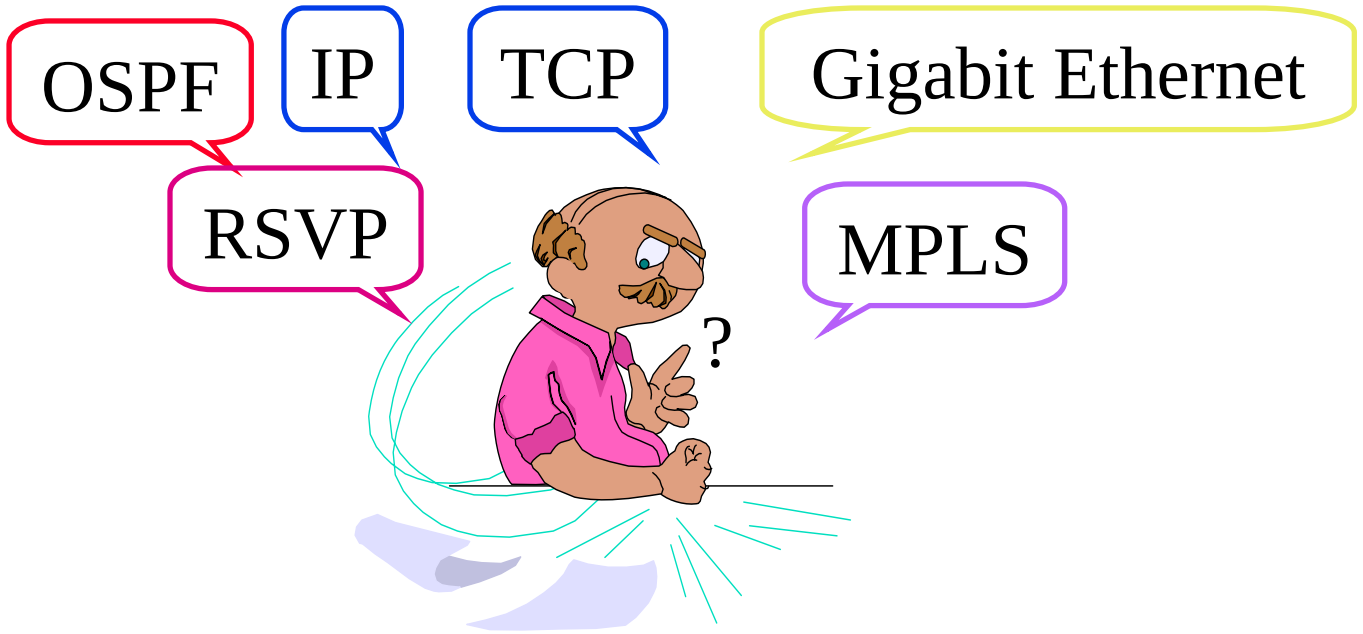
Computer Networking and Internet Protocols: A Comprehensive Introduction



Raj Jain
Professor of Computer Science and Engineering
Washington University in Saint Louis
jain@acm.org
http://www.cse.wustl.edu/~jain/

- IP: Addressing, forwarding, IPv6, TCP
- Ethernet
- Quality of Service (QoS): RSVP
- Multi-protocol Label Switching (MPLS)
- Route Discovery Protocols : RIP, OSPF, BGP
- Wireless networking
- Optical networking

1. Introduction to TCP/IP

- TCP/IP Reference Model
- Internet Protocol (IP)
- Forwarding an IP Datagram
- IP Datagram Format
- IPv6 Enhancements
- Domain Name Service
- TCP: Key Features
- User Datagram Protocol (UDP)

2. Ethernet

- Carrier Sense Multiple Access with Collision Detection (CSMA/CD)
- IEEE 802 Address Format
- Interconnection Devices
- Distance-B/W Principle
- Gigabit Ethernet
- Spanning Tree
- 10Gbps Ethernet PHYs
- Metro Ethernet Services

3. Quality of Service (QoS)

- ATM QoS and Issues
- Integrated Services and RSVP
- Differentiated Services:
  Expedited and Assured Forwarding
- Subnet Bandwidth Manager (SBM)
- COPS Protocol for Policy
- IEEE 802.1D Model

4. MPLS

- Routing vs Switching
- Label Stacks
- Label Distribution Protocol (LDP)
- RSVP Extensions
- Traffic Engineering
- Traffic Trunks
- Traffic Engineering Extensions to OSPF and IS-IS

5. Routing Protocols

- Building Routing Tables
- Routing Information Protocol Version 1 (RIP V1)
- RIP V2
- OSPF
- BGP and IDRP.

6. Wireless Networks

- Recent advances in wireless PHY
- WiMAX Broadband Wireless Access
- Cellular Telephony Generations
- WiMAX vs LTE
- 4G: IMT-Advanced
- 700 MHz

7. Optical Networks

- Recent DWDM Records
- OEO vs OOO Switches
- More Wavelengths
- Ultra-Long Haul Transmission
- Passive Optical Networks
- IP over DWDM: MPλS, GMPLS
- Free Space Optical Comm
- Optical Packet Switching

Day 1: Schedule (Tentative)

- 10:00-10:15 Course Introduction
- 10:15-11:30 Internet Protocol (IP), IPv6
- 11:30-11:45 *Coffee Break*
- 11:45-1:15 DNS, TCP
- 1:15-2:00 *Lunch Break*
- 2:00-3:15 Metro Ethernet
- 3:15-3:30 *Coffee Break*
- 3:30-5:00 Quality of Service

Day 2: Schedule (Tentative)

- 10:00-11:00 MPLS, MPLS-TE
- 11:00-11:15 *Coffee Break*
- 11:15-12:15 Routing Protocols
- 12:15-1:00 *Lunch Break*
- 1:00-2:15 Wireless Networking
- 2:15-2:30 *Coffee Break*
- 2:30-4:00 Optical Networking

Pre-Test

- Check if you know the difference between:
  - Private addresses and public addresses
  - Class C vs Class A addresses
  - Extension header vs base header
  - Distance vector vs link state routing
  - Inter-domain vs intra-domain routing
  - Universal vs multicast bit
  - Spanning tree vs IS-IS
  - UBR vs ABR
  - DiffServ vs IntServ
  - RSVP vs LDP
  - CDMA vs OFDMA
  - OOO vs OEO optical switching
  - MPLS vs GMPLS
  - Routing vs switching

Pre-Test (Cont)

- If you checked more than 7 items, you may not gain much from this course.
- If you checked only a few or none, don't worry. This course will cover all this and much more.

Disclaimers

- This course covers a lot of topics
- These topics are normally taught in 3 quarter-courses
- Fundamental and basics will be covered
- You will need to read RFC's for detailed info
- This course has been designed specifically for you. Please feel free to ask questions, make comments, agree or disagree.
- More discussion ⇒ More relevant topics

Student Questionnaire

- Name (Optional): _____
- Computer networking courses taken:

  _____

- Telecom Networking background:

  _____

- What do you want covered in this course:

