New Internet and Networking Technologies and Their Application on Computational Sciences

COSCI 2004

Dai Hoc Bach Khoa, Ho Chi Minh City, Vietnam, March 3, 2004

C. Pham

University Lyon, France

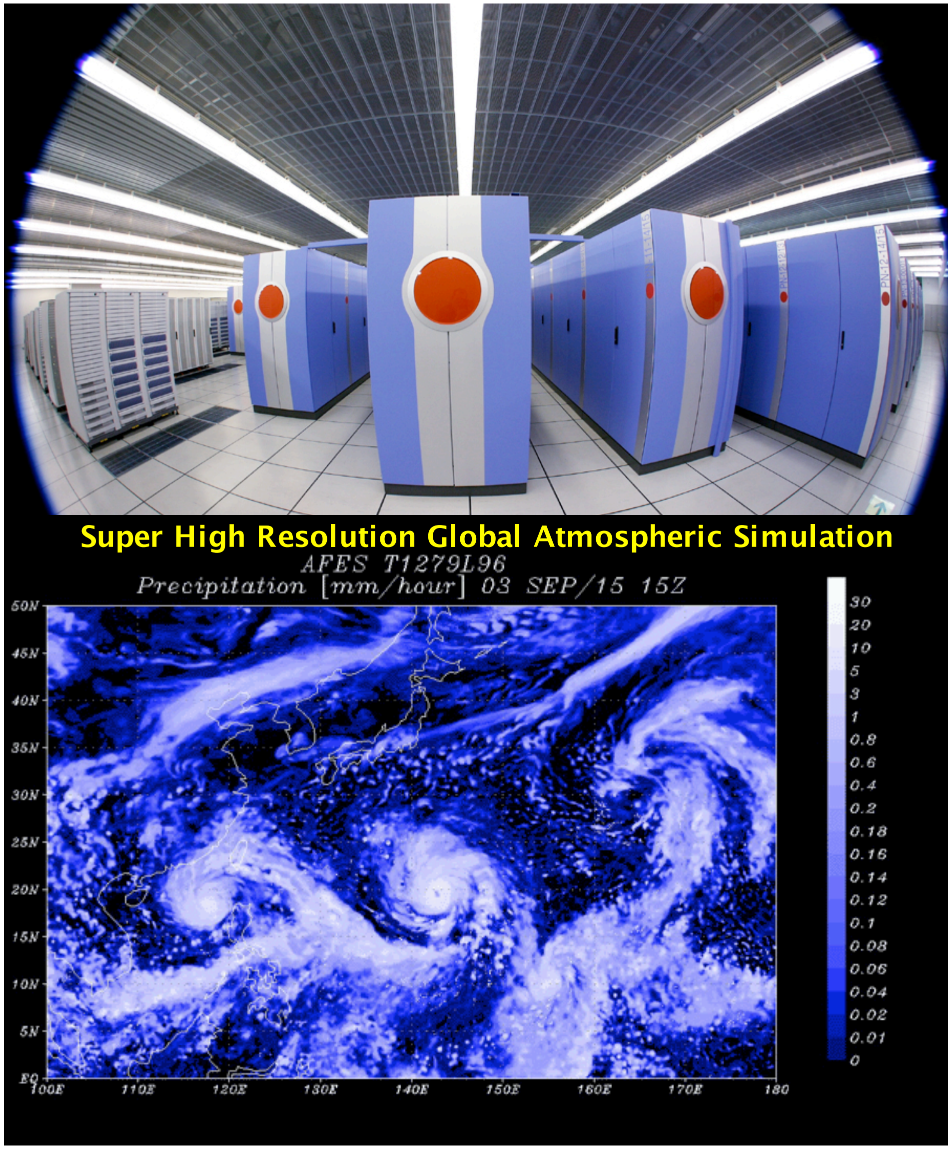
LIP (CNRS-INRIA-ENS-UCBL)

Computational Sciences

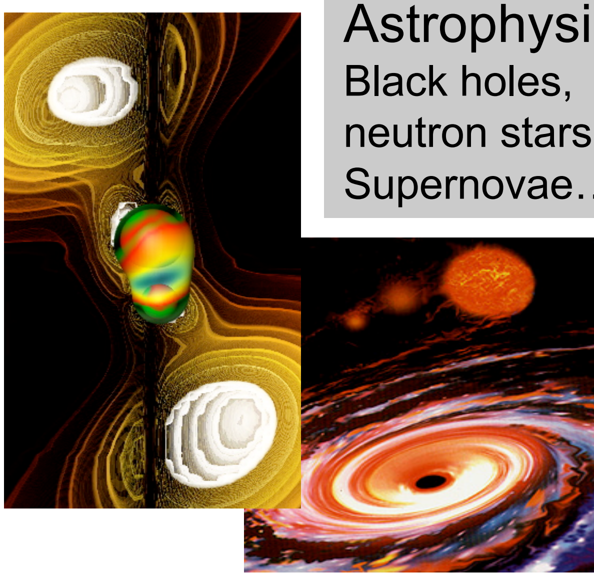
- Use of computers to solve complex problems
  - Modelling techniques
  - Simulation tehniques
  - Analytic & Mathematic methods
  - …
- Large problems require huge amount of processing power: supercomputers, high-performance clusters, etc.

Earth simulator: #1 TOP500

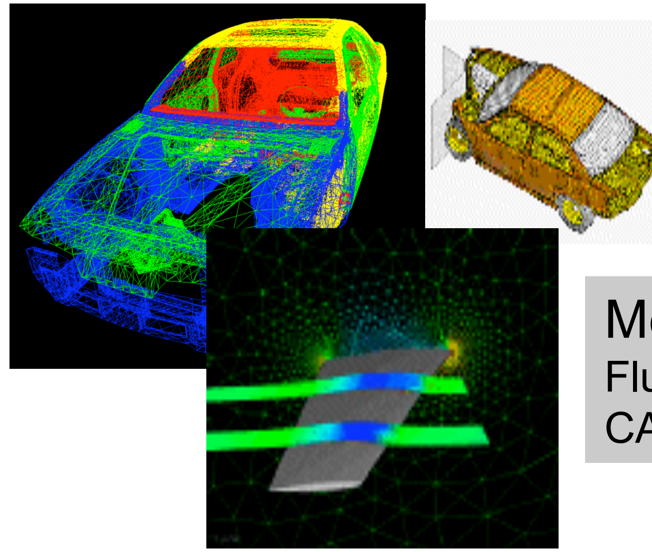
- Intensive numerical simulations
- Ex: Super High Resolution Global Atmospheric Simulation



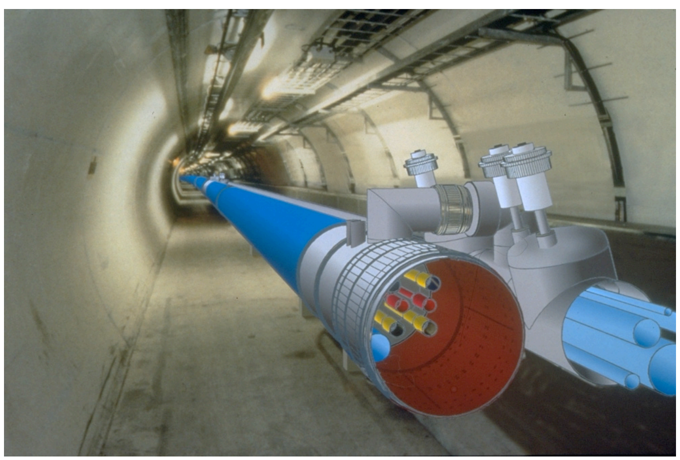
A large variety of applications



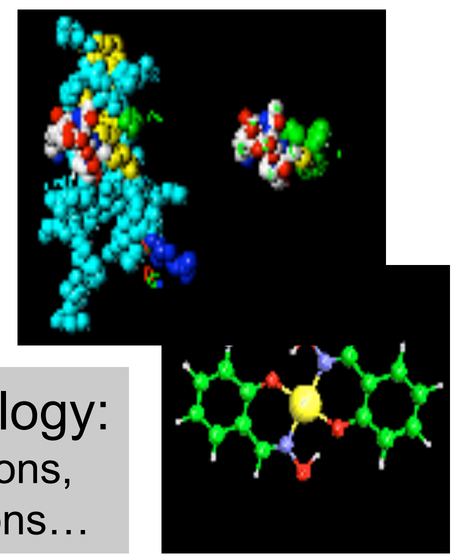
Astrophysics:
Black holes,
neutron stars,
Supernovae…



Mechanics:
Fluid dynamic,
CAD, simulation.



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

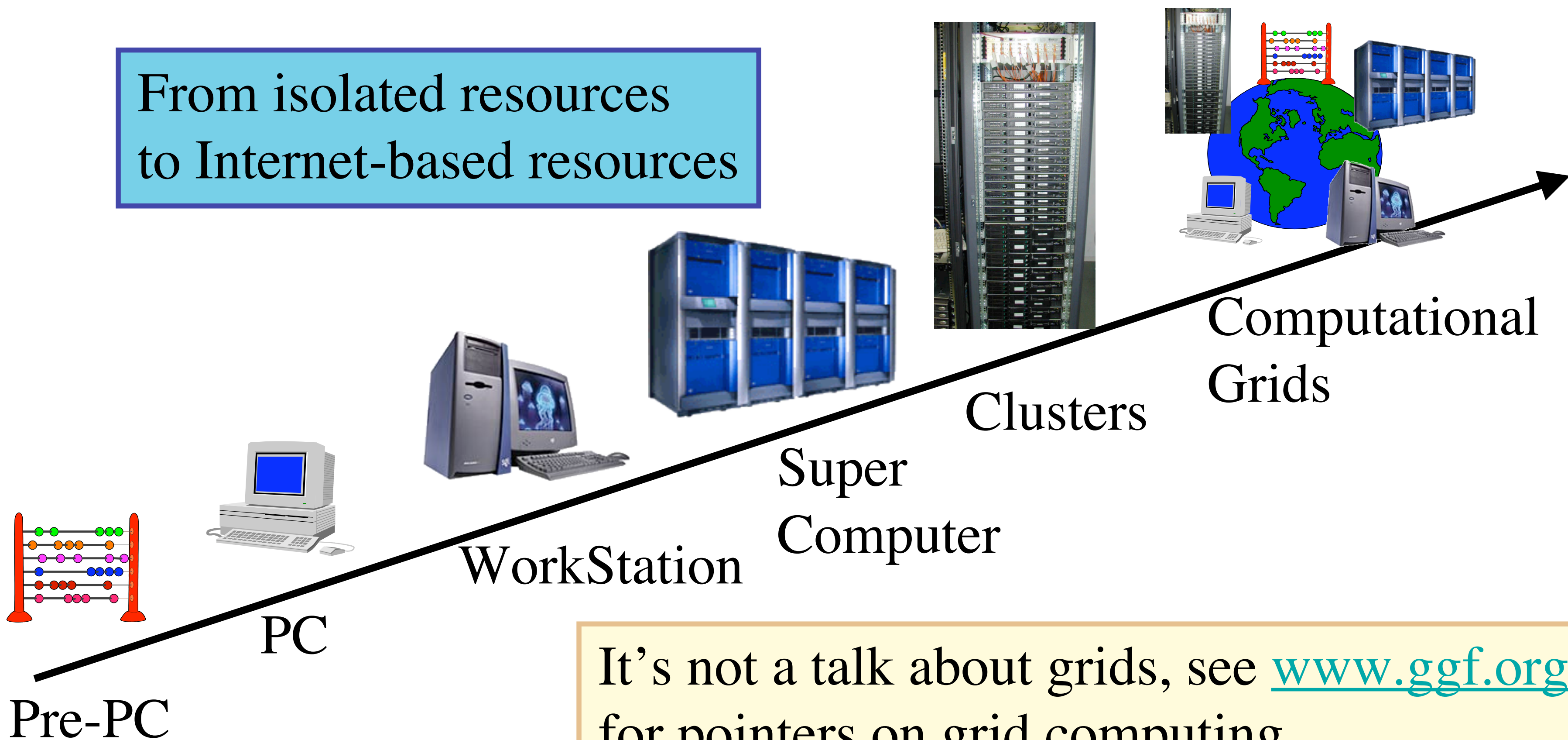


Chemistry&biology:
Molecular simulations,
Genomic simulations…

This talk is about…

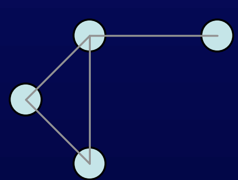
- How the Internet revolution could be beneficial to computational sciences

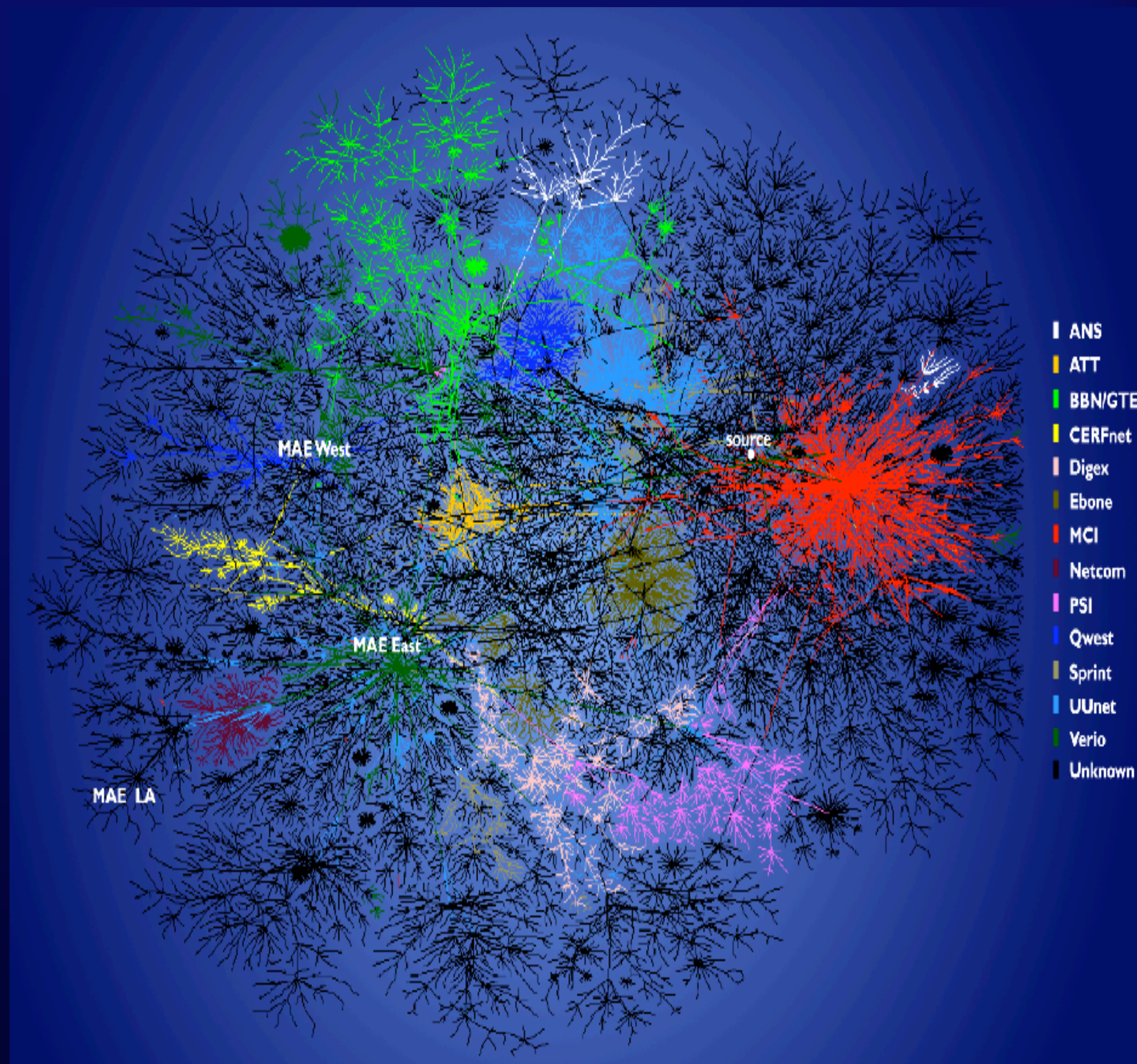
From isolated resources
to Internet-based resources



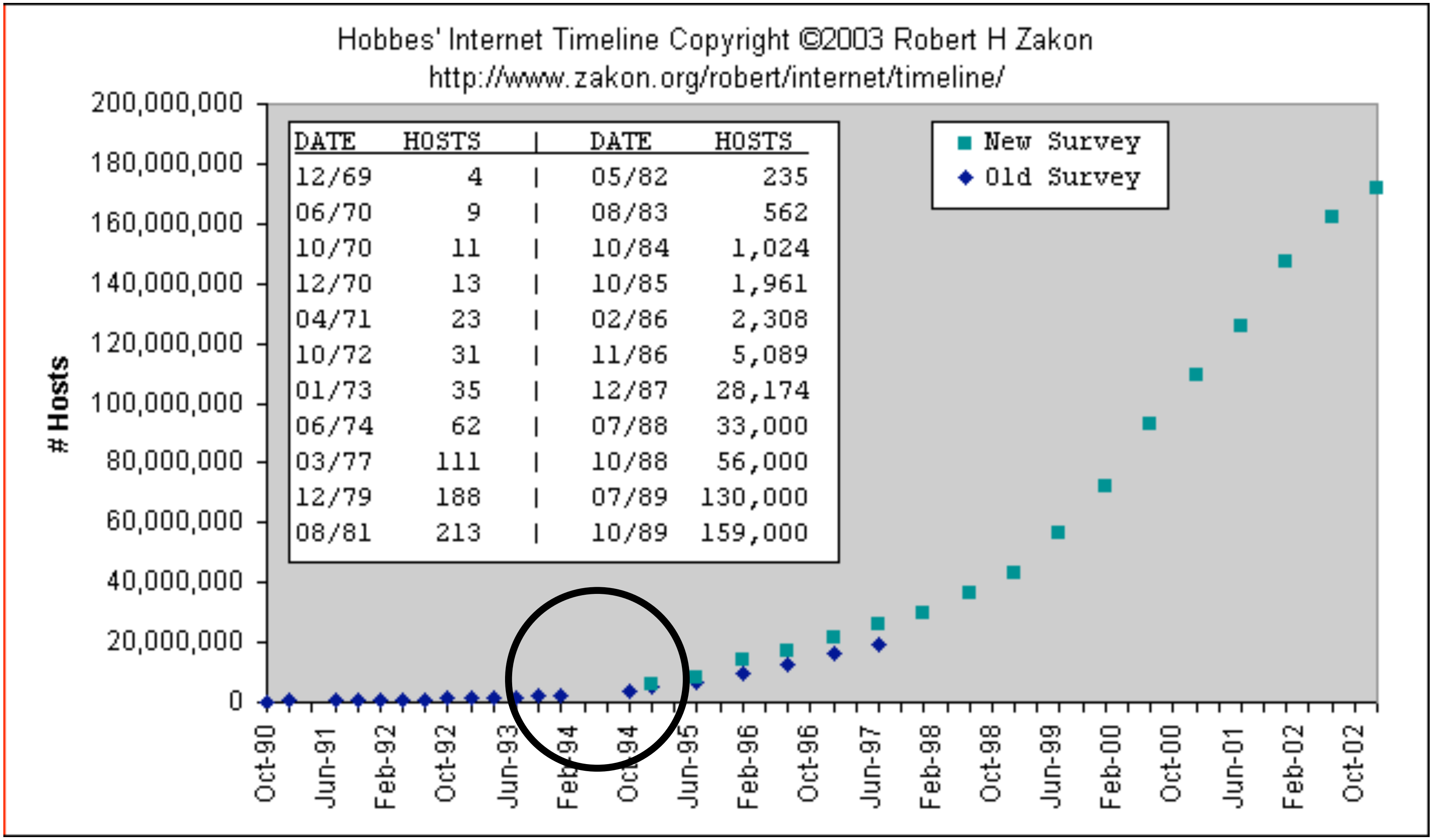
It's not a talk about grids, see www.ggf.org
for pointers on grid computing

