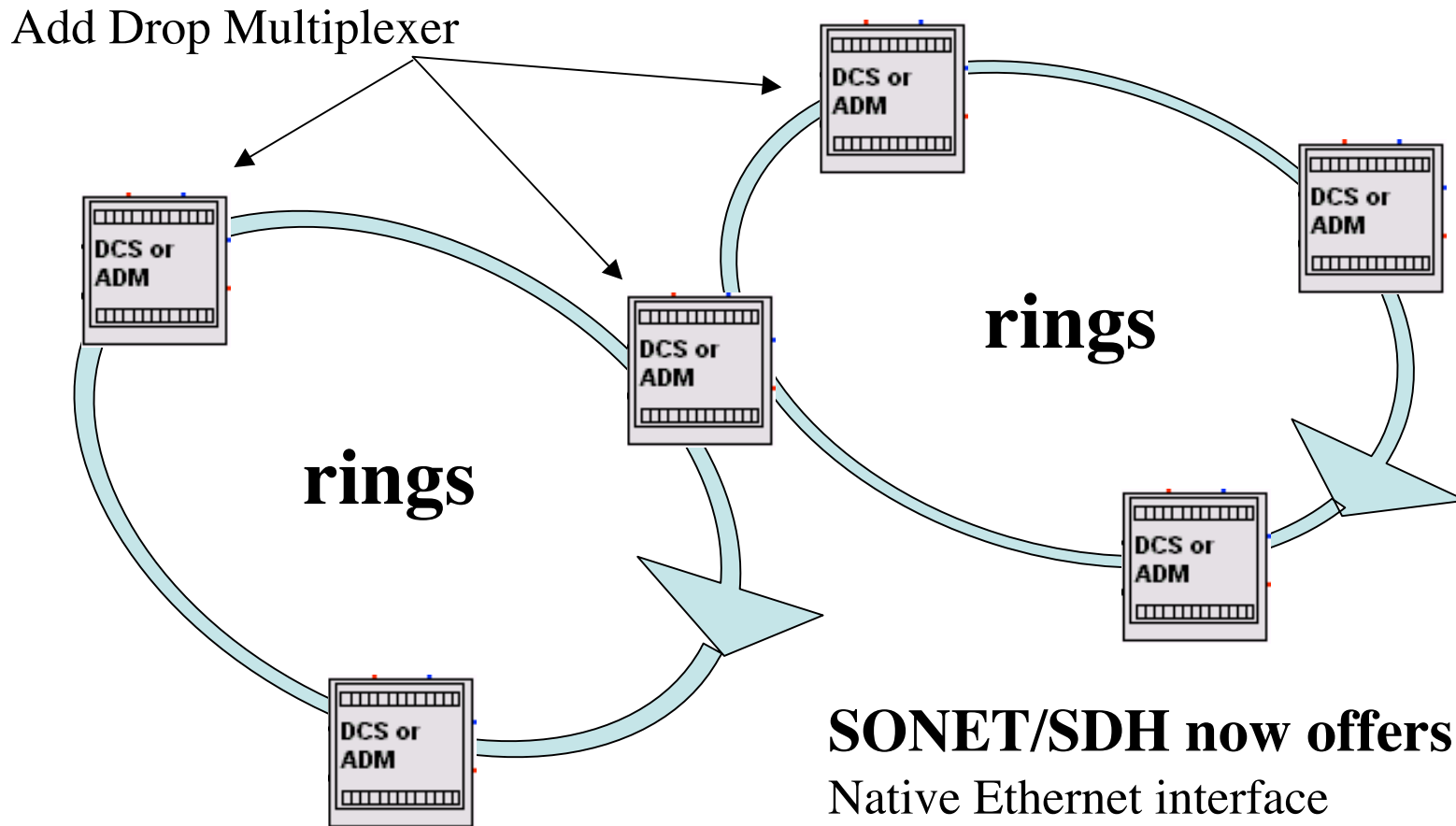


The SONET frame

- ❑ Basic frame length is 810 bytes (TDM)
 - ❑ Sent every 125us, raw throughput of 51.84 Mbits/s (STS-1)
 - ❑ Better seen as a block with 90 columns and 9 lines



SONET/SDH transport network infrastructure

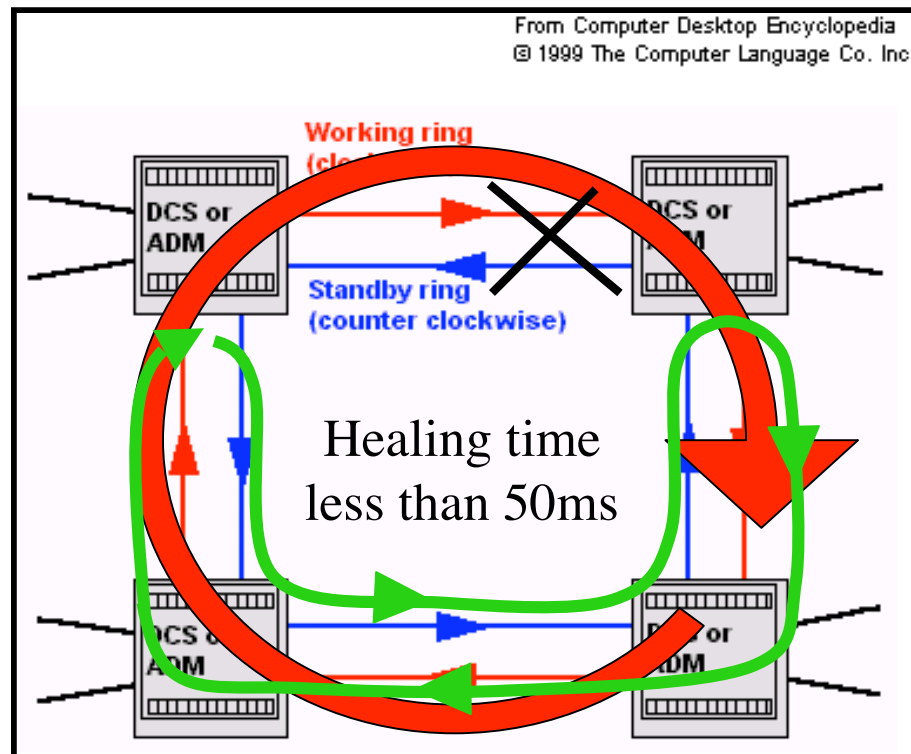


SONET/SDH now offers
Native Ethernet interface
Generic Framing Procedure
Virtual Concatenation

SONET/SDH and resiliency

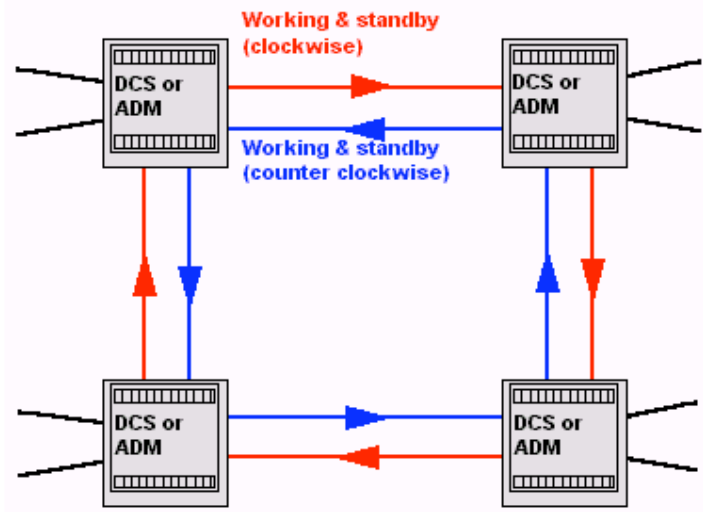
- ❑ SONET/SDH has built-in fault-tolerant features with multiple rings
- ❑ Ex: simple case

DCS
(Digital Cross-Connects)



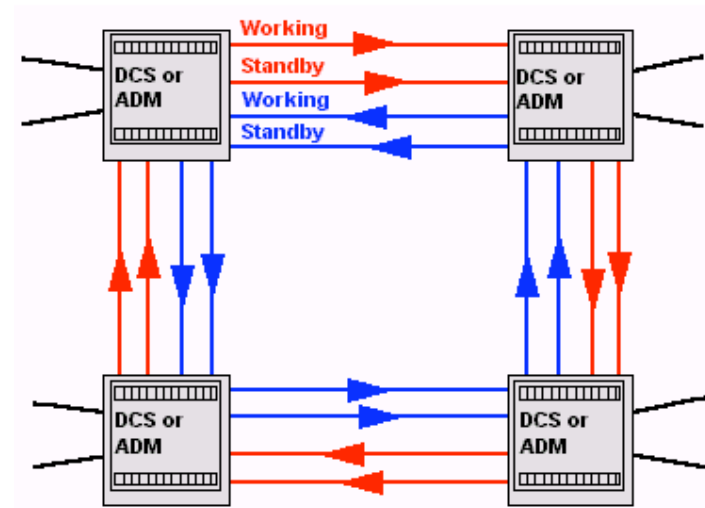
High availability in SONET/SDH networks

From Computer Desktop Encyclopedia
© 1999 The Computer Language Co. Inc.



bi-directional

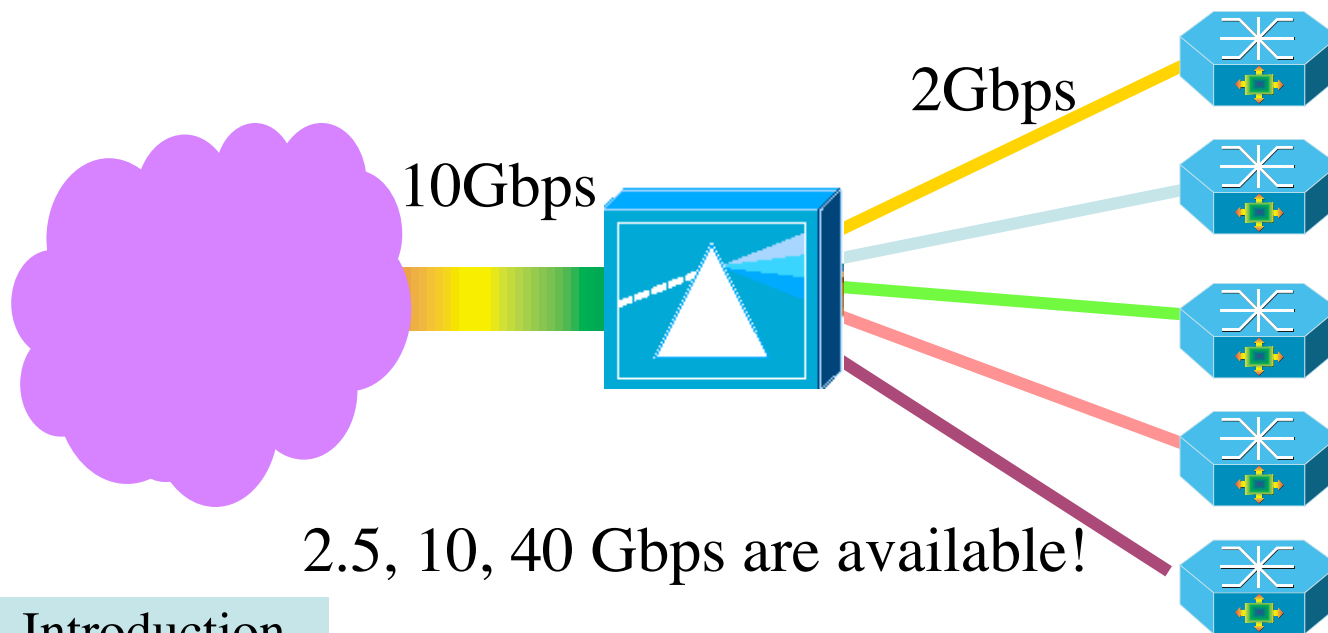
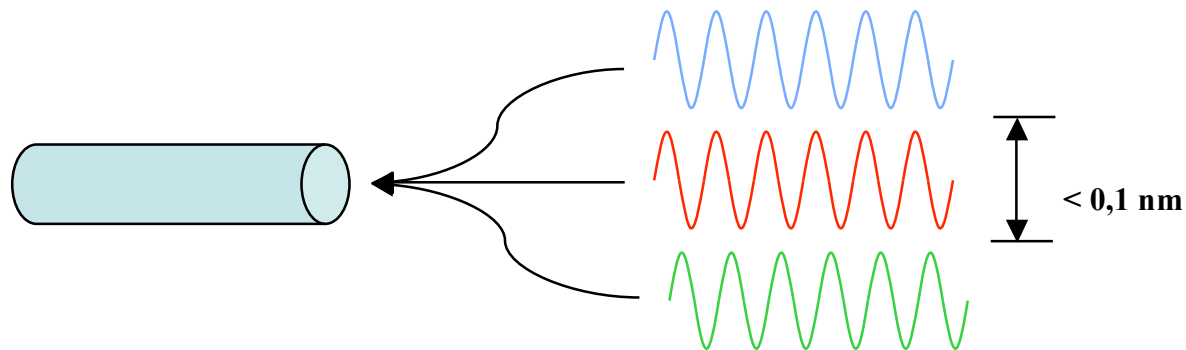
From Computer Desktop Encyclopedia
© 1999 The Computer Language Co. Inc.



Found in most operators' networks

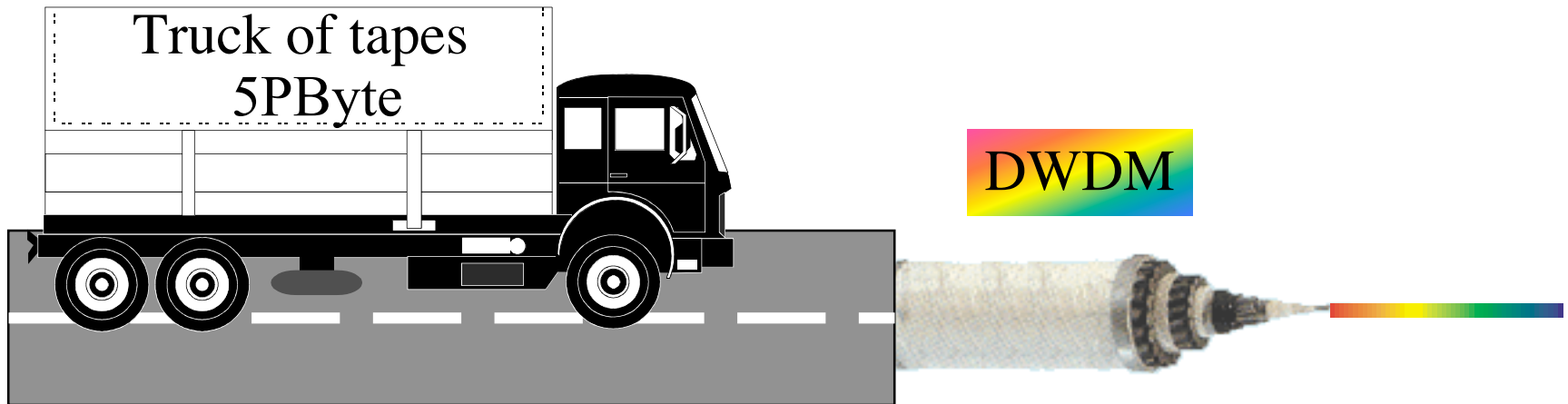
DWDM, bandwidth for free?

DWDM: Dense Wavelength Division Multiplexing



From Computer Desktop Encyclopedia
Reproduced with permission.
© 2001 Metromedia Fiber Network

The information highways



NEWS of Dec 15th, 2004

- 3 A throughput of 1.28 Tbits/s has been achieved on a 430kms regular monomode fiber between France Telecom and Deutsch Telecom using 8 DWDM channels (EU project TOPRATE)

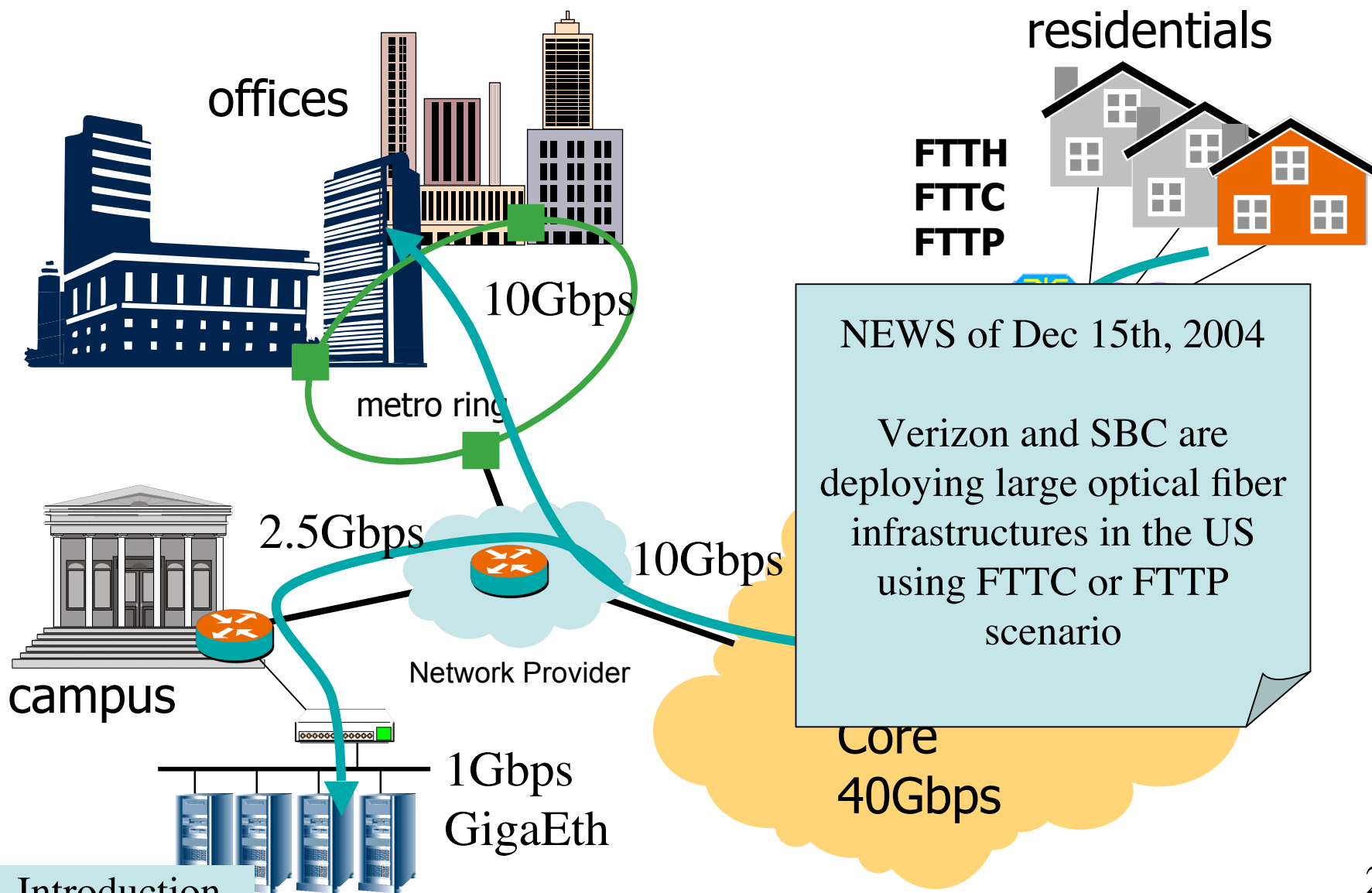
Revisiting the truck of tapes

(18 of 18)

Consider one fiber

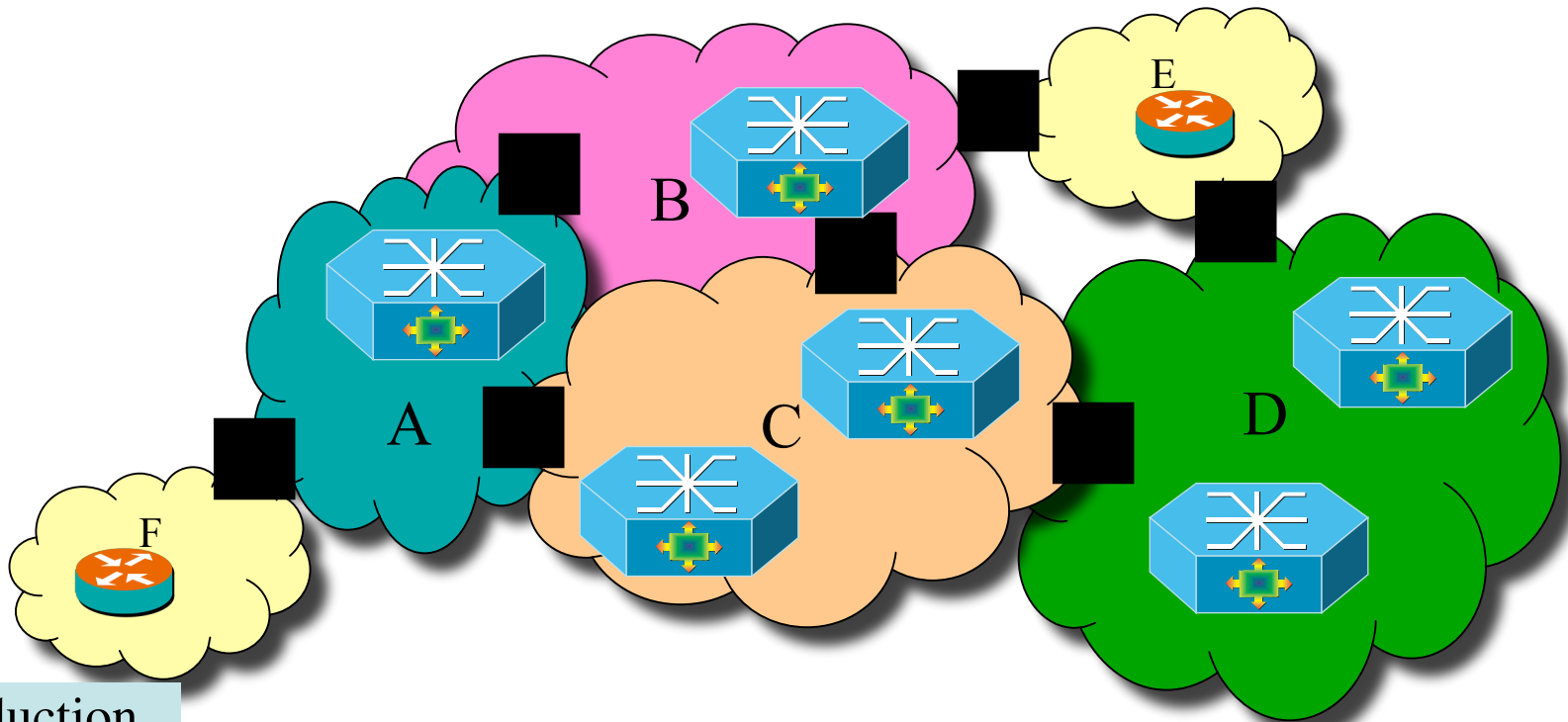
- Current technology allows for 320 λ in one of the frequency bands
- Each λ has a bandwidth of 40 Gbit/s
- Transport: $320 * 40 * 10^9 / 8 = 1600$ GByte/sec
- Take a 10 metric ton truck
- One tape contains 50 Gbyte, weights 100 gr
- Truck contains $(10000 / 0.1) * 50$ Gbyte = 5 PByte
- Truck / fiber = 5 PByte / 1600 GByte/sec = 3125 s \approx one hour
- For distances further away than a truck drives in one hour (50 km) minus loading and handling 100000 tapes **the fiber wins!!!**

Fibers everywhere?



Operator's infrastructure

- ❑ Backbones are optical: OC48 (2.5Gbps), OC192 (10Gbps), OC768 (40Gbps) soon
- ❑ New technologies deployed by operators, POPs available worldwide



New applications on the information highways

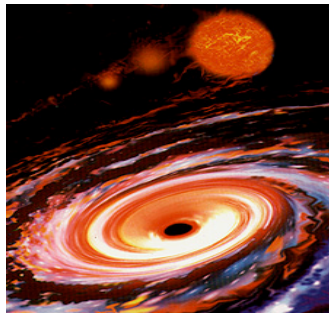
Think about...