  _____

  _____

  _____

  _____

Introduction to TCP/IP

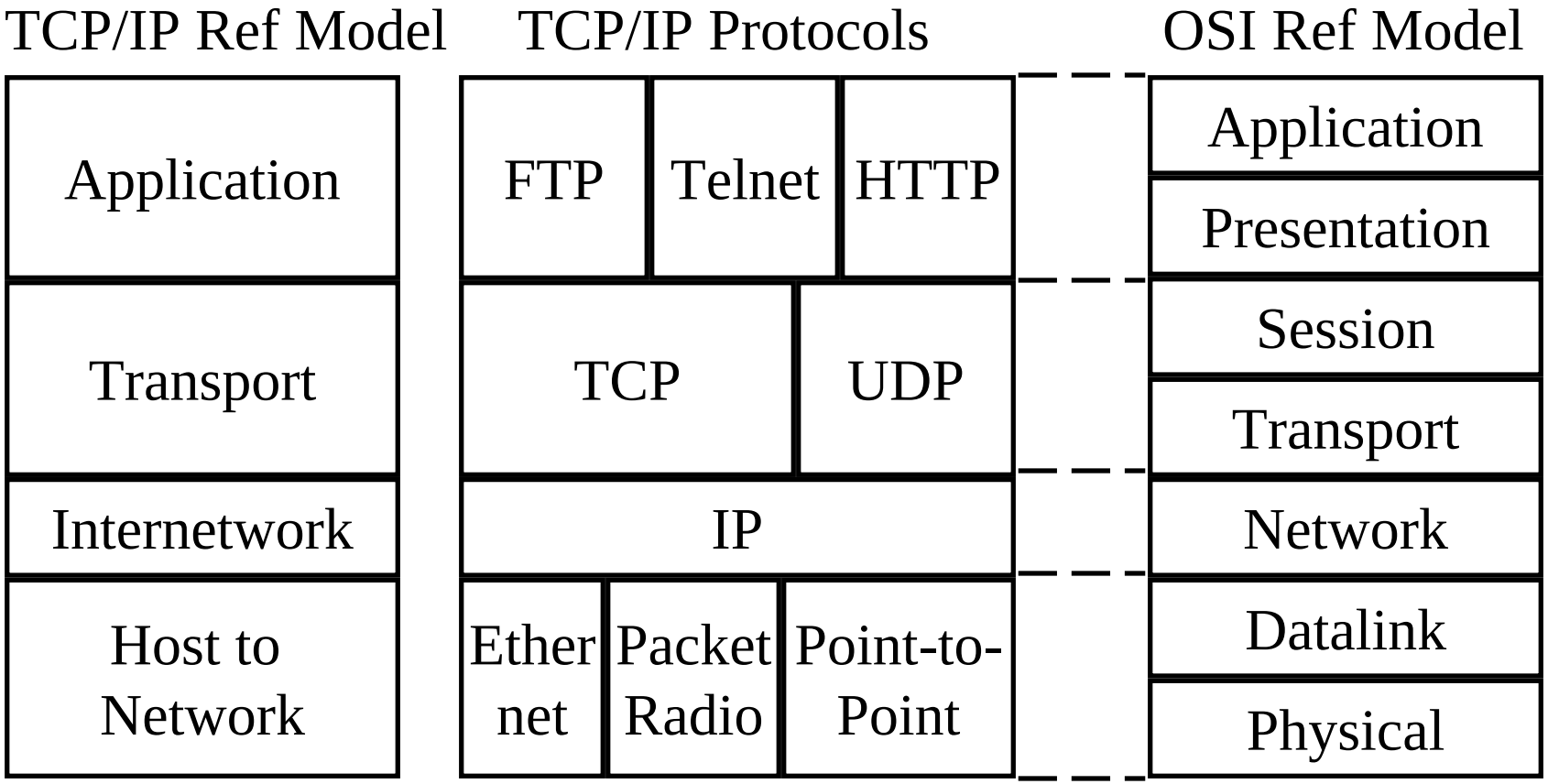
Raj Jain
Professor of Computer Science and Engineering
Washington University in Saint Louis
Saint Louis, MO, USA
jain@acm.org
http://www.cse.wustl.edu/~jain/

Overview

1. TCP/IP Reference Model
2. Internet Protocol (IP)
3. Forwarding an IP Datagram
4. IP Datagram Format
5. IPv6 Enhancements
6. Domain Name Service
7. TCP: Key Features
8. User Datagram Protocol (UDP)

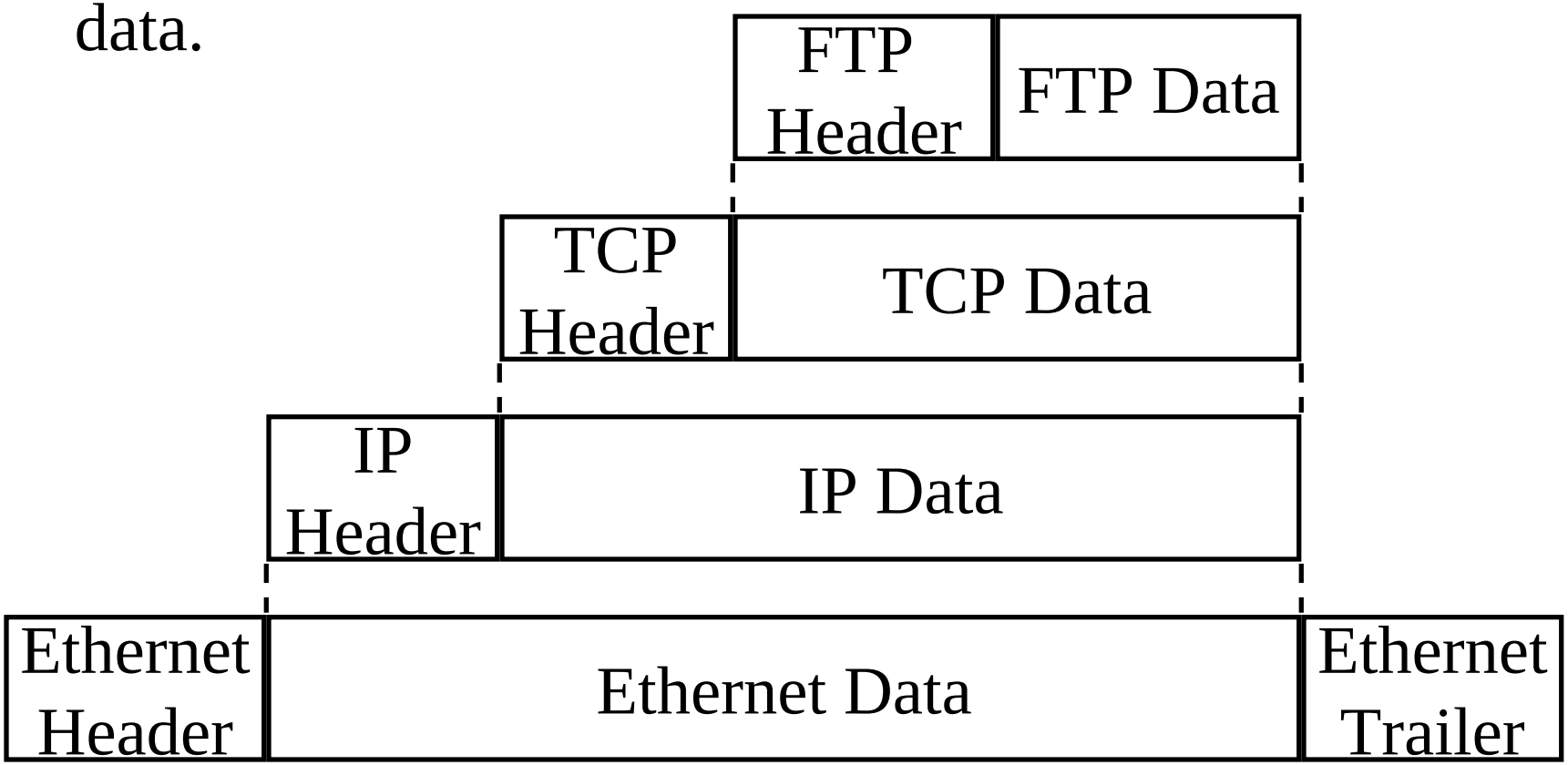
TCP/IP Reference Model

- TCP = Transport Control Protocol
- IP = Internet Protocol (Routing)



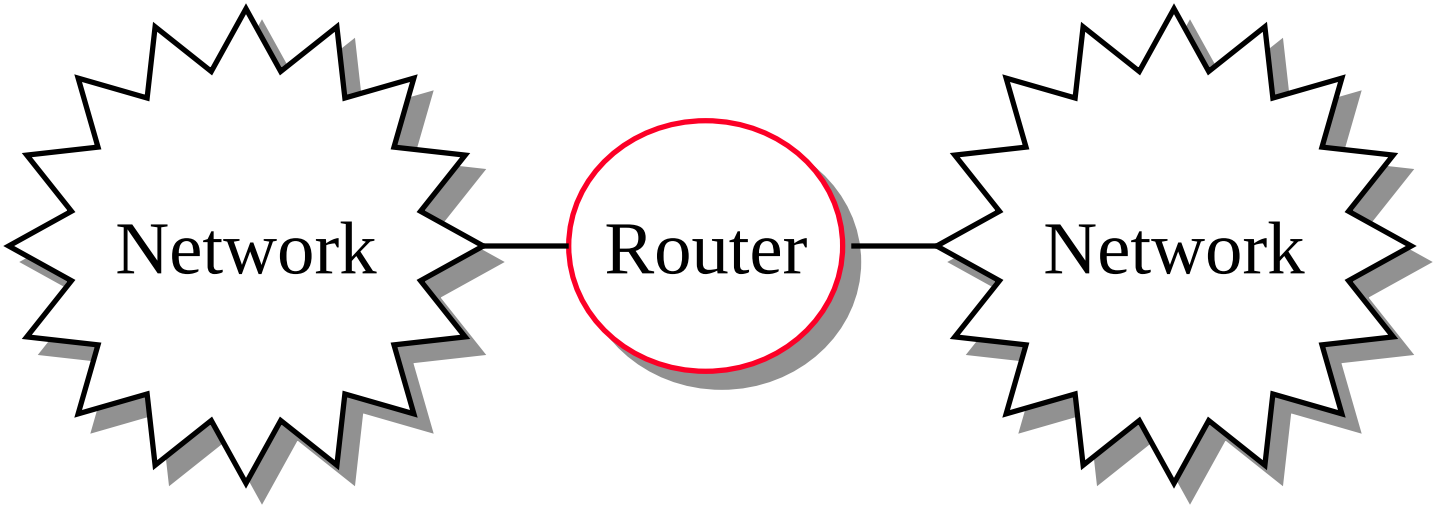
Layered Packet Format

- Nth layer control info is passed as N-1th layer data.