The big-bang of the Internet

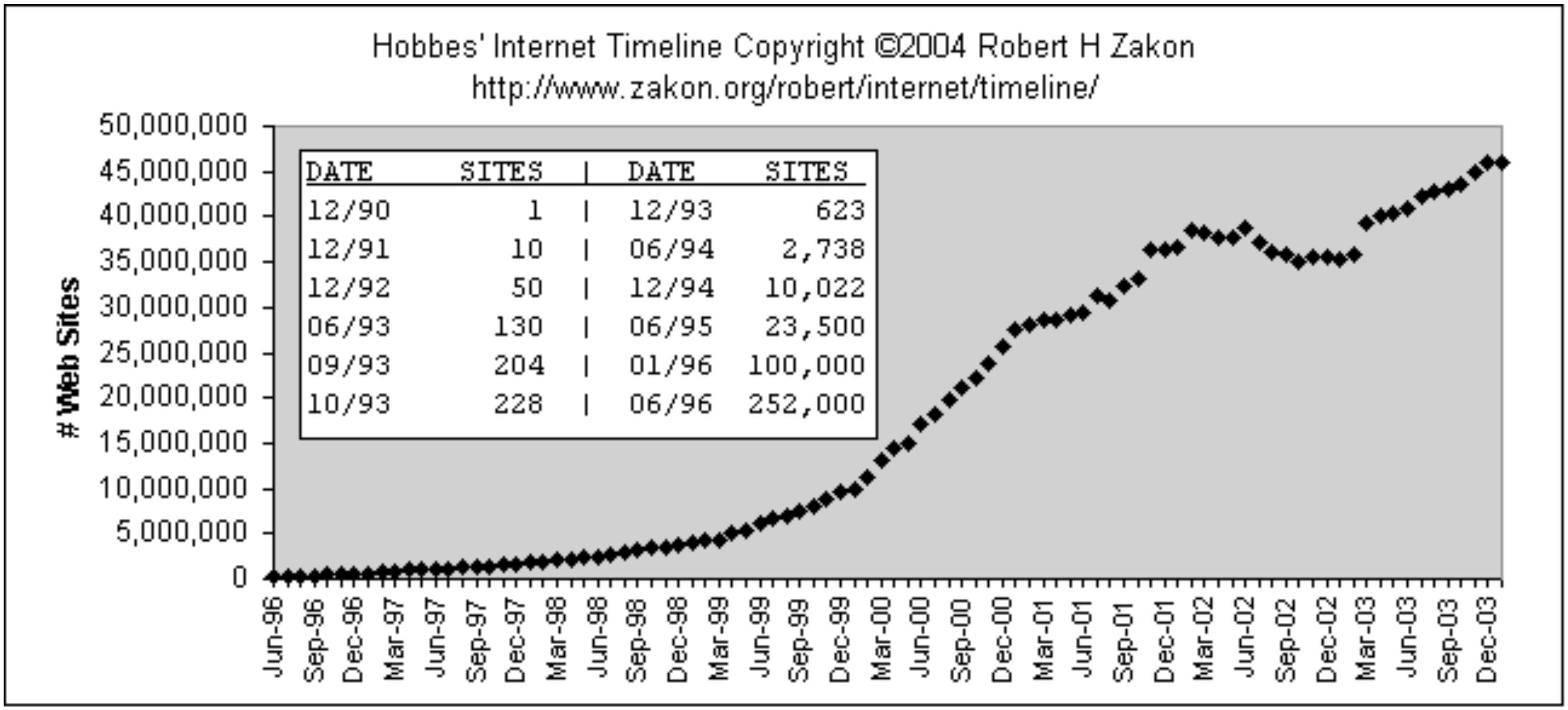




Internet host



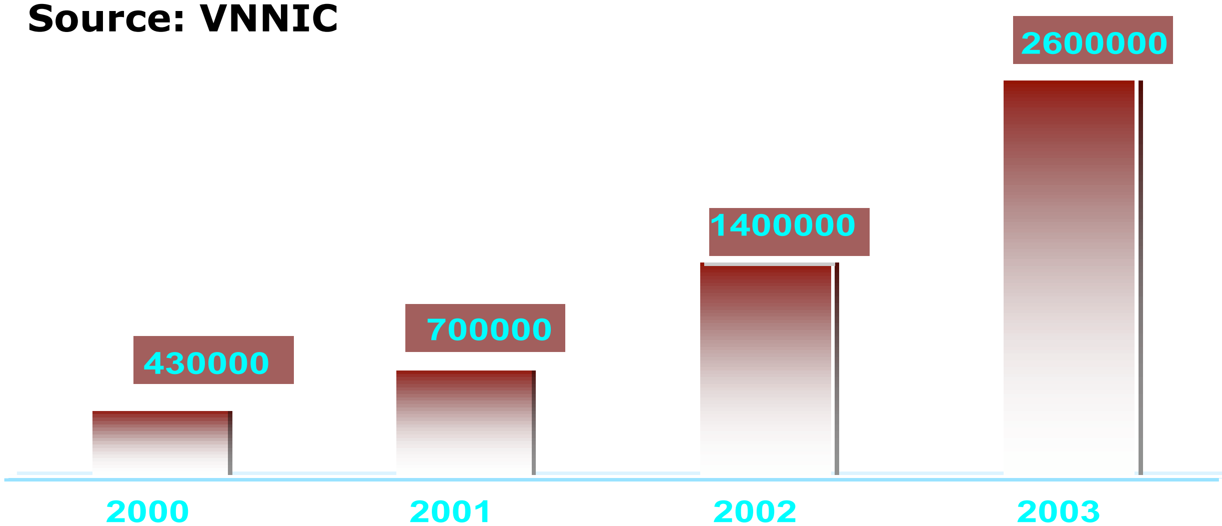
www.web-the-big-bang.org



The Internet in Vietnam

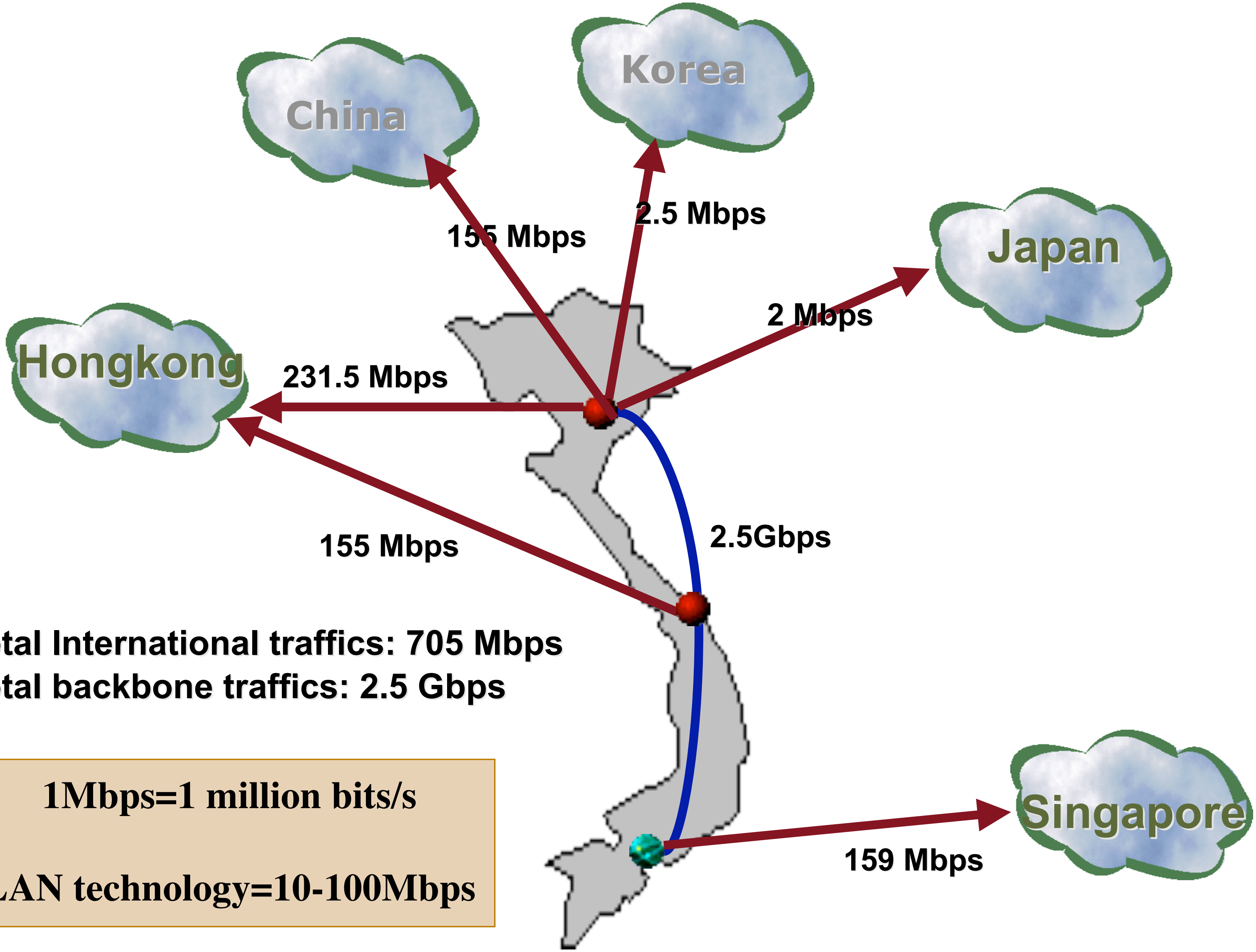
Year	2000	2001	2002	oct2003
Subscribers	103,751	166,616	350,000	650,654
Users	430,000	700,000	1.4 mil	2.6 mil
Penetration rate	0.5%	0.9%	1.7%	3.2%

Source: VNNIC



Jan 2004

868,059
3.5 mil
4.31%



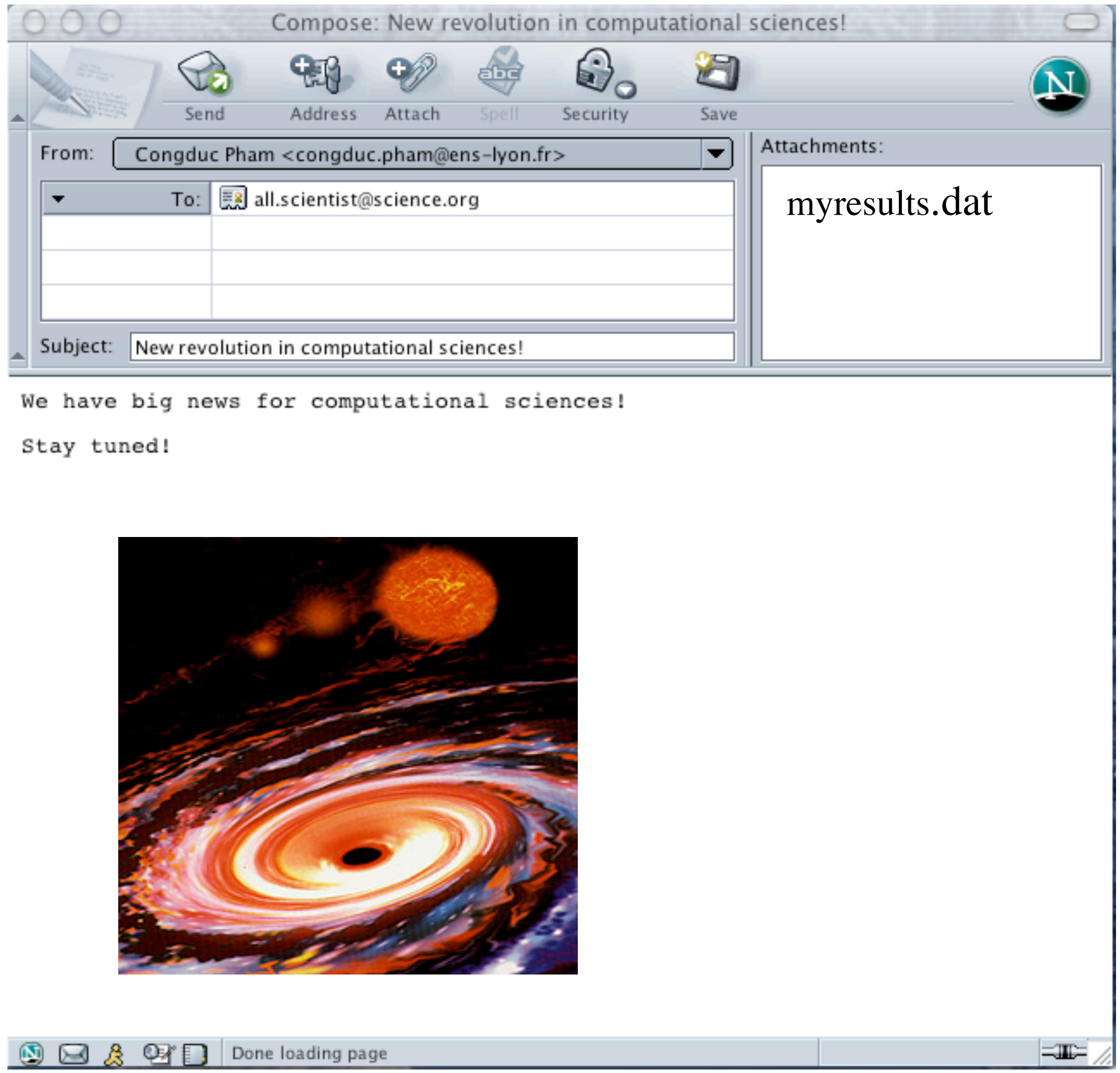
Total International traffics: 705 Mbps
Total backbone traffics: 2.5 Gbps

1Mbps=1 million bits/s

LAN technology=10-100Mbps

Internet usage: e-mail…

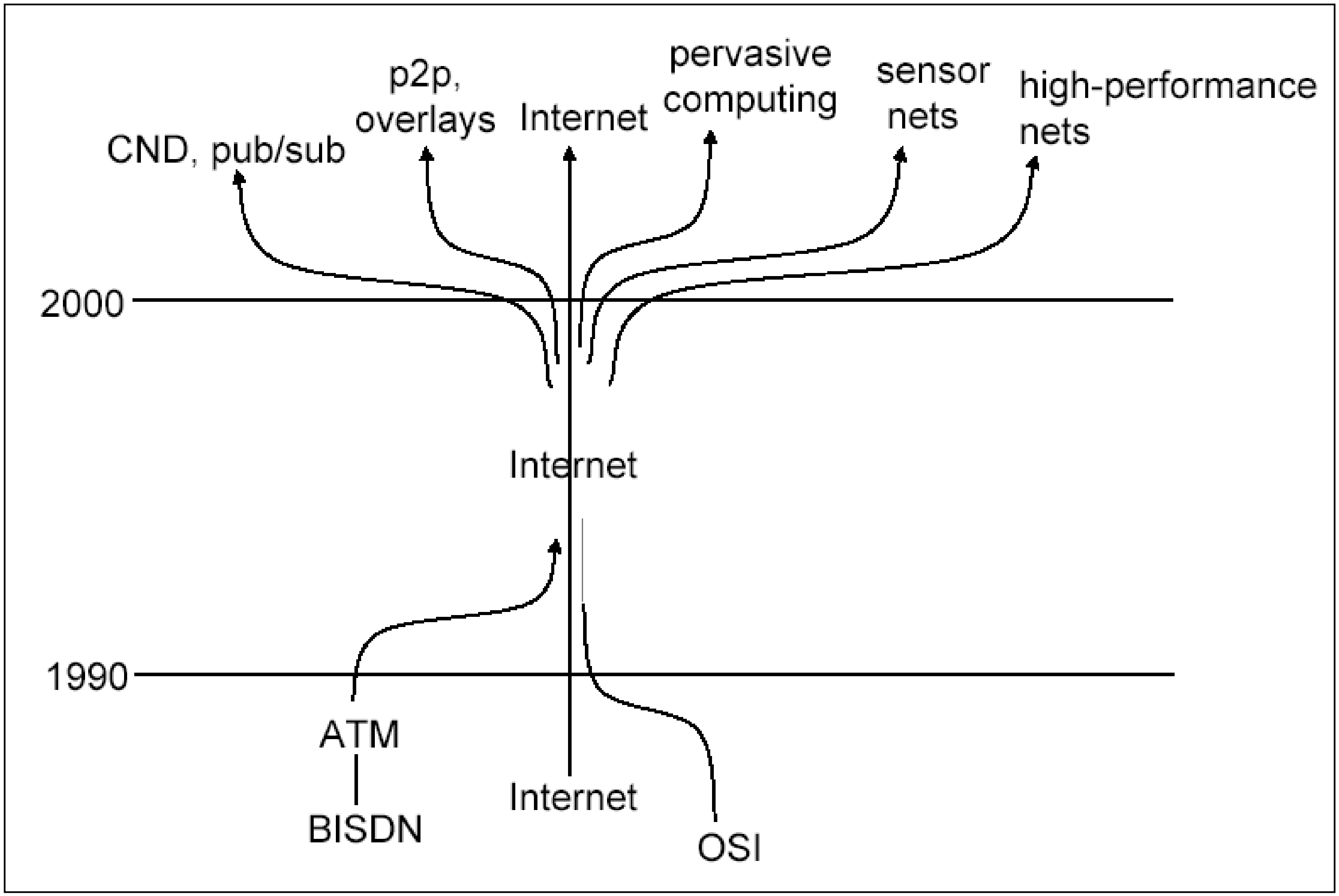
- Convenient way to communicate in an informal manner
- Attachments as a easy way to exchange data files, images…



...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
  - ...

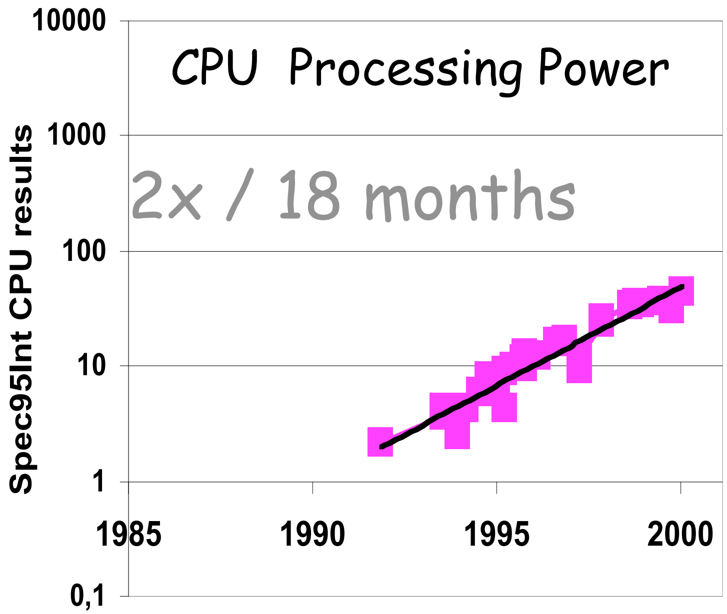
Towards all IP



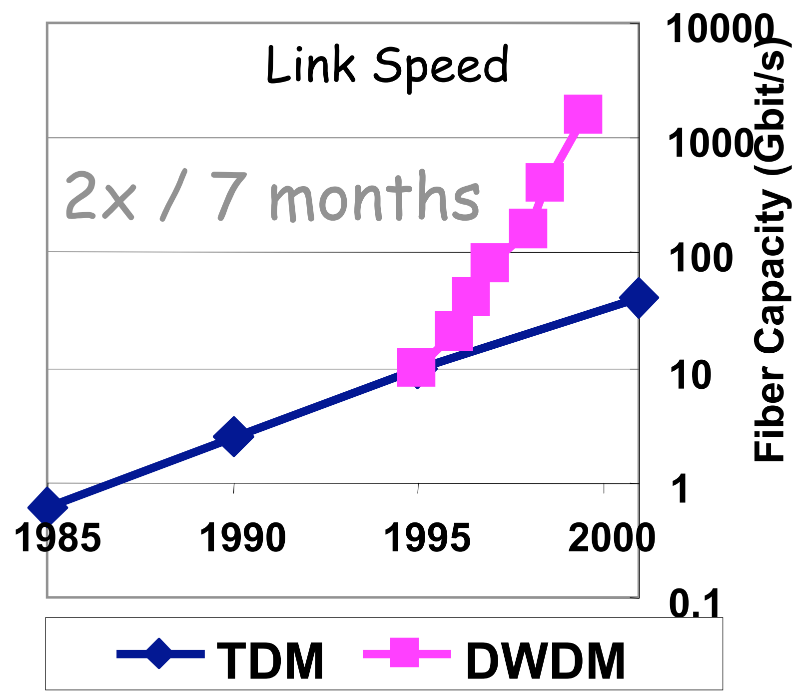
From Jim Kurose

A whole new world for IP

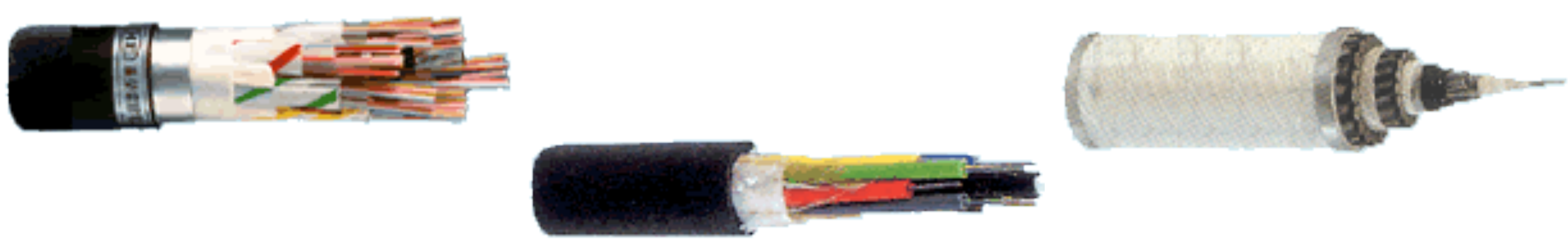
The optical revolution



From McKeown

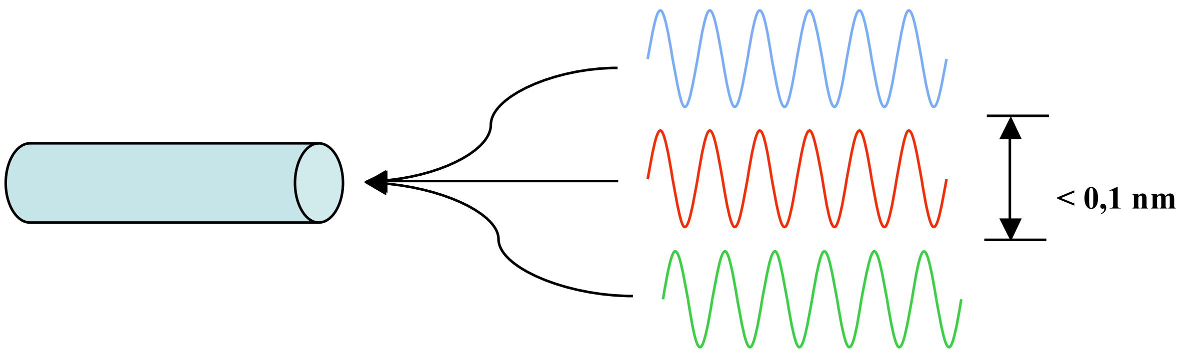


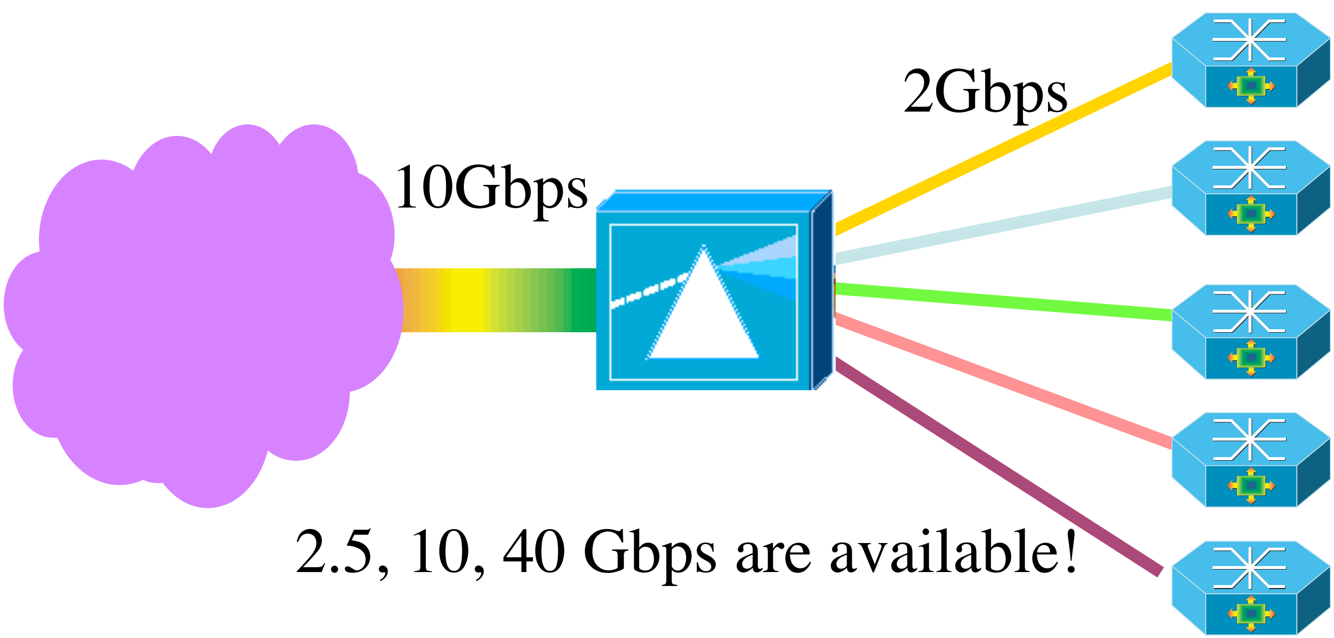
Demand: about 111 million km of cabled optical fiber / year



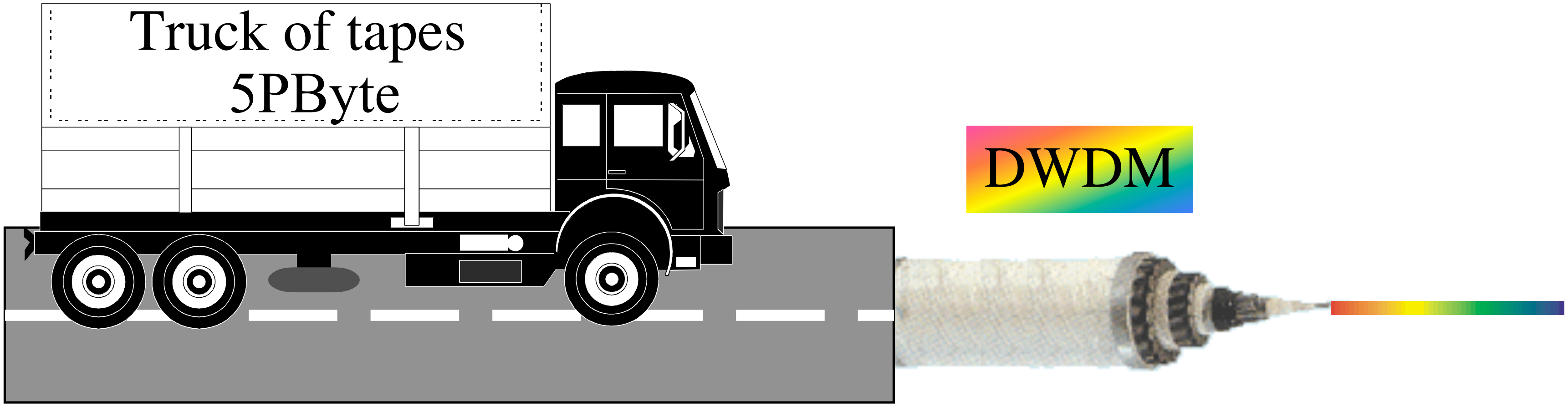
DWDM, bandwidth for free?