- video-conferencing
- video-on-demand
- interactive TV programs
- remote archival systems
- tele-medecine
- virtual reality, immersion systems
- high-performance computing, grids
- distributed interactive simulations



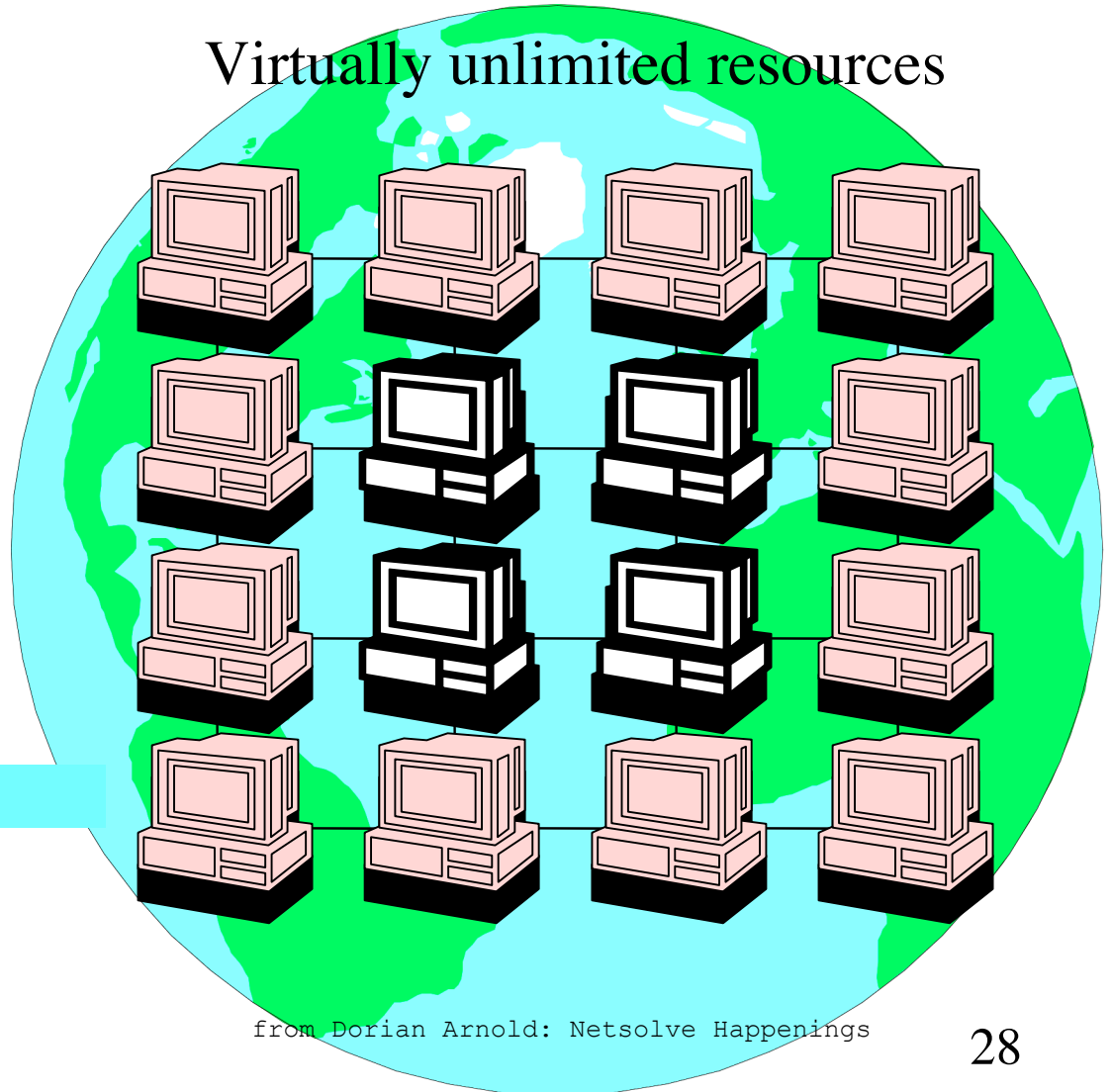
Computational grids

user application



1PFlops

Virtually unlimited resources

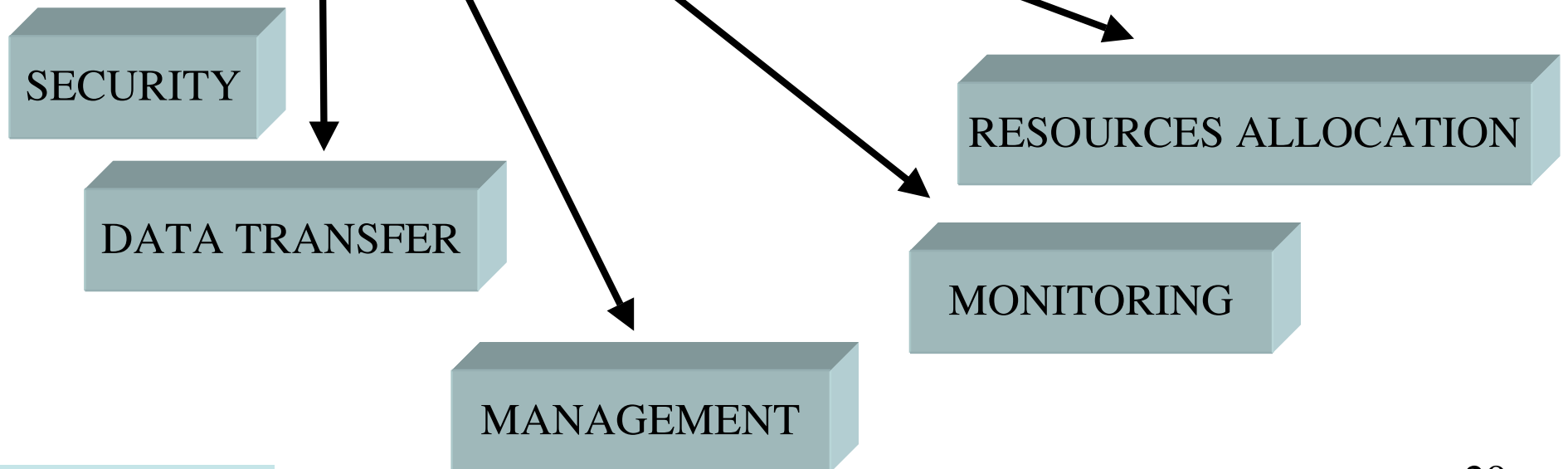


from Dorian Arnold: Netsolve Happenings

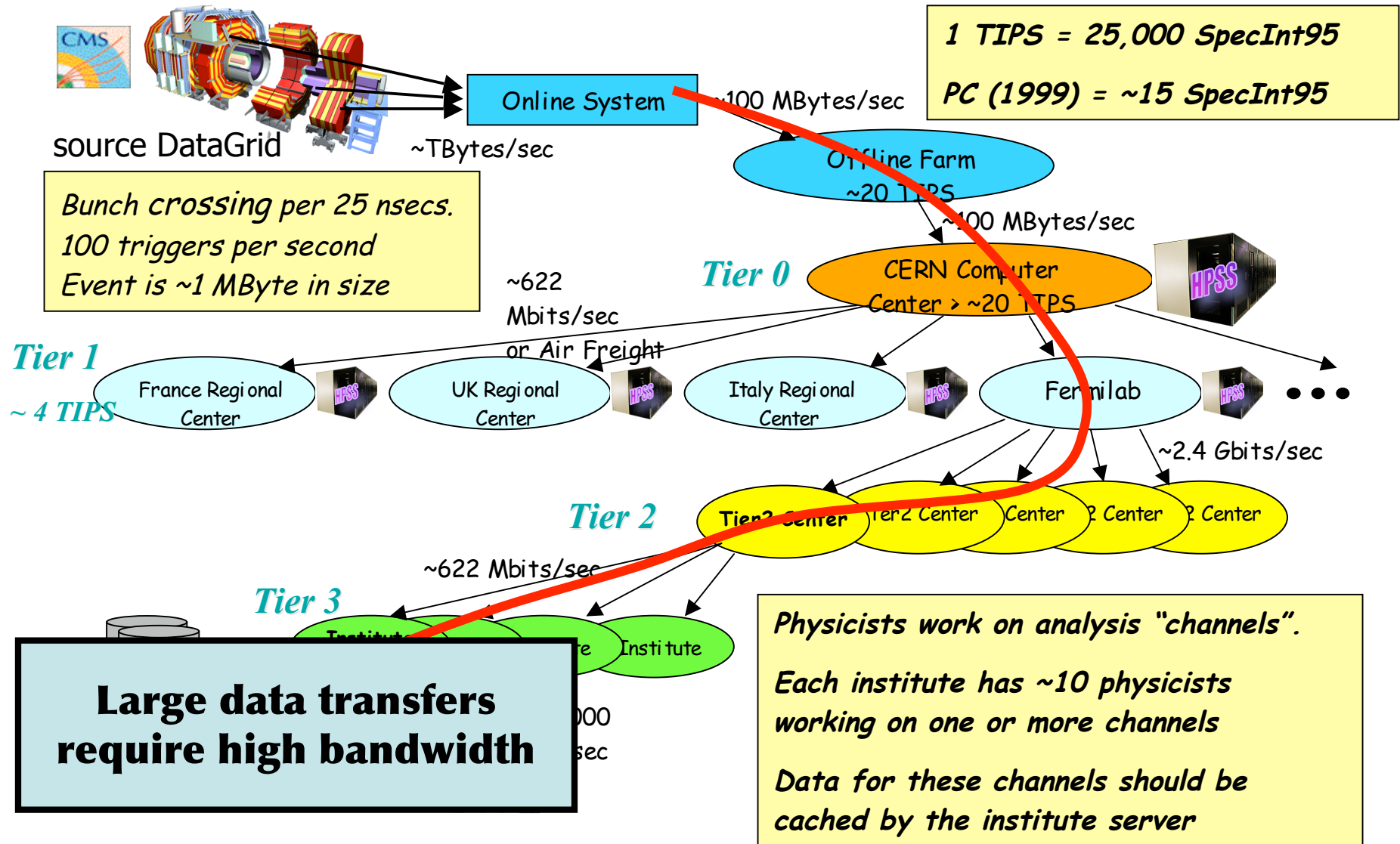
One Grid definition

The Grid

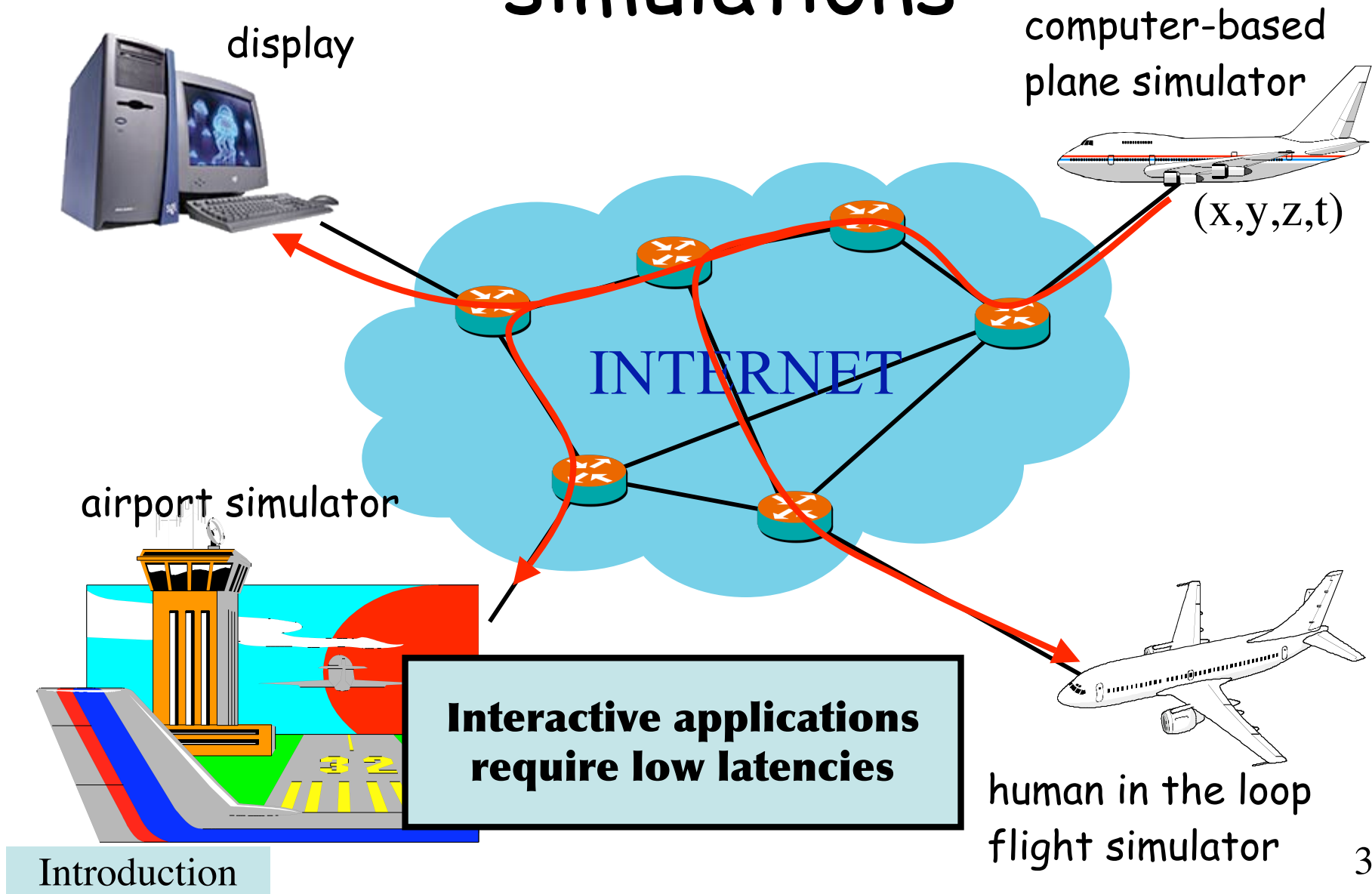
“A Grid is a collection of distributed computing resources over **network** that appear to an user or an application as one large virtual computing system”



Distributed Databases



Wide-area interactive simulations



In search for the perfect grid

For me, as a user, a computational grid should be:

Easy to use

Fast & Performant

Reliable

Transparent

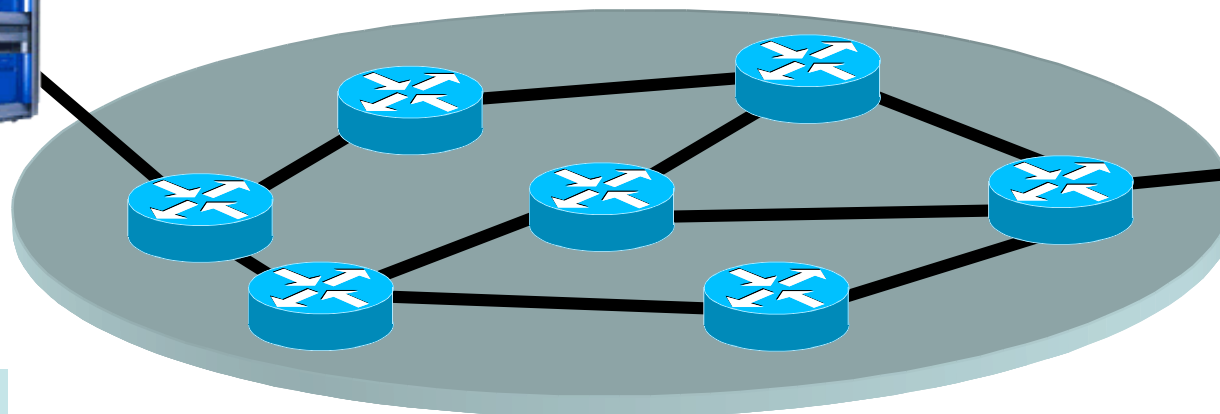


Networking issues in the Grid

Manages network resources (link, routers, bandwidth) to offer reliability and guaranteed/predictable performances

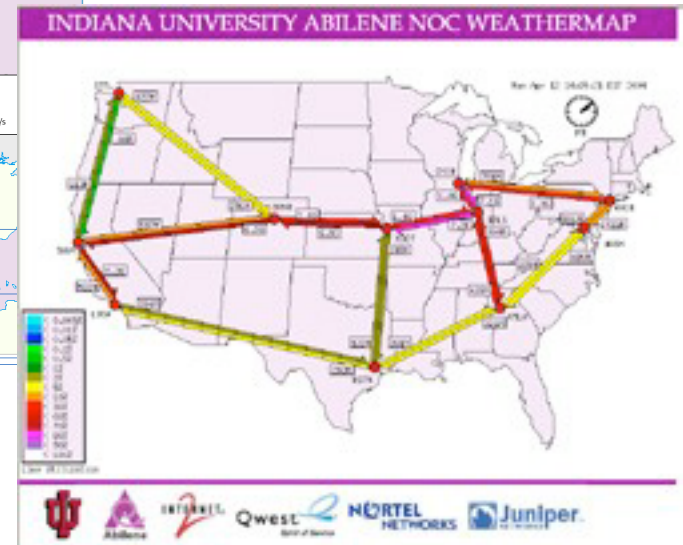
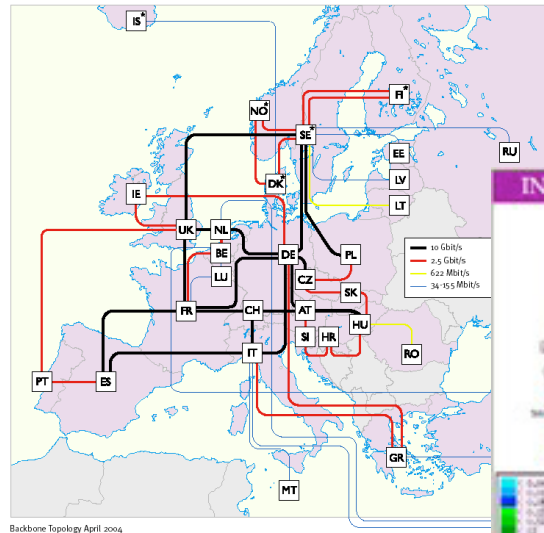
Optimizes communication protocols to offer full/optimal utilization of network resources

Deploys new technologies to offer new value-added/efficient communication features

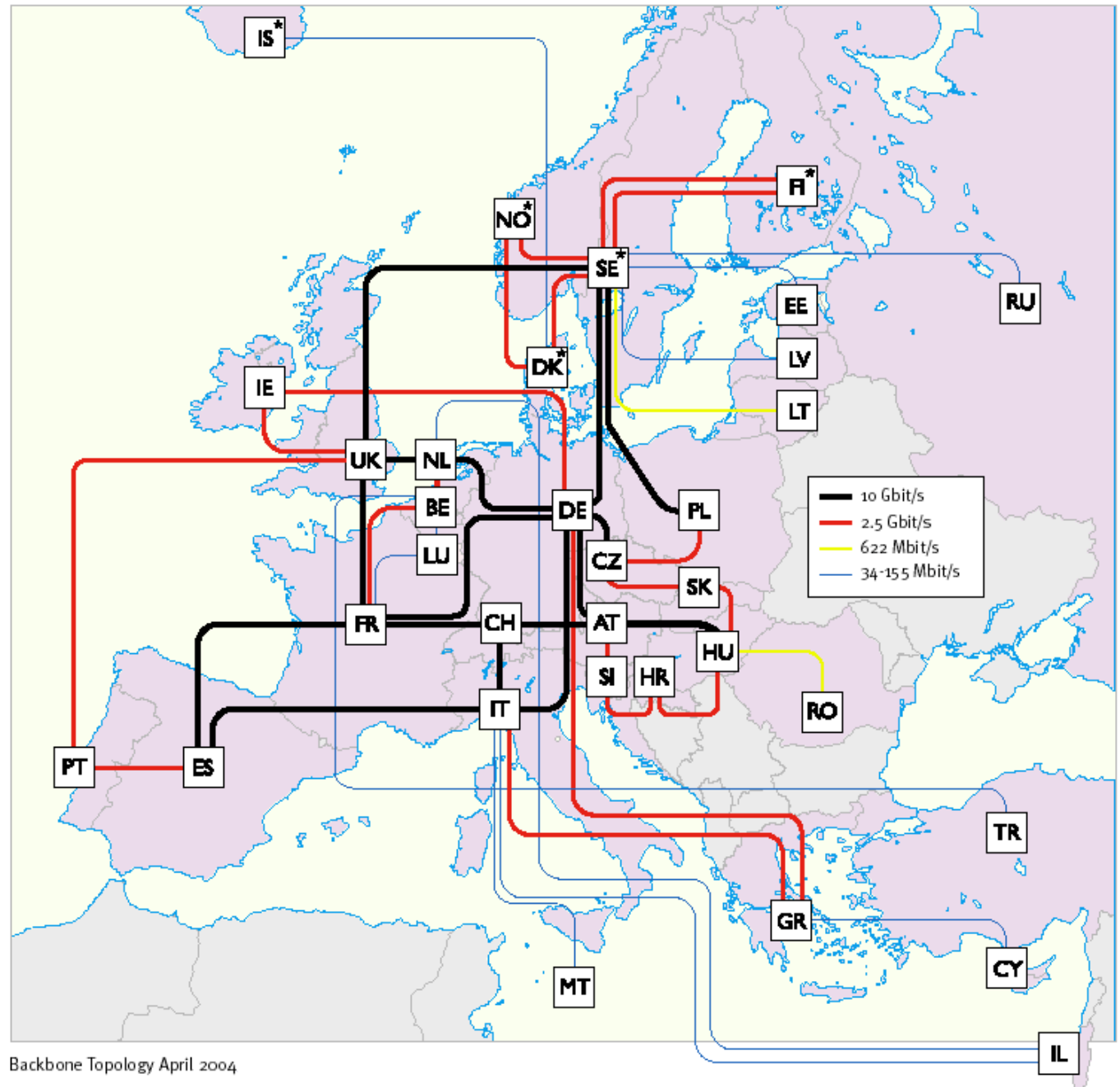


The new networks

- vBNS
- Abilene
- SUPERNET
- DREN
- CA*NET
- GEANT
- DATATAG
- ...much more to come!



GEANT



Backbone Topology April 2004

Limitations of the current Internet

- ❑ Bandwidth
 - ❑ Raw bandwidth is not a problem: DWDM
 - ❑ Provisioning bandwidth on demand is more problematic
- ❑ Latency
 - ❑ Mean latencies on Internet is about 80-160ms
 - ❑ Bounding latencies or ensuring lower latencies is a problem
- ❑ End-to-end performances
 - ❑ Links are getting faster and faster!
 - ❑ Why my FTP is still going so slow?
- ❑ Communication models
 - ❑ Only unicast communications are well-defined: TCP, UDP
 - ❑ Multi-parties communication models are lacking

Application people come from Venus, Networking people come from Mars

Application guys

The network is a cloud.

Only see TCP, IP and sometimes routing protocols

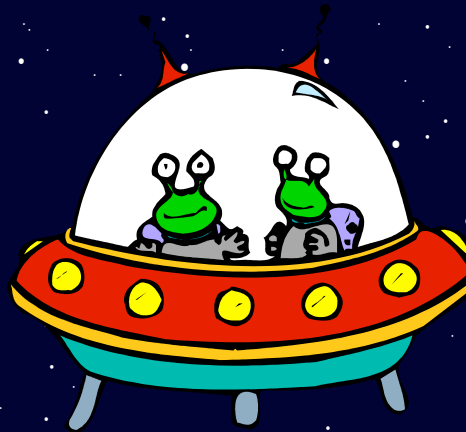
Will use what is available and working!



Networking guys

Don't care about applications!

If any applications then must be mainly FTP and web traffic!



Middleware guys

GGF GHPN

will make a bridge between 3 communities?