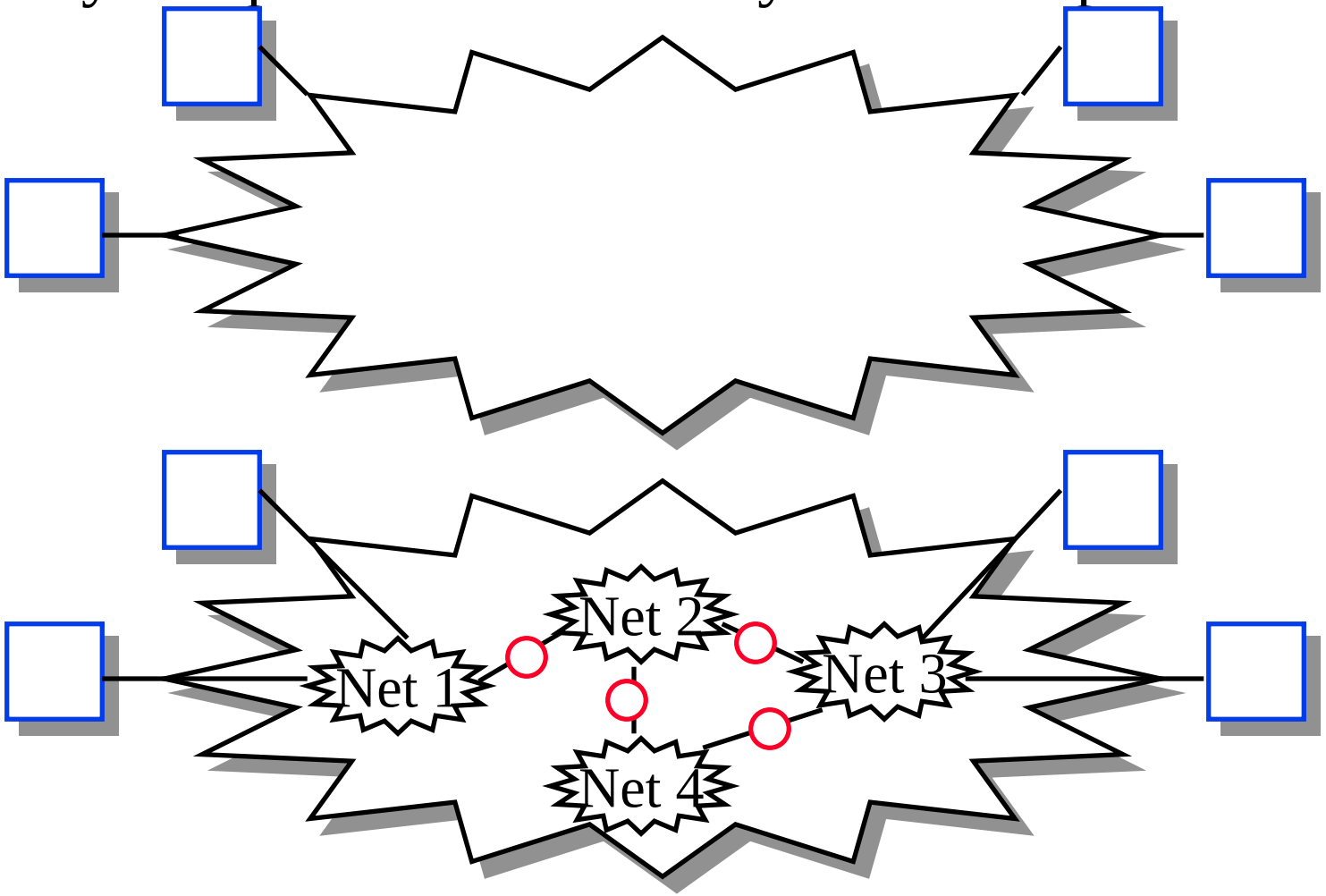
Internetworking

- Inter-network = Collection of networks
  Connected via routers



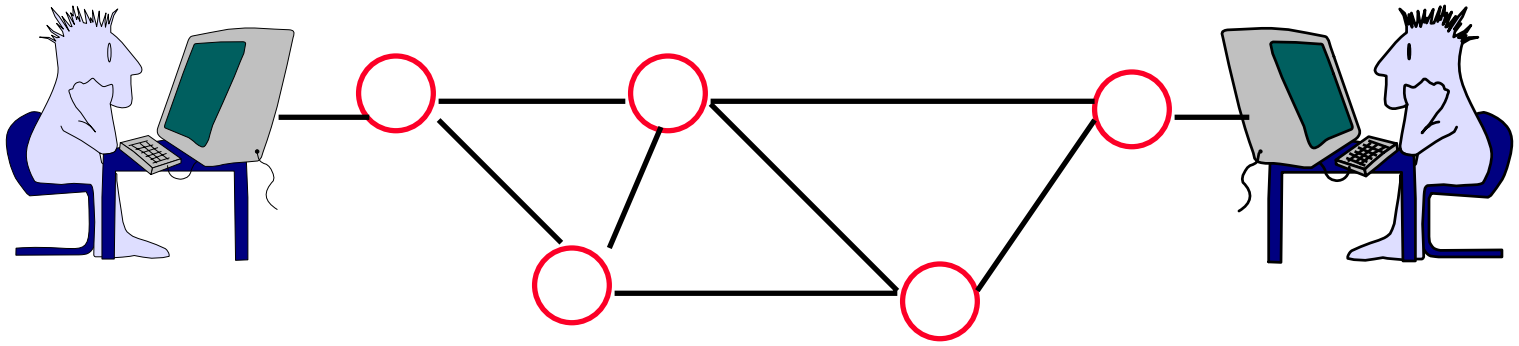
Internet = Collection of Networks

- Any computer can talk to any other computer



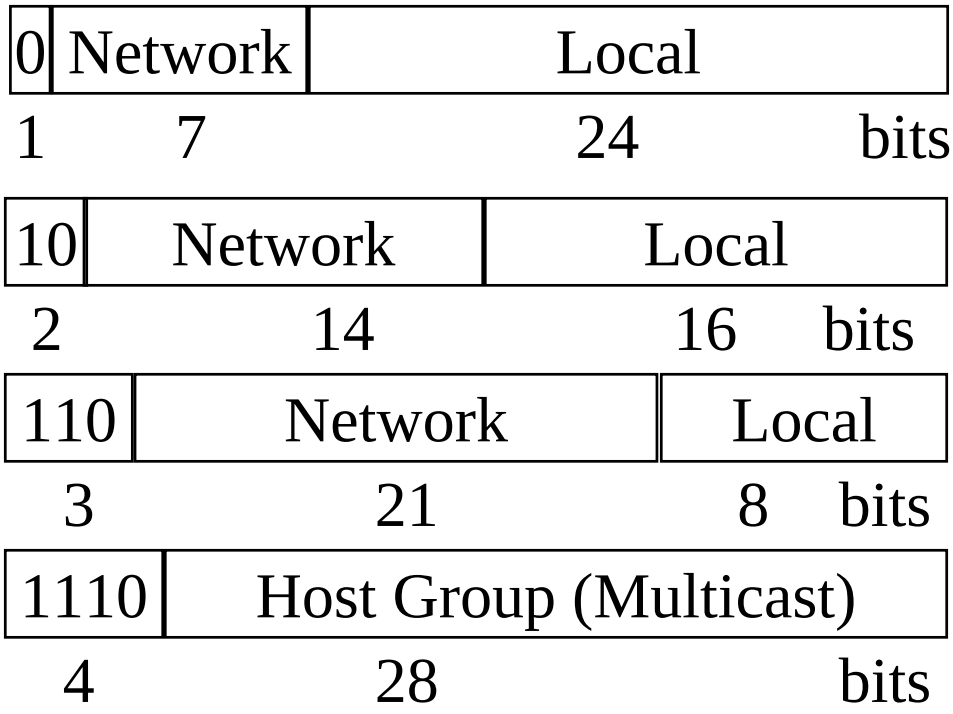
Internet Protocol (IP)