DWDM: Dense Wavelength Division Multiplexing

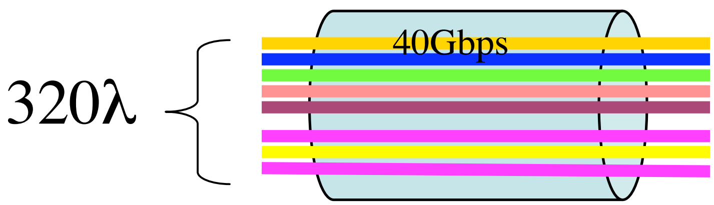




The information highways



1600 Gbyte/s



Revisiting the truck of tapes

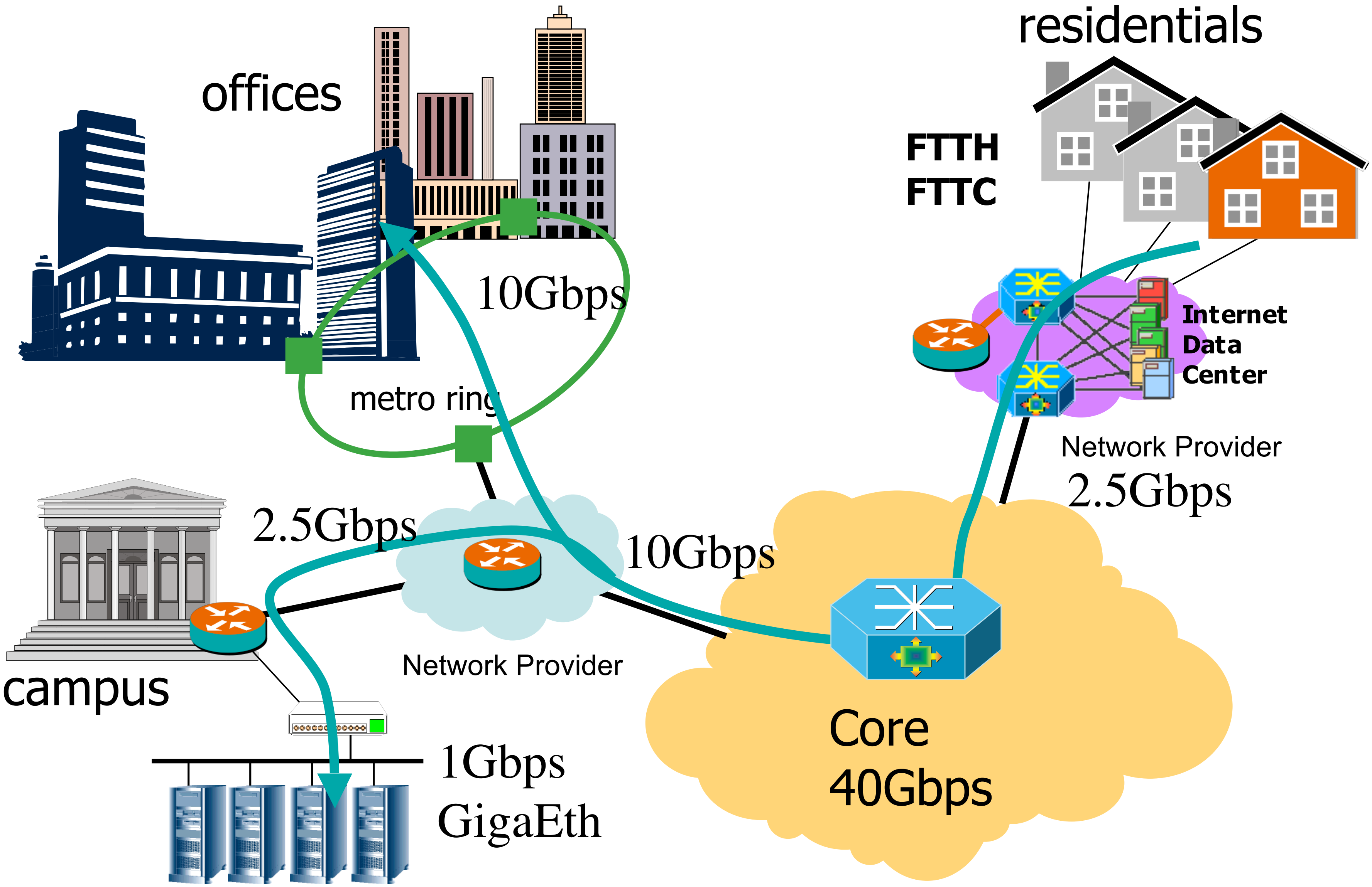
(18 of 18)

Consider one fiber

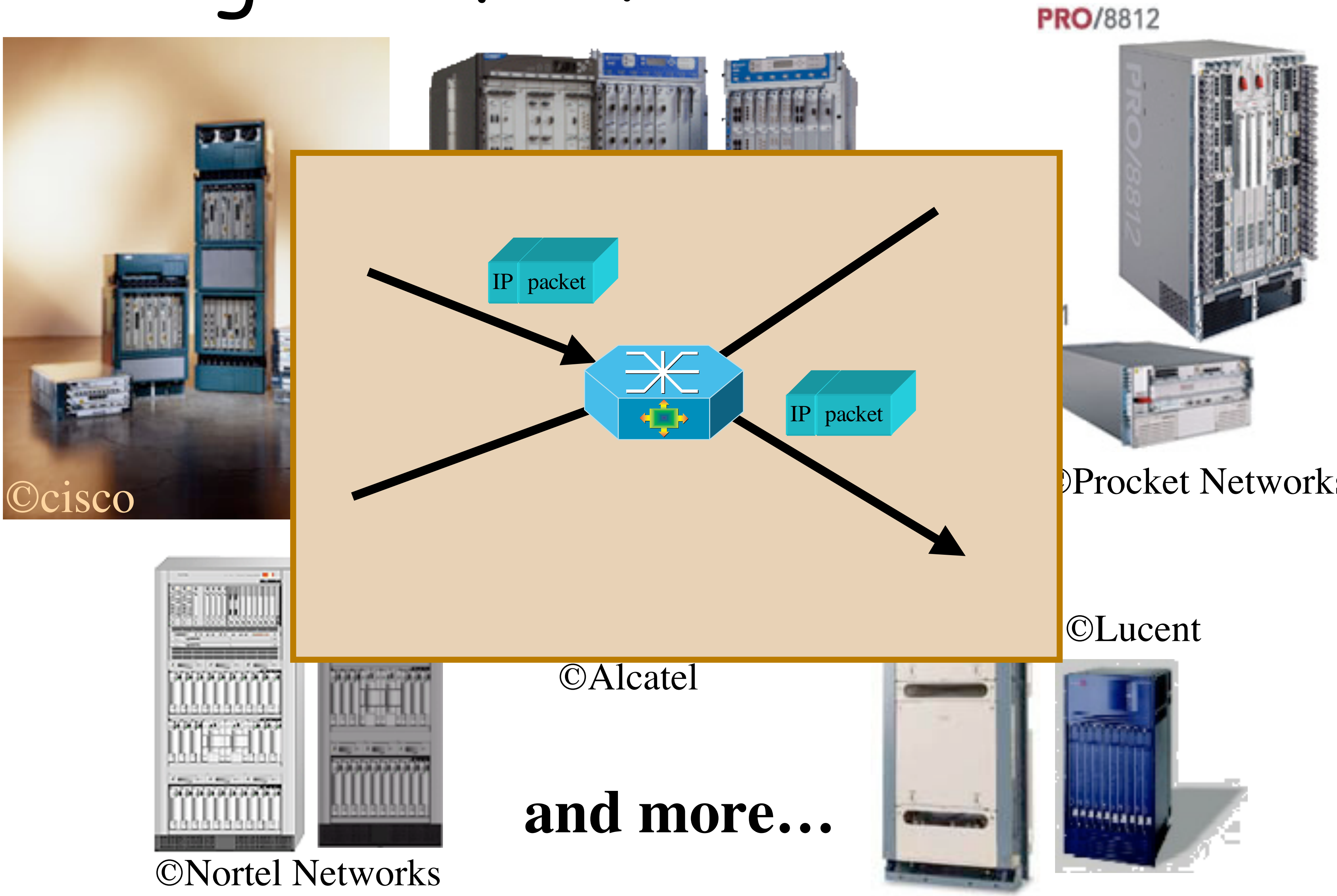
- Current technology allows for 320 λ in one of the frequency bands
- Each λ has a bandwidth of 40 Gbit/s
- Transport: $320 * 40*10^9 / 8 = 1600$ GByte/sec
- Take a 10 metric ton truck
- One tape contains 50 Gbyte, weights 100 gr
- Truck contains (10000 / 0.1) * 50 Gbyte = 5 PByte
- Truck / fiber = 5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

Example from A. Tanenbaum, slide from Cees De Laat

Fibers everywhere?



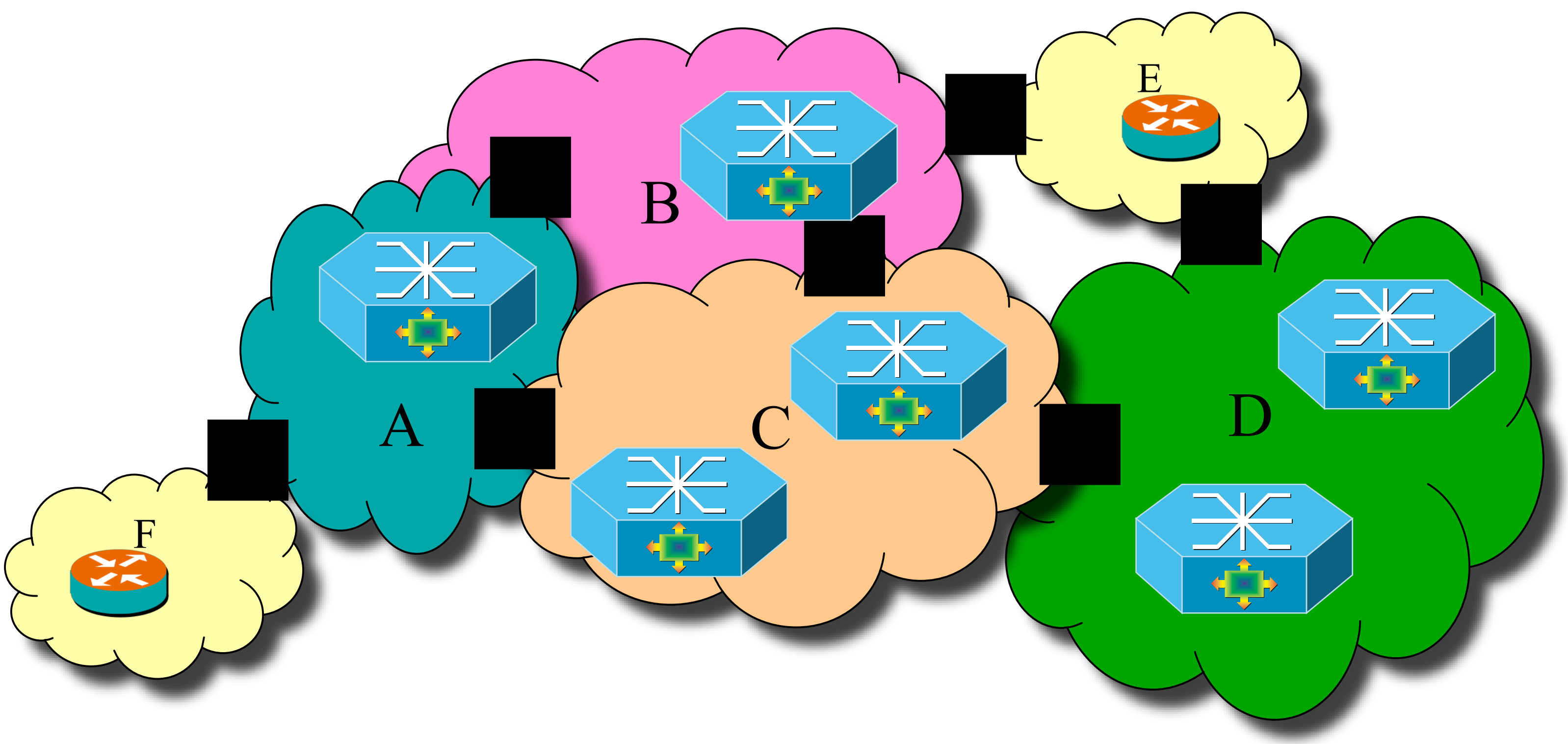
High Performance Routers



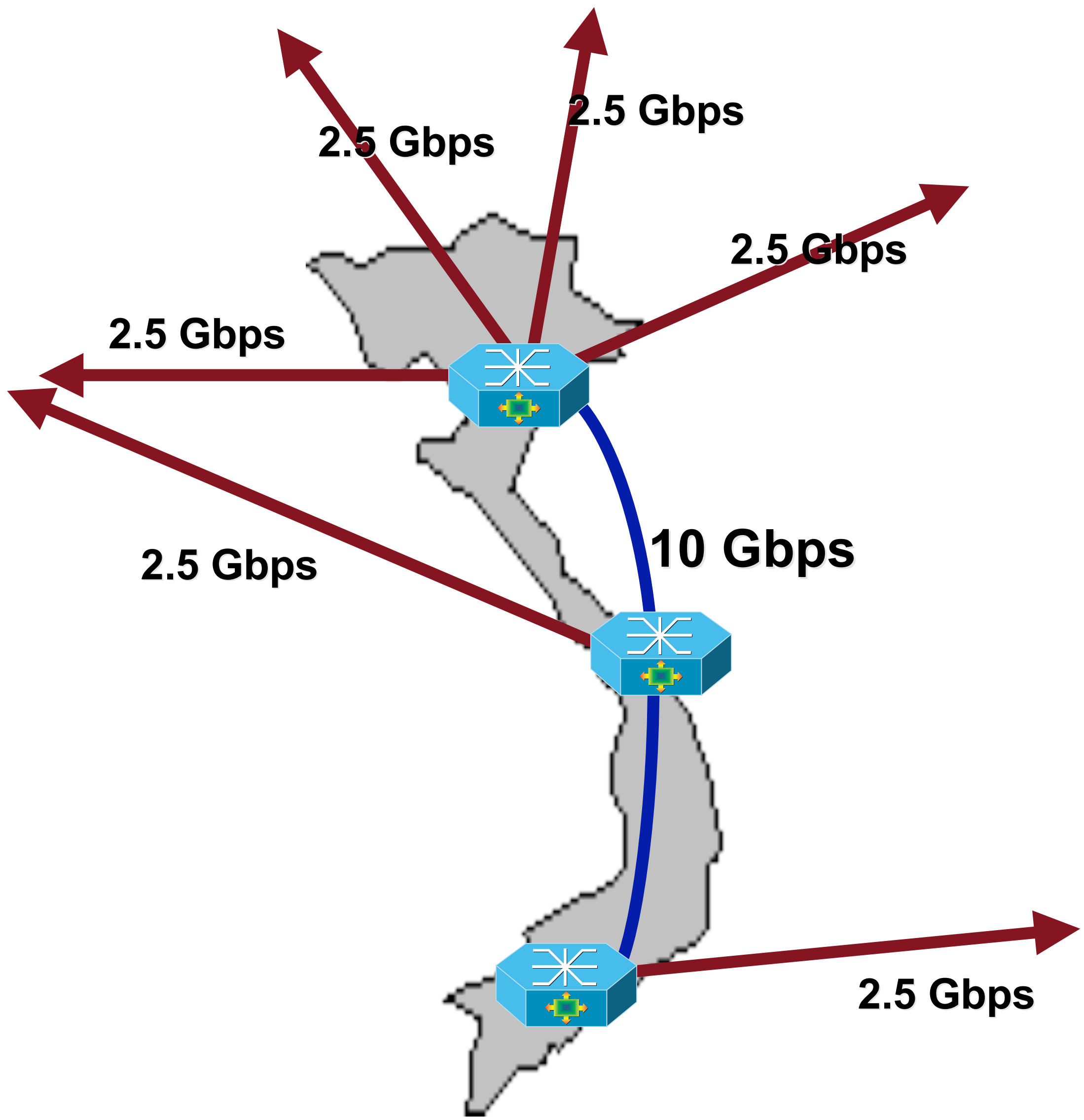
and more…

Operator's infrastructure

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



In a near future?