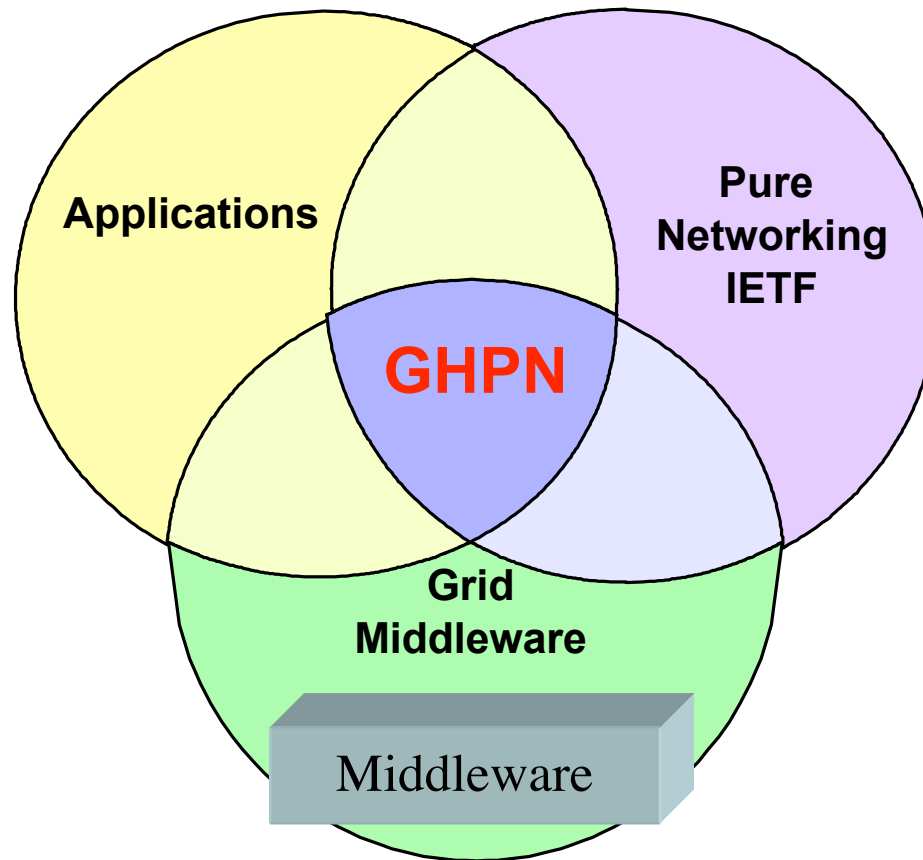
Application guys

The network is a cloud.

Only see TCP, IP and sometimes routing protocols

Will use what is available and working!

Introduction



Networking guys

Don't care about applications!

If any applications then must be mainly FTP and web traffic!

Only problems!

New technologies addressed in this talk

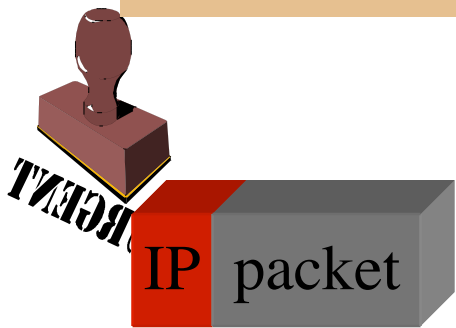
- ❑ More Quality of Service: **Differentiated Services**, who pays more gets more!
- ❑ **Bandwidth provisioning**: MPLS for virtual circuit in the core networks
- ❑ **Beyond TCP**: fast transport protocols for very high-speed networks
- ❑ **Multicast**: enhancing the communication model

Revisiting the *same service* *for all* paradigm

NEW
CHAPTER



Enhancing the best-effort service



**Introduce
Service Differentiation**

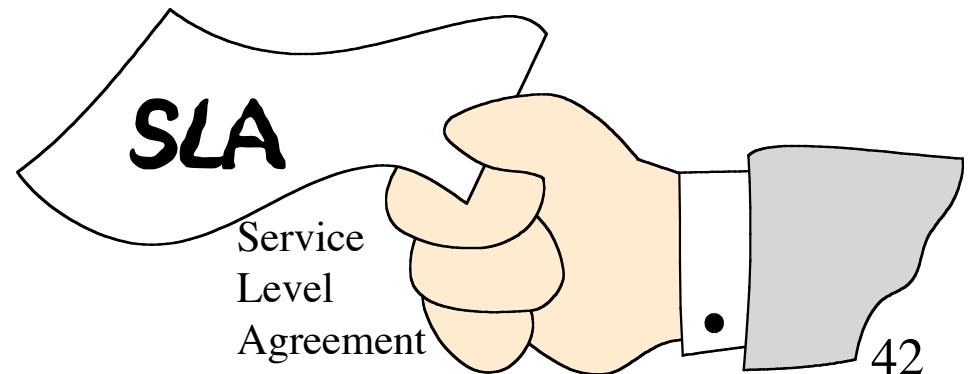
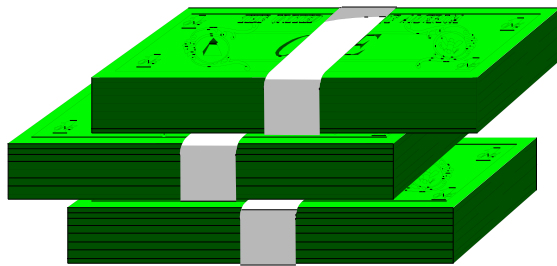


DiffServ

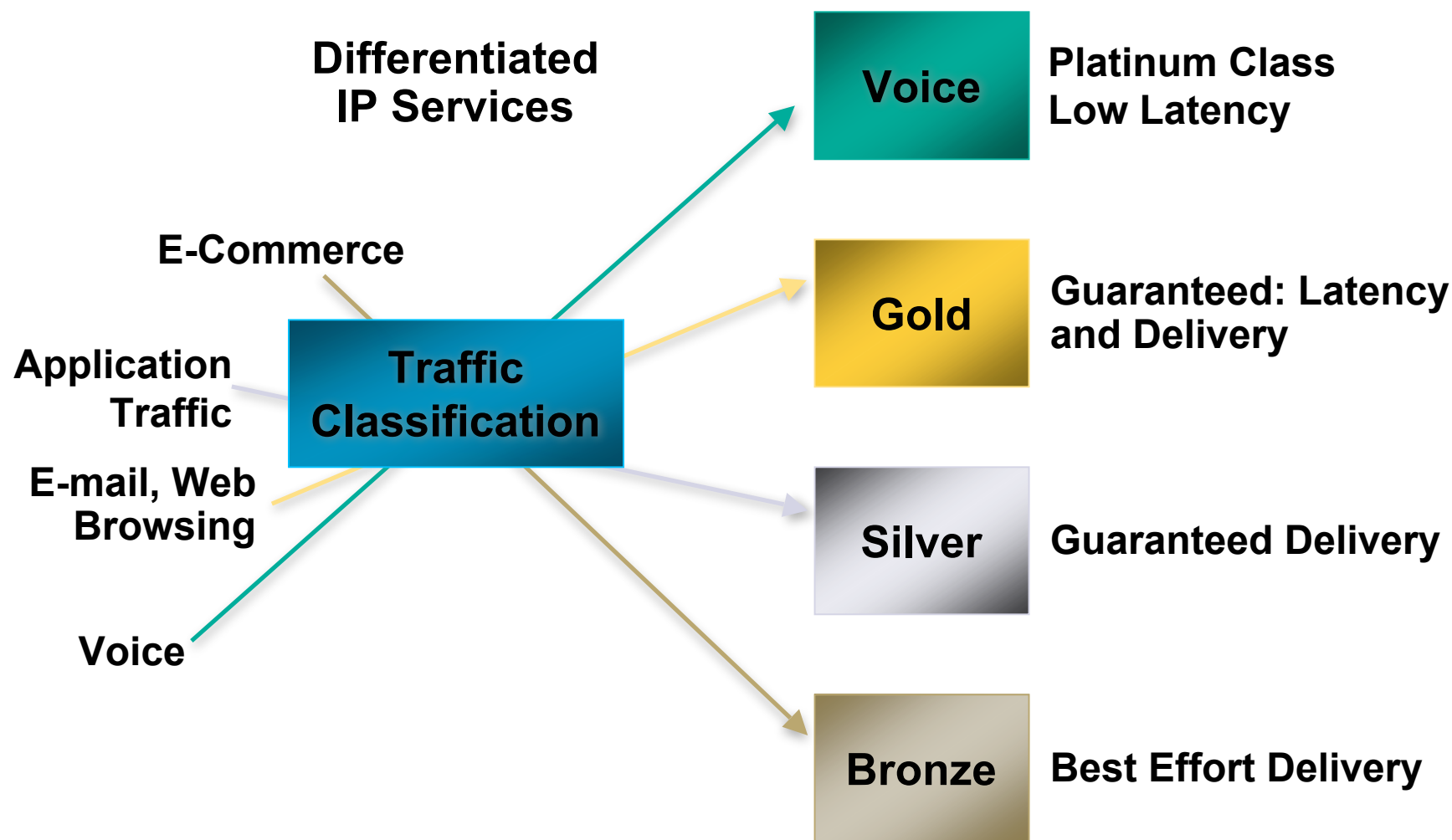
Service Differentiation

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

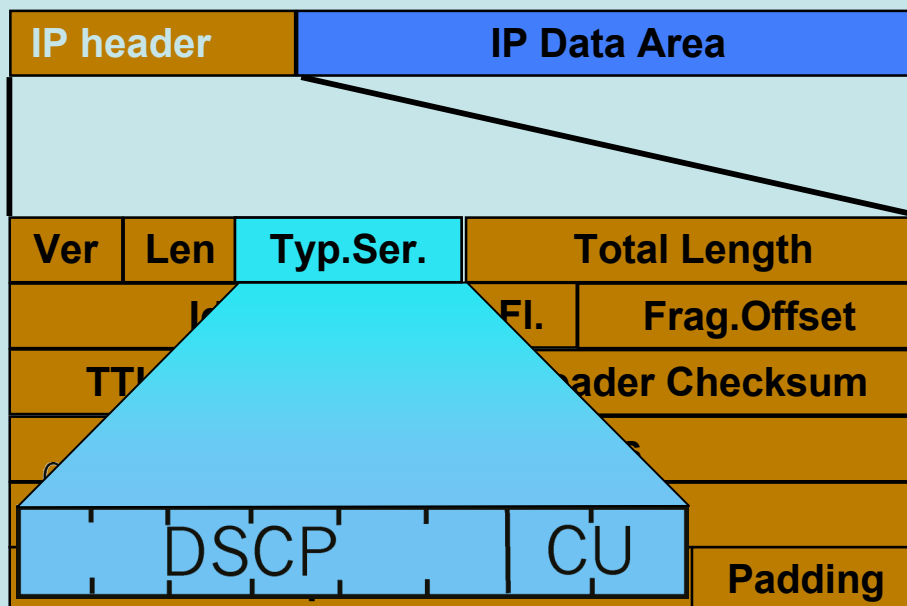
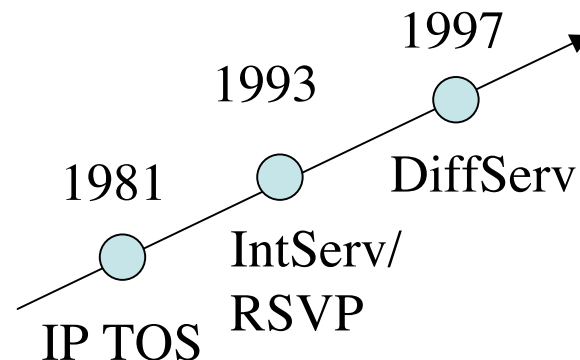


Flow 1
Flow 2
Flow 3
Flow 4
...

10Gbps=2.4Mpps
with 512-byte packets

**Stateful approaches
scalable
at gigabit rates**

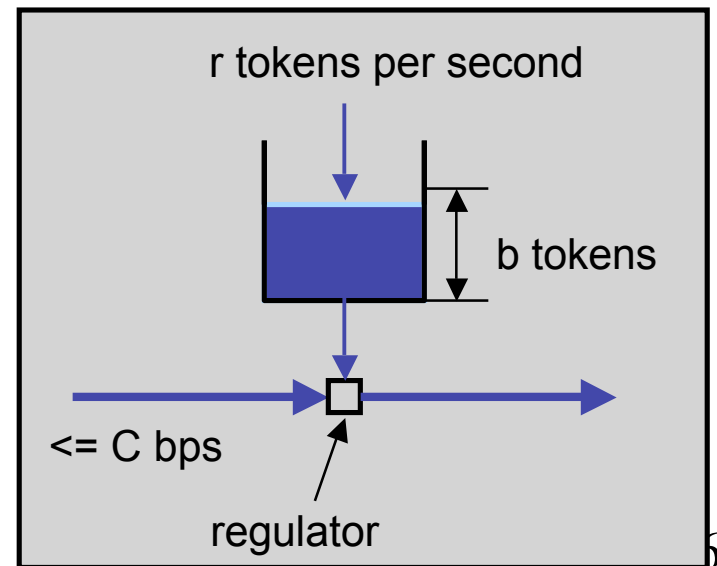
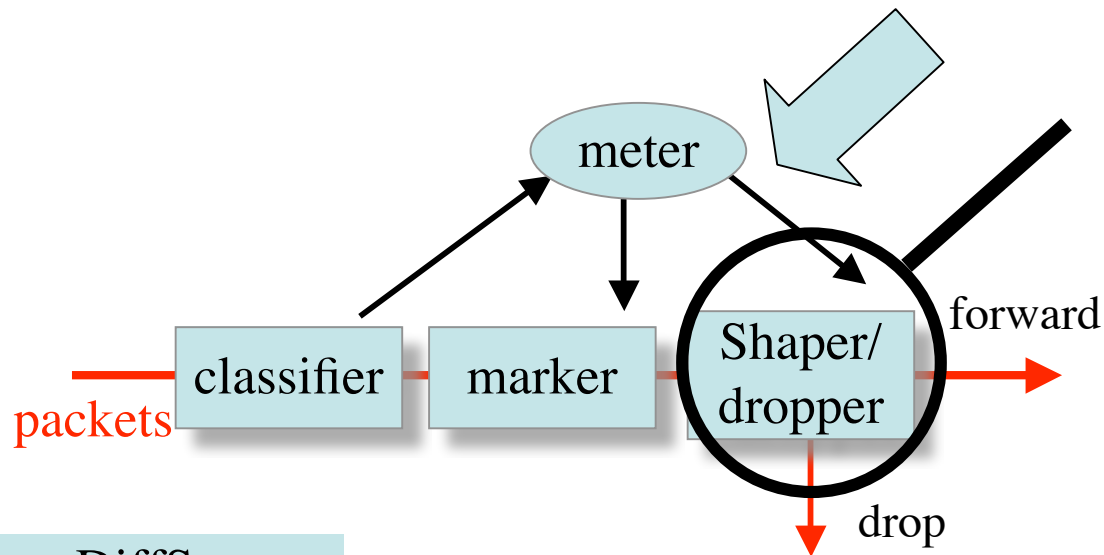
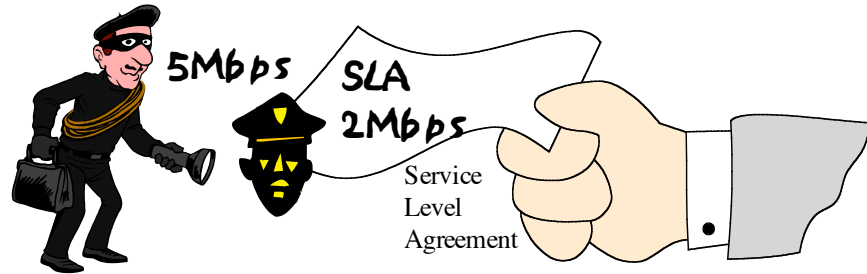
6 bits used for Differentiated Service Code Point (DSCP) and determine PHB that the packet will receive



DiffServ

Traffic Conditioning

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

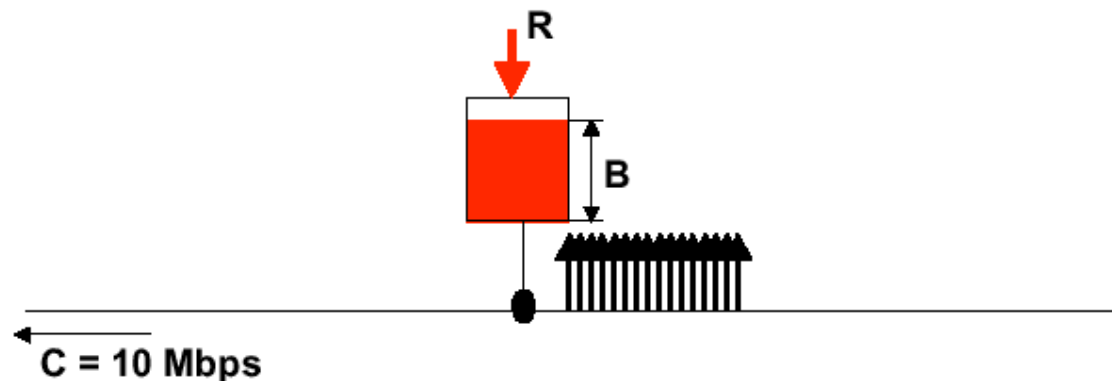


DiffServ

Token Bucket (1)

Example

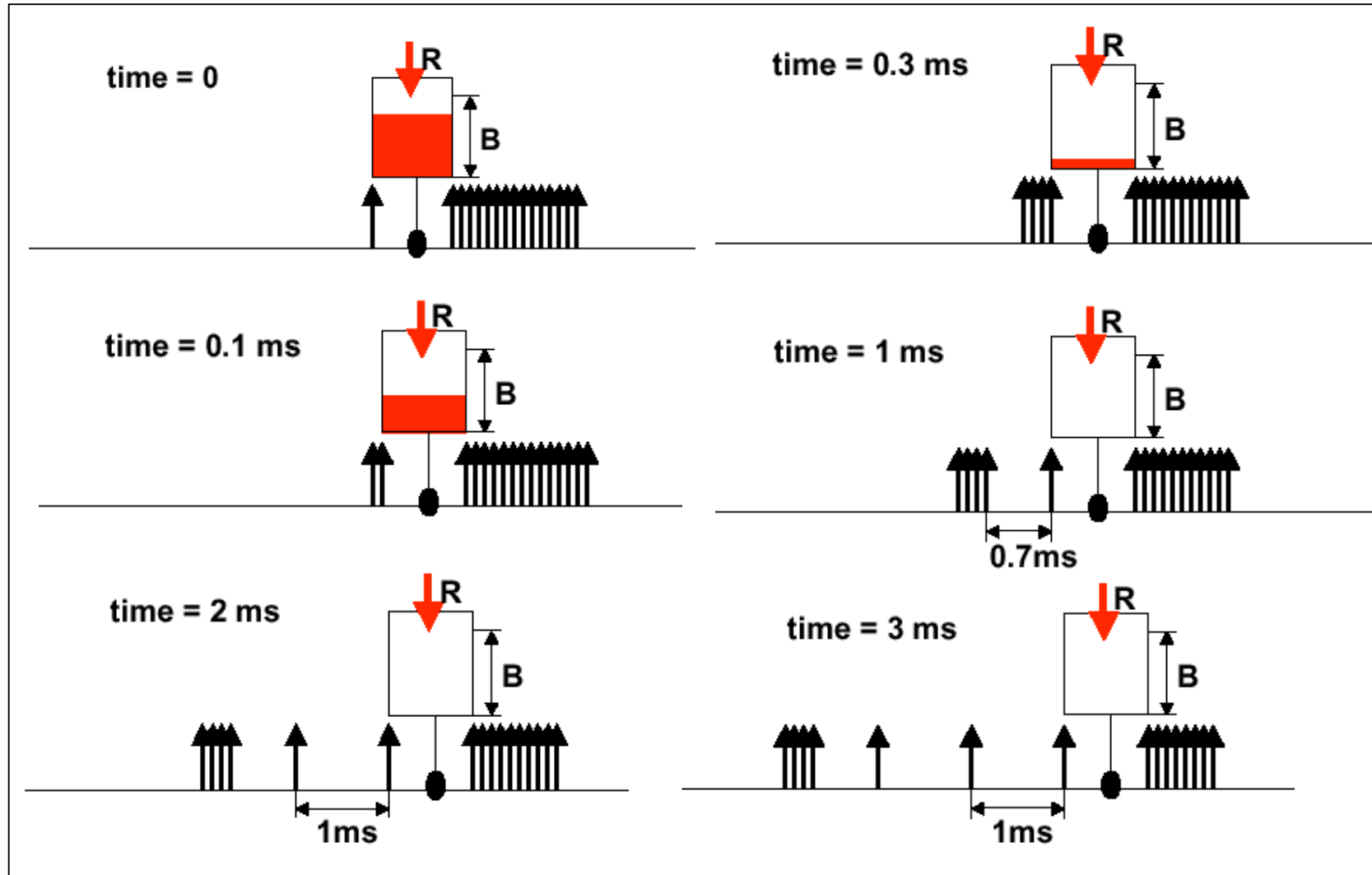
- $B = 4000$ bits, $R = 1$ Mbps, $C = 10$ Mbps
- Packet length = 1000 bits
- Assume the bucket is initially full and a “large” burst of packets arrives



istoica@cs.cmu.edu

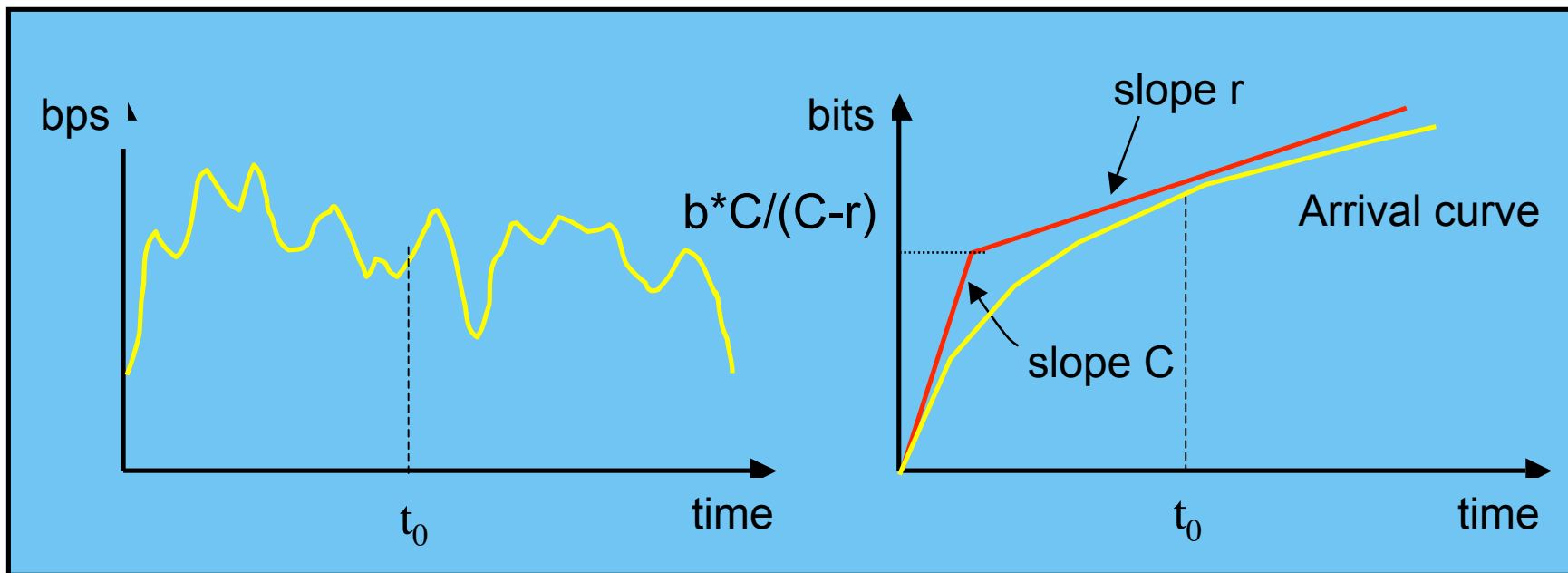
Token Bucket (2)