- Layer 3 protocol that *forwards* datagrams across internet
- Uses routing tables prepared by routing protocols, e.g., Open Shortest Path First (OSPF), Routing Information Protocol (RIP)
- Connectionless service
  vs connection-oriented (circuits)

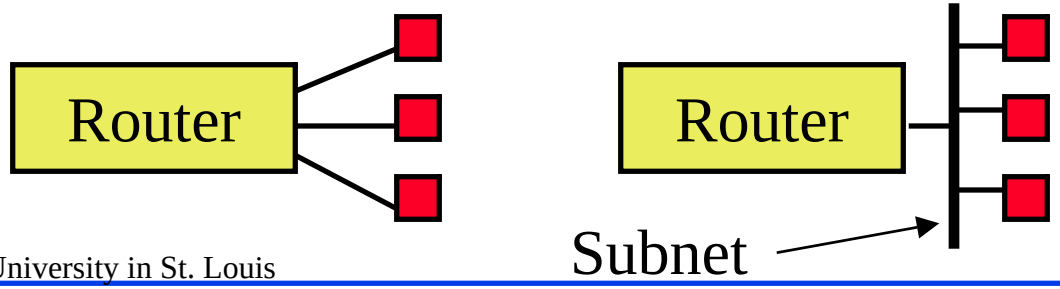


IP Address

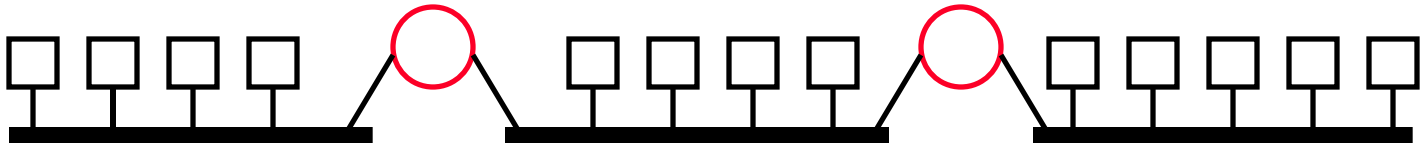
- Class A: (1+3 bytes)
- Class B: (2+2 bytes)
- Class C: (3+1 bytes)
- Class D:



- Local = Subnet + Host (Variable length)

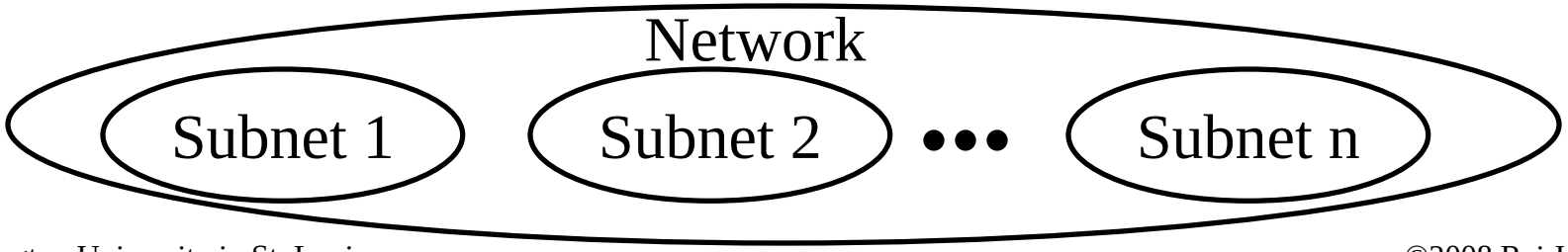


Subnetting

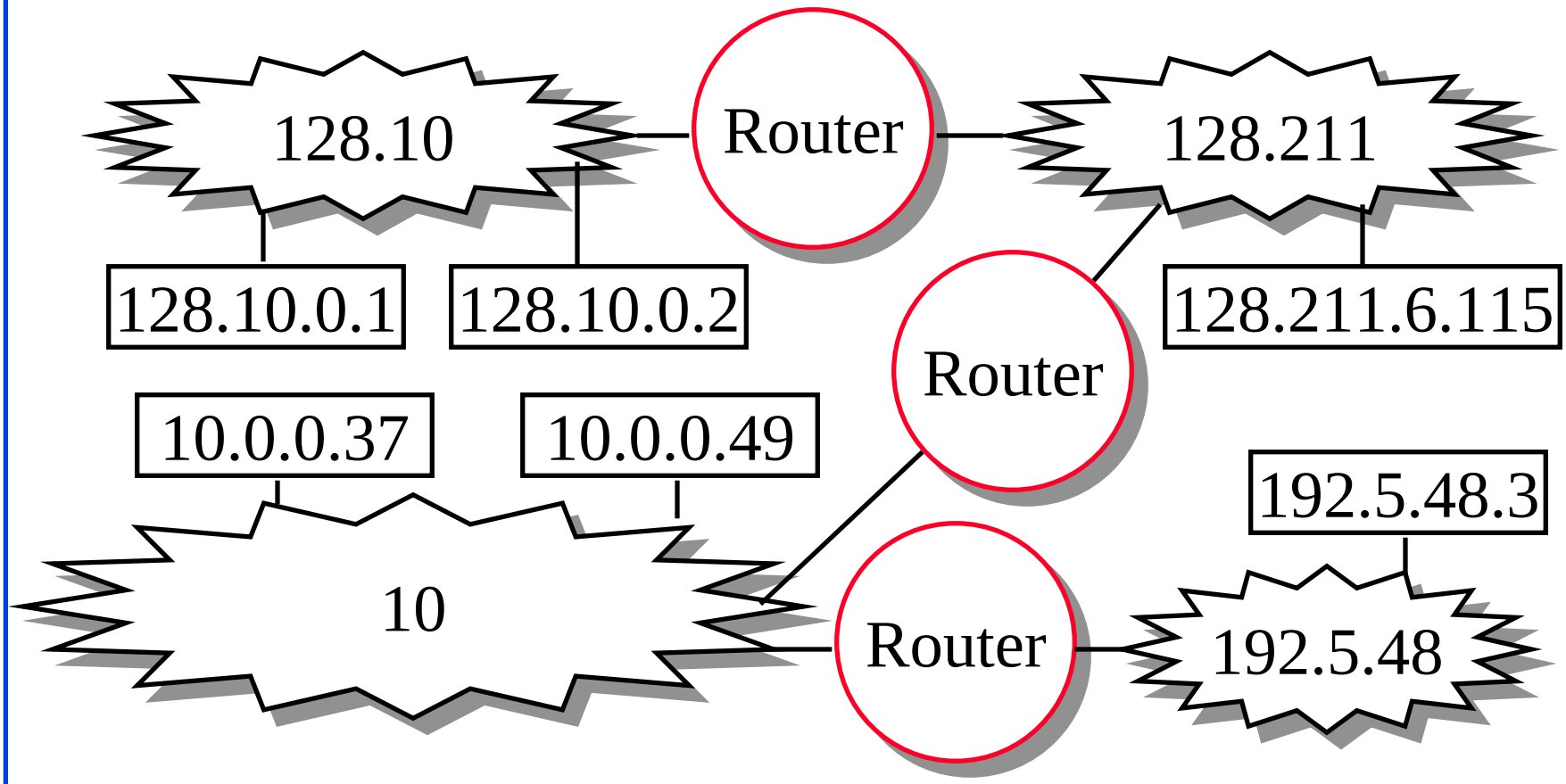


- All hosts on a subnetwork have the same prefix. Position of the prefix is indicated by a "subnet mask"
- Example: First 23 bits = subnet

  Address: 10010100 10101000 00010000 11110001
  Mask: 11111111 11111111 11111110 00000000
  .AND. 10010100 10101000 00010000 00000000



An Addressing Example



- All hosts on a network have the same network prefix

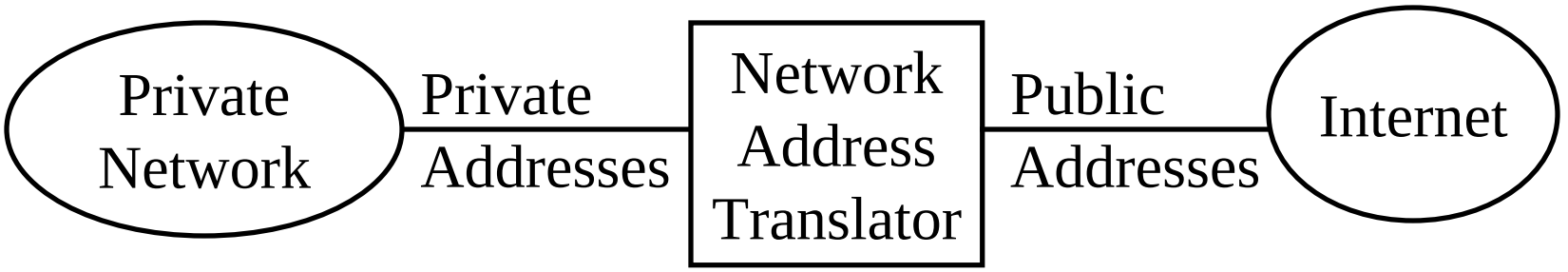
Fig 14.6

Special IP Addresses

- All-0 host suffix ⇒ Network Address
- All-0s ⇒ This computer
  (In some old networks: 0.0.0.0 = broadcast. Not used.)
- All-0s network ⇒ This network.
  E.g., 0.0.0.2 = Host 2 on this network
- All-1 host suffix ⇒ All hosts on the destination net (directed broadcast),
  All-0 host suffix ⇒ Berkeley directed broadcast address
- All-1s ⇒ All hosts on this net (limited broadcast)
  ⇒ Subnet number cannot be all 1
- 127.*.*.* ⇒ Looback through IP layer