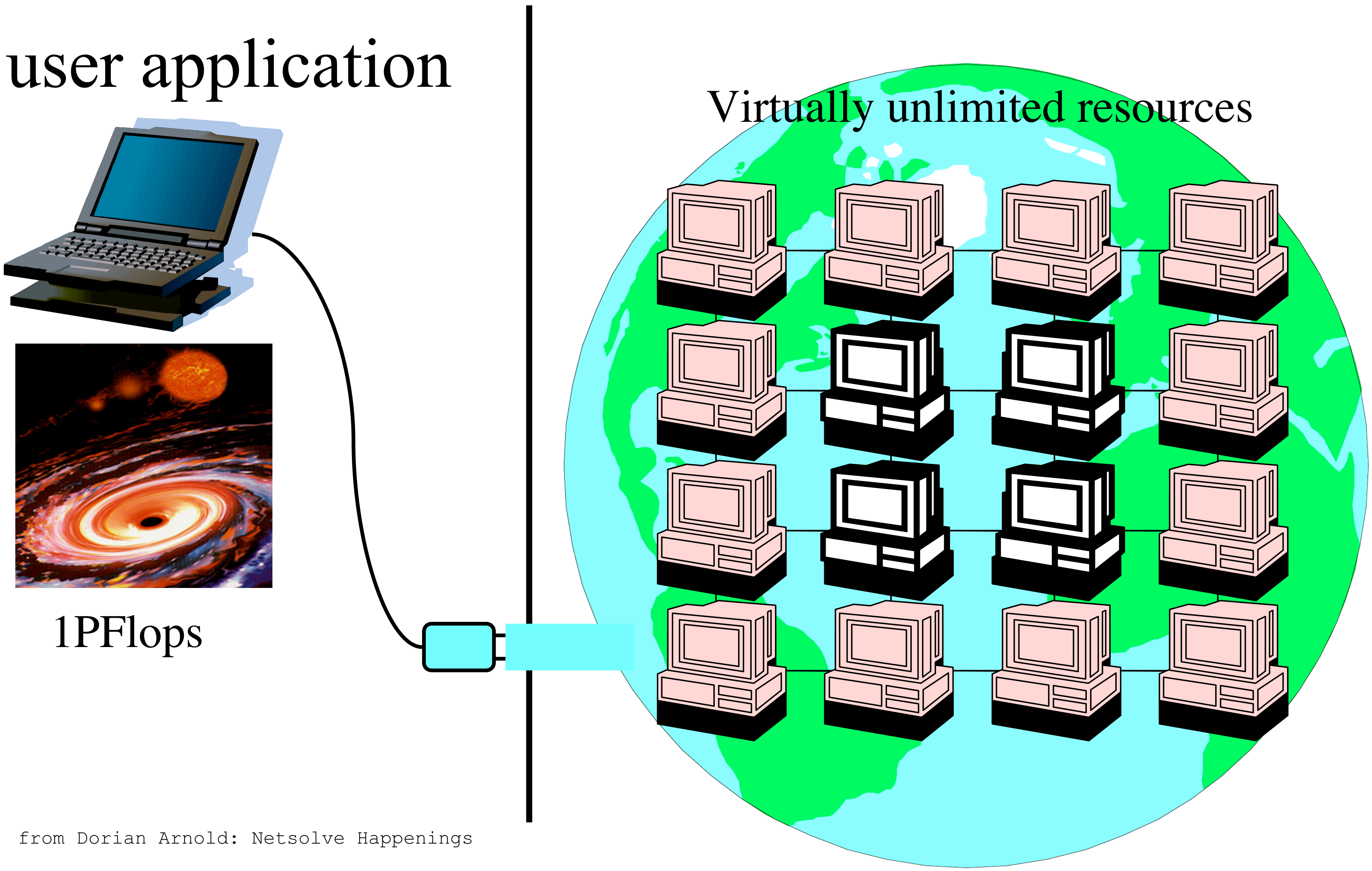


New applications on the information highways

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

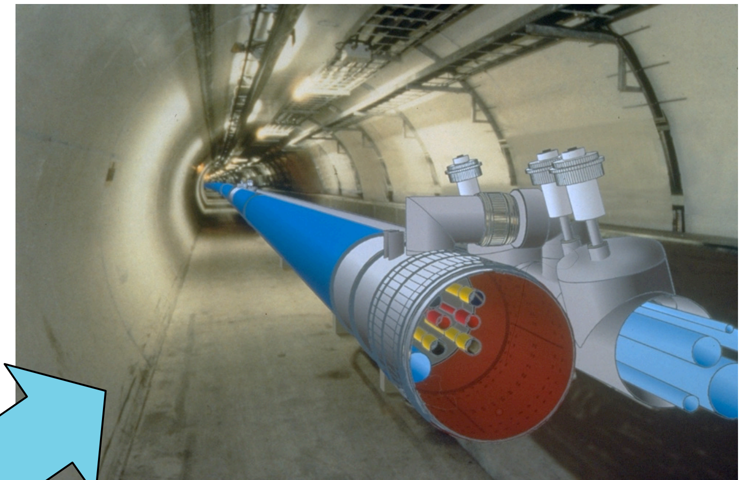
Computational grids



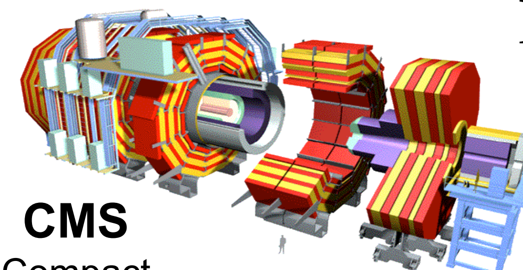
from Dorian Arnold: Netsolve Happenings

High Energy Physics at CERN

LEP

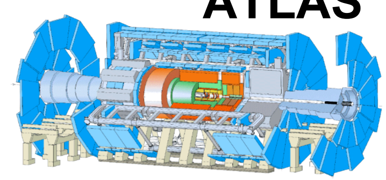


LHC



CMS
Compact
Muon
Solenoid

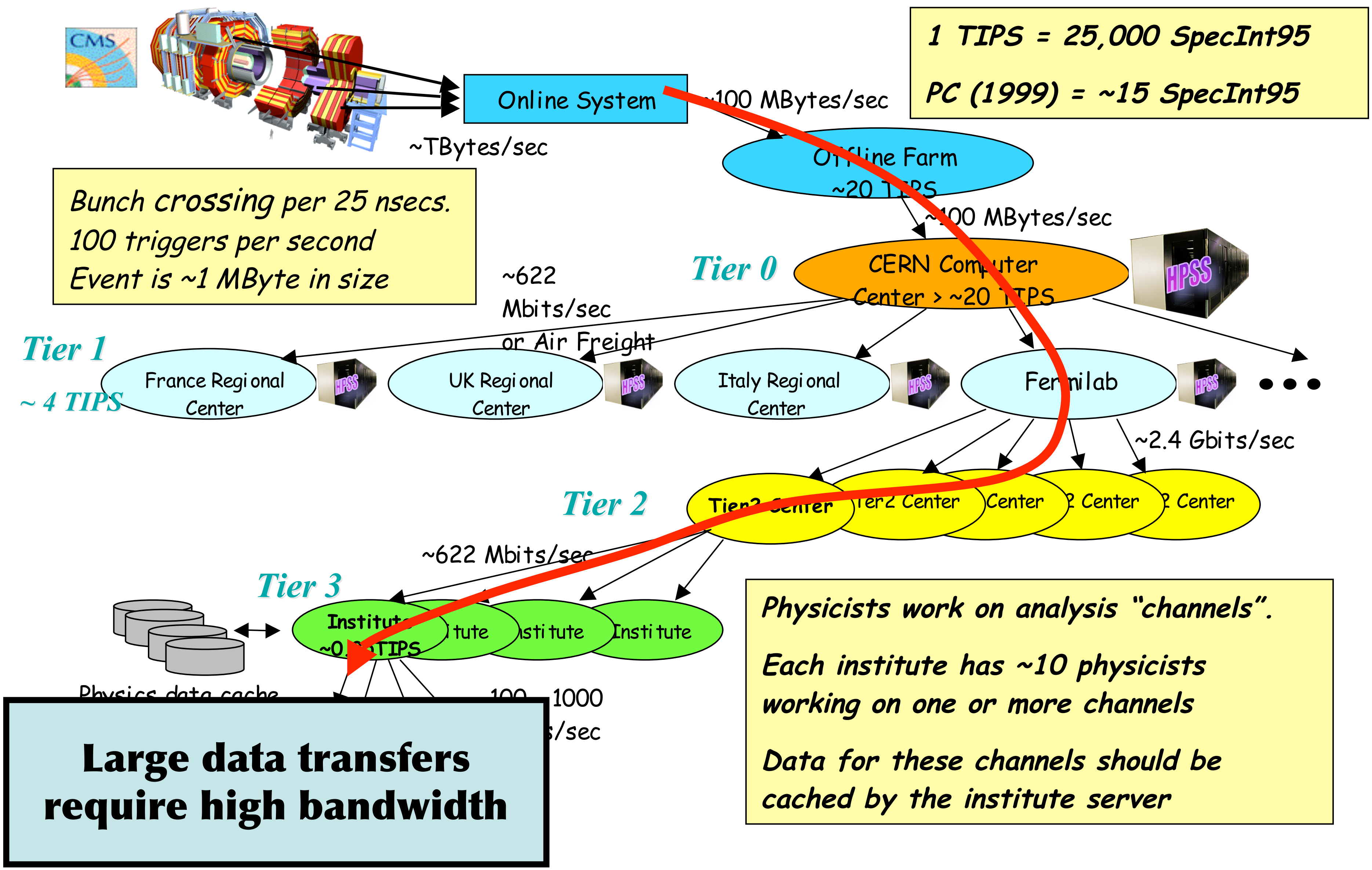
ATLAS



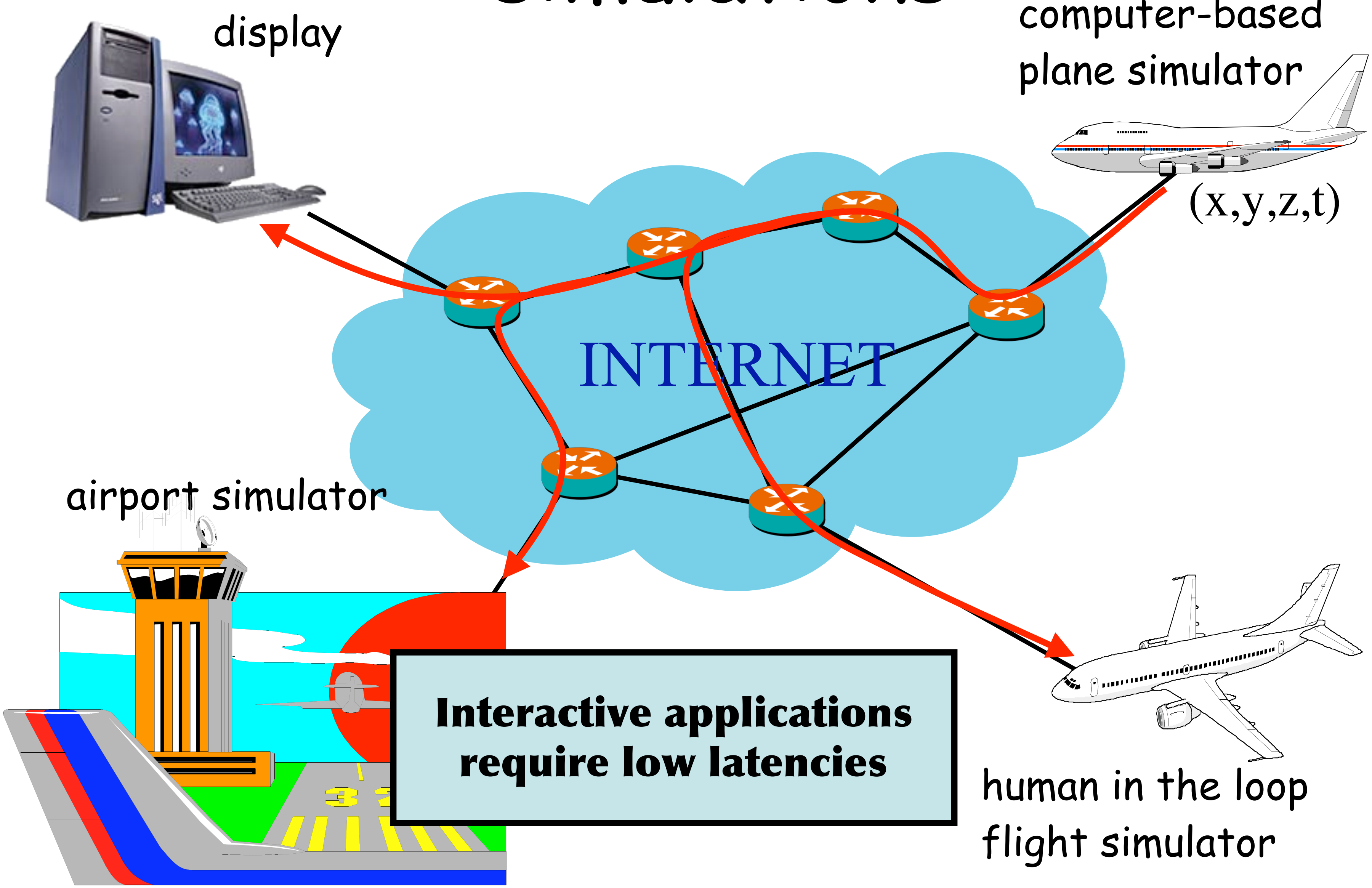
Images from EDG (DataGrid) project

3.5 Petabytes/year ≈ 10^9 events/year

Distributed Databases



Wide-area interactive simulations



Interactive applications require low latencies

Limitations of the current Internet

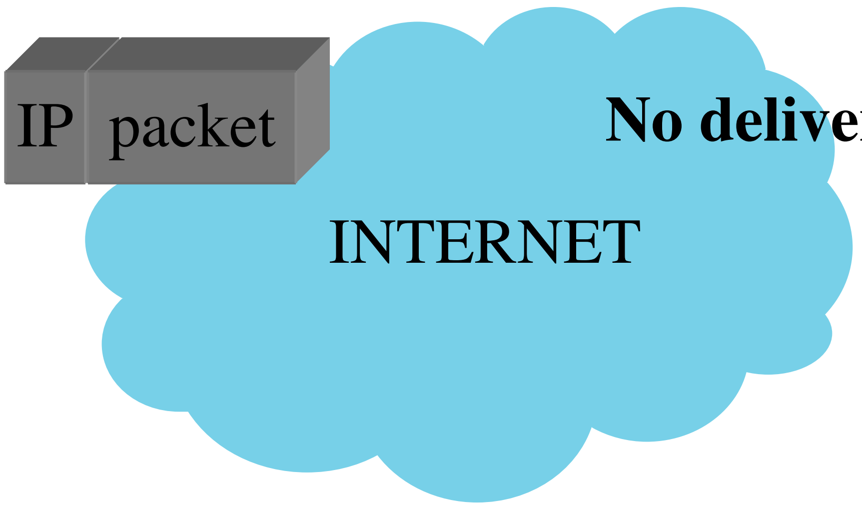
- Bandwidth
  - Raw bandwidth is not a problem: DWDM
  - Provisioning bandwidth on demand is more problematic
- Latency
  - Mean latencies on Internet is about 80-160ms
  - Bounding latencies or ensuring lower latencies is a problem
- Loss rate
  - Loss rate in backbone is very low
  - End-to-End loss rates, at the edge of access networks are much higher
- Communication models
  - Only unicast communications are well-defined: UDP, TCP
  - Multi-parties communication models are slow to be deployed

New technologies addressed in this talk

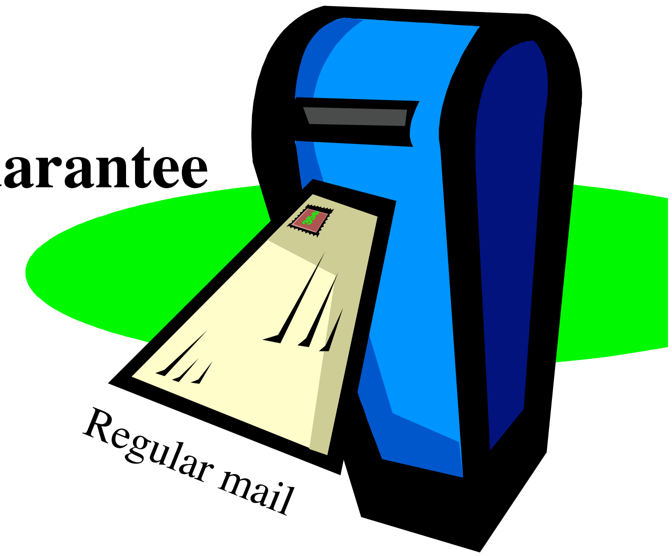
- More Quality of Service: Differentiated Services, who pays more gets more!
- Bandwidth provisioning: MPLS for virtual circuit in the core networks
- Multicast: enhancing the communication model

NEW CHAPTER

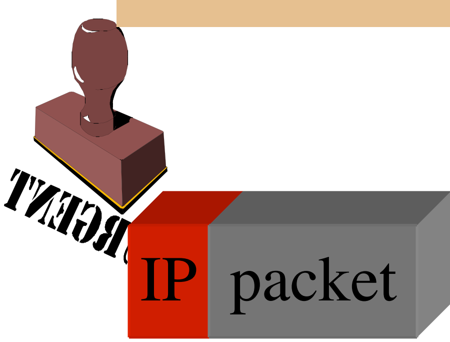
Revisiting the *same service for all* paradigm



No delivery guarantee



Enhancing the best-effort service

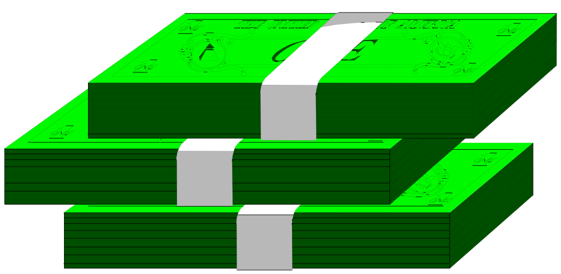


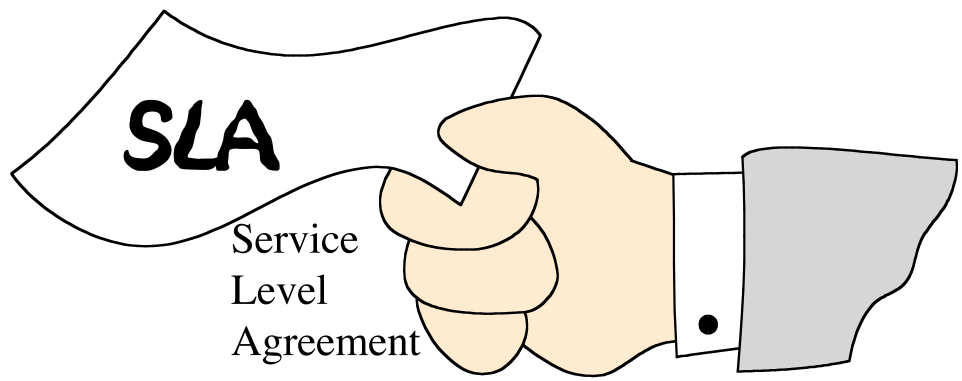
Introduce
Service Differentiation

Service Differentiation

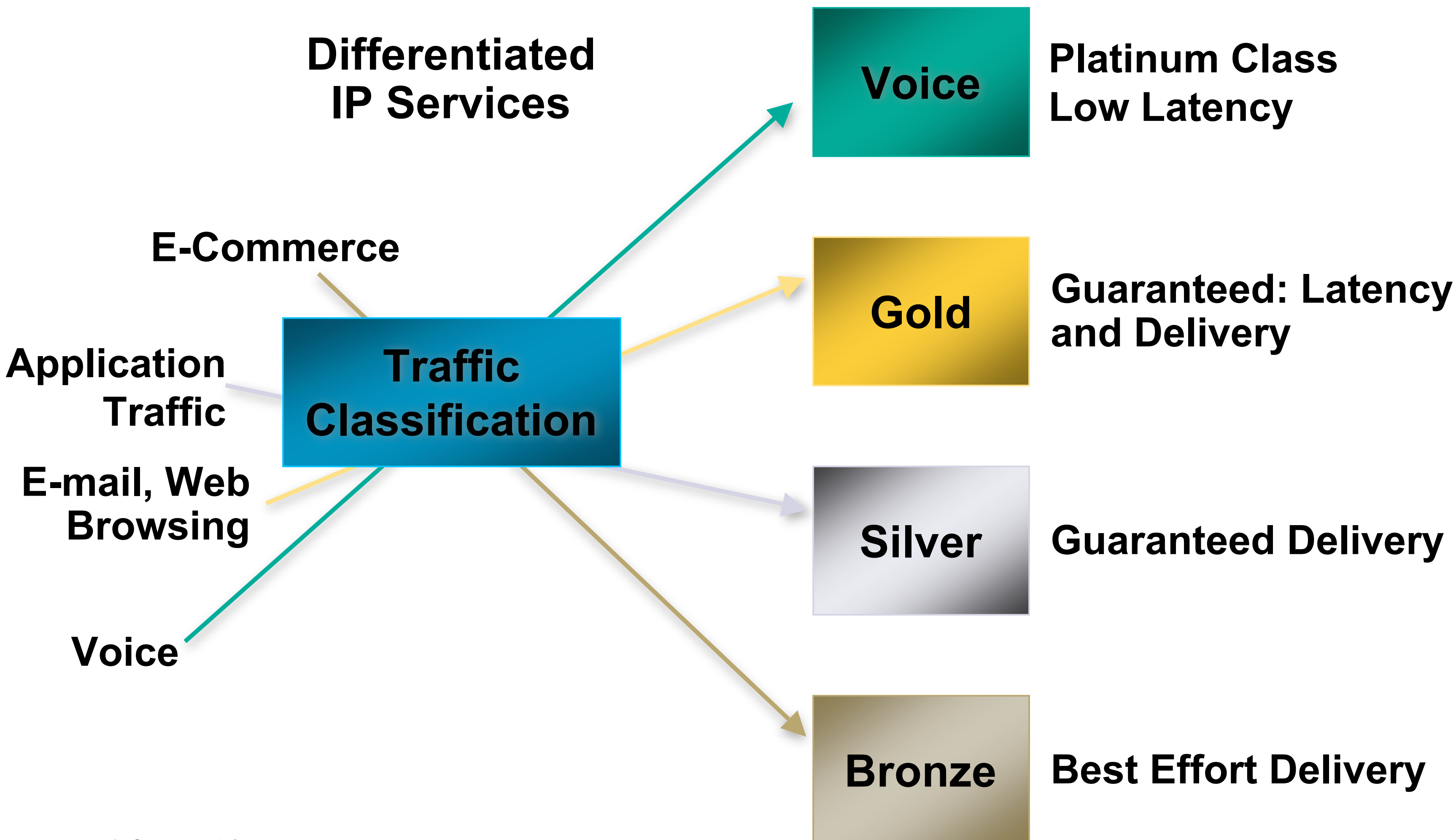
The real question is to choose which packets shall be dropped. The first definition of differential service is something like "not mine."
-- Christian Huitema

- 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



Borrowed from Cisco

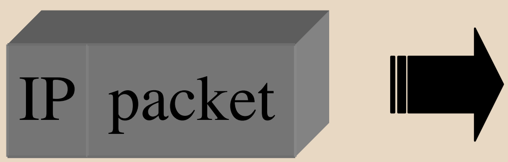
Design Goals/Challenges

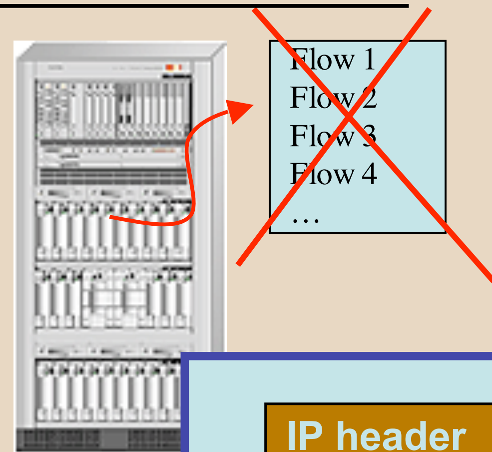
- Ability to charge differently for different services
- No per flow state or per flow signaling
- All policy decisions made at network boundaries
  - Boundary routers implement policy decisions by tagging packets with appropriate priority tag
- Traffic policing at network boundaries
- Deploy incrementally,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

IP implementation: DiffServ

RFC 2475

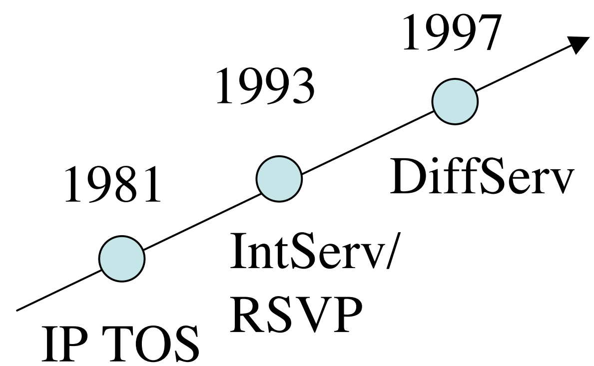
<u>No per flow state in the core</u>

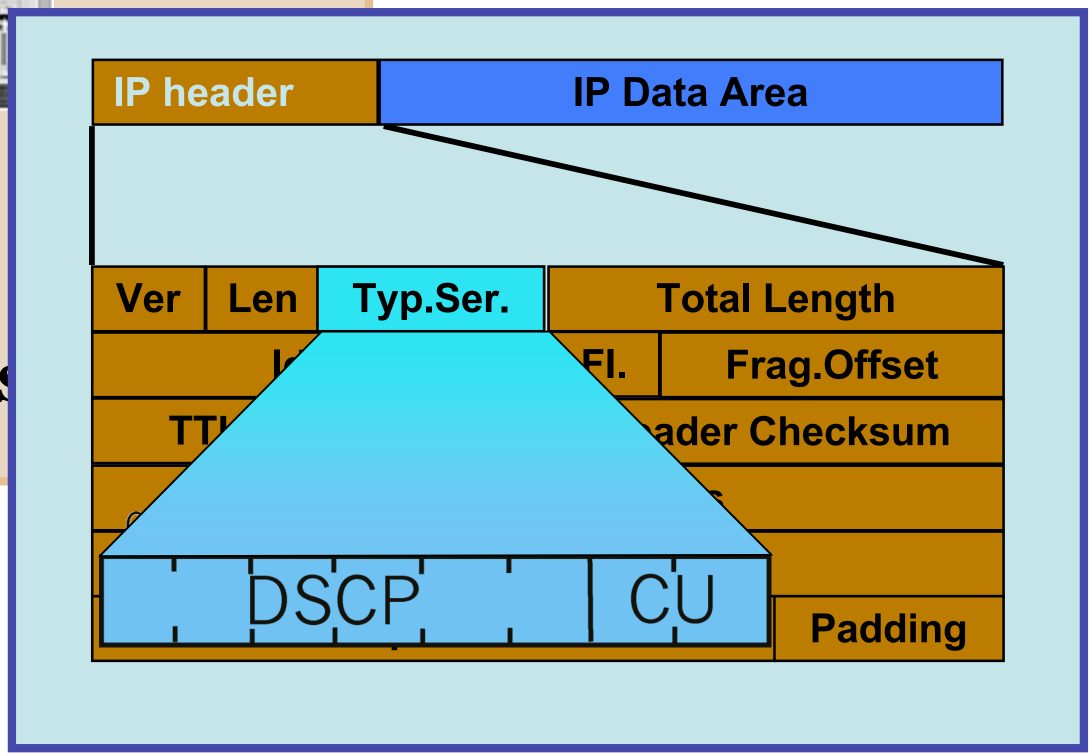




10Gbps=2.4Mpps
with 512-byte packets

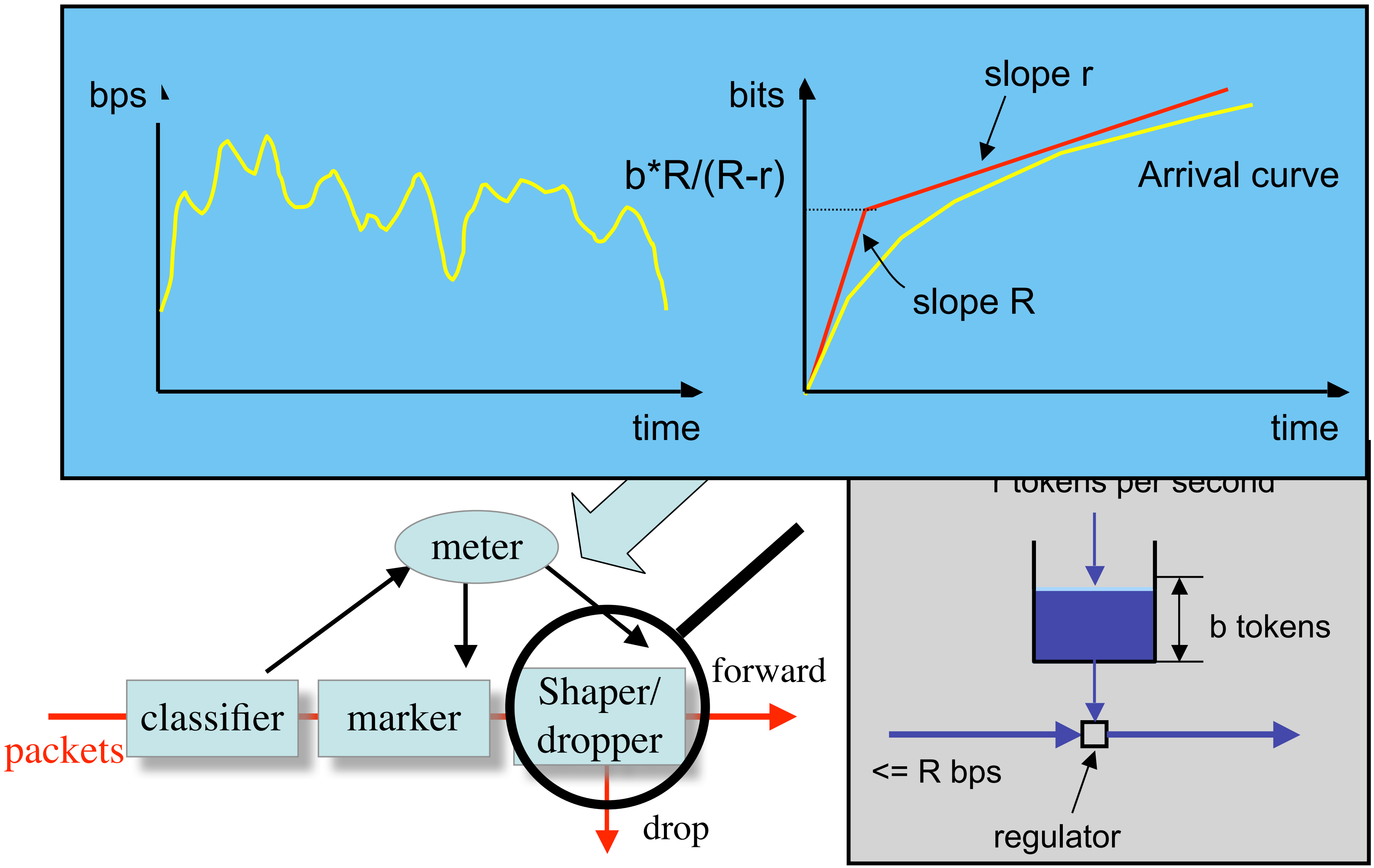
Stateful approaches
scalable
at gigabit rates