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

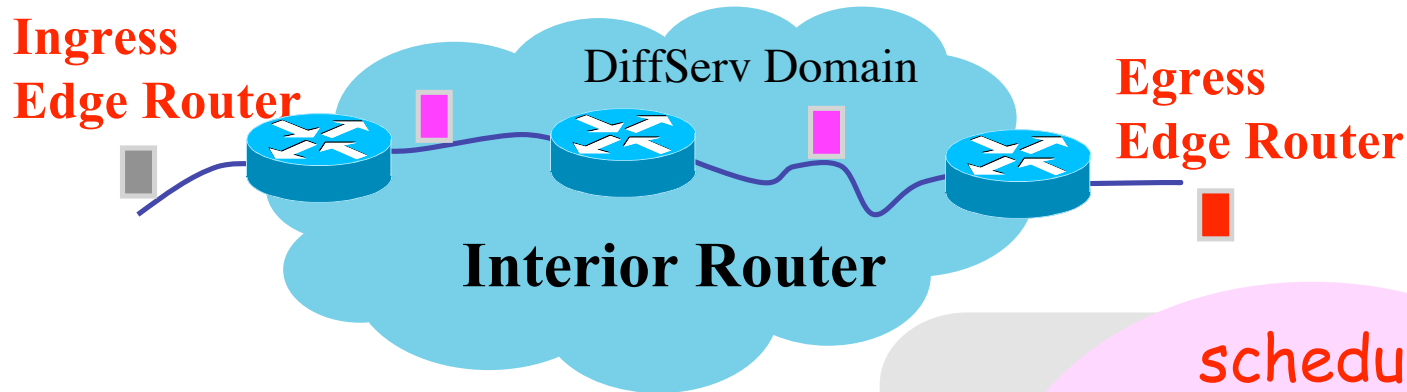


Token Bucket for traffic characterization

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



Differentiated Architecture

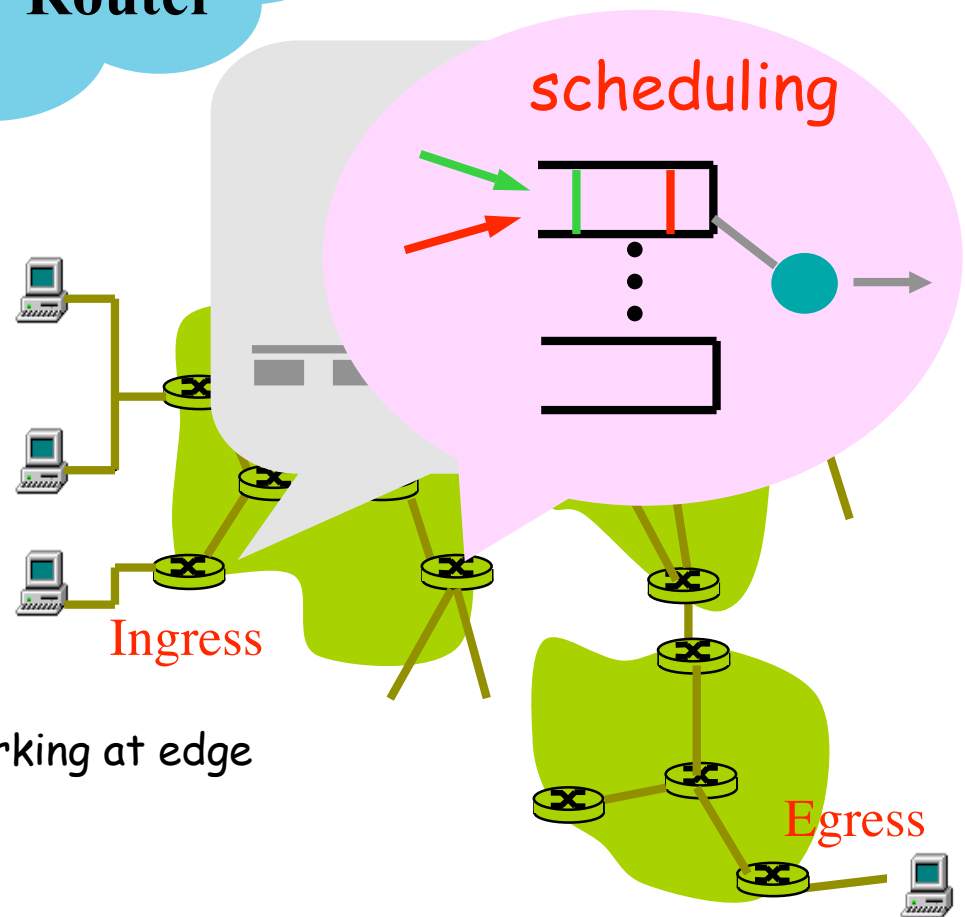


Marking:

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

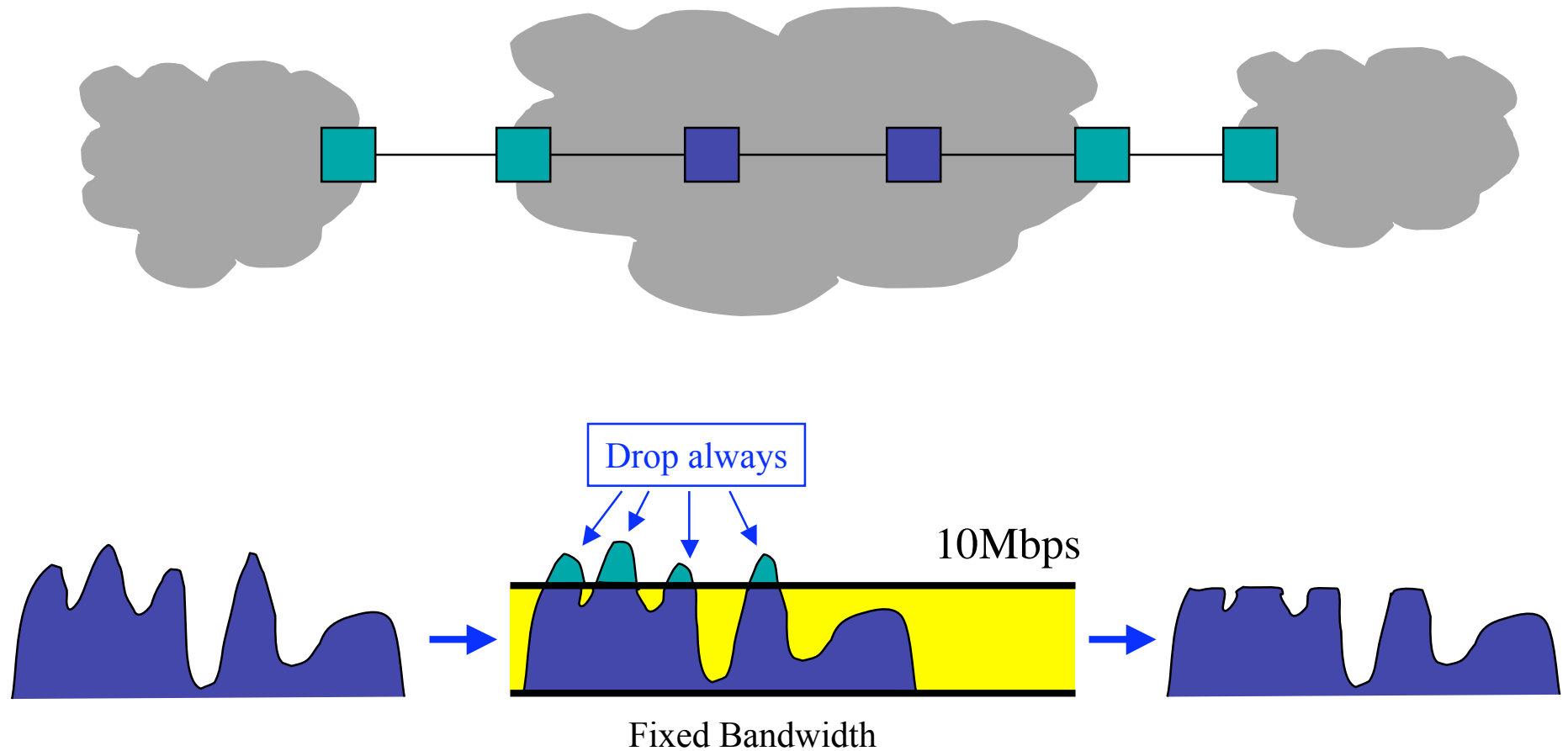
❑ Expedited Forwarding (EF, premium):

- ❑ departure rate of packets from a class equals or exceeds a specified rate (logical link with a minimum guaranteed rate)
- ❑ Emulates leased-line behavior

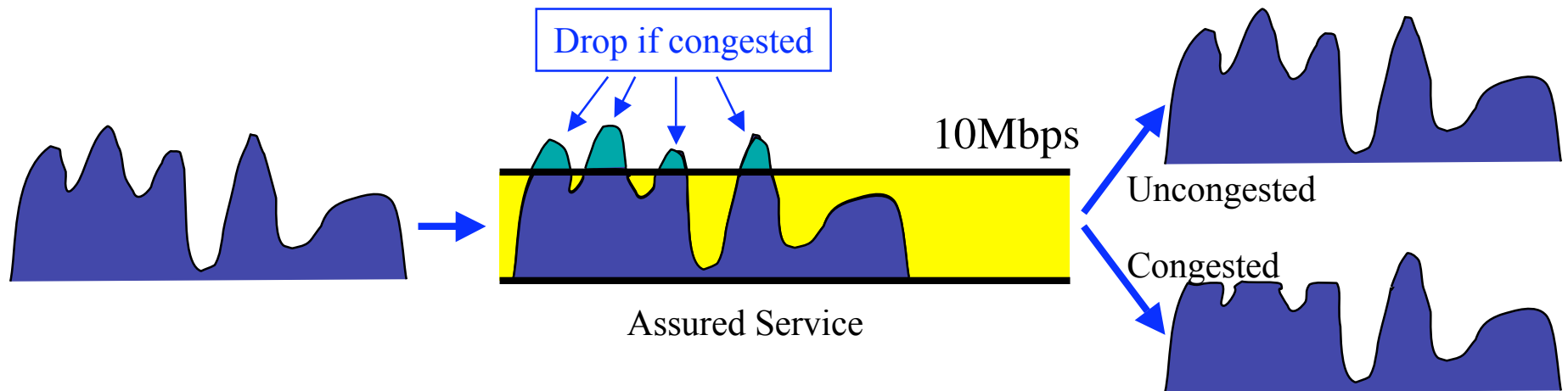
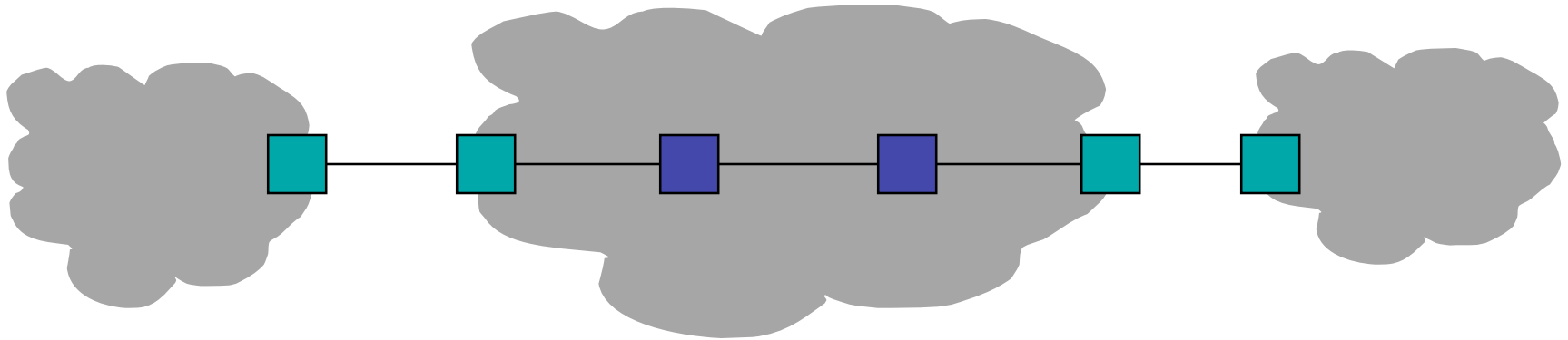
❑ Assured Forwarding (AF):

- ❑ 4 classes, each guaranteed a minimum amount of bandwidth and buffering; each with three drop preference partitions
- ❑ Emulates frame-relay behavior

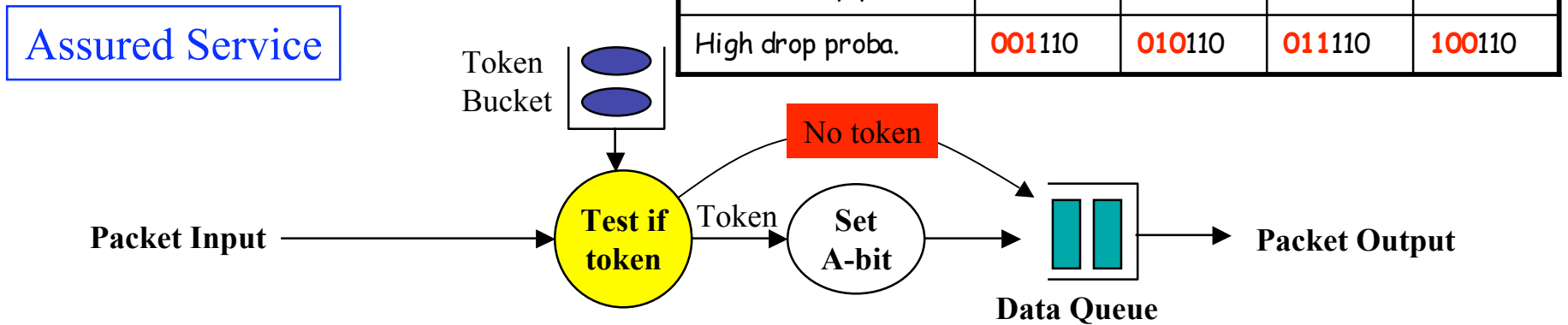
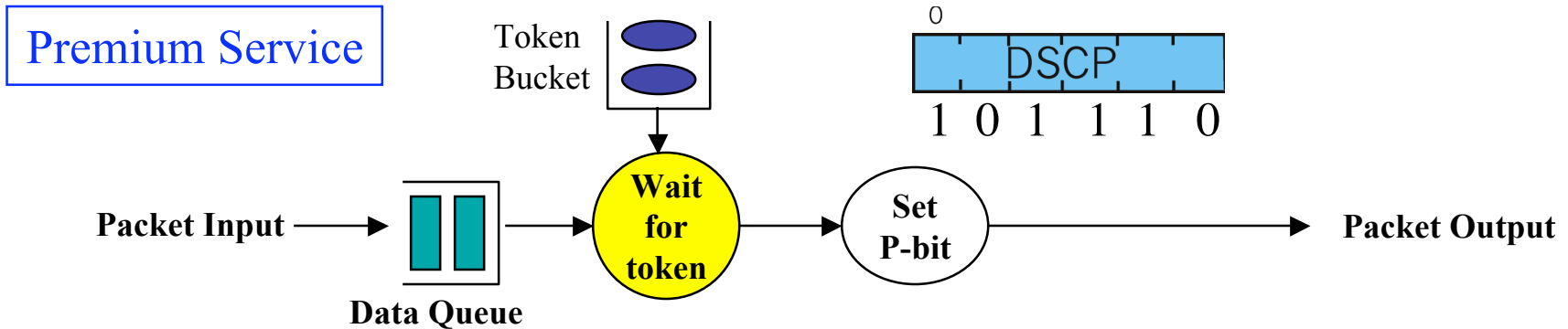
Premium Service Example



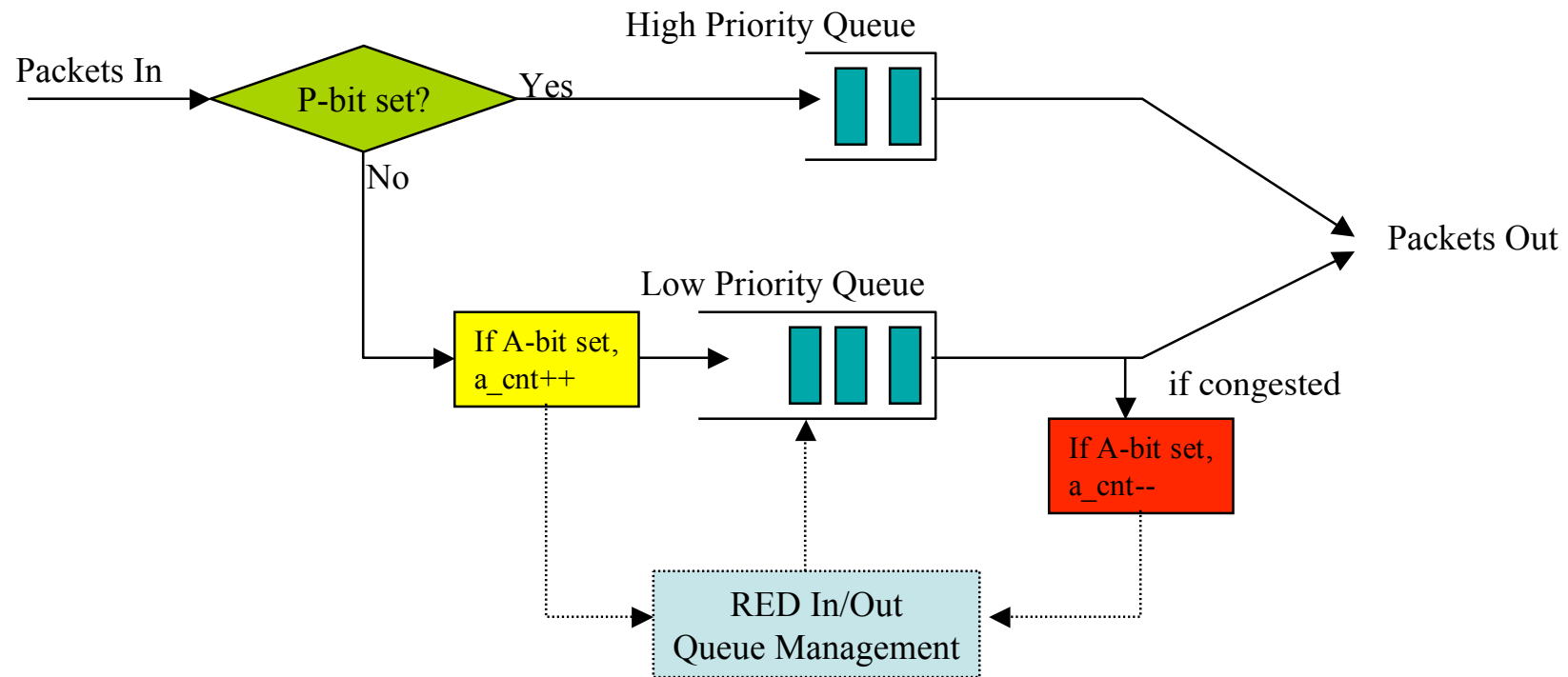
Assured Service Example



Border Router Functionality

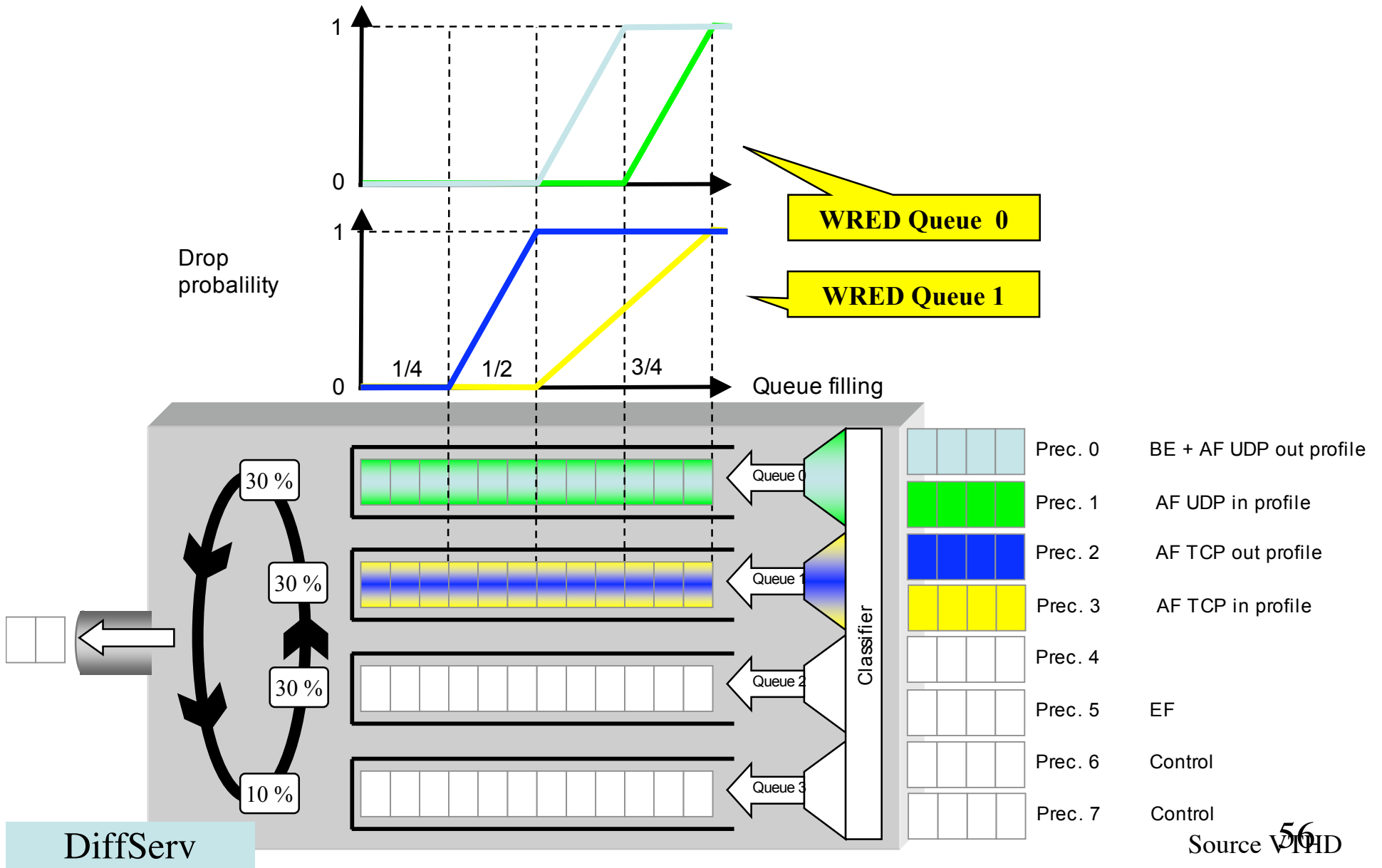


Internal Router Functionality

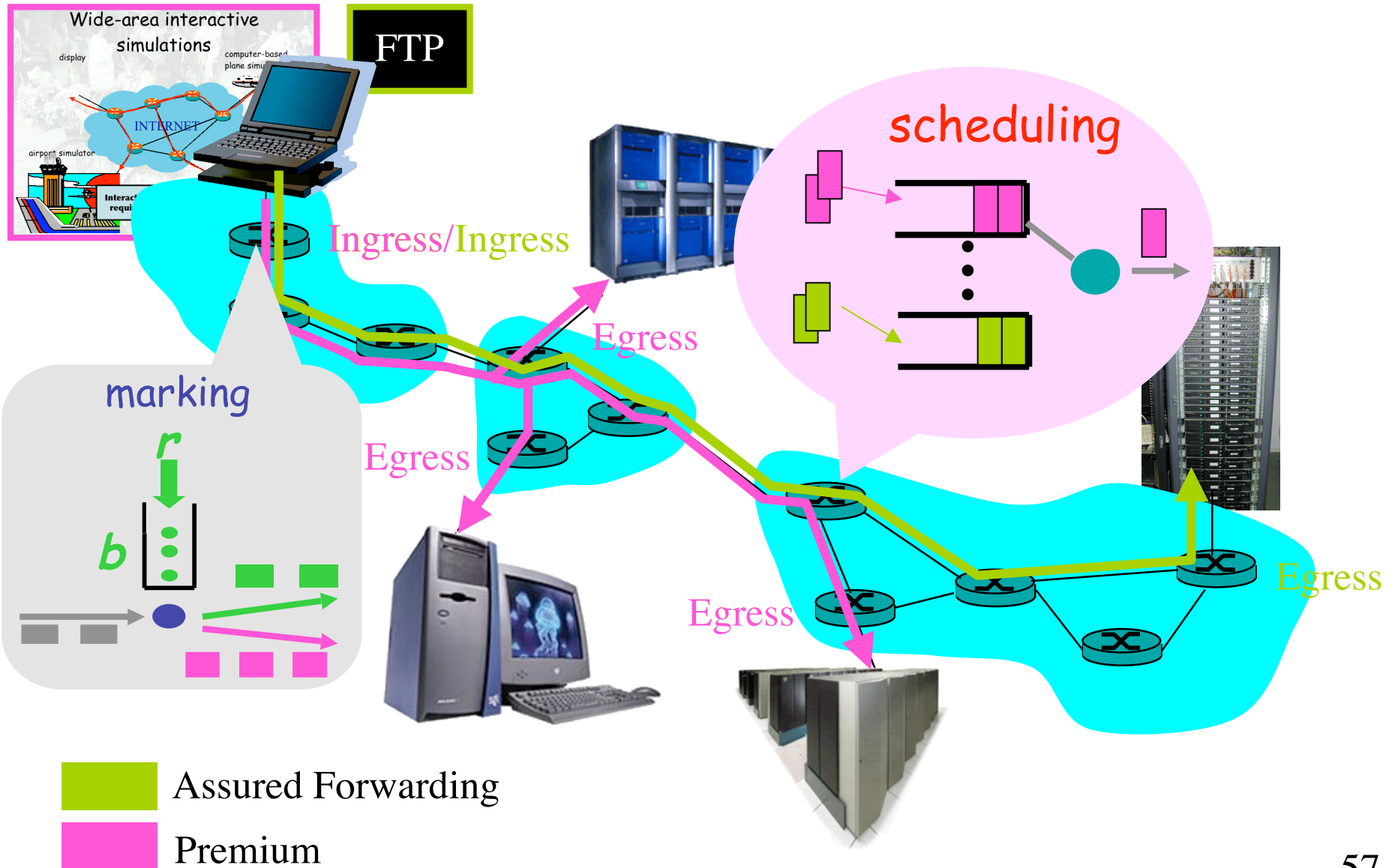


A DSCP codes aggregates, not individual flows
No state in the core
Should scale to millions of flows