Private Addresses

- Any organization can use these inside their network Can't go on the internet. [RFC 1918]
- 10.0.0.0 - 10.255.255.255 (10/8 prefix)
- 172.16.0.0 - 172.31.255.255 (172.16/12 prefix)
- 192.168.0.0 - 192.168.255.255 (192.168/16 prefix)



Forwarding an IP Datagram

- Delivers datagrams to destination network (subnet)
- Routers maintain a "routing table" of "next hops"
- Next Hop field does not appear in the datagram

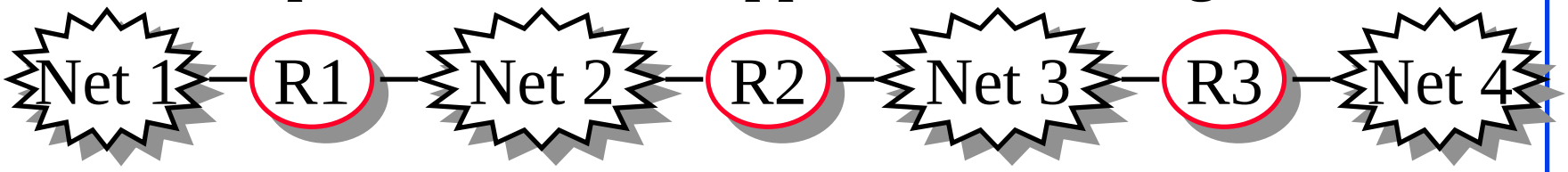


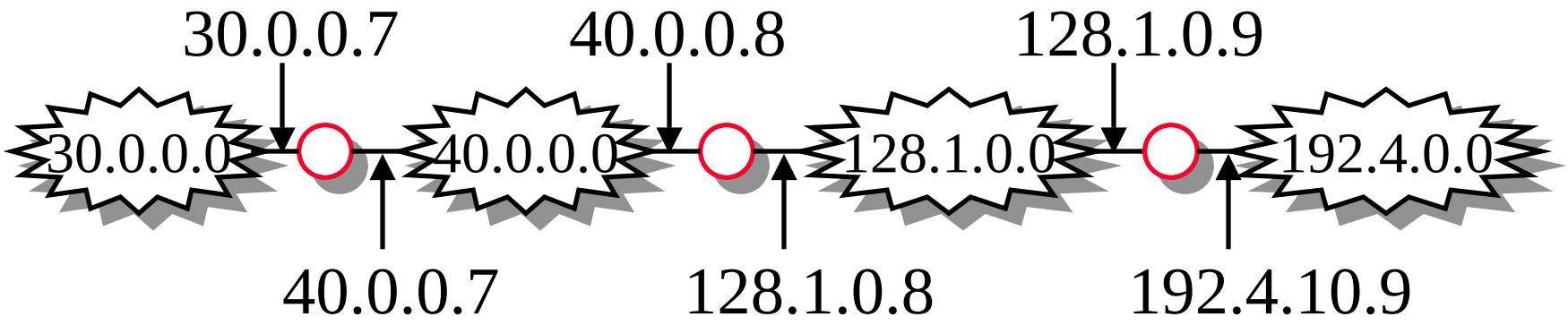
Table at R2:

Destination	Next Hop
Net 1	Forward to R1
Net 2	Deliver Direct
Net 3	Deliver Direct
Net 4	Forward to R3

Fig 16.2

IP Addresses and Routing Table Entries

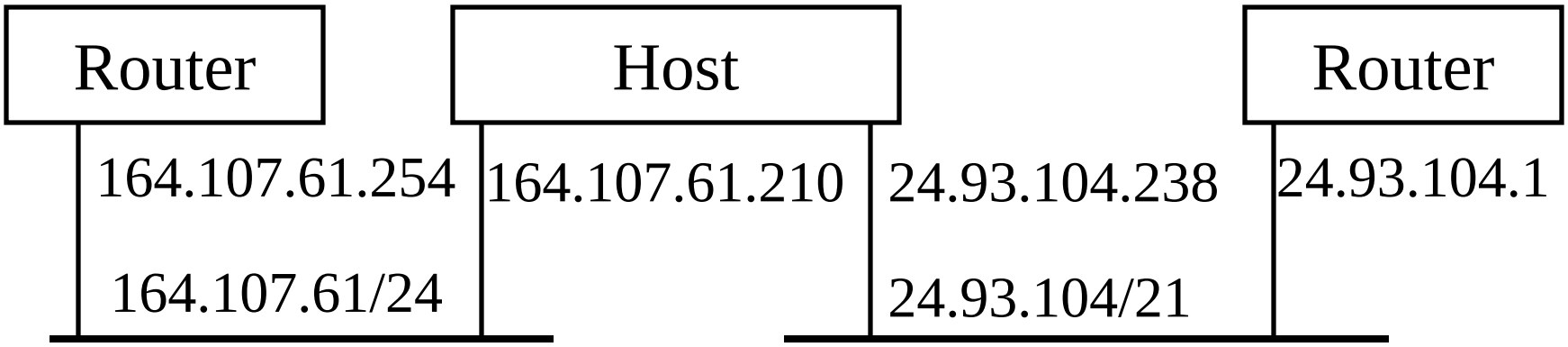
- IF ((Mask[i] & Destination Addr) = = Destination[i])
  Forward to NextHop[i]



Destination	Mask	Next Hop
30.0.0.0	255.0.0.0	40.0.0.7
40.0.0.0	255.0.0.0	Deliver direct
128.1.0.0	255.255.0.0	Deliver direct
192.4.10.0	255.255.255.0	128.1.0.9

Fig 16.3

Sample Routing Table



Network-Address	Netmask	Gateway-Address	Interface	Metric
0.0.0.0	0.0.0.0	24.93.104.1	24.93.107.238	1
24.93.104.0	255.255.248.0	24.93.107.238	24.93.107.238	1
24.93.107.238	255.255.255.255	127.0.0.1	127.0.0.1	1
24.255.255.255	255.255.255.255	24.93.107.238	24.93.107.238	1
127.0.0.0	255.0.0.0	127.0.0.1	127.0.0.1	1
128.146.0.0	255.255.0.0	164.107.61.254	164.107.61.210	1
164.107.61.0	255.255.255.0	164.107.61.210	164.107.61.210	1
164.107.61.210	255.255.255.255	127.0.0.1	127.0.0.1	1
164.107.255.255	255.255.255.255	164.107.61.210	164.107.61.210	1
224.0.0.0	224.0.0.0	24.93.107.238	24.93.107.238	1
224.0.0.0	224.0.0.0	164.107.61.210	164.107.61.210	1
255.255.255.255	255.255.255.255	164.107.61.210	164.107.61.210	1