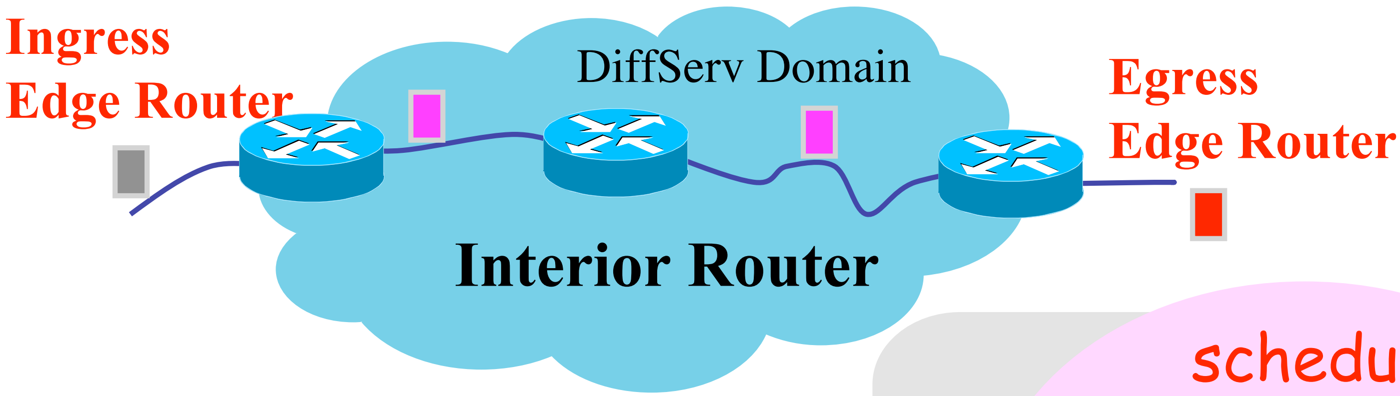


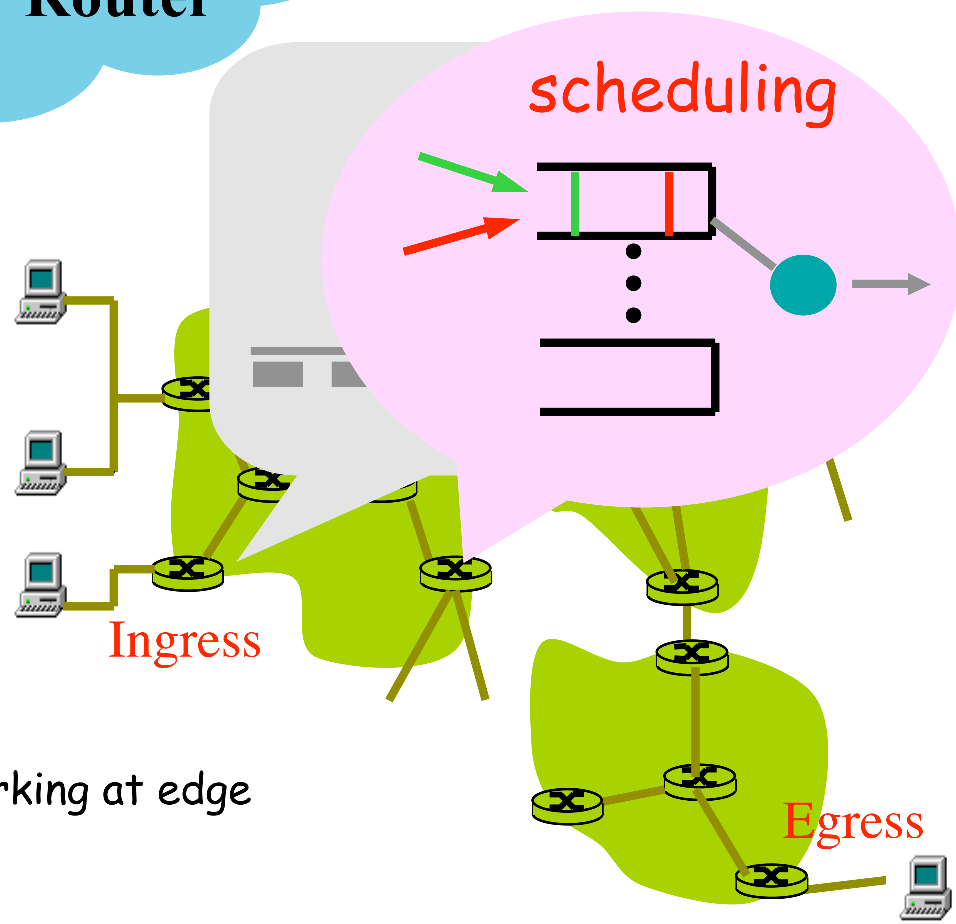
6 bits used for Differentiated Service Code Point (DSCP) and determine PHB that the packet will receive

Traffic Conditioning



Differentiated Architecture





Marking:

per-flow traffic management

marks packets as in-profile and out-profile

Per-Hop-Behavior (PHB):

per class traffic management

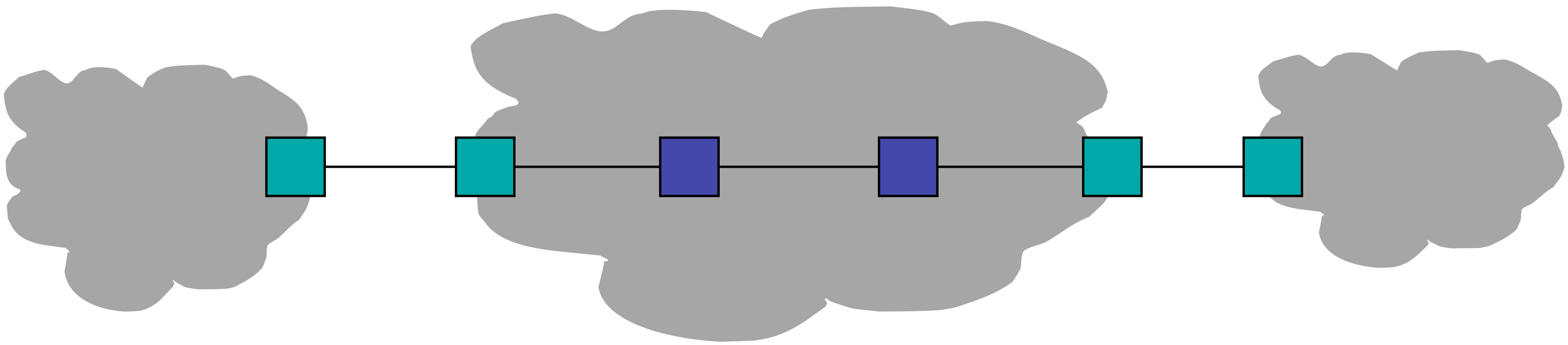
buffering and scheduling based on marking at edge

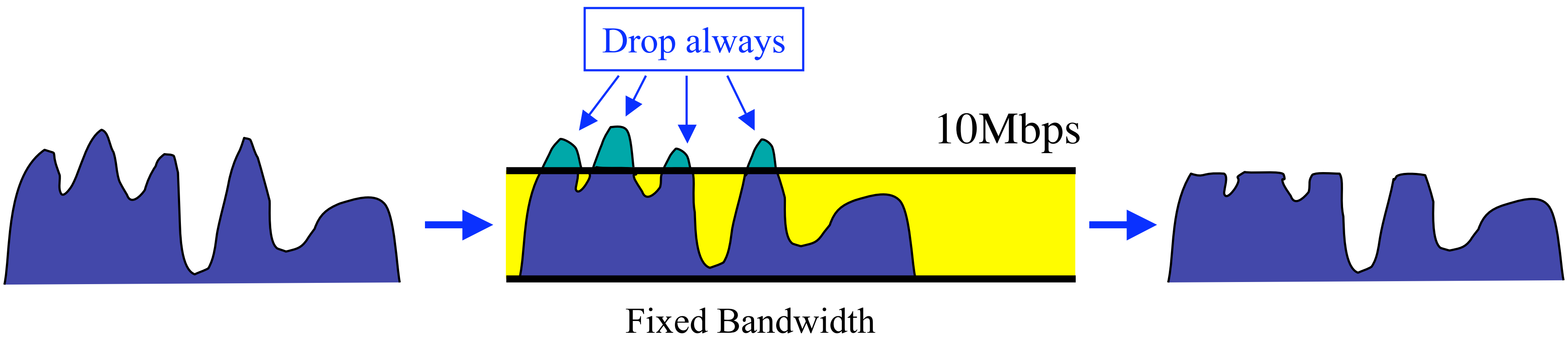
preference given to in-profile packets

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

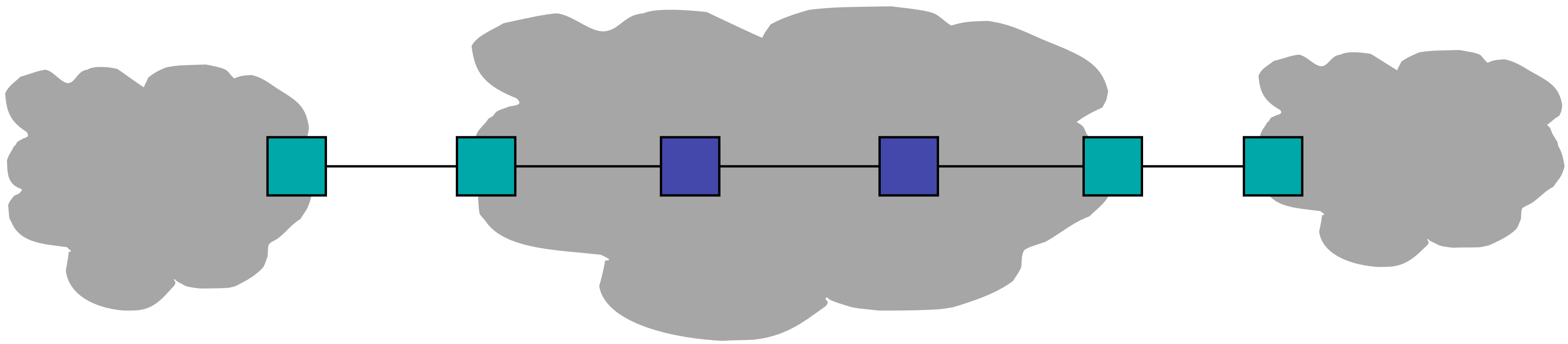
Premium Service Example

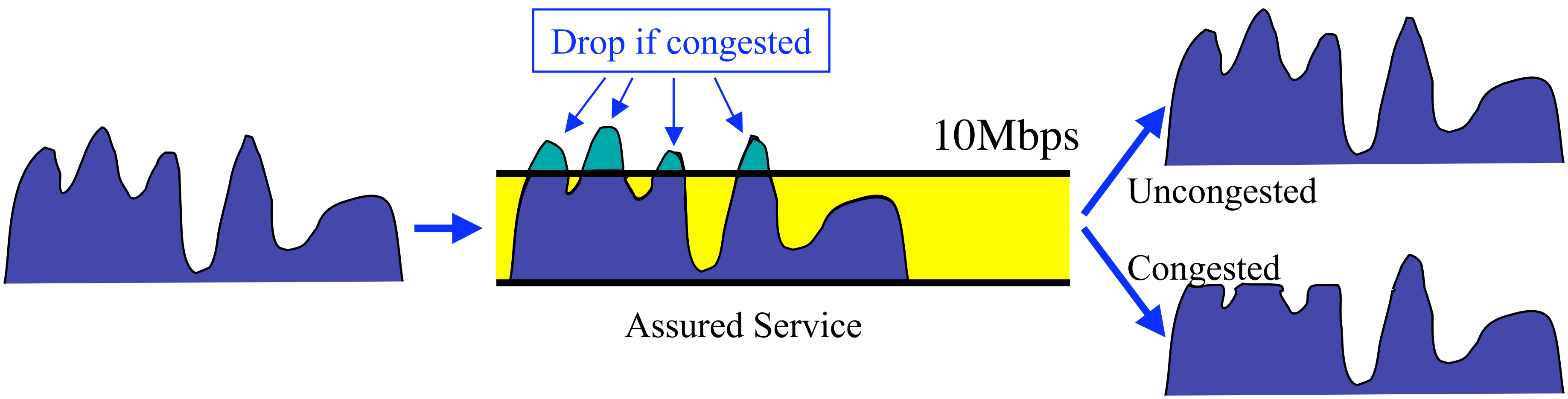




source Gordon Schaffee

Assured Service Example

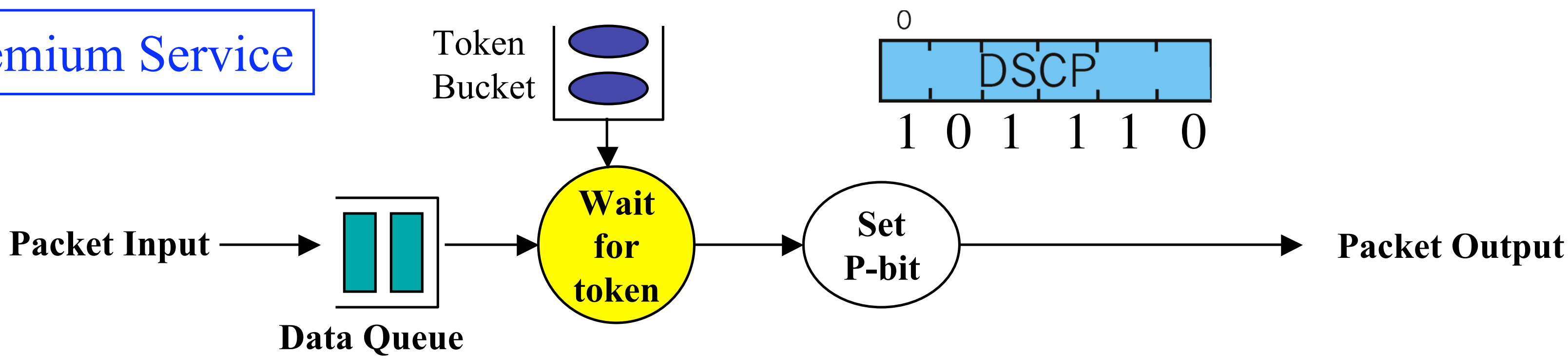




source Gordon Schaffee

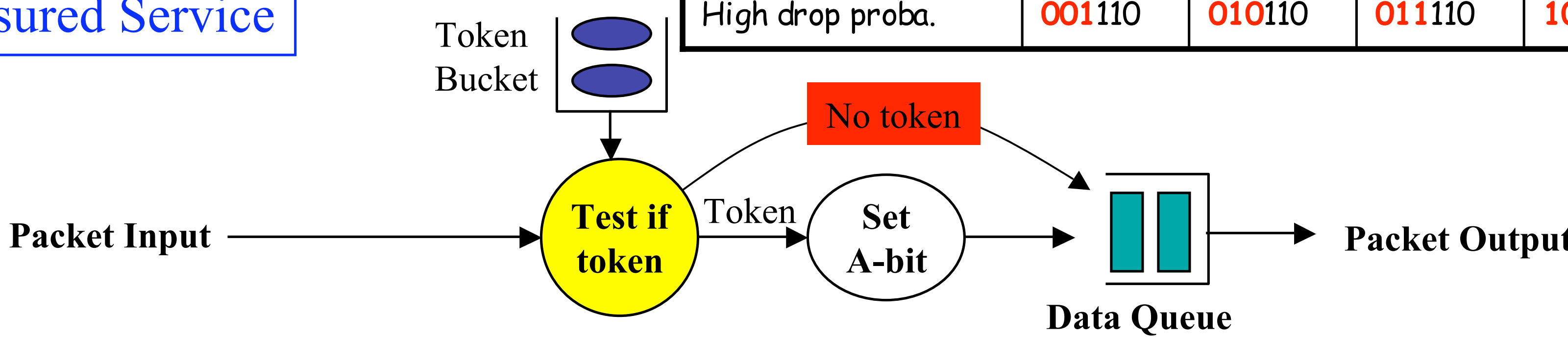
Border Router Functionality

Premium Service



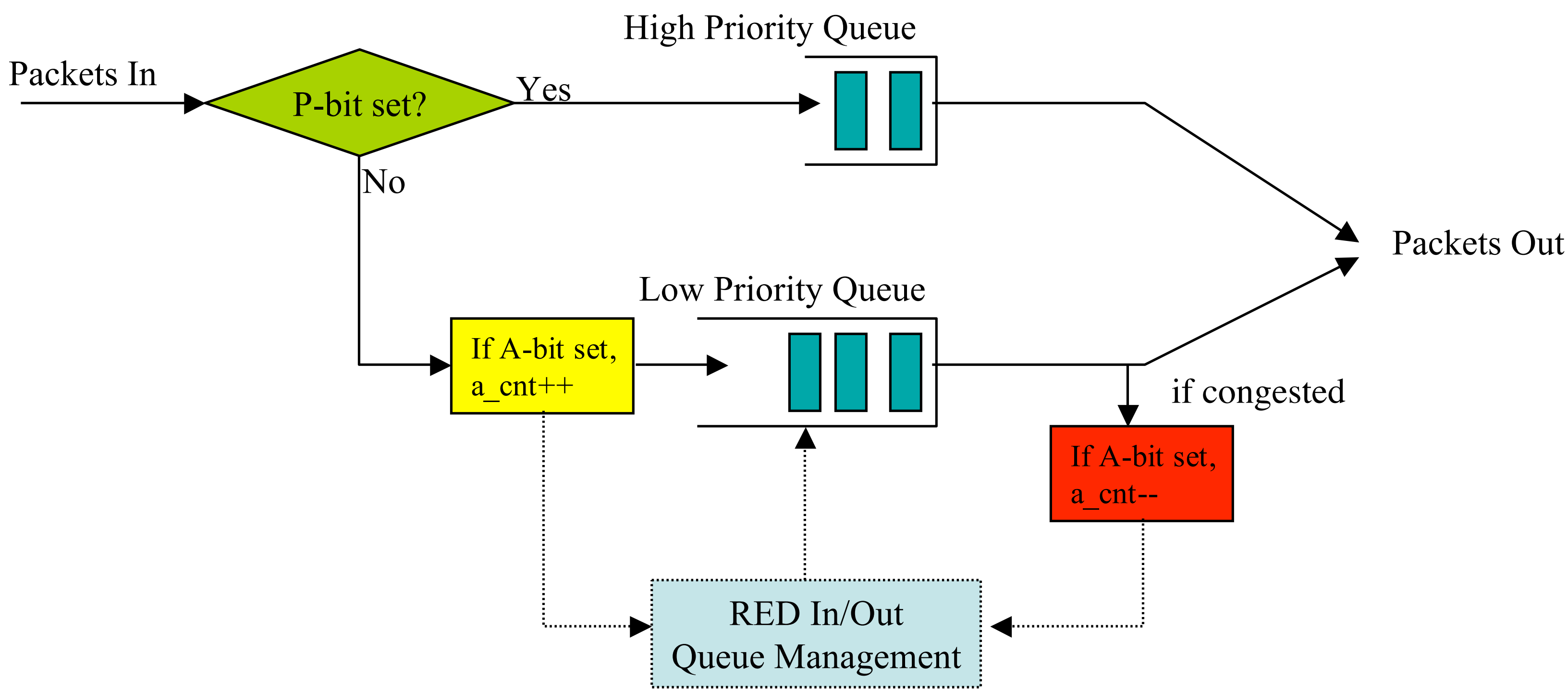
DSCP	Class 1	Class 2	Class 3	Class 4
Low drop probability	001010	010010	011010	100010
Medium drop proba.	001100	010100	011100	100100
High drop proba.	001110	010110	011110	100110