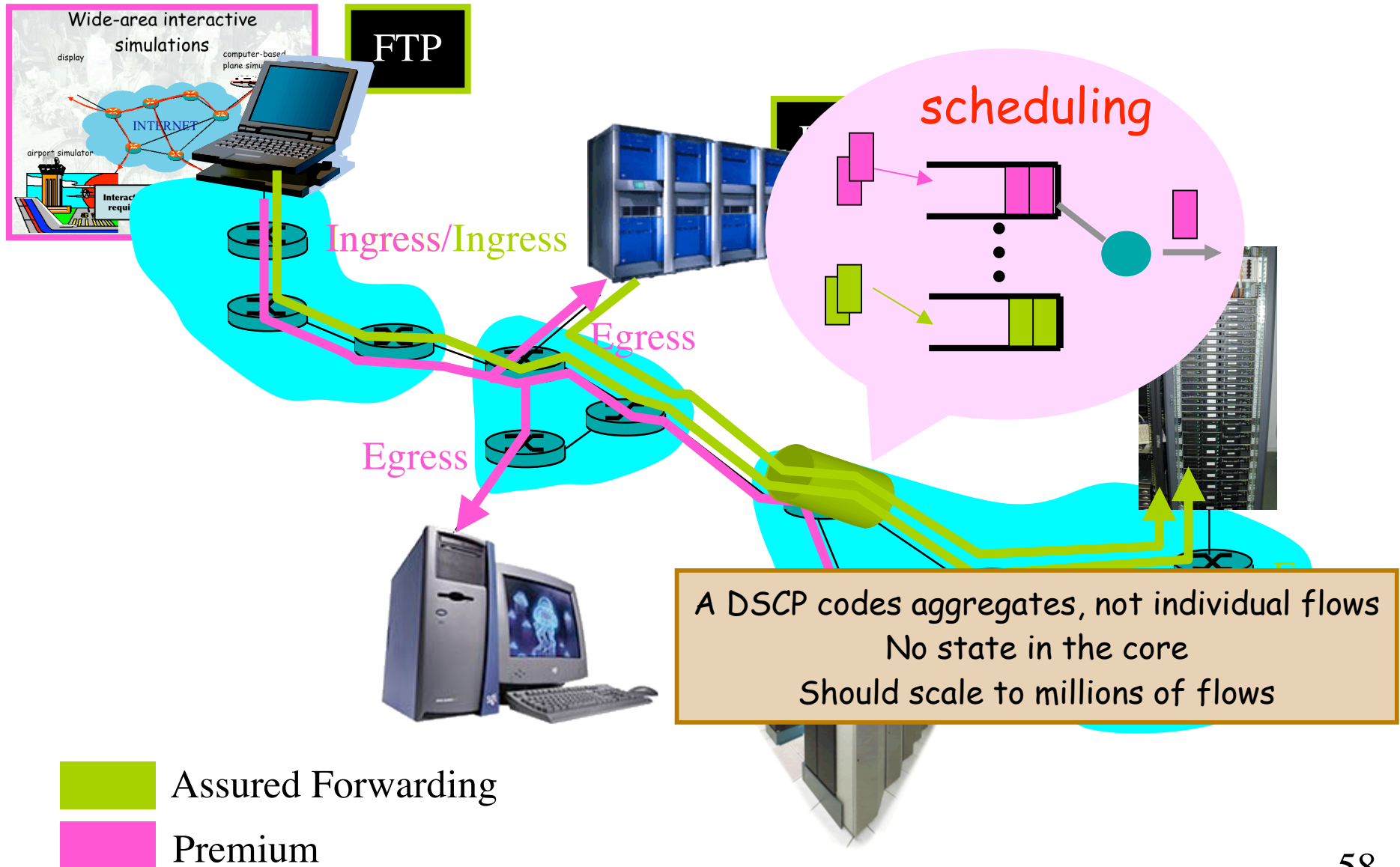
Practical realization



DiffServ for grids

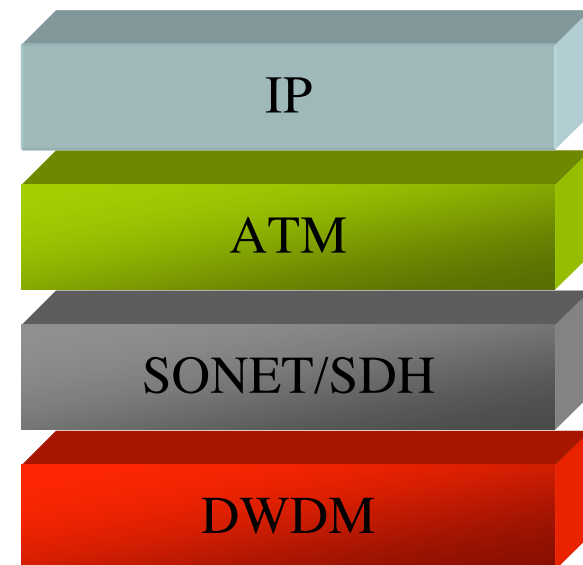


DiffServ for grids (con't)



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

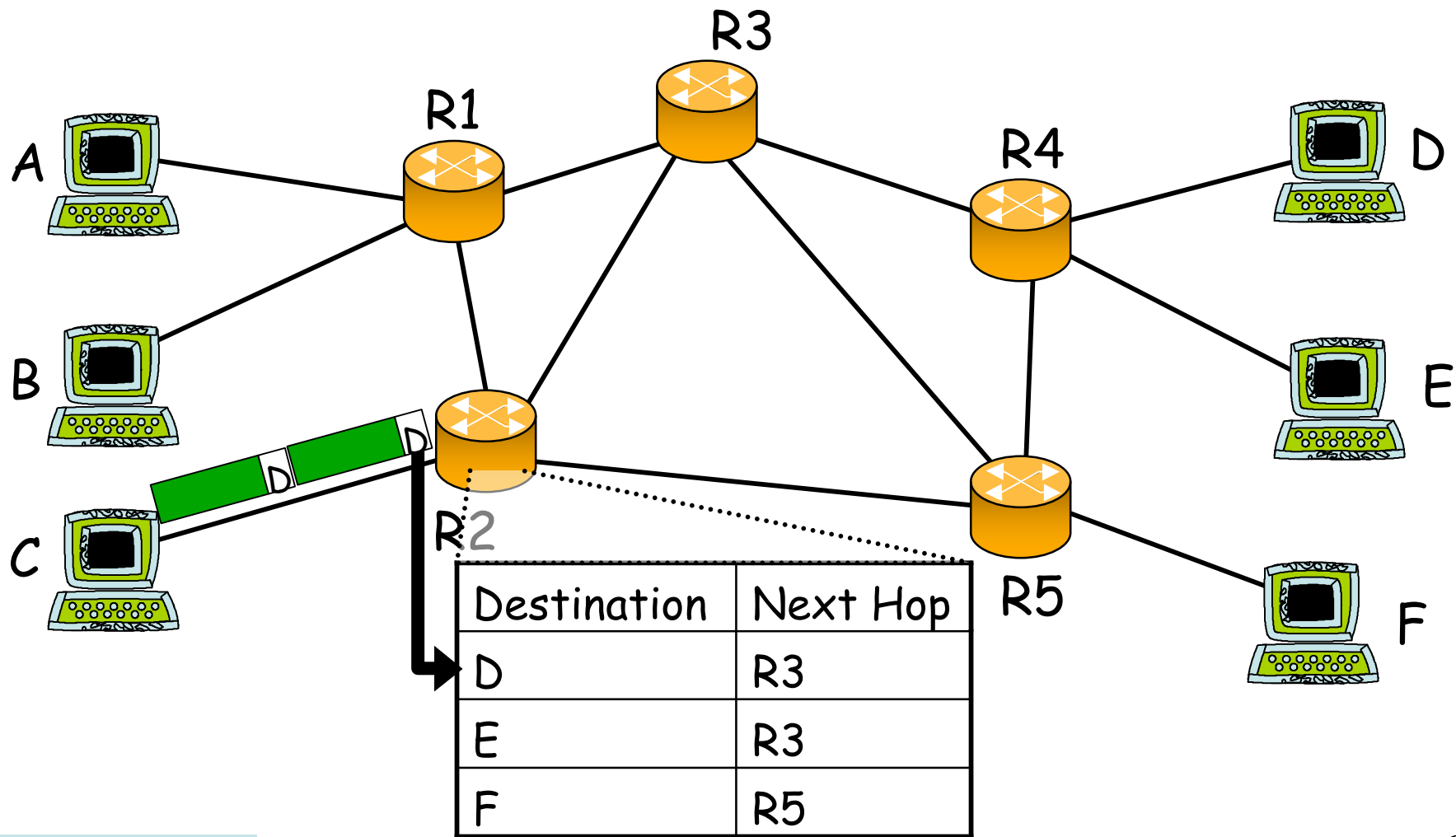


Provider's view

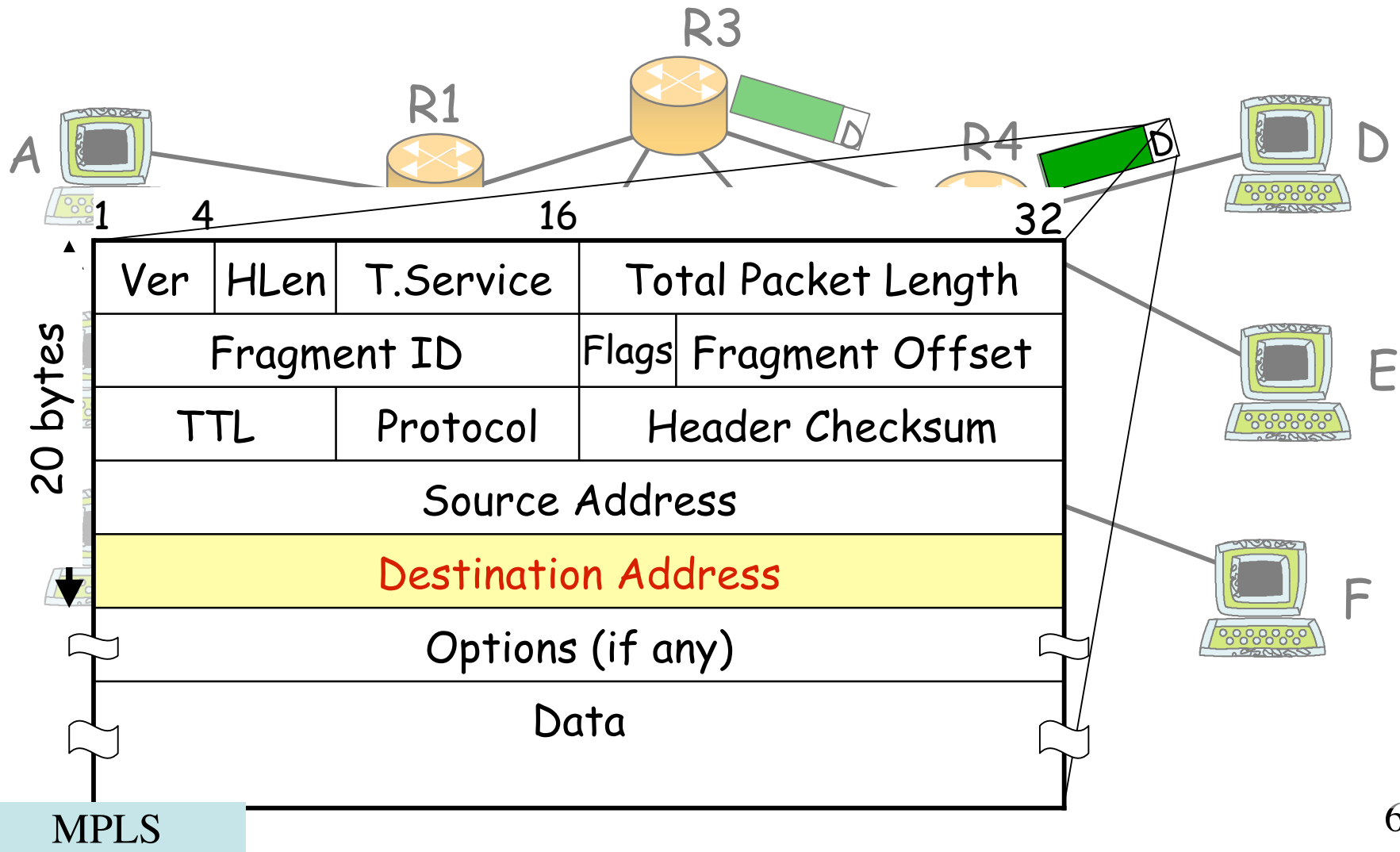


Today's setting time is several weeks/months!
We want to set dynamic links within hours

Review of IP routing



Review of IP routing

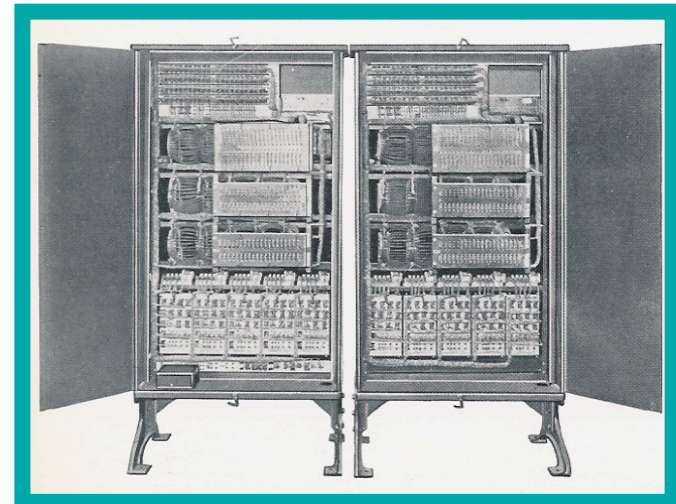


Review of telephone network



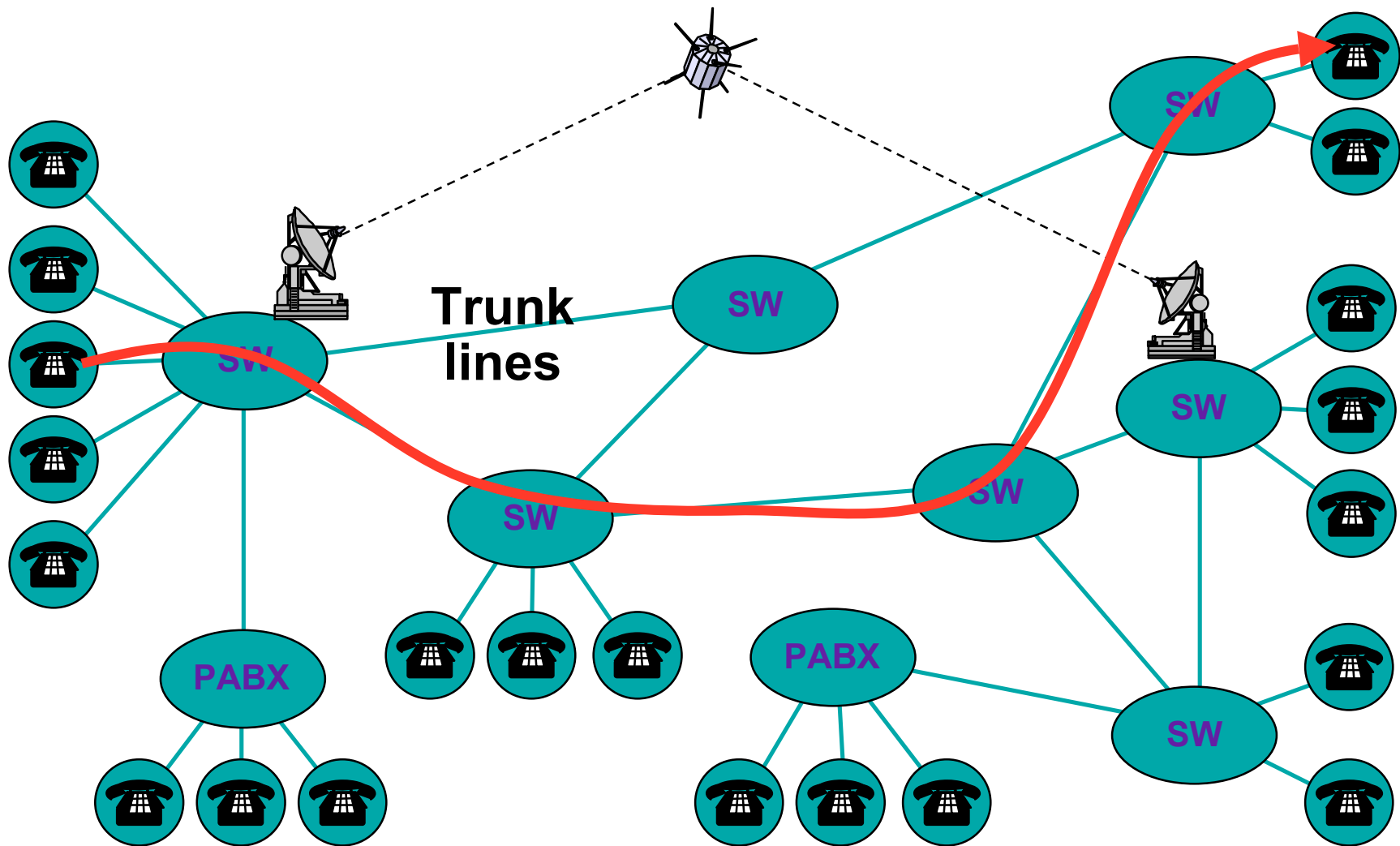
*First automatic Branch Exchange
Almond B. Strowger, 1891...*

**Signaling replaces the
operator**



Source J. Tiberghien, VUB

The telephone circuit view



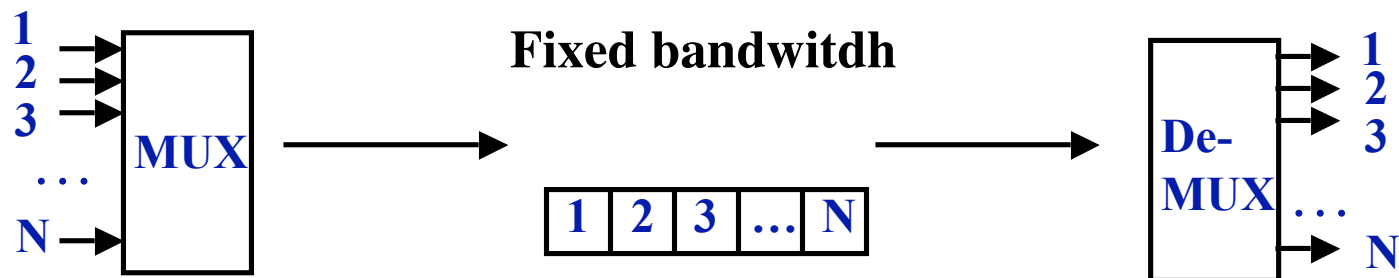
Advantages of circuits

- ❑ 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

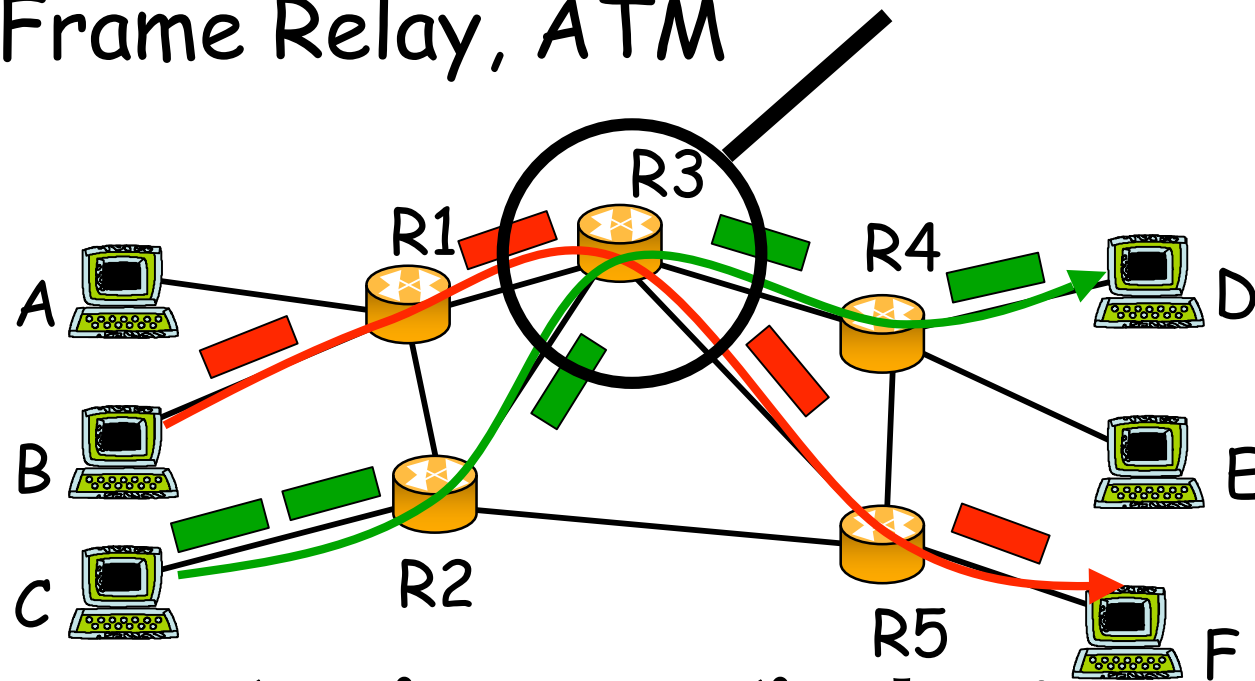


1 sample every 125us gives a 64Kbits/s channel

Back to virtual circuits

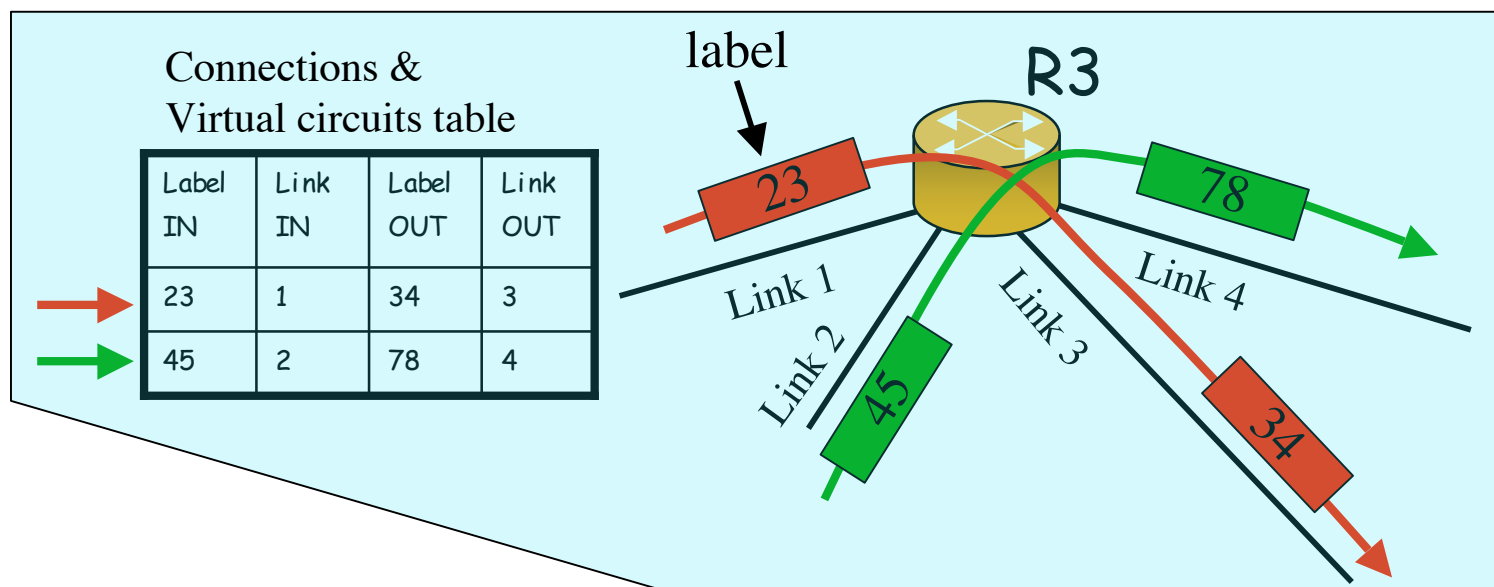
- Virtual circuit refers to a connection oriented network/link layer: e.g. X.25, Frame Relay, ATM