IP Datagram Format

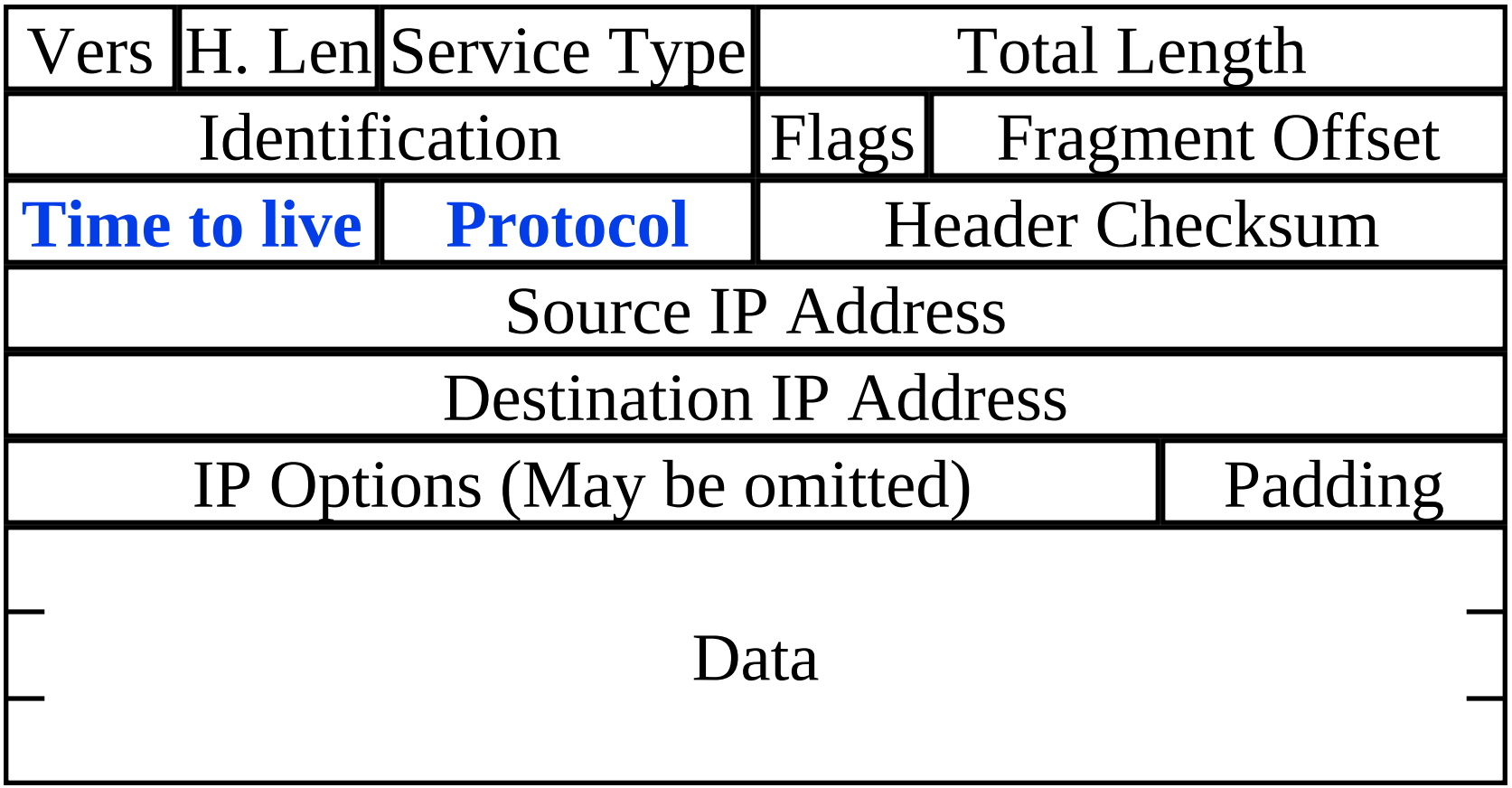


Fig 16.4

IP Header Format

- Version (4 bits)
- Internet header length (4 bits): in 32-bit words.
  Min header is 5 words or 20 bytes.
- Type of service (8 bits): Reliability, precedence, delay, and throughput
- Total length (16 bits): header + data in bytes
  Total must be less than 64 kB.
- Identifier (16 bits): Helps uniquely identify the datagram during its life for a given source, destination address

IP Header (Cont)

- Flags (3 bits): More flag - used for fragmentation
  No-fragmentation
  Reserved
- Fragment offset (13 bits): In units of 8 bytes
- Time to live (8 bits): Specified in router hops
- Protocol (8 bits): Next level protocol to receive the data
- Header checksum (16 bits): 1's complement sum of all 16-bit words in the header

IP Header (Cont)

- Source Address (32 bits): Original source.
  Does not change along the path.
- Destination Address (32 bits): Final destination.
  Does not change along the path.
- Options (variable): Security, source route, record route, stream id (used for voice) for reserved resources, timestamp recording
- Padding (variable):
  Makes header length a multiple of 4
- Data (variable): Data + header $\leq$ 65,535 bytes

Maximum Transmission Unit

- Each subnet has a maximum frame size
  Ethernet: 1518 bytes
  FDDI: 4500 bytes
  Token Ring: 2 to 4 kB
- Transmission Unit = IP datagram (data + header)
- Each subnet has a maximum IP datagram length: MTU

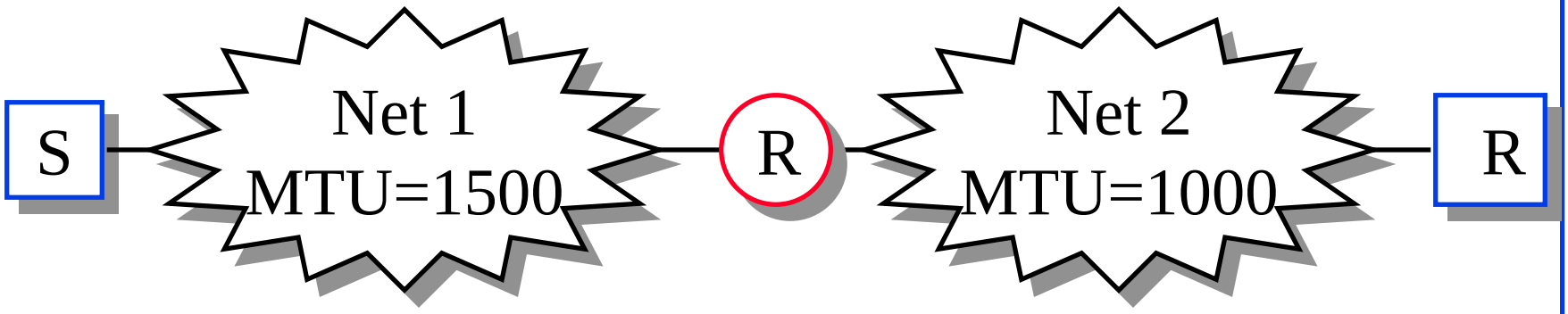


Fig 17.3

IPv6 Enhancements

1. Expanded address space: 128 bit
2. Address auto-configuration: Dynamic assignment
3. Increased addressing flexibility: Anycast + Multicast
4. Improved option mechanism: Extension Headers
   - Improved speed and simplified router processing
5. Support for resource allocation
   - Replaces type of service
   - Labeling of packets to particular traffic flow

Colon-Hex Notation

- **Dot-Decimal**: 127.23.45.88
- **Colon-Hex:**
  FEDC:0000:0000:0000:3243:0000:0000:ABCD
  - Can skip leading zeros of each word
  - Can skip one sequence of zero words, e.g.,
    FEDC::3243:0000:0000:ABCD
    ::3243:0000:0000:ABCD
  - Can leave the last 32 bits in dot-decimal, e.g.,
    ::127.23.45.88
  - Can specify a prefix by /length, e.g.,
    2345:BA23:0007::/50

Local-Use Addresses

- Link Local: Not forwarded outside the link, FE:80::xxx

10 bits	n bits	118-n
1111 1110 10	0	Interface ID

- Site Local: Not forwarded outside the site, FE:C0::xxx

10 bits	n bits	m bits	118-n-m bits
1111 1110 11	0	Subnet ID	Interface ID

- Provides plug and play

Extension Headers

Most extension headers are examined only at destination

1. Hop-by-Hop Options
2. Fragmentation: All IPv6 routers can carry 536 Byte payload
3. Routing: Loose or tight source routing
4. Destination Options

Extension Header (Cont)

- Only Base Header:

Base Header Next = TCP	TCP Segment

- Only Base Header and One Extension Header:

Base Header Next = Routing	Route Header Next = TCP	TCP Segment

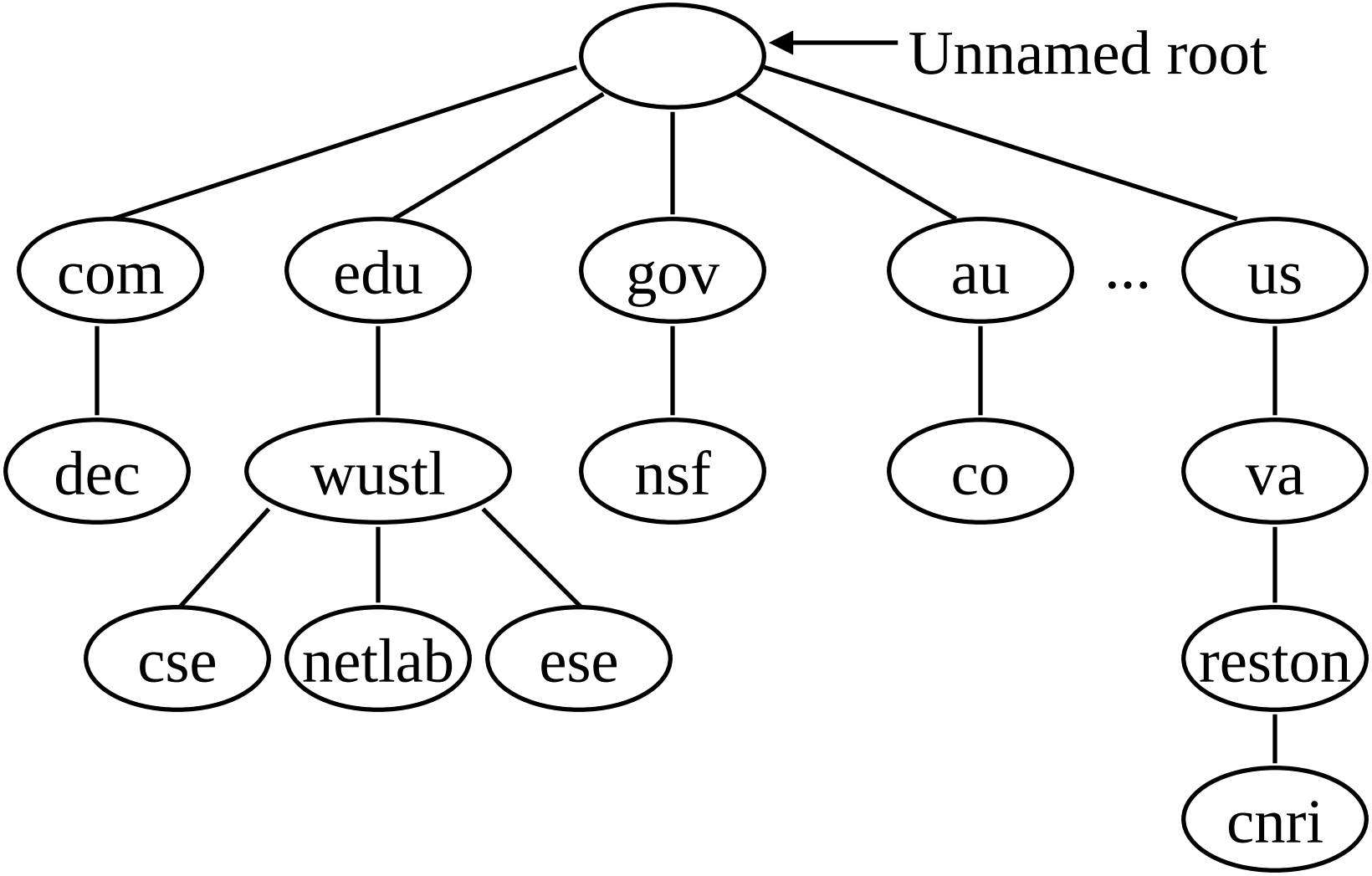
- Only Base Header and Two Extension Headers:

Base Header Next = Hop	Hop Header Next = Routing	Routing Header Next = TCP	TCP Segment

Domain Name Service

- Computers use addresses
- Humans cannot remember IP addresses
  $\Rightarrow$ Need names
  Example, Liberia for 164.107.51.28
- Simplest Solution: Each computer has a unique name and has a built in table of name to address translation
- Problem: Not scalable
- Solution: DNS (Adopted in 1983)
- Hierarchical Names: Liberia.cse.wustl.edu

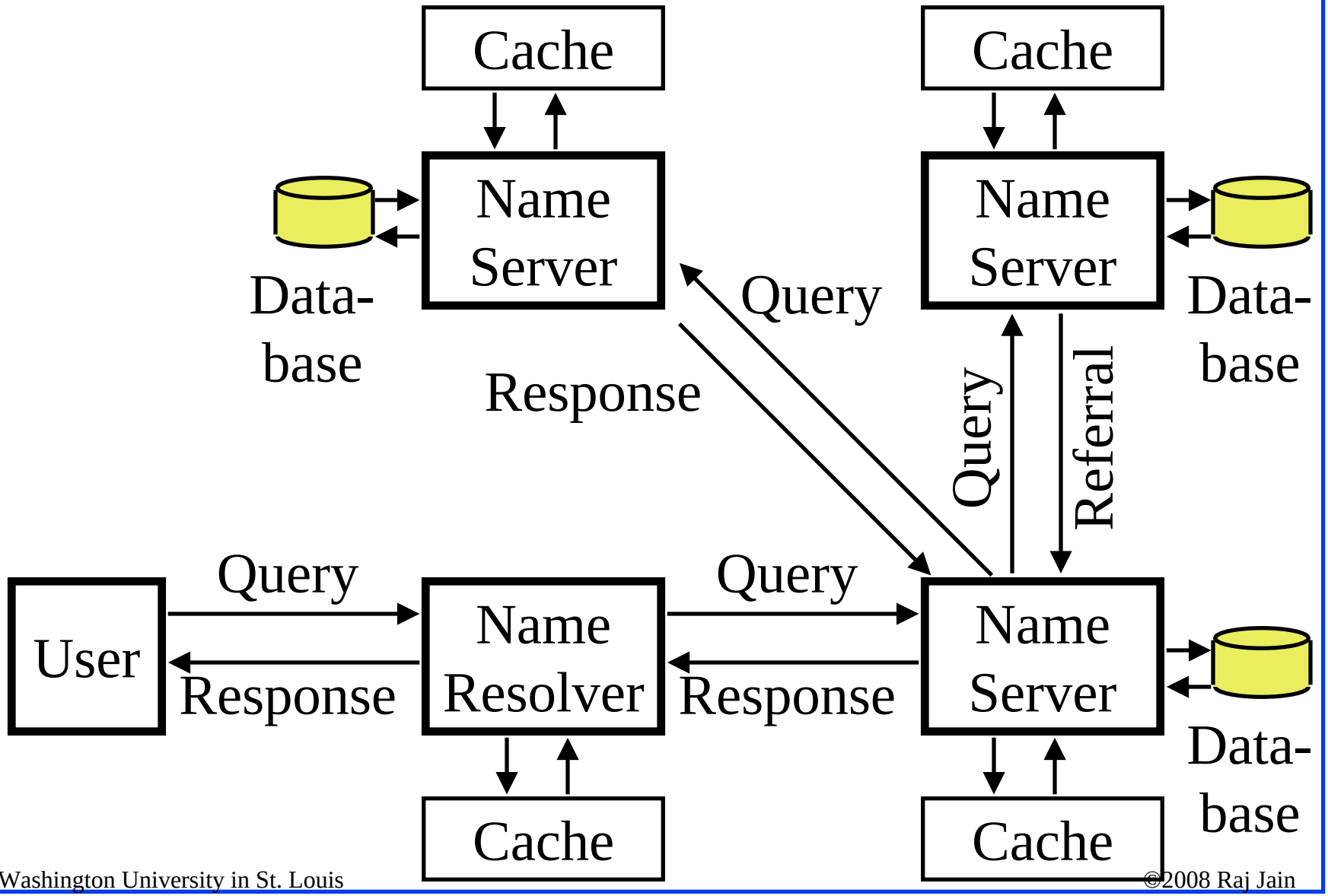
Name Hierarchy



Name Hierarchy

- Unique domain suffix is assigned by Internet Assigned Number Authority (IANA)
- The domain administrator has complete control over the domain
- No limit on number of sub-domains or number of levels
- computer.site.division.company.com
  computer.site.subdivision.division.company.com
- Name space is not related to physical interconnection, e.g., math.wustl and cse.wustl could be on the same floor or in different cities