Assured Service



source Gordon Schaffee, modified by C. Pham

Internal Router Functionality



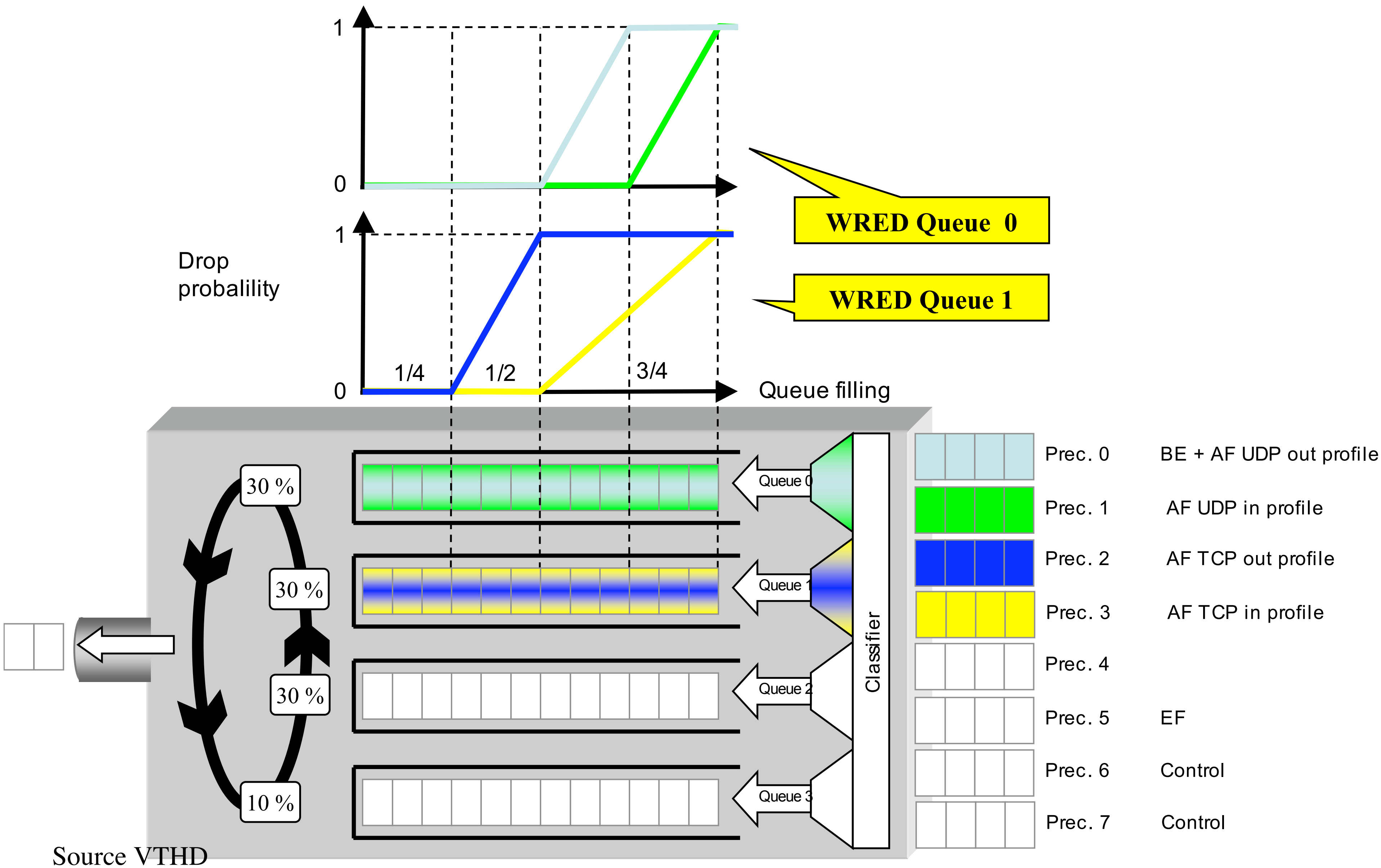
A DSCP codes aggregates, not individual flows

No state in the core

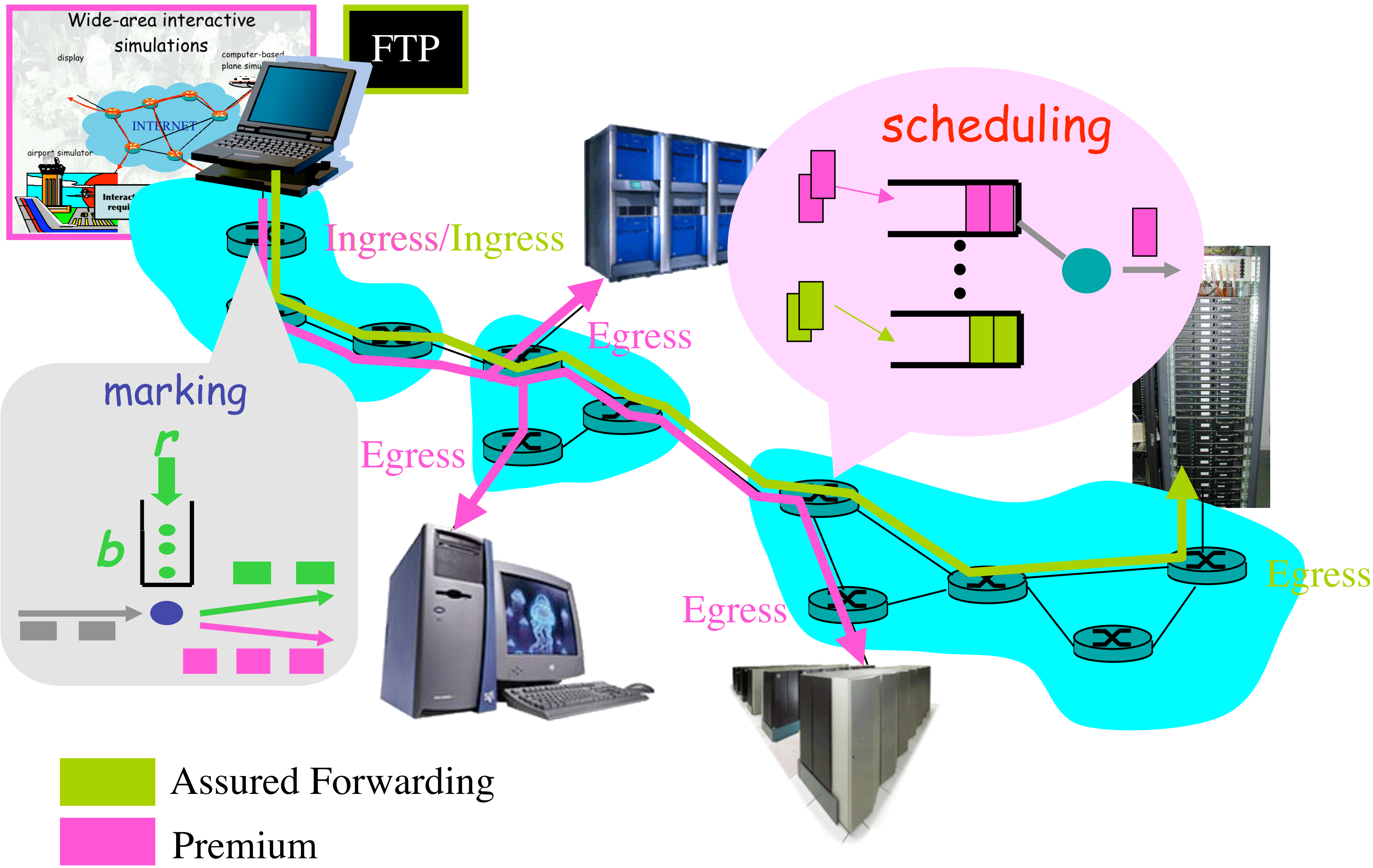
Should scale to millions of flows

source Gordon Schaffee, modified by C. Pham

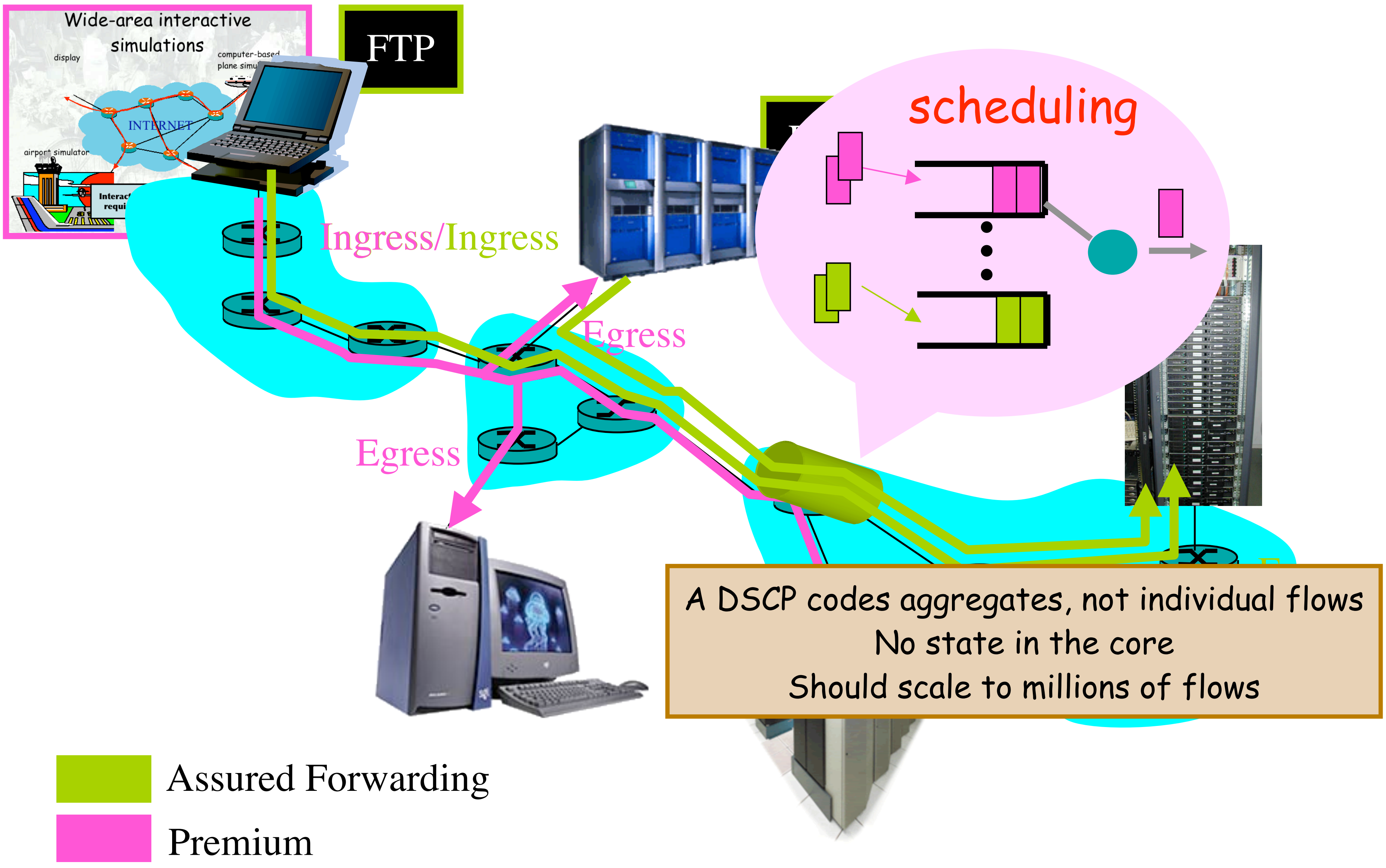
Practical realization



DiffServ for grids



DiffServ for grids (con't)

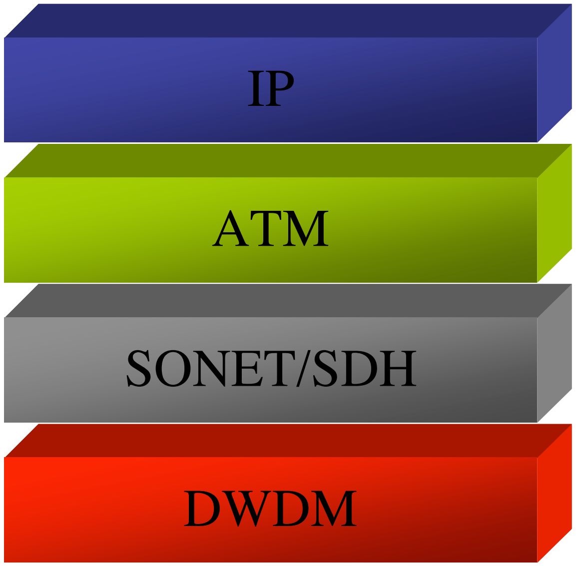


N
E
W

C
H
A
P
T
E
R

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

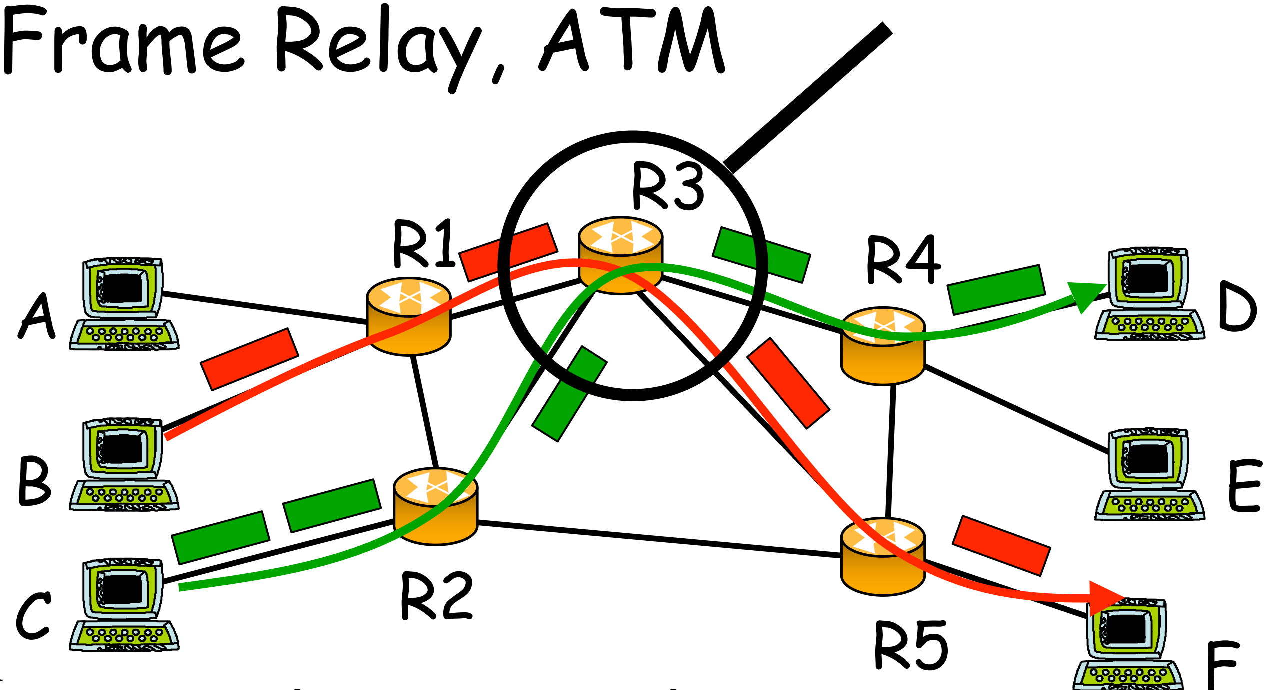


Provider's view

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

Back to virtual circuits

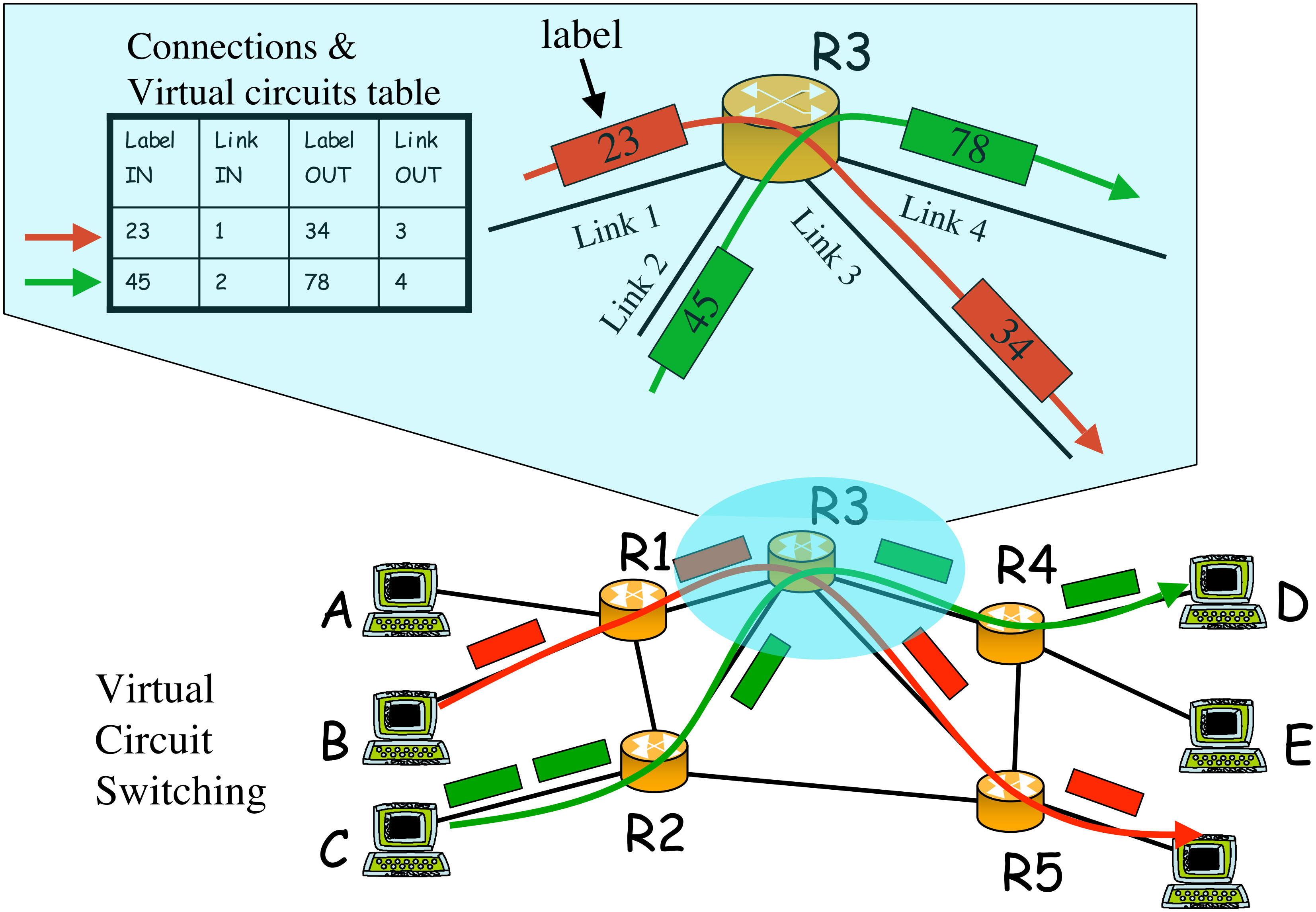
- Virtual circuit refers to a connection oriented network/link layer: e.g. X.25, Frame Relay, ATM



Virtual Circuit Switching: a path is defined for each connection

IP is connectionless!

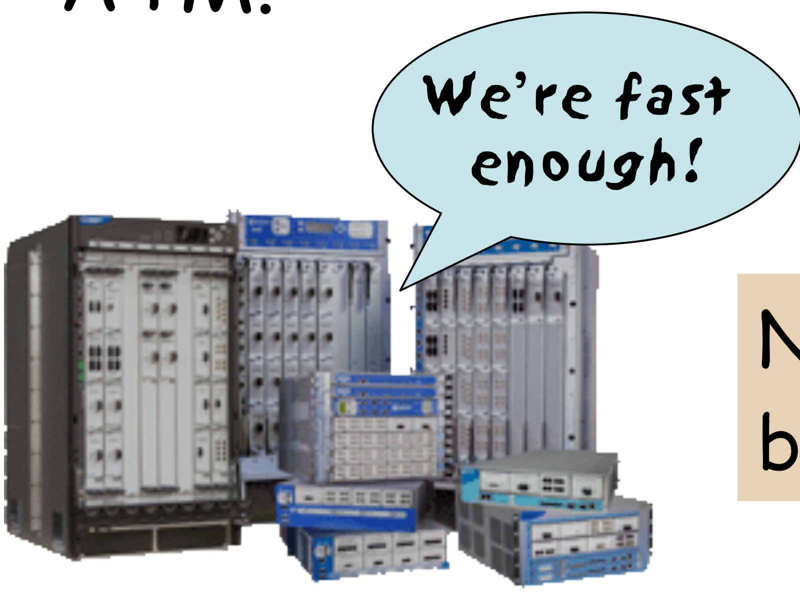
Virtual circuit explained



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

Why virtual circuit?

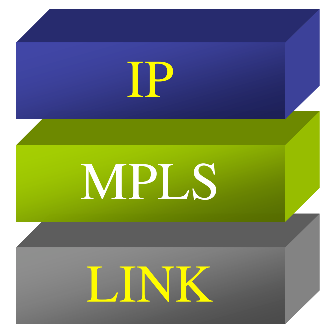
- Initially to speed up router forwarding tasks: X.25, Frame Relay, ATM.



Now: Virtual circuits for bandwidth provisioning!