Virtual Circuit Switching:
a path is defined for each connection

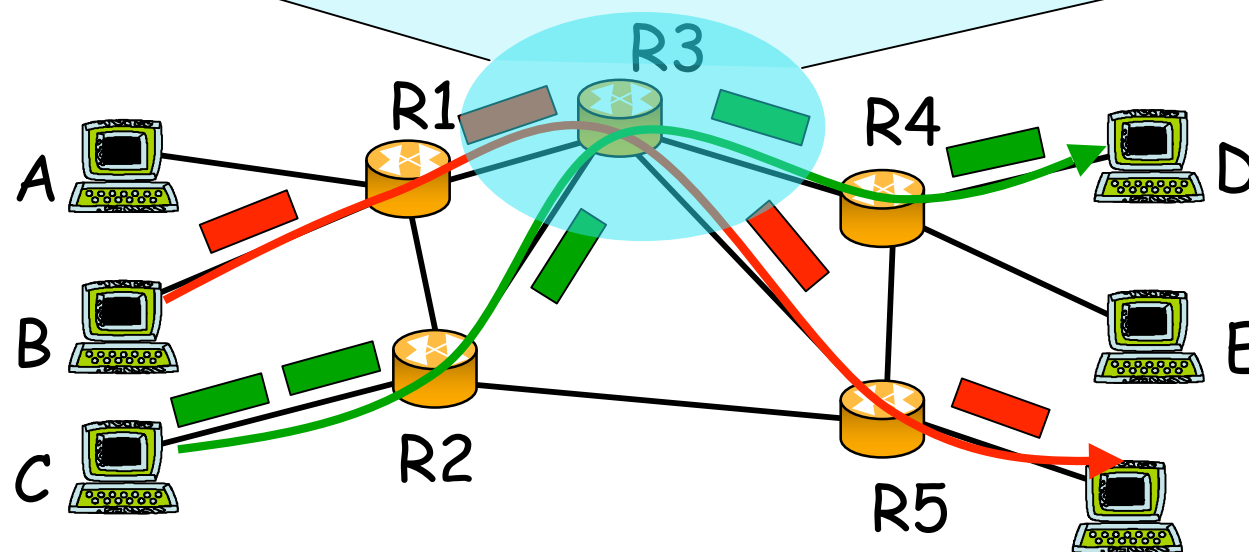


But IP is connectionless!

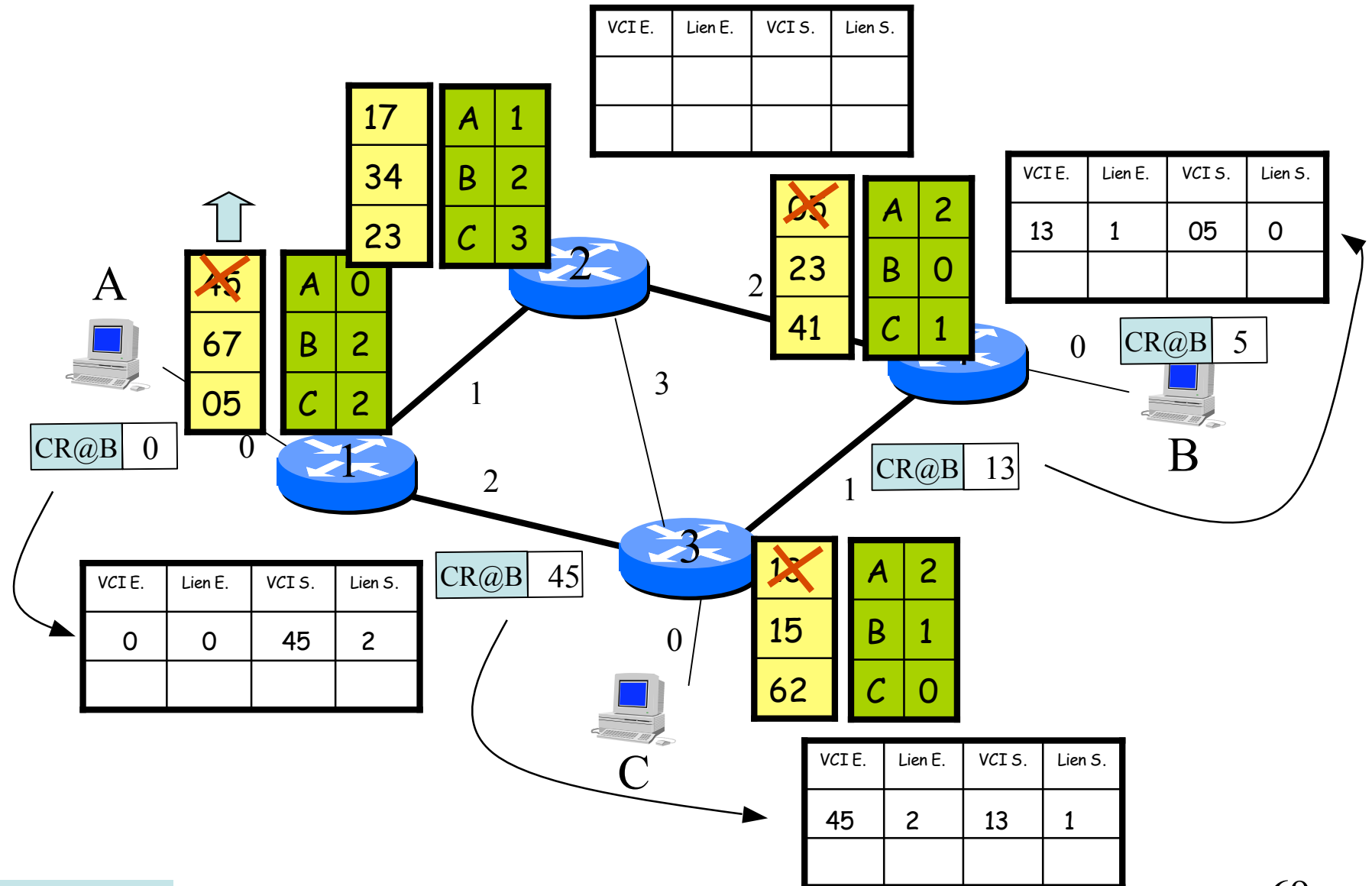
Virtual circuit principles



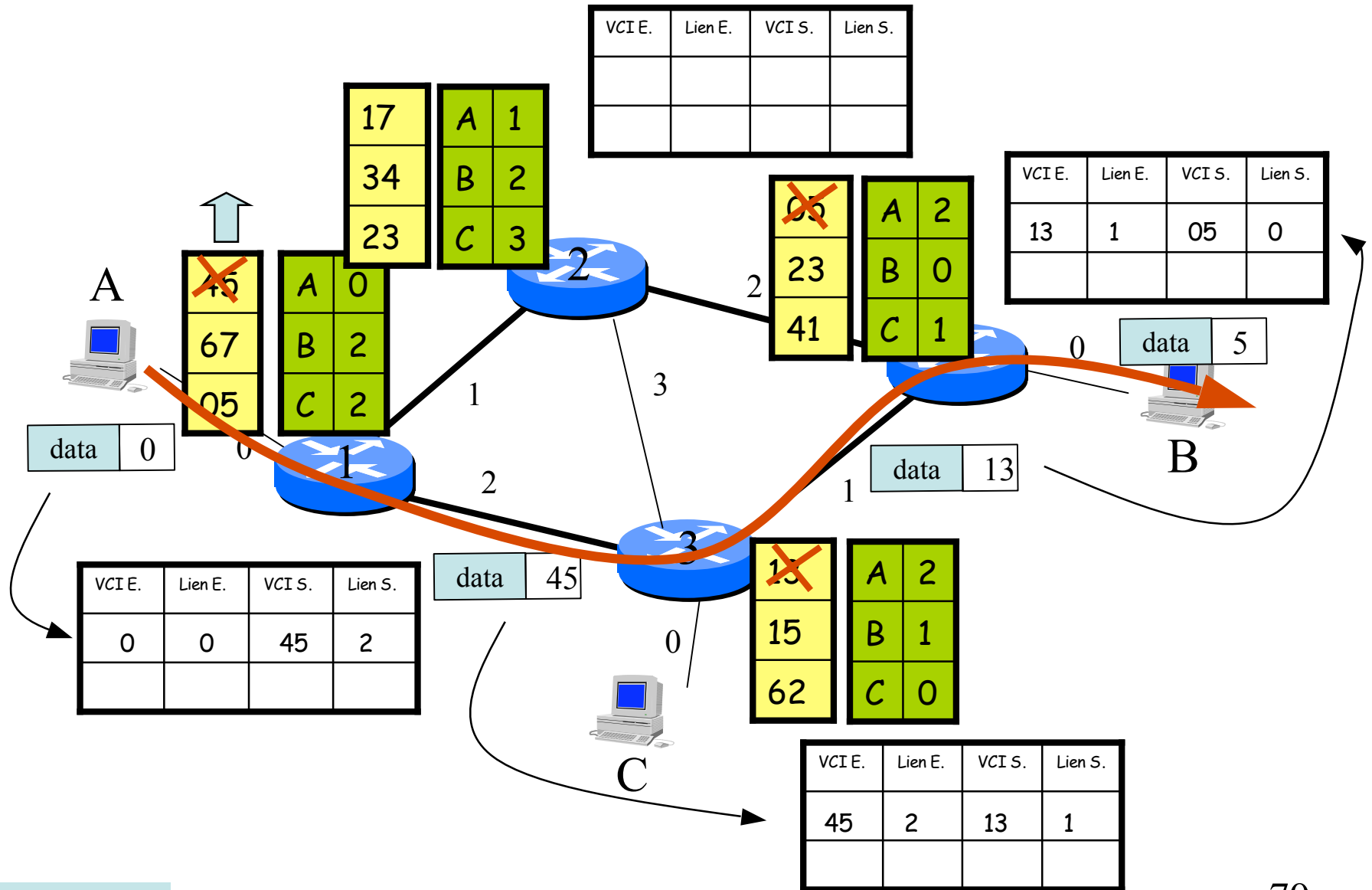
Virtual
Circuit
Switching



End-to-end operation (1)



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

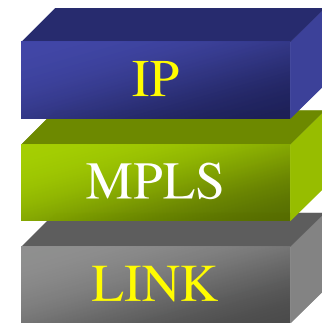
- Multi-Protocol Label Switching

- Fast: use label switching → LSR



- Multi-Protocol: above link layer, below network layer

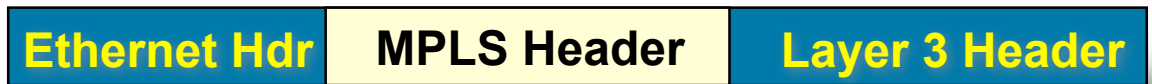
- Facilitate traffic engineering



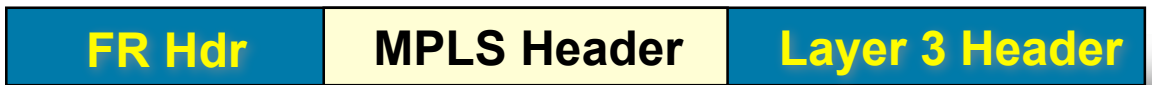
PPP Header(Packet over SONET/SDH)



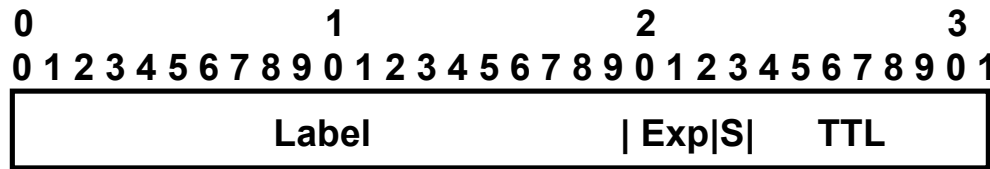
Ethernet



Frame Relay



Label structure



Label = 20 bits

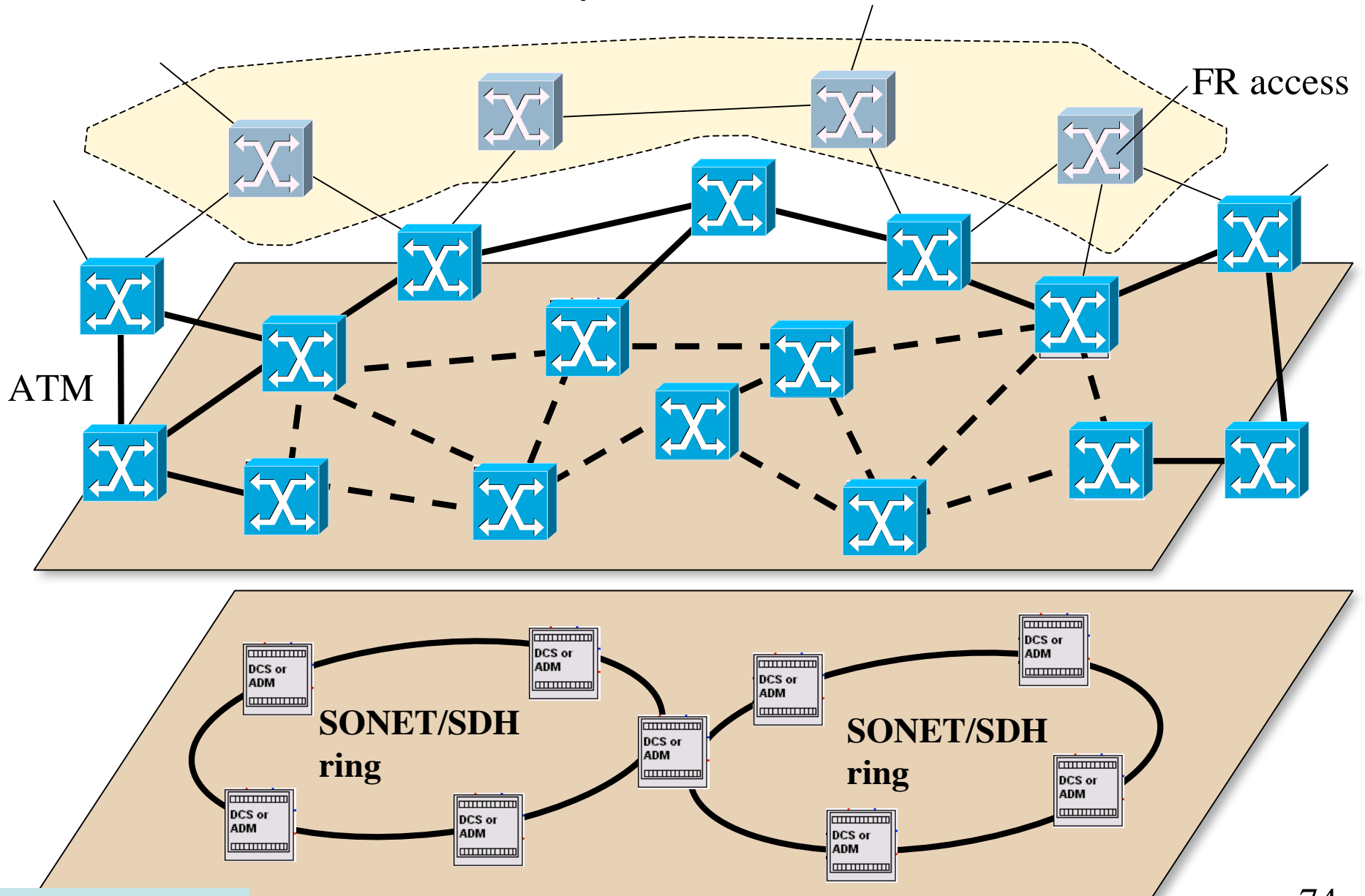
Exp = Experimental, 3 bits

S = Bottom of stack, 1bit

TTL = Time to live, 8 bits

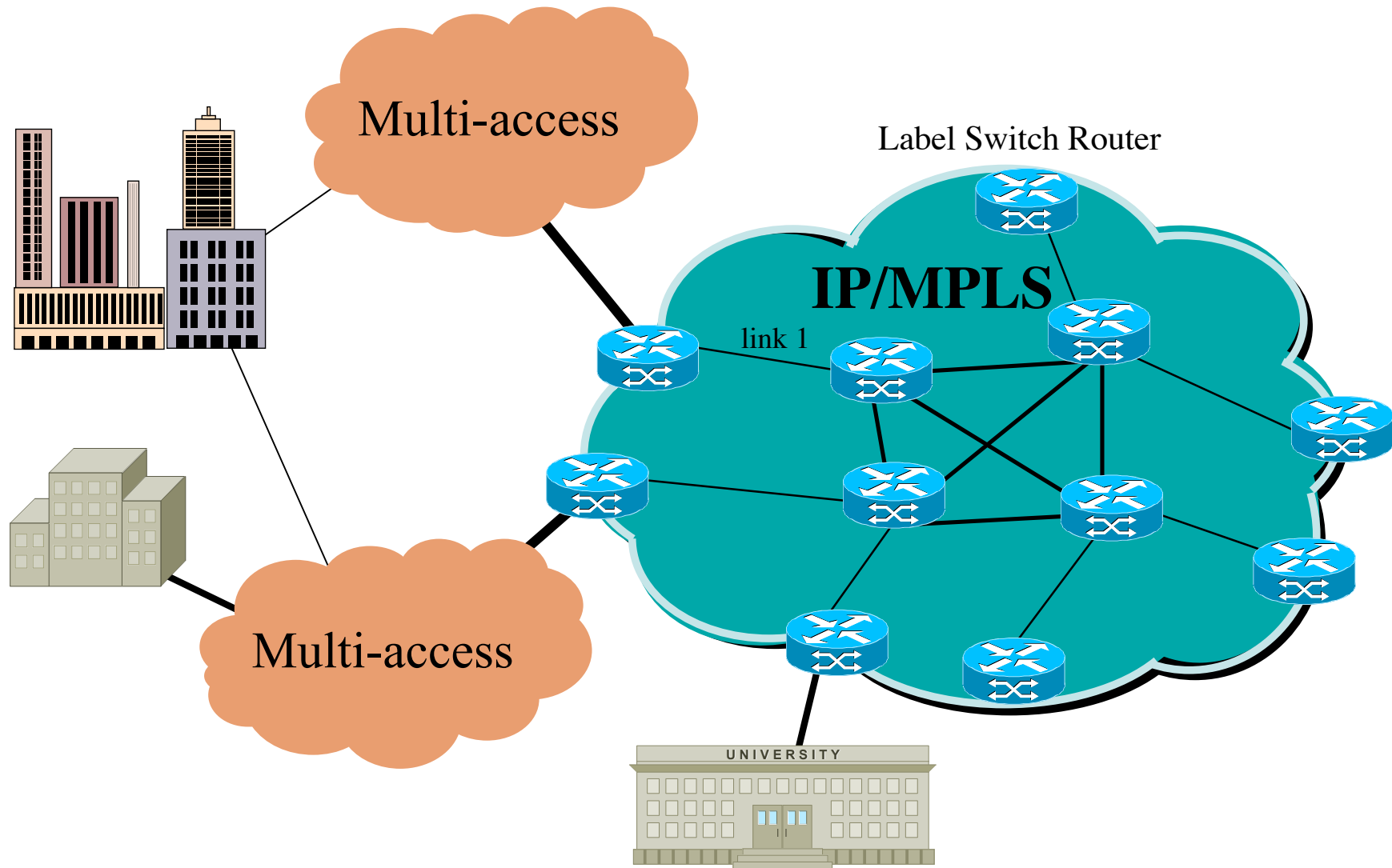
- ❑ More than one label is allowed -> Label Stack
- ❑ MPLS LSRs always forward packets based on the value of the label at the top of the stack

From multilayer networks...