Name Resolution

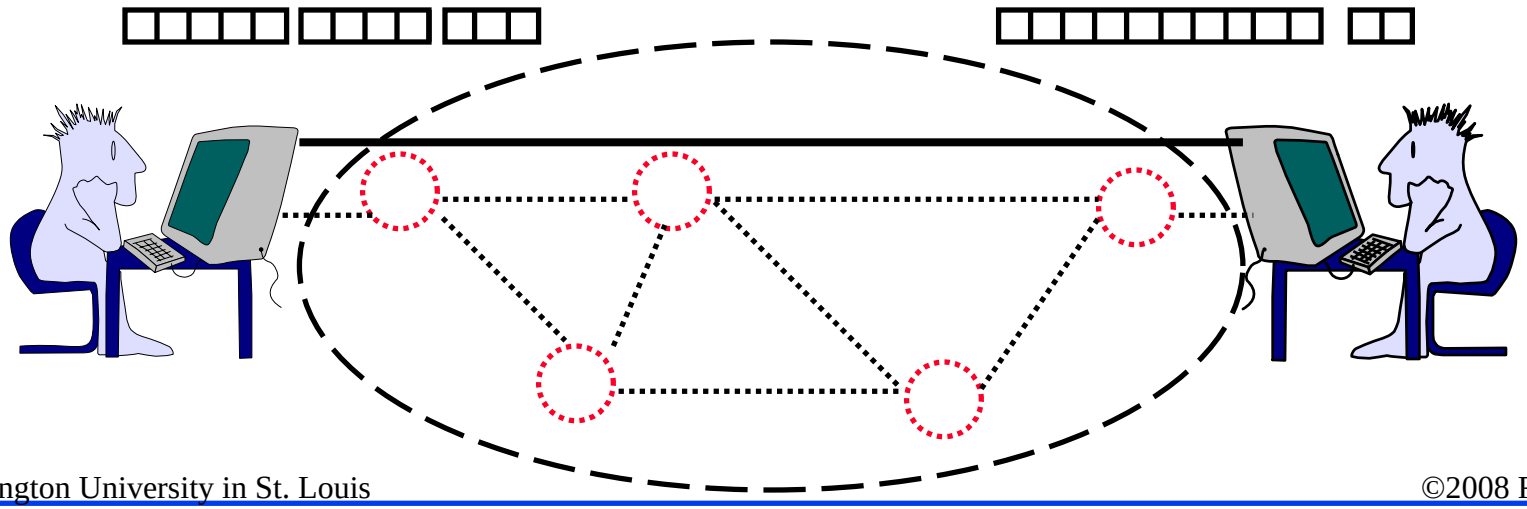


Name Resolution (Cont)

- Each computer has a name resolver routine, e.g., gethostbyname in UNIX
- Each resolver knows the name of a local DNS server
- Resolver sends a DNS request to the server
- DNS server either gives the answer, forwards the request to another server, or gives a referral
- Referral = Next server to whom request should be sent
- Servers respond to a full name only
  However, humans may specify only a partial name
  Resolvers may fill in the rest of the suffix, e.g.,
  Liberia.cis = Liberia.cis.wustl.edu

TCP: Key Features

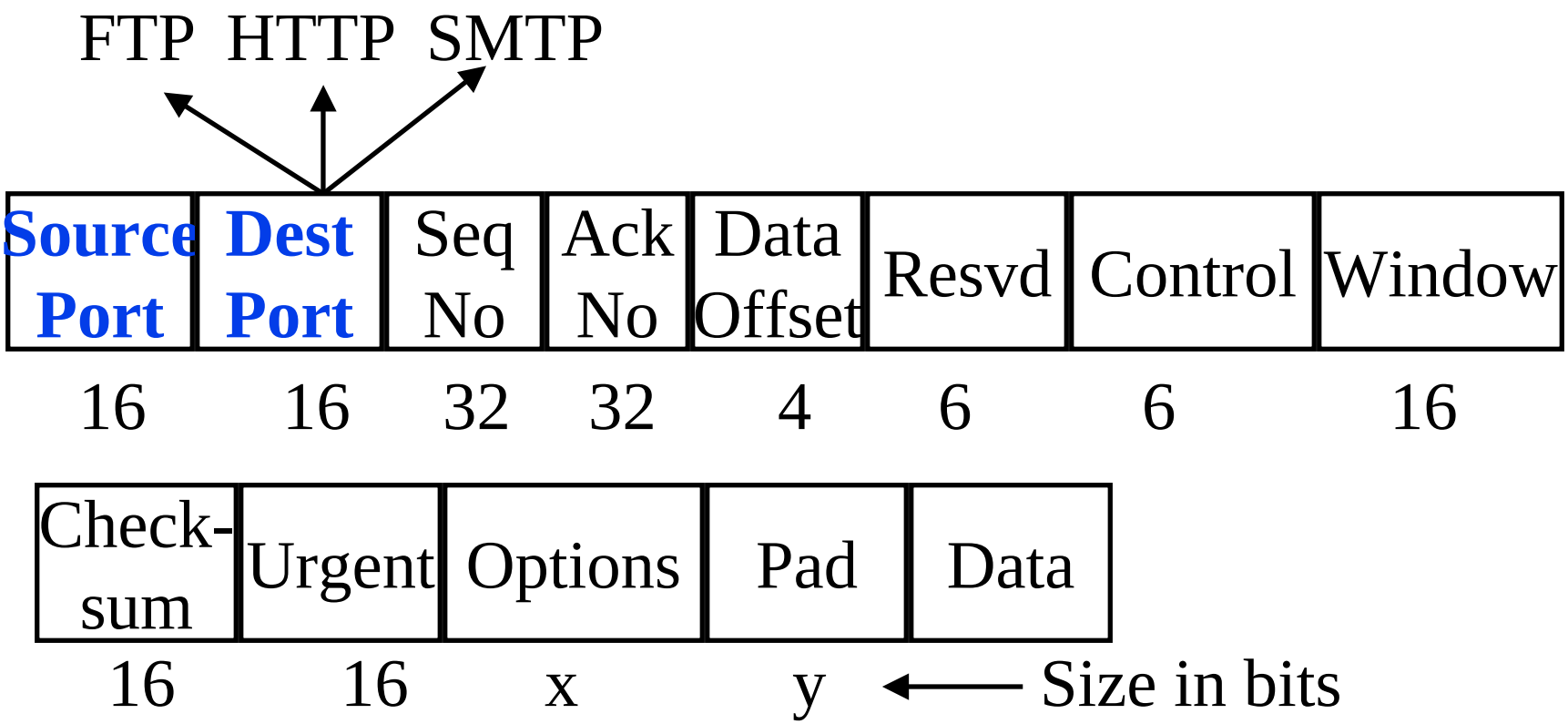
- Point-to-point communication: **Two** end-points
- **Connection** oriented. Full duplex communication.
- **Reliable** transfer: Data is delivered in order
  Lost packets are retransmitted.
- **Stream** interface: Continuous sequence of octets



Transport Control Protocol (TCP)

- Key Services:
  - Send: Please send when convenient
  - Data stream push: Please send it all now, if possible.
  - Urgent data signaling: Destination TCP! please give this urgent data to the user
    (Urgent data is delivered in sequence. Push at the should be explicit if needed.)
  - Note: Push has no effect on delivery.
    Urgent requests quick delivery

TCP Header Format

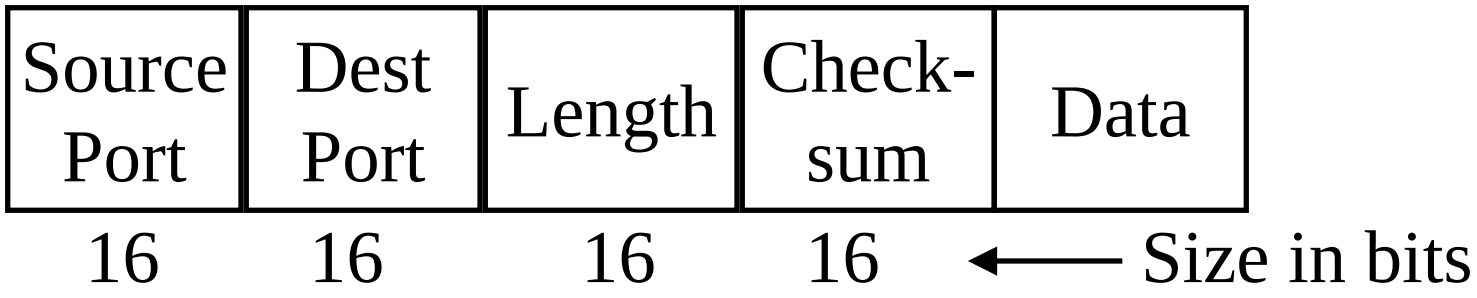


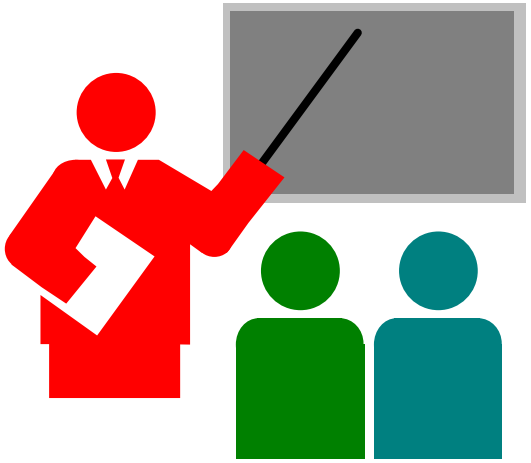
TCP Header

- Source Port (16 bits): Identifies source user process 20 = FTP, 23 = Telnet, 53 = DNS, 80 = HTTP, ...
- Destination Port (16 bits