MPLS

- 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

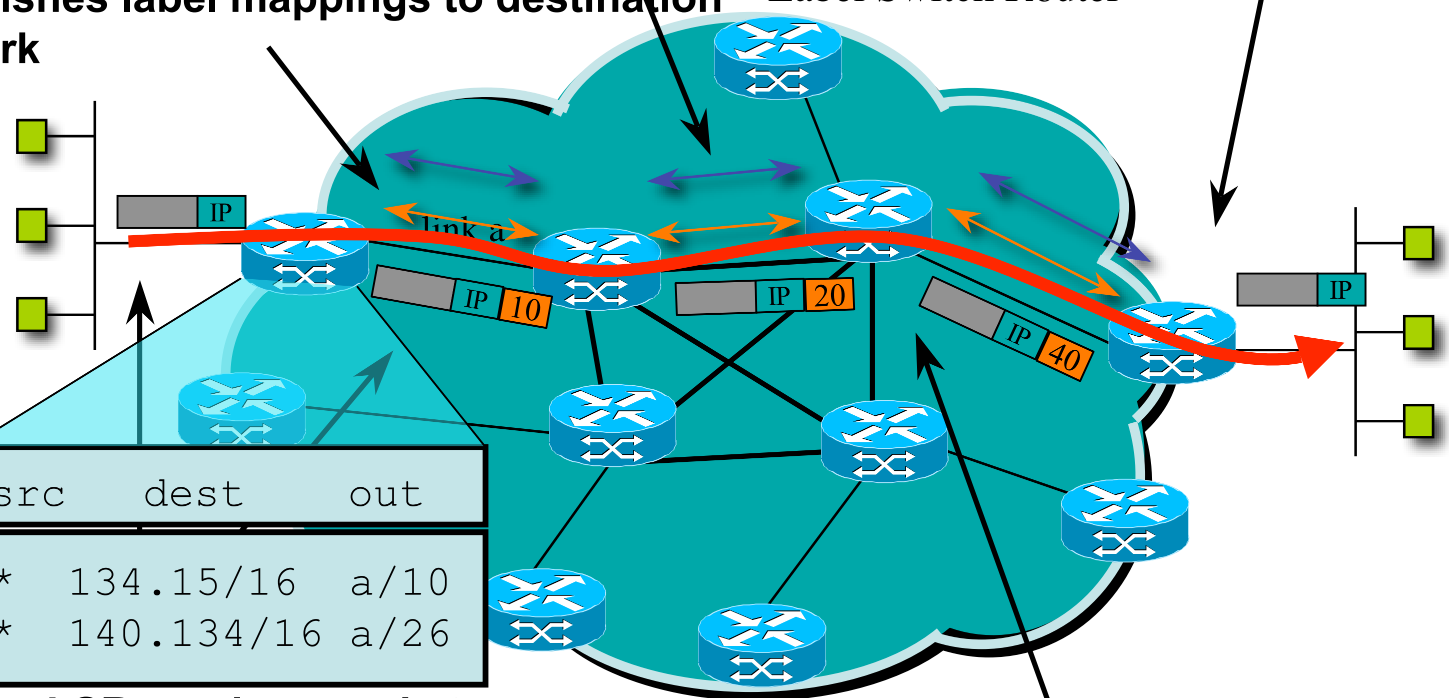
MPLS operation

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

Label Switch Router



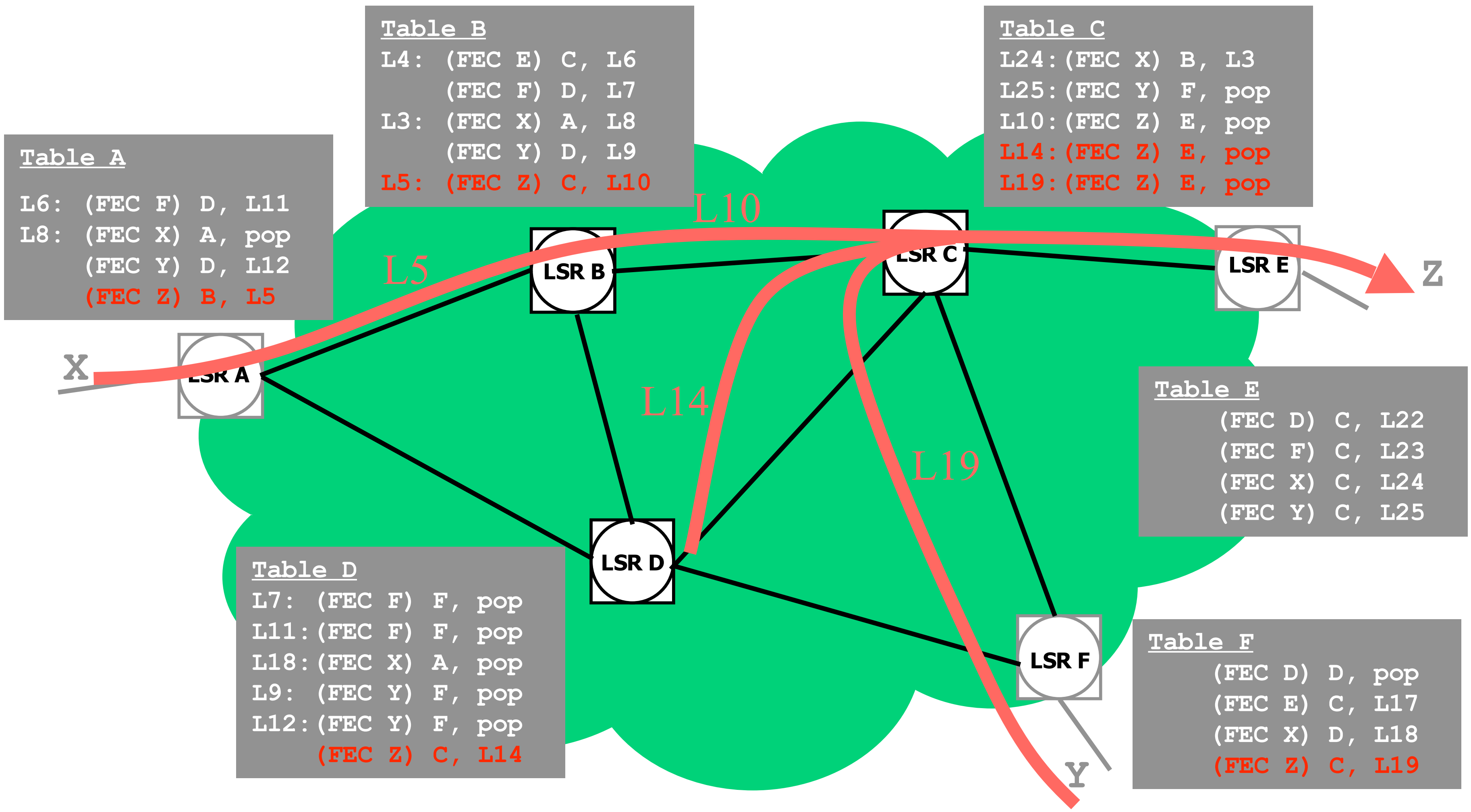
src	dest	out
*	134.15/16	a/10
*	140.134/16	a/26

2. Ingress LSR receives packet and “label”s packets

3. LSR forwards packets using label switching

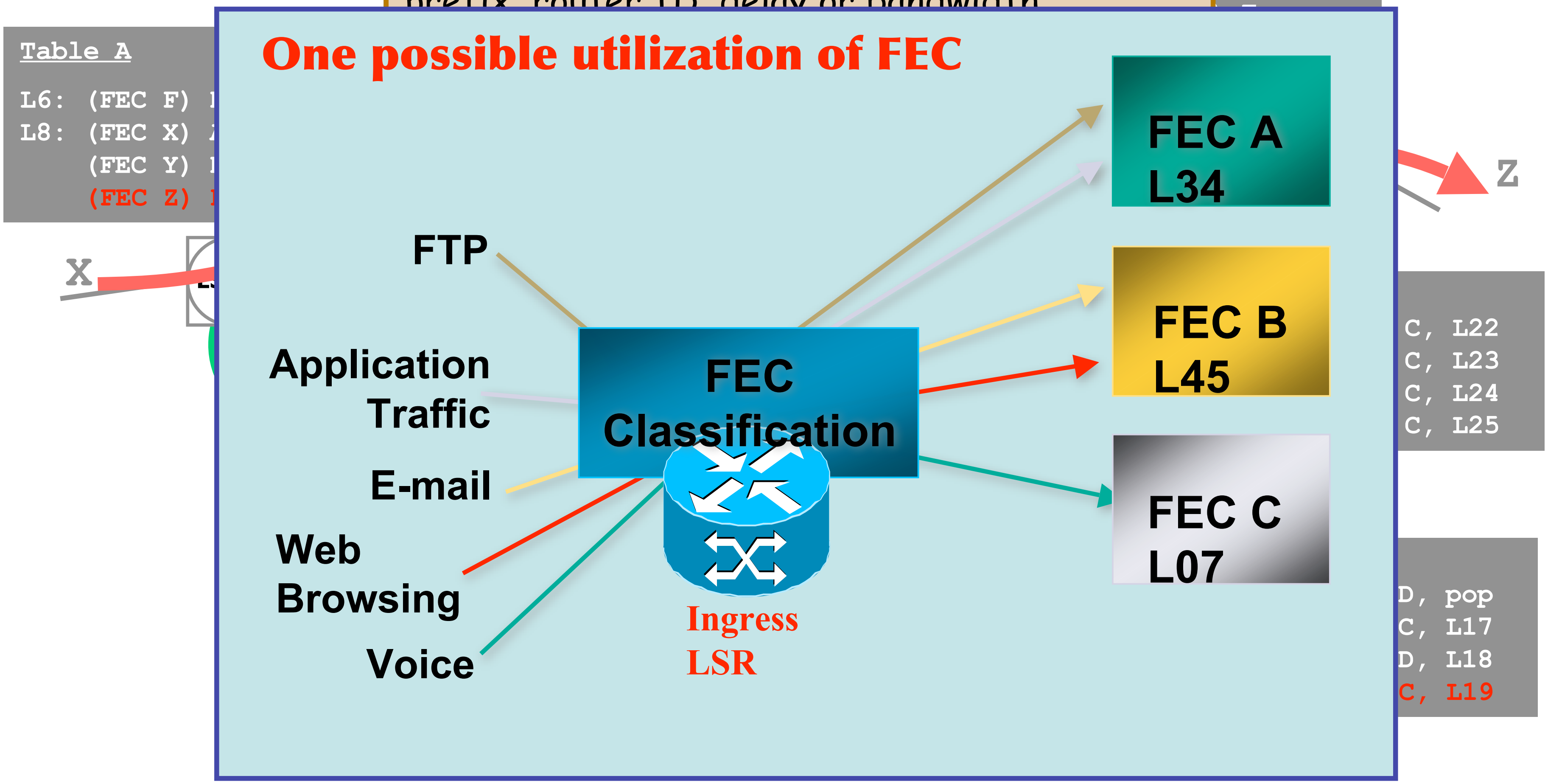
Source Yi Lin, modified C. Pham

Forwarding Equivalent Class: high-level forwarding criteria



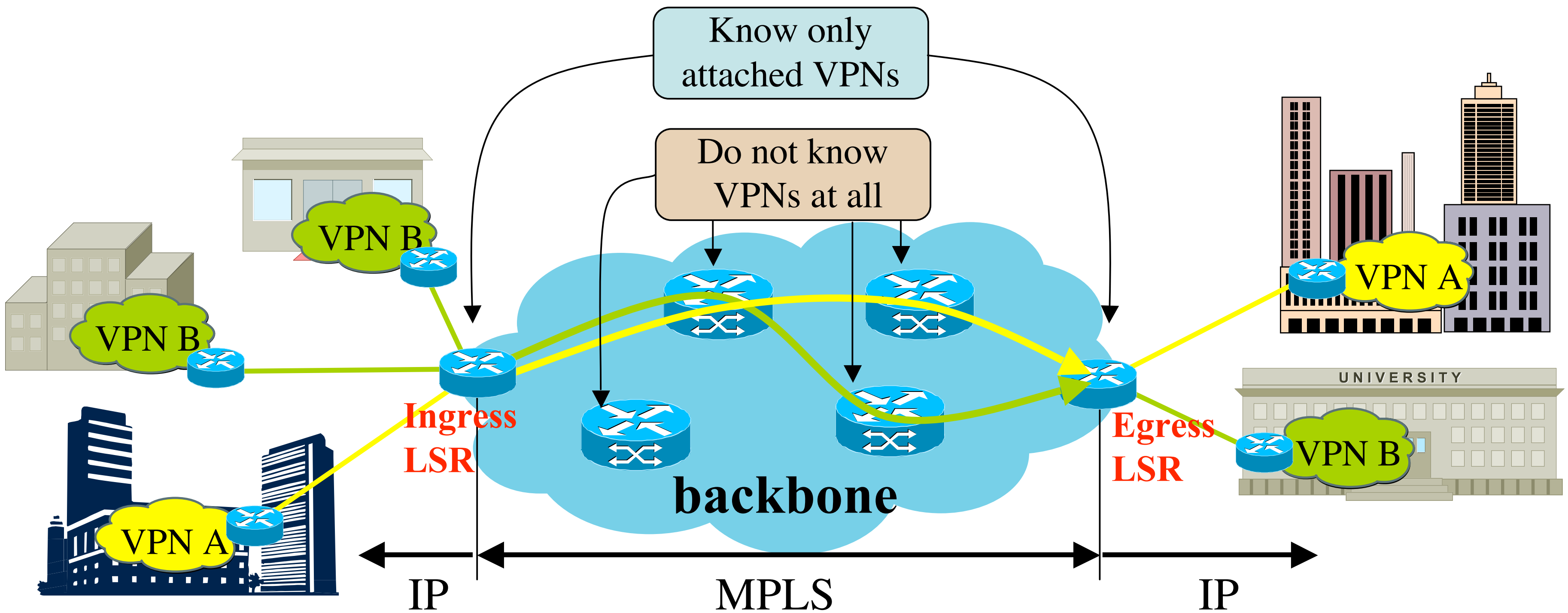
Forwarding Equivalent Class

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

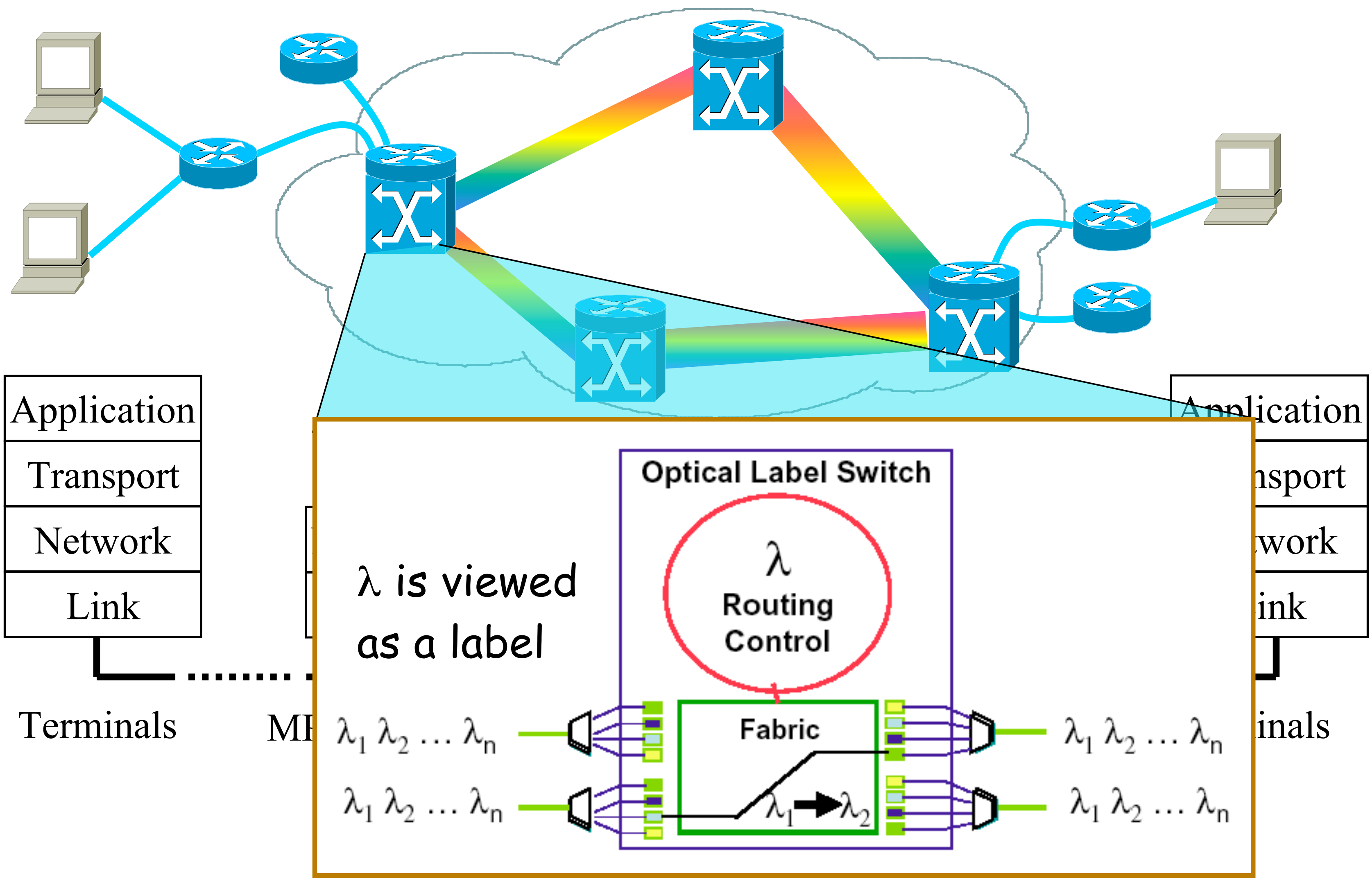


MPLS & VPN

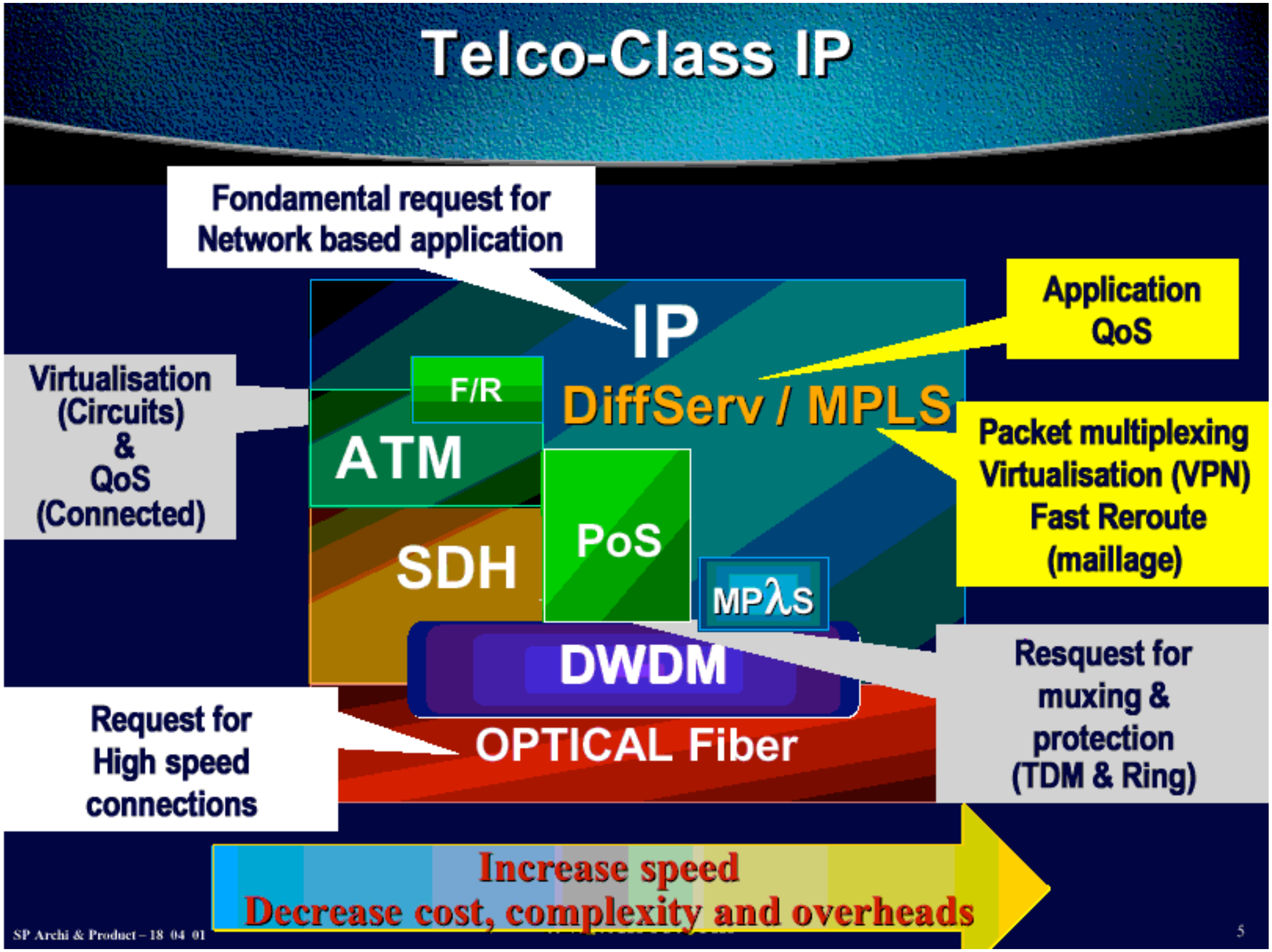
- Virtual Private Networks: build a secure, confidential communication on a public network infrastructure using routing, encryption technologies and controlled accesses
- MPLS reduces VPN complexity by reducing routing information needed at provider's routers