MPLS

...to IP/MPLS networks

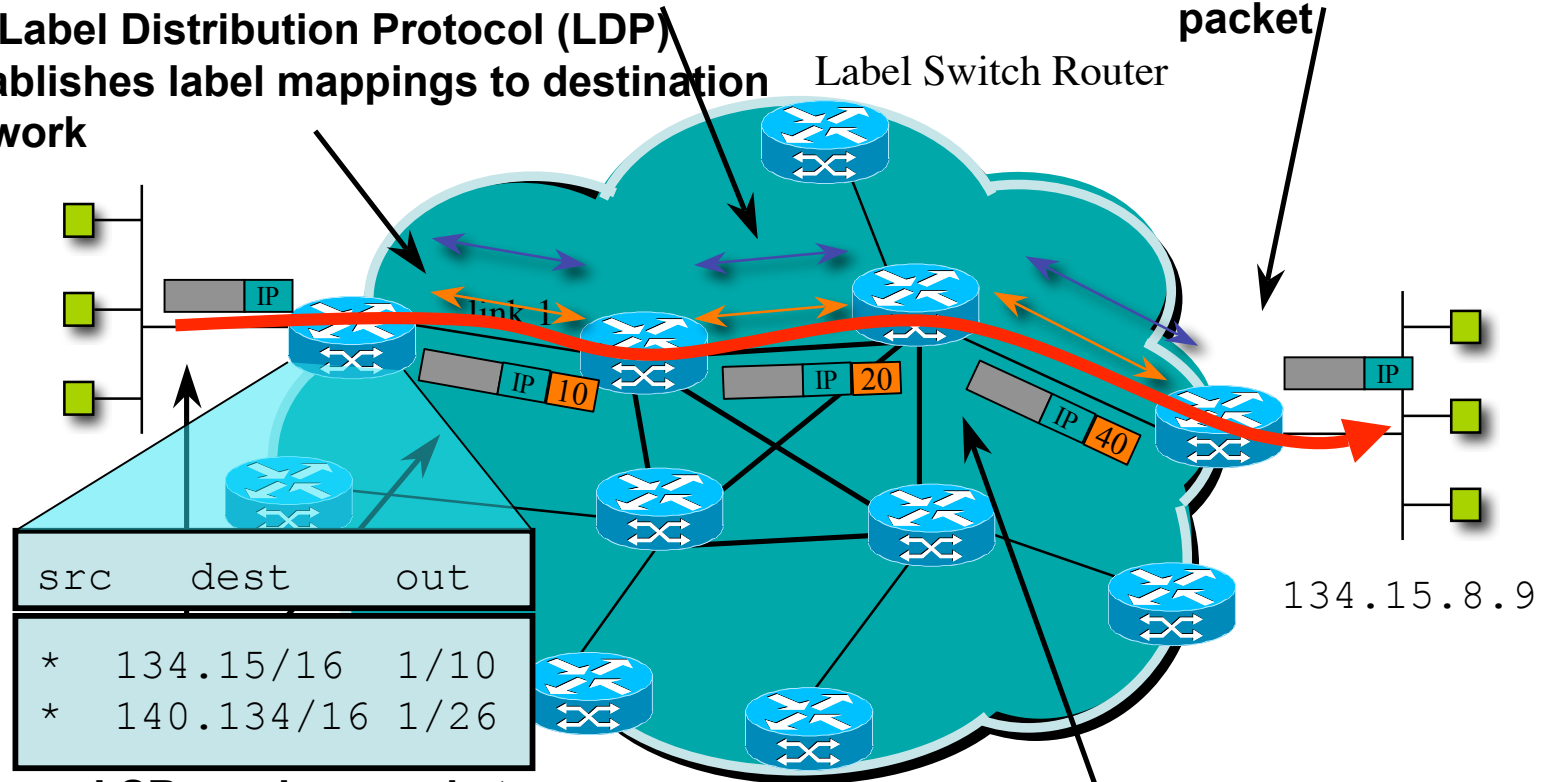


MPLS operation

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

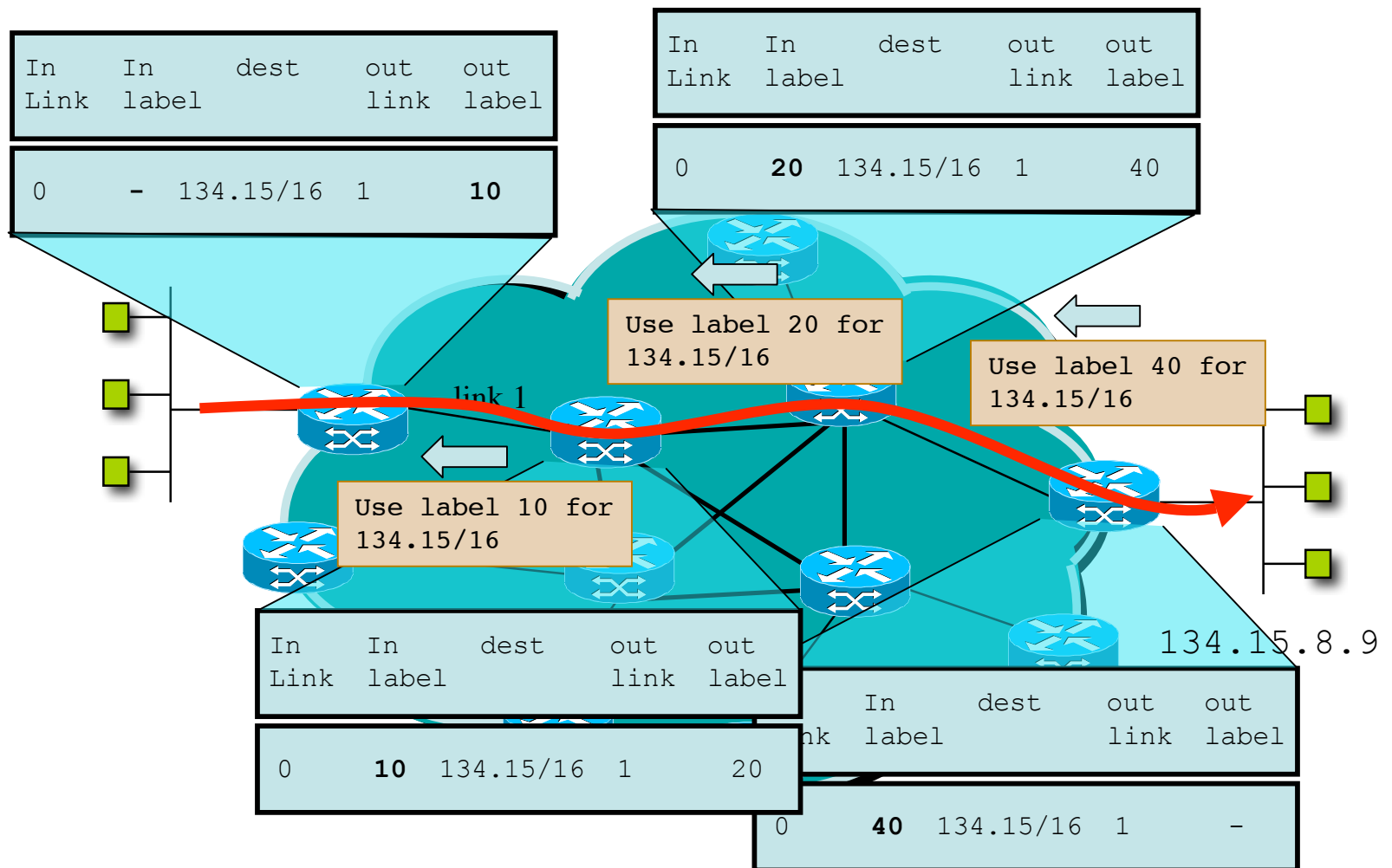


2. Ingress LSR receives packet and "label"s packets

3. LSR forwards packets using label switching

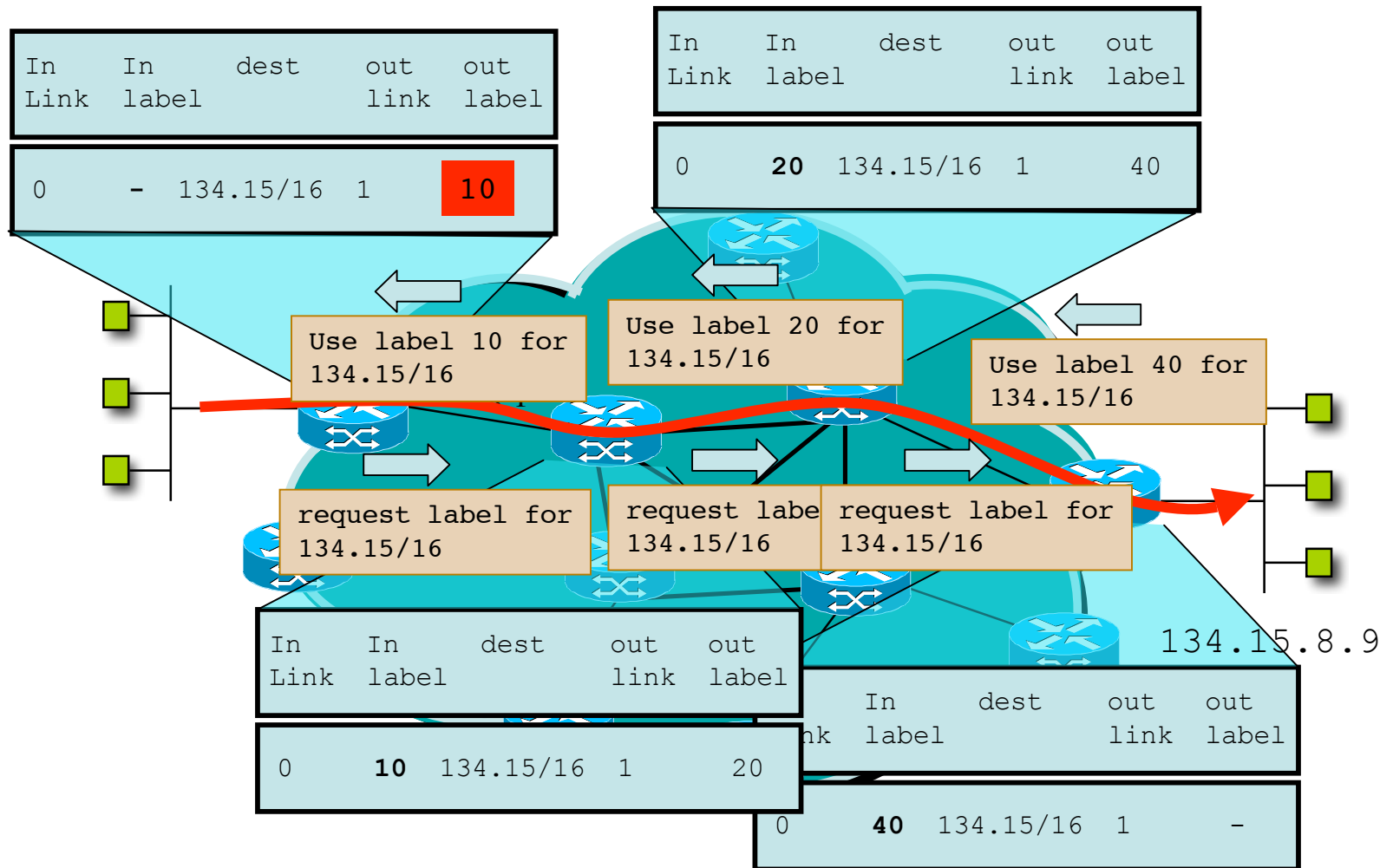
Source Yi Lin, modified C. Pham

Label Distribution



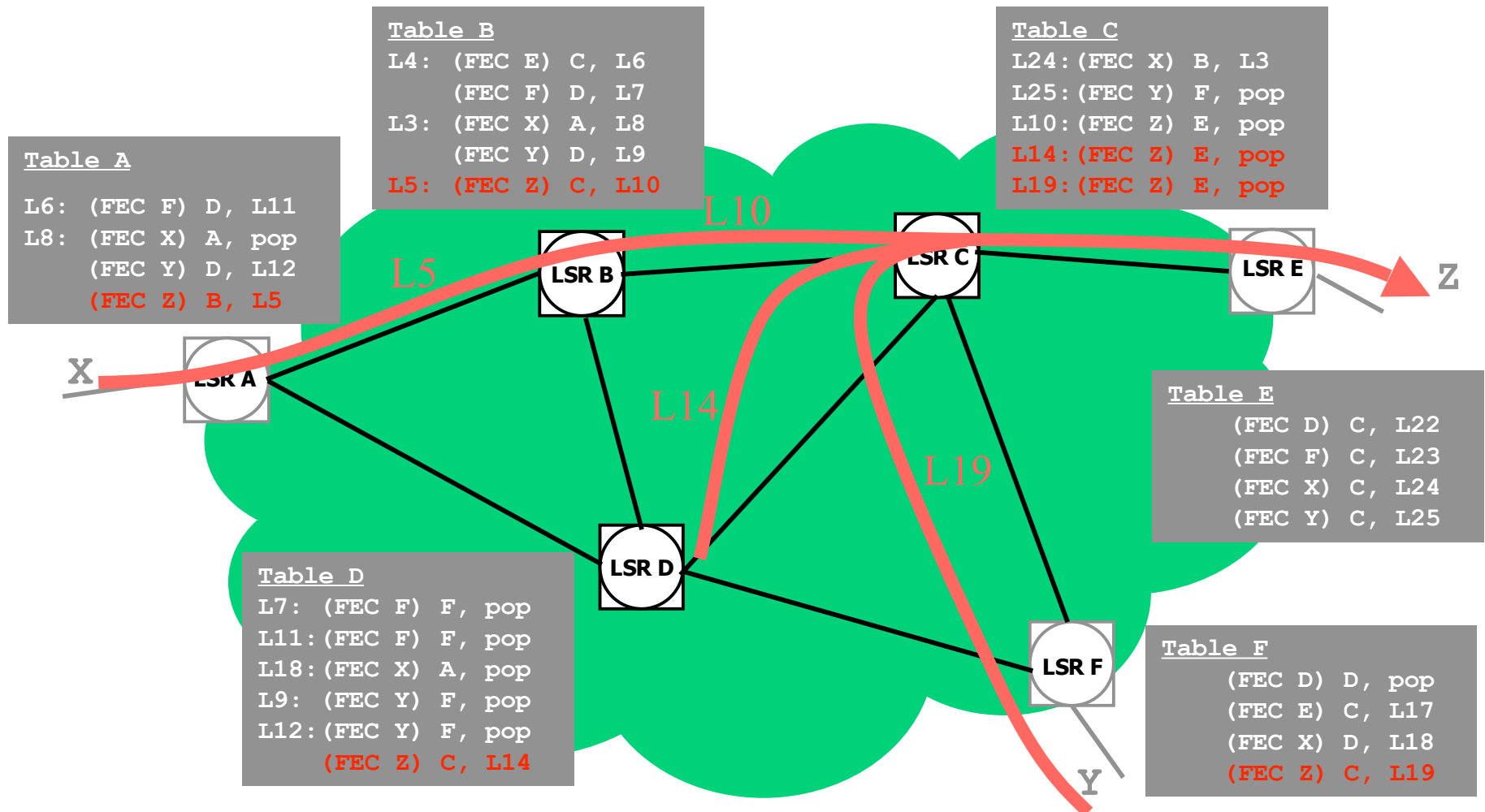
Unsolicited downstream label distribution

Label Distribution (con't)



On-demand downstream label distribution

Forwarding Equivalent Class: high-level forwarding criteria



Forwarding Equivalent Class

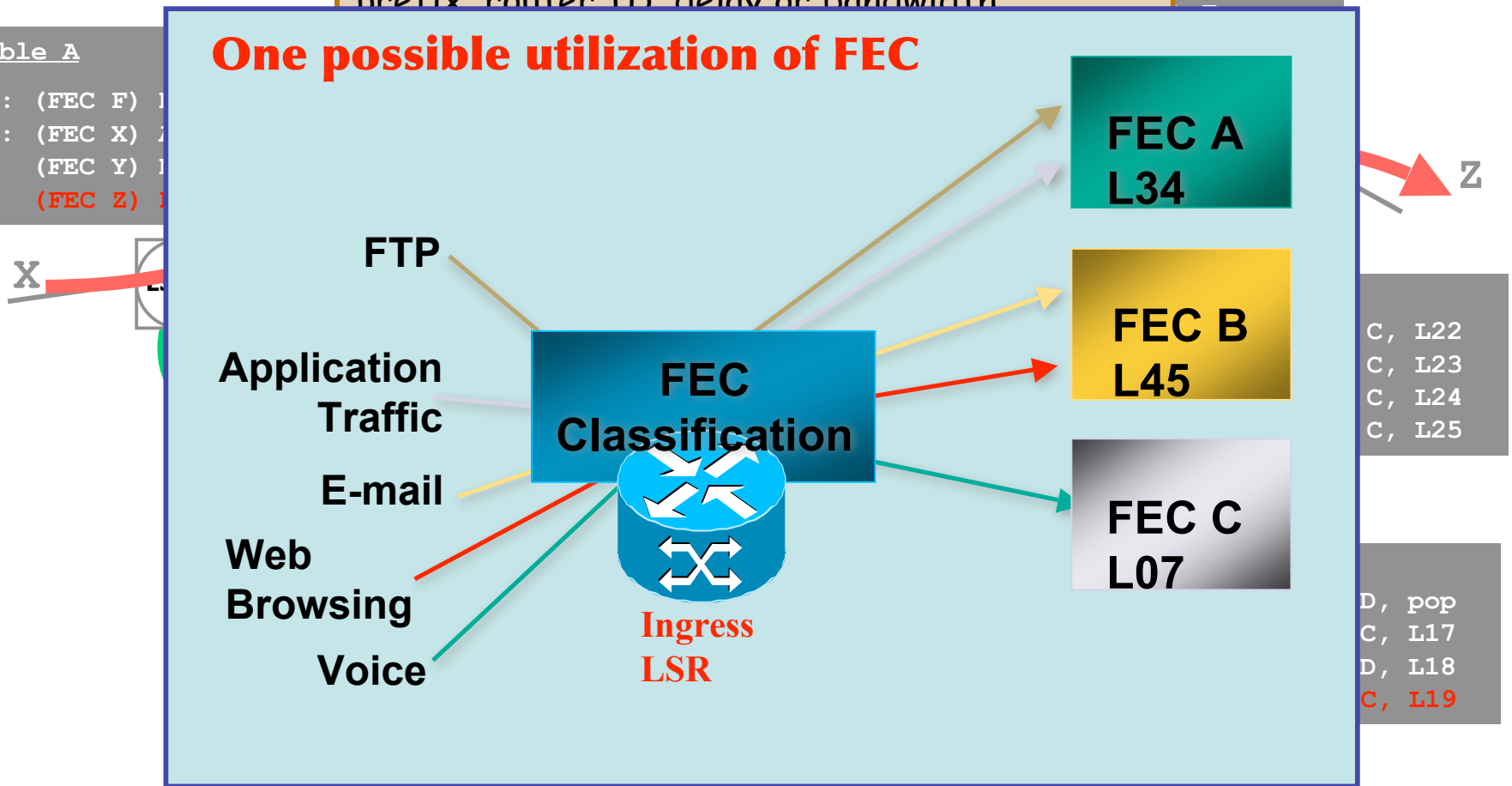
A FEC aggregates a number of individual flows with the same characteristics: IP prefix, router ID, delay or bandwidth

B, L3
F, pop

Table A

L6: (FEC F)
L8: (FEC X)
(FEC Y)
(FEC Z)

One possible utilization of FEC



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, CR-LDP

- Used in traffic engineering

❑ BGP

- External labels (VPN)

❑ PIM

- For multicast states label mapping

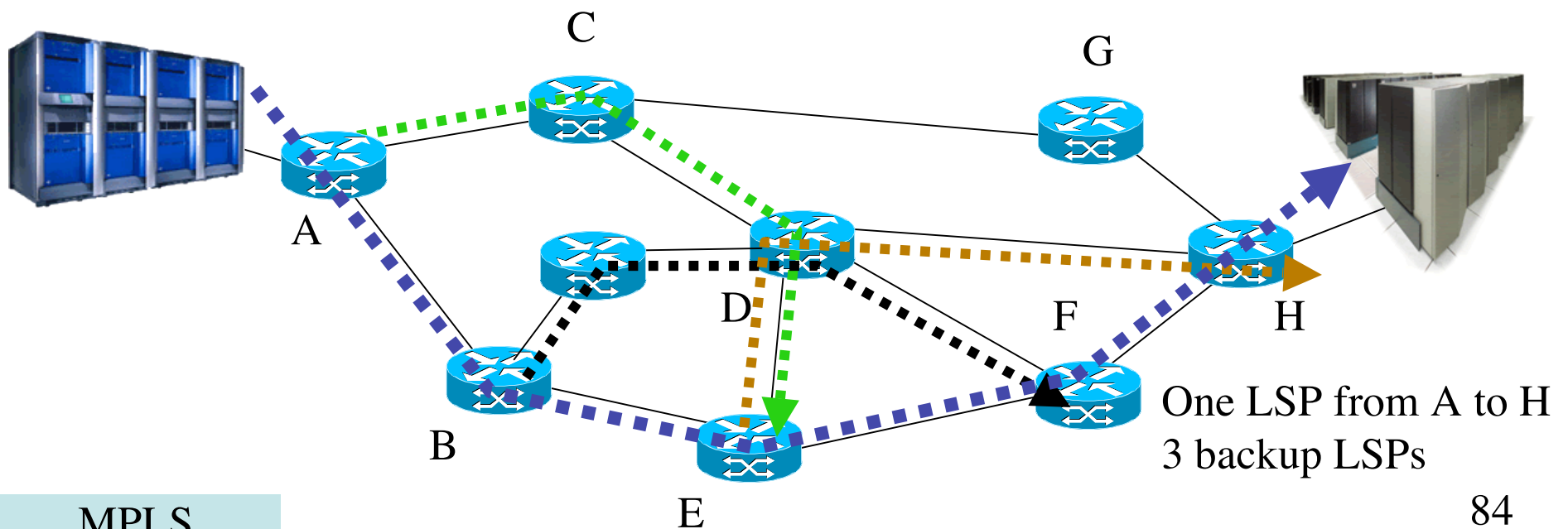
MPLS FastReroute

- ❑ 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

Backup LSPs

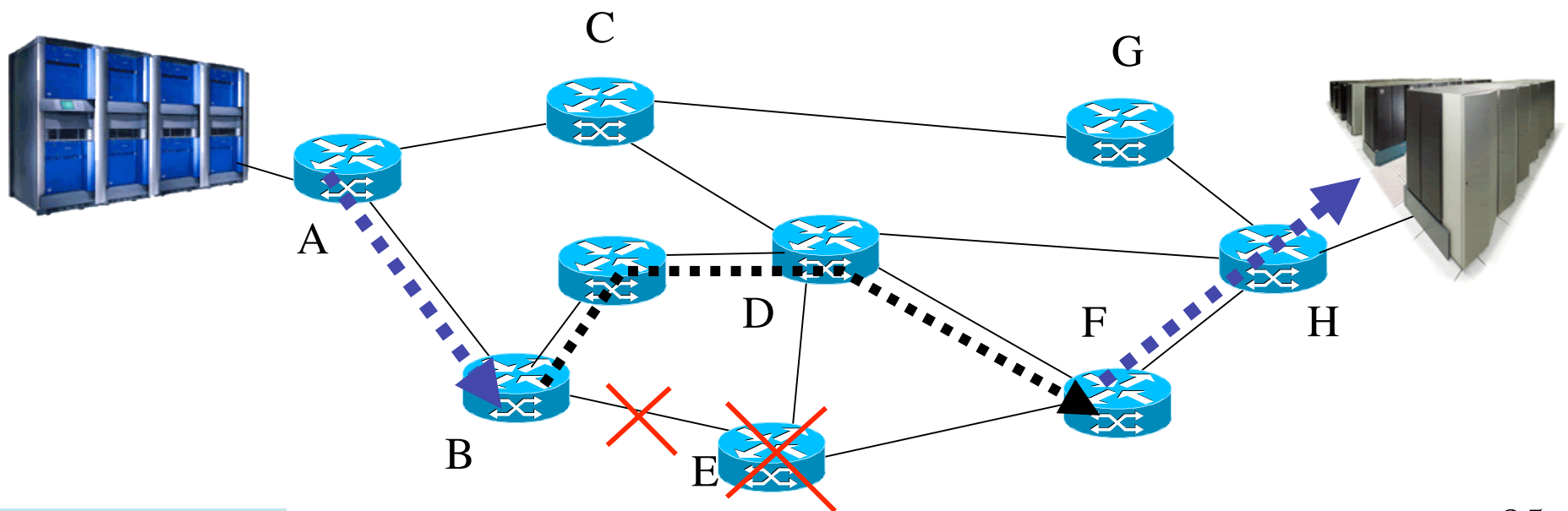
- One-to-one

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



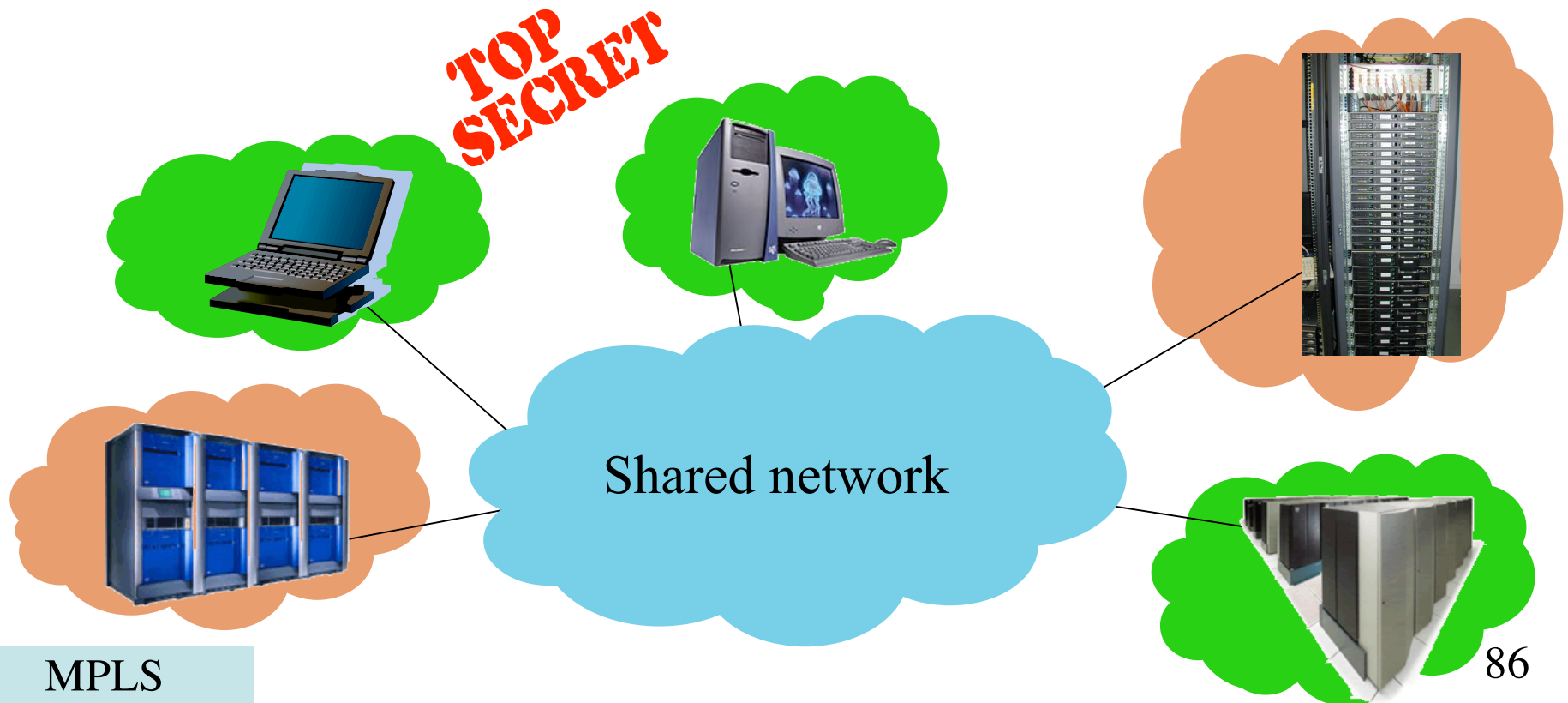
Recovery on failures

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



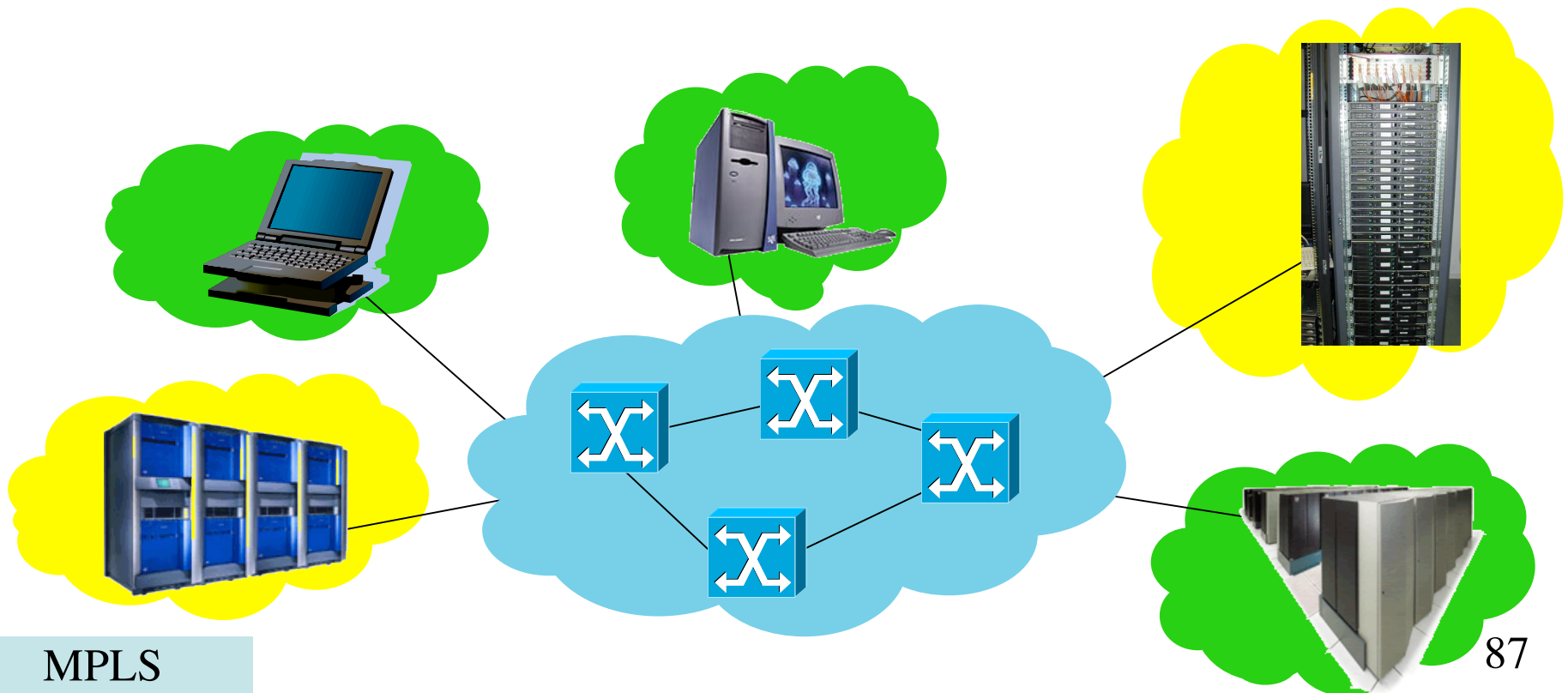
Virtual Private Networks

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



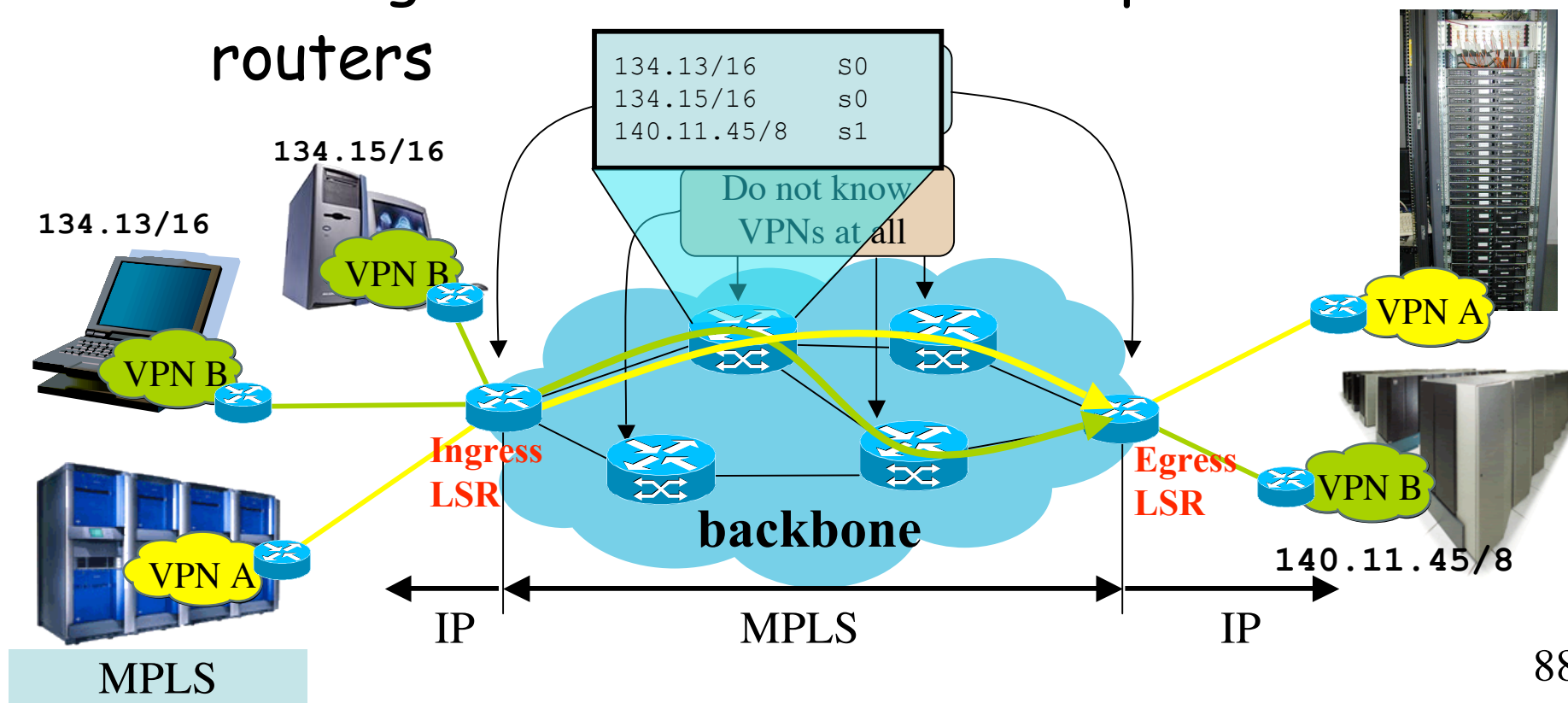
The traditional way of VPN

- ❑ Uses leased lines, Frame Relay/ATM infrastructures...

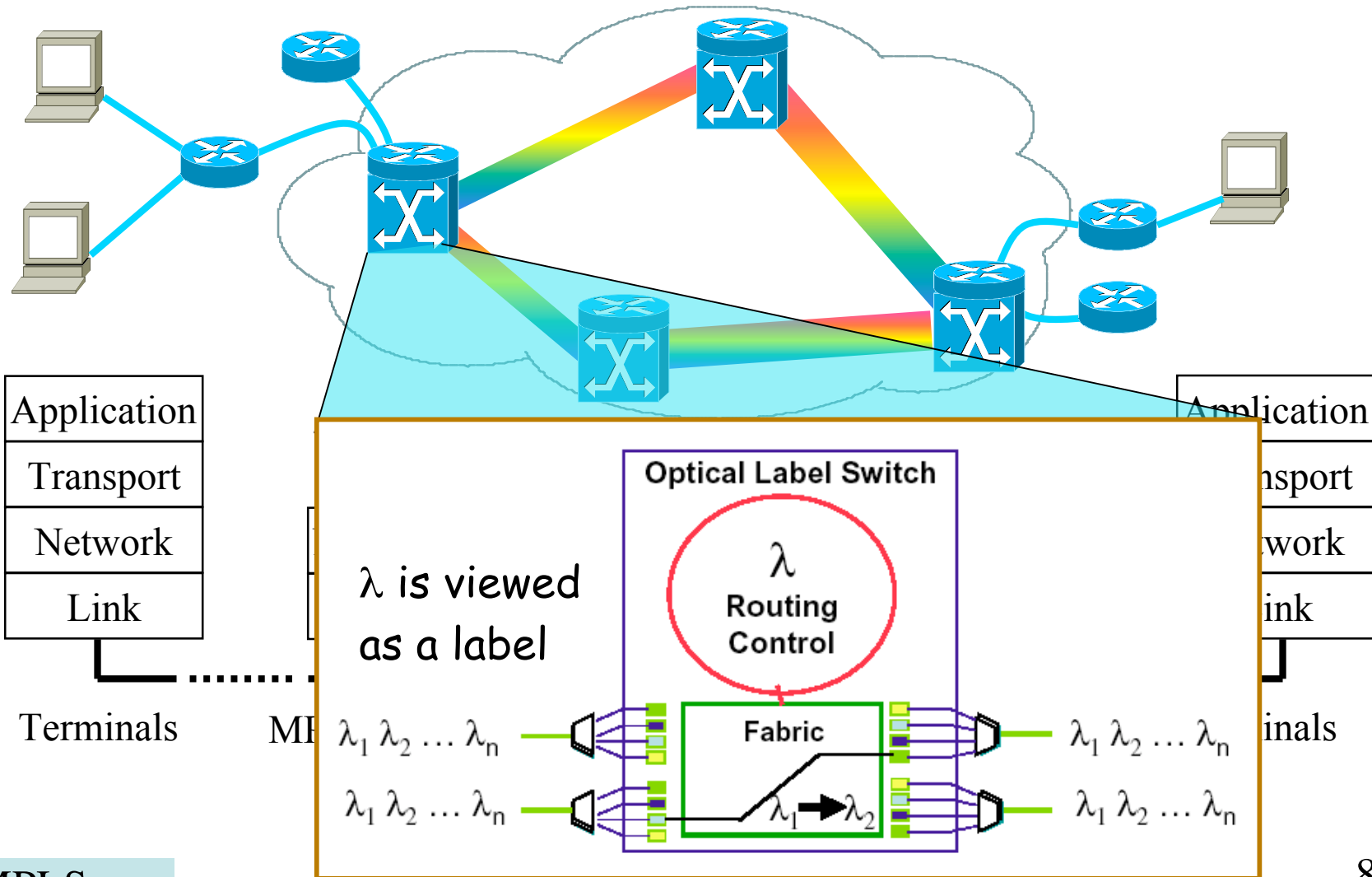


IP/MPLS & VPN

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

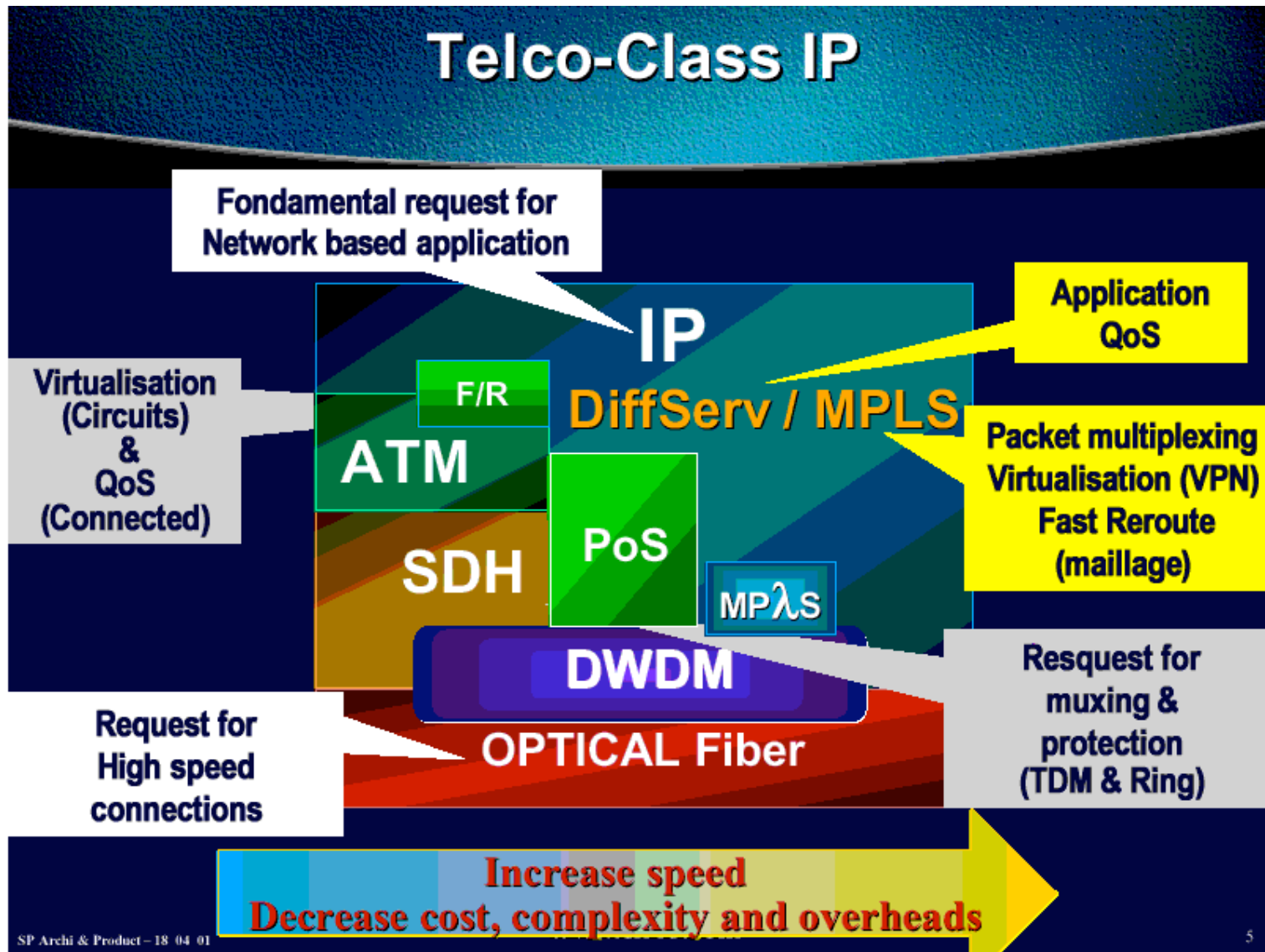


MP λ S: MPLS+optical



Towards IP/MPLS/DWDM

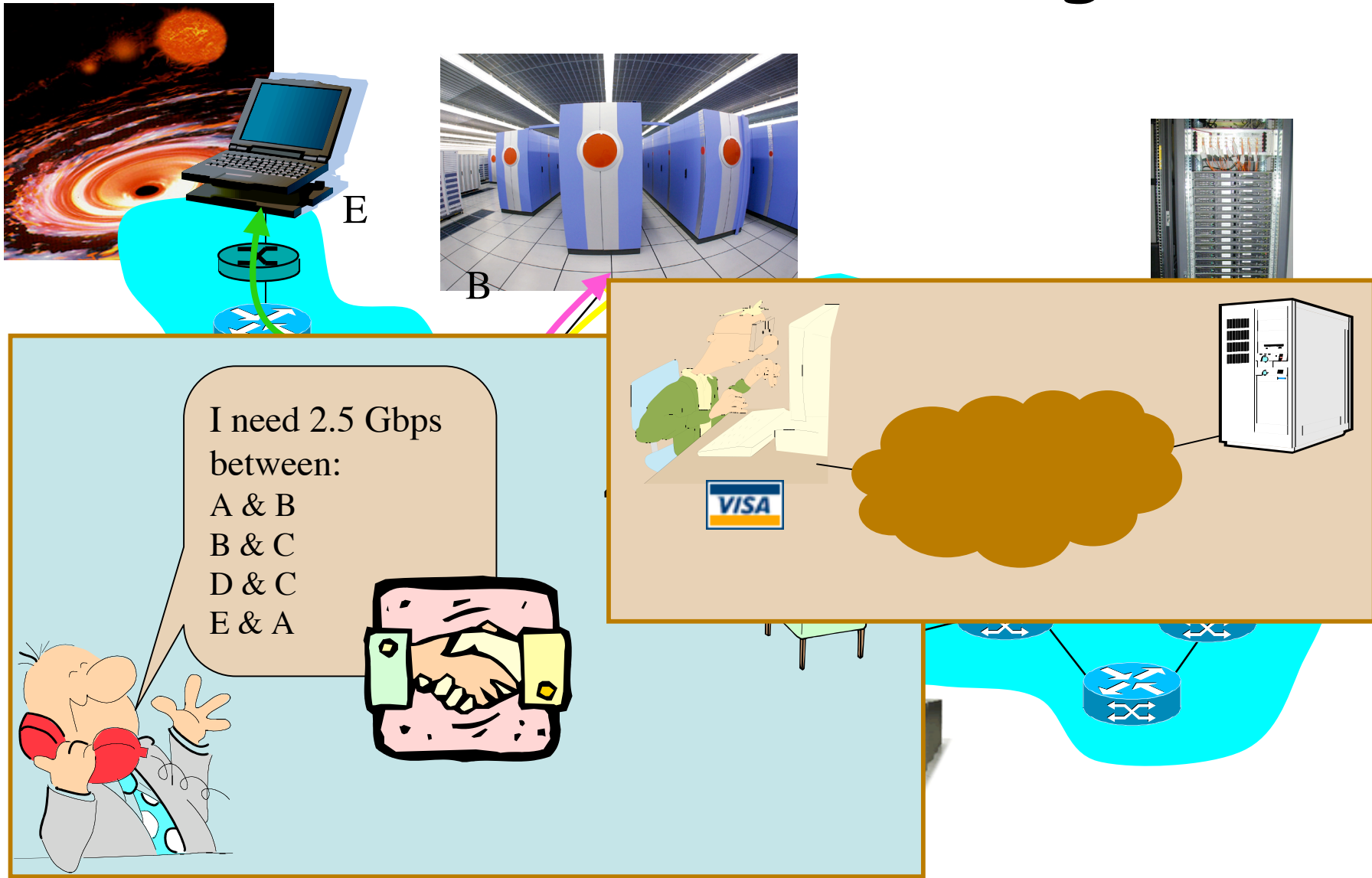
From cisco



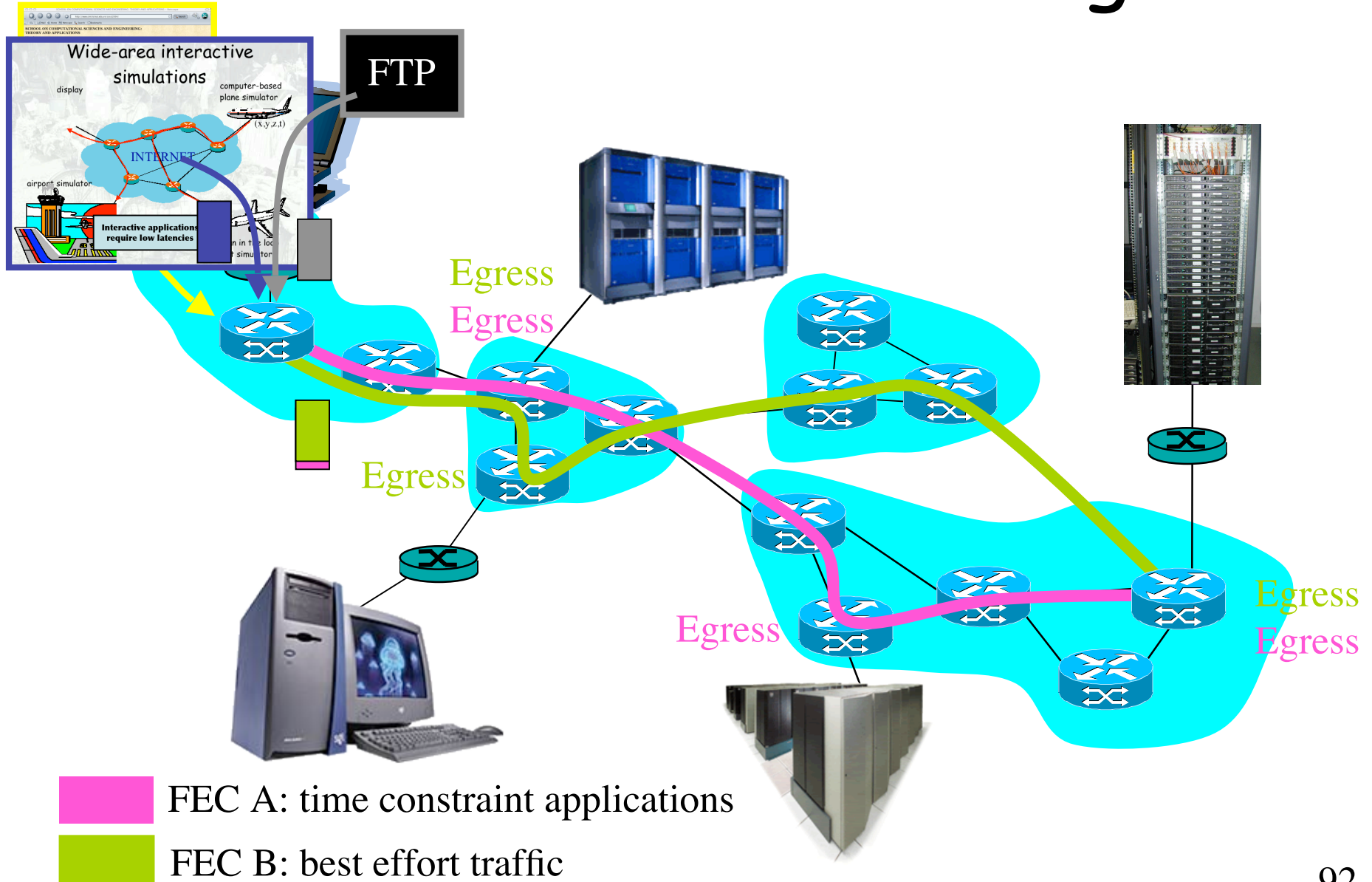
SP Archi & Product - 18 04 01

5

Ex: MPLS circuits on grids



Ex: MPLS FEC for the grid



End of part 1, go to part 2