MPλS: MPLS+optical

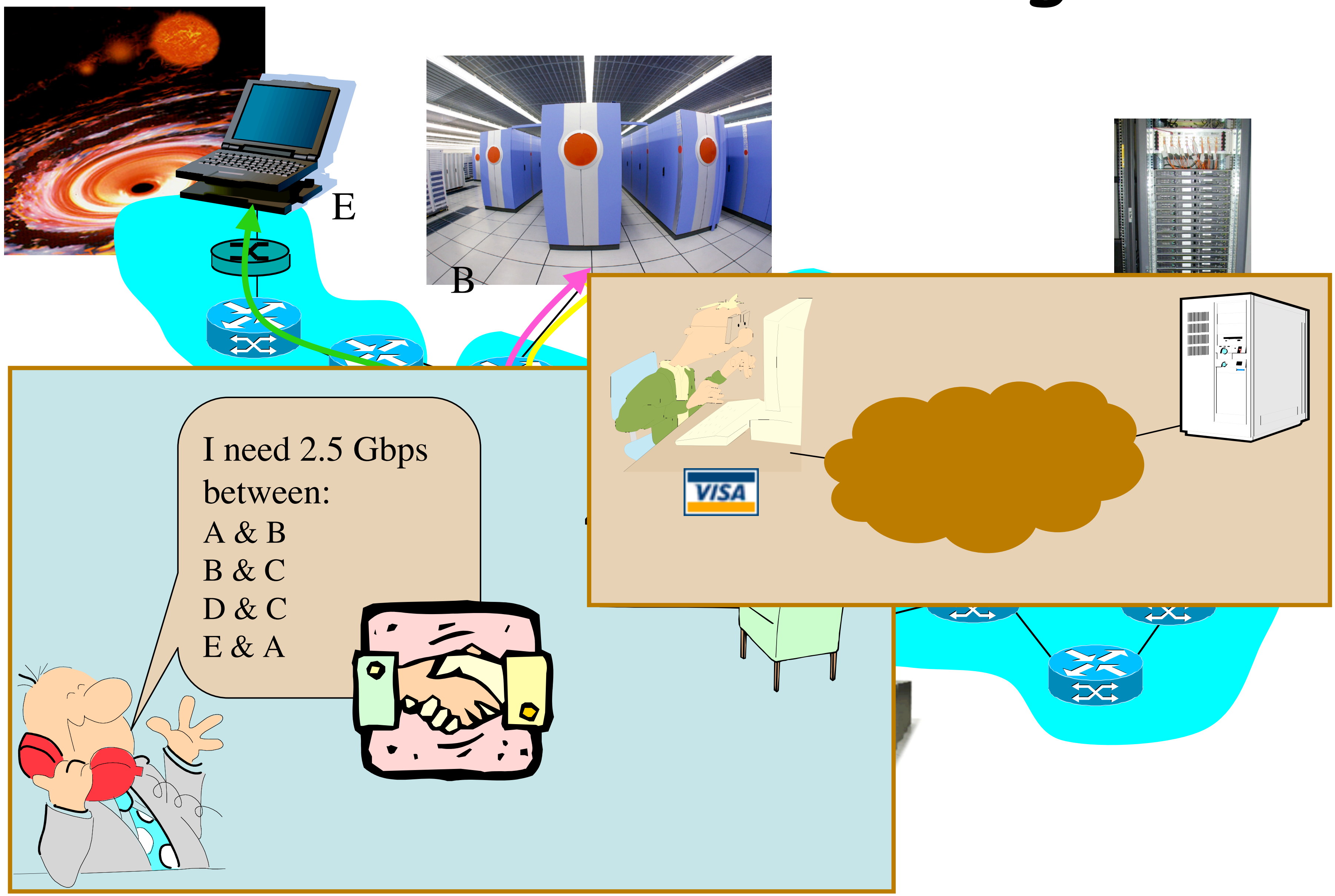


Towards IP/MPLS/DWDM

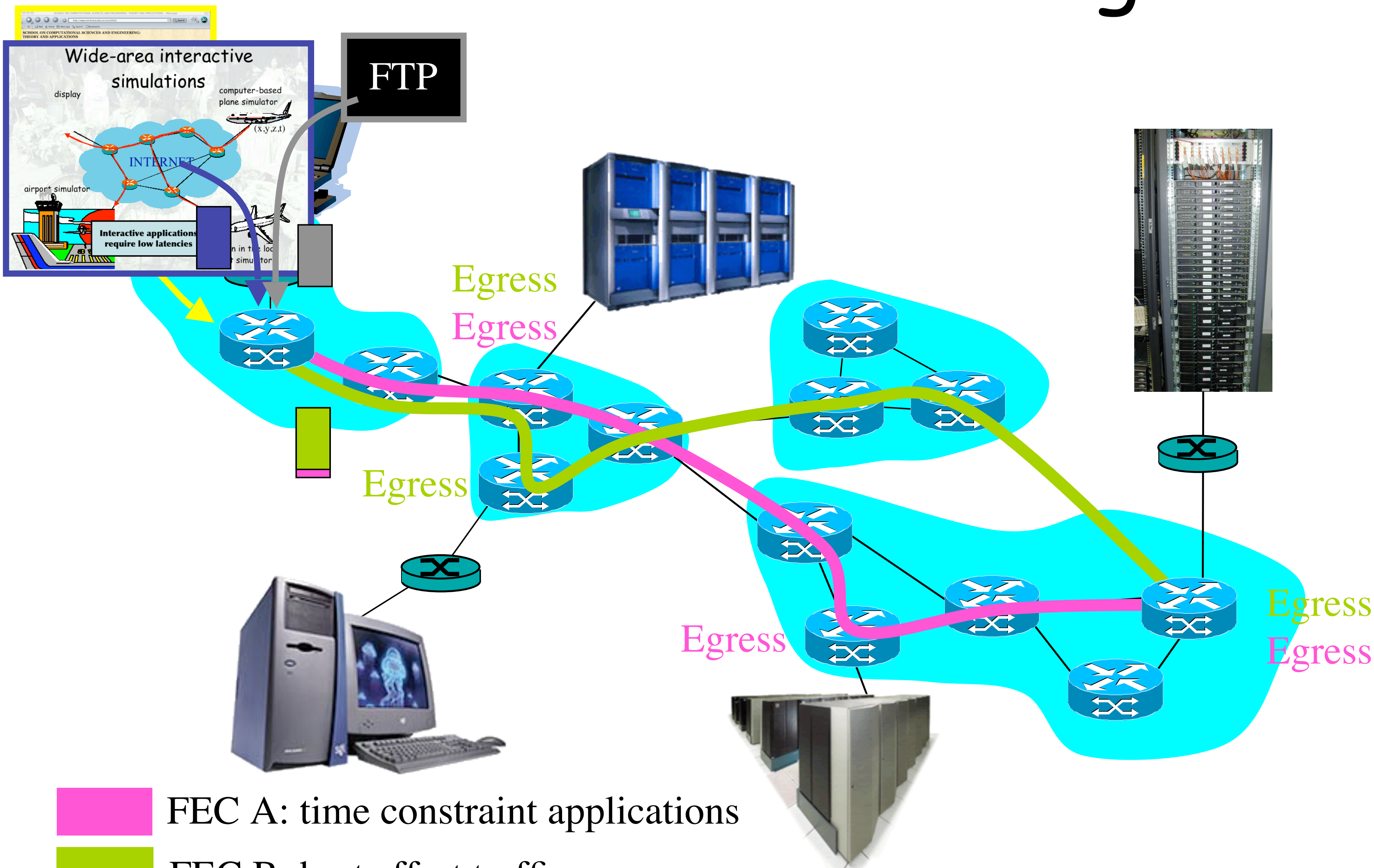


From cisco

Ex: MPLS circuits on grids



Ex: MPLS FEC for the grid

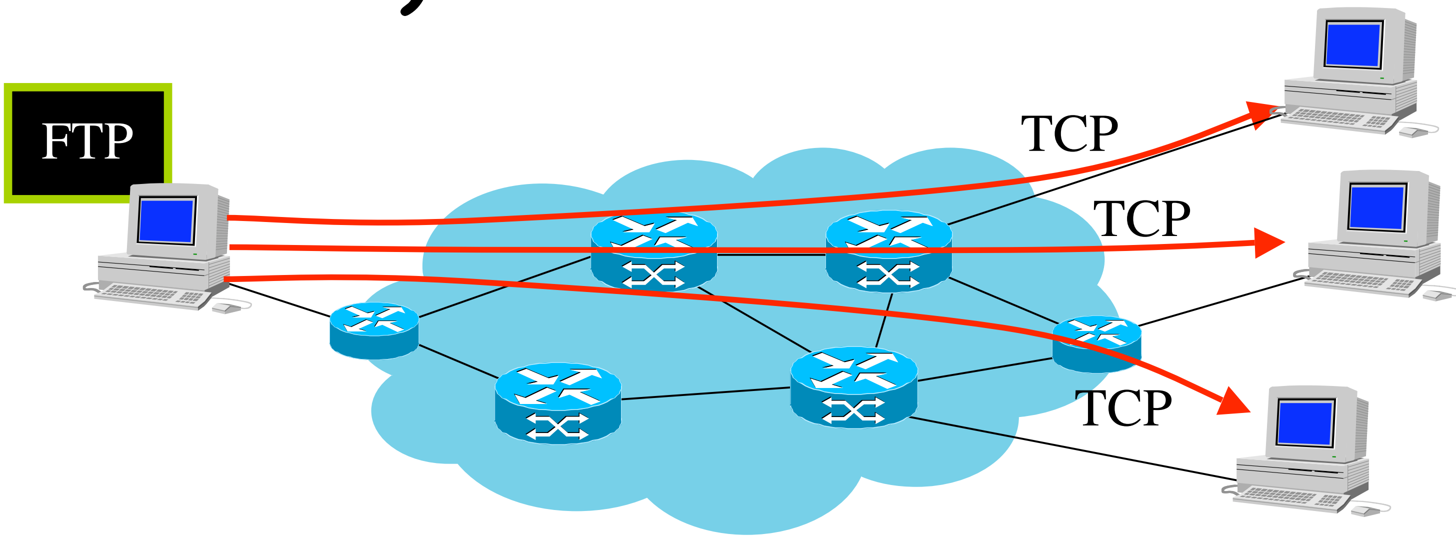


FEC A: time constraint applications

FEC B: best effort traffic

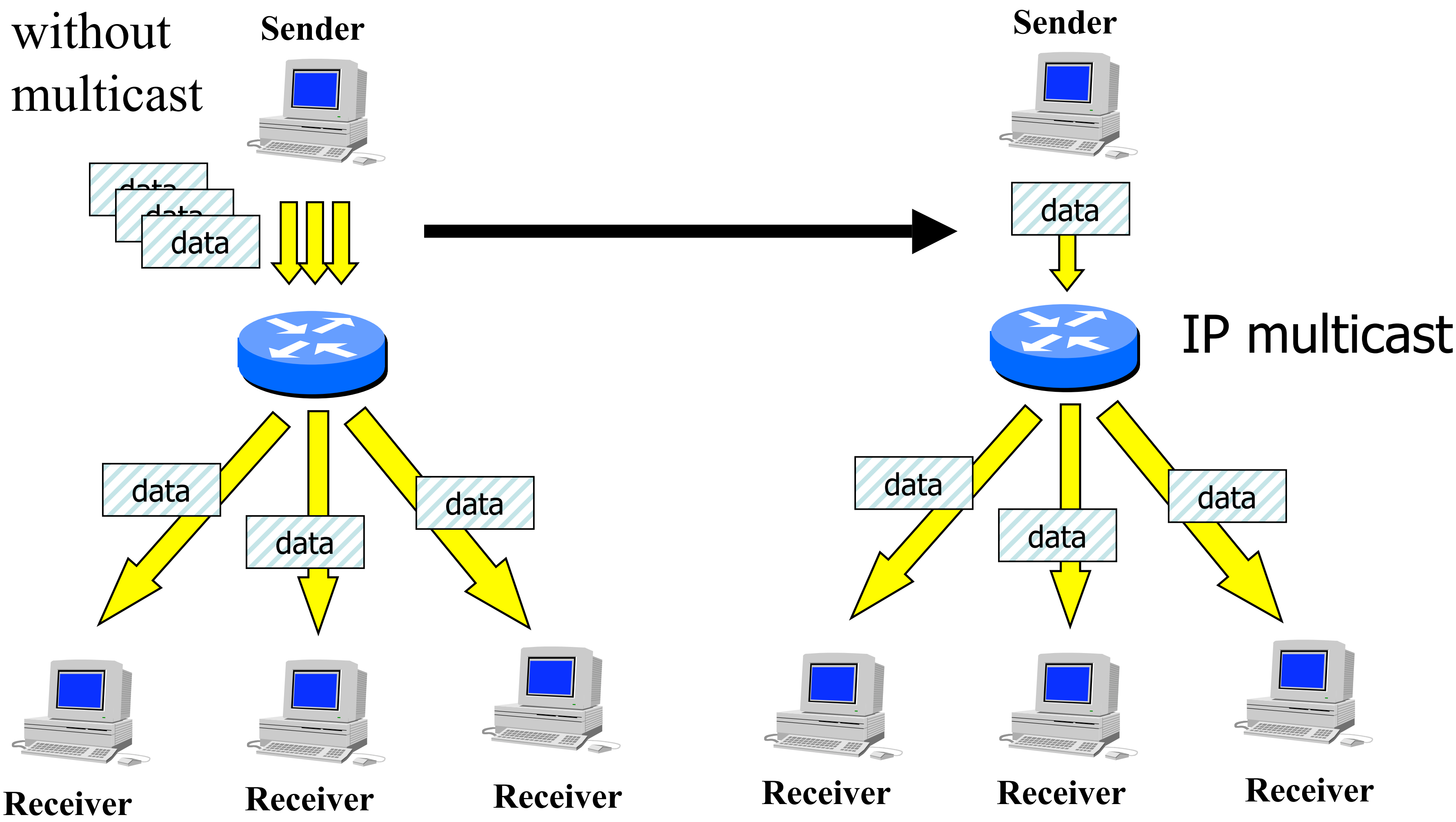
NEW CHAPTER

Unicast, the current (Internet) communication model



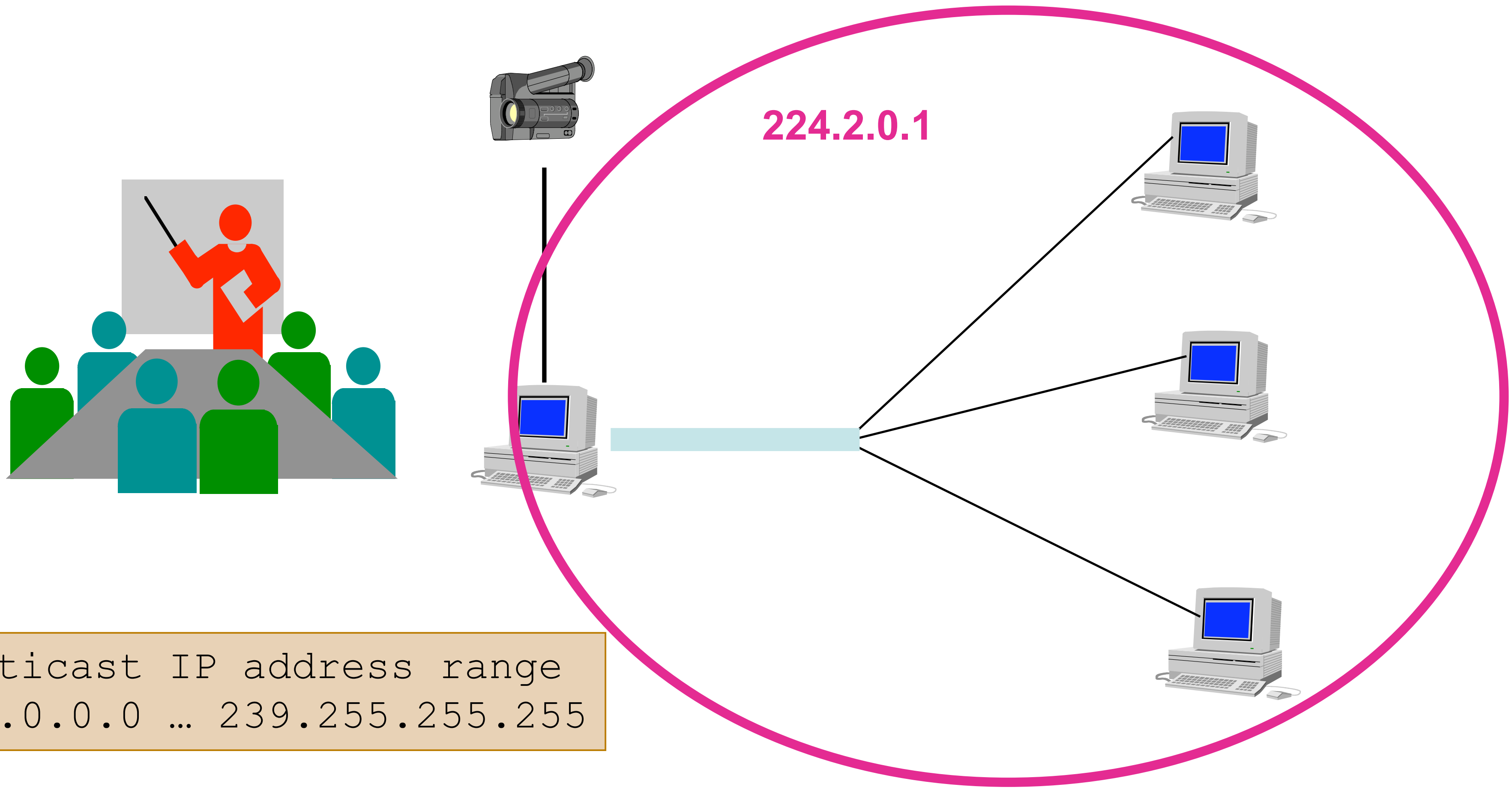
- There are applications that naturally need multi-destination communication model
  - Collaborative works
  - Visio-conferencing
  - Software distribution
  - Video-on-Demand
  - Virtual Reality
  - Distributed Simulation

From unicast to multicast



Multicast in example

The user's perspective

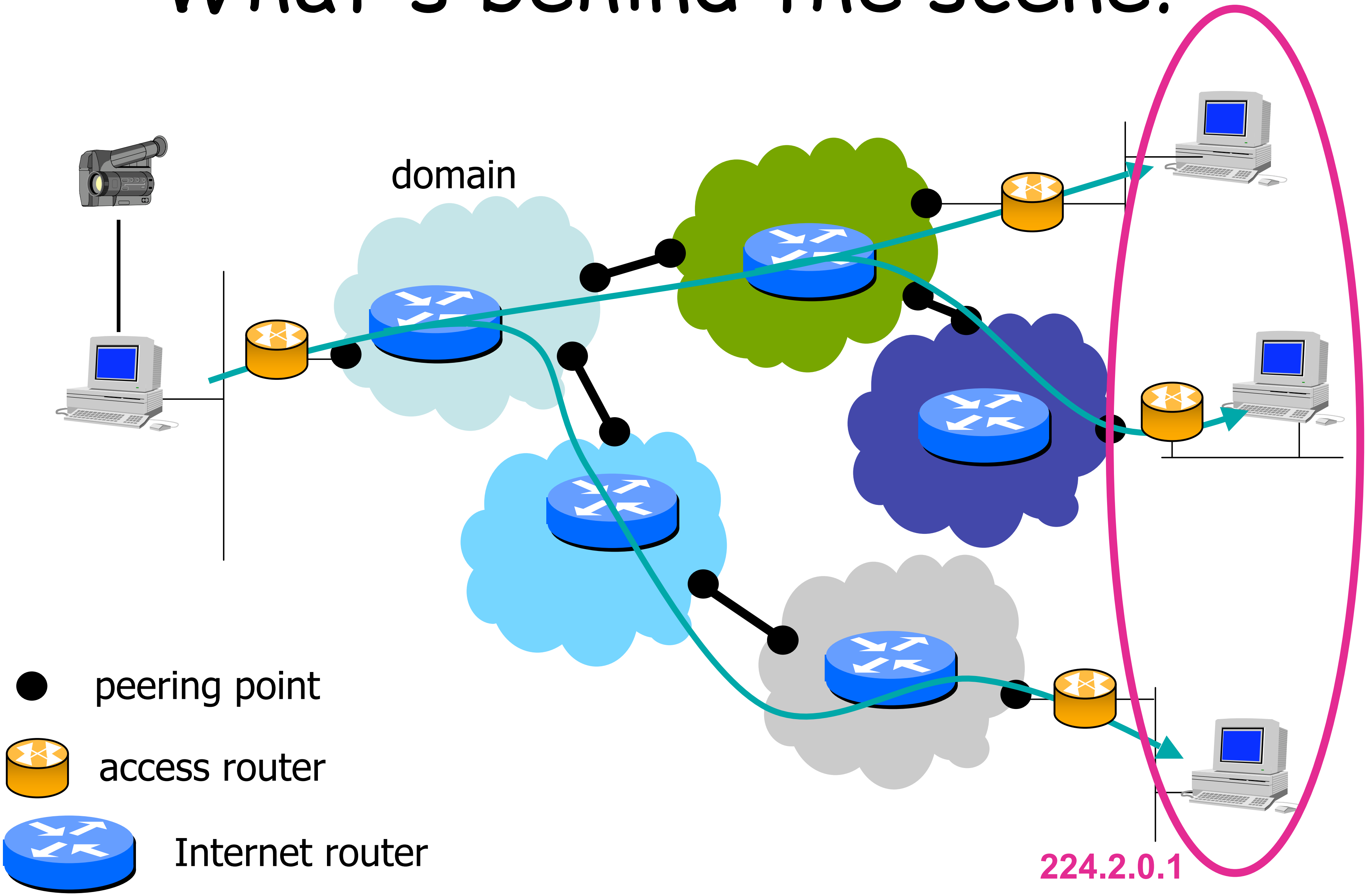


Multicast IP address range
224.0.0.0 … 239.255.255.255

from UREC, http://www.urec.fr

Multicast address group 224.2.0.1

What's behind the scene?



IP multicast TODO list

- ✓ Receivers must be able to subscribe to groups, need group management facilities
- ✓ A communication tree must be built from the source to the receivers
- ✓ Branching points in the tree must keep multicast state information
- ✓ Inter-domain routing must be reconsidered for multicast traffic
- ✓ Need to consider non-multicast clouds

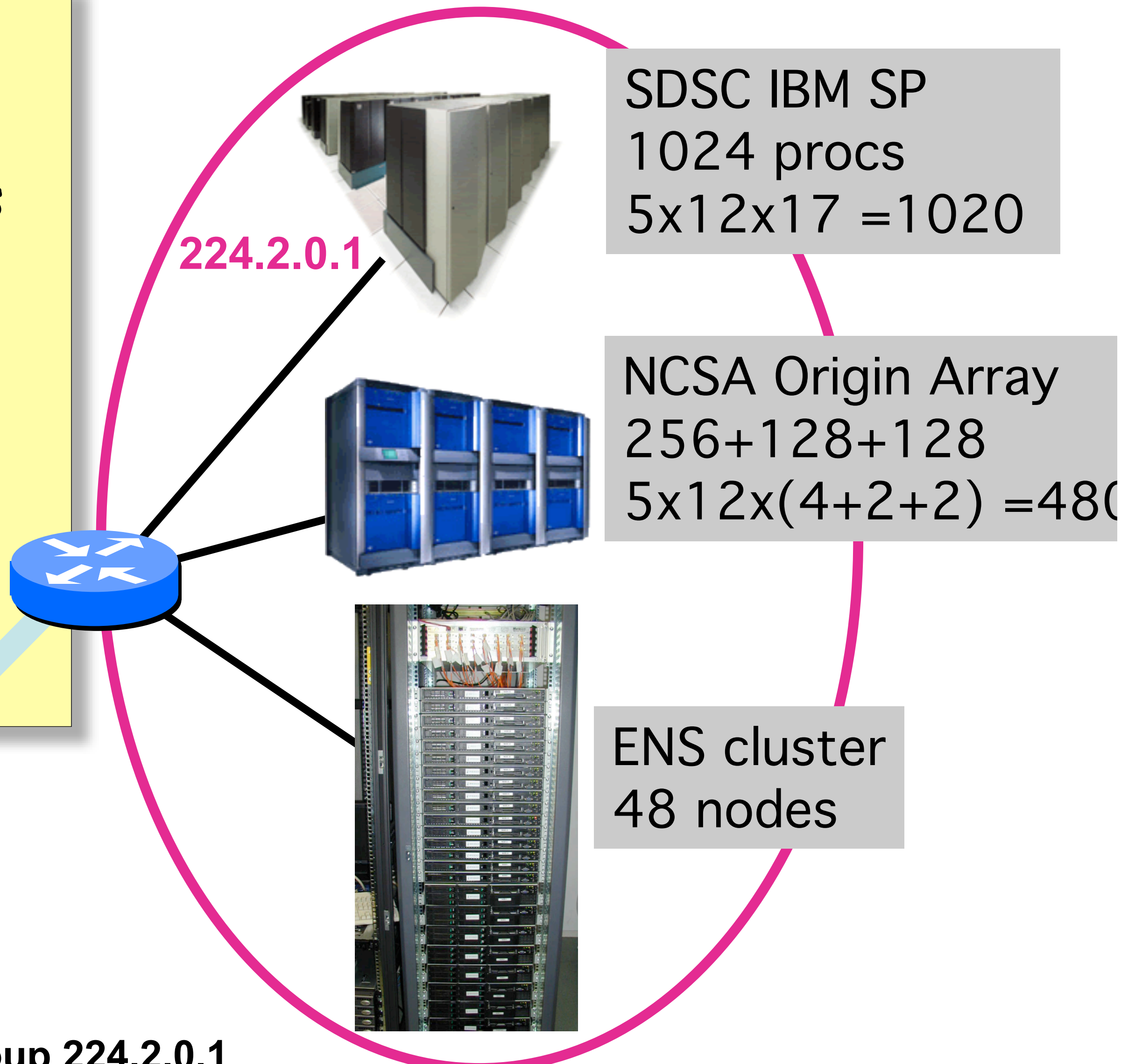
Ex: Reliable multicast on grids

Data replications

Code & data transfers, interactive job submissions

Data communications for distributed applications (collective & gather operations, sync. barrier)

Databases, directories services



Multicast address group 224.2.0.1

Conclusions

- There's a lot more technologies going on that have impact on computational science
  - Pure optical networks, broadband wireless
  - Peer-to-Peer, Overlays
  - Web services…
- The future will be all connected, all IP, anytime, anywhere, for more…

...fun in computational sciences!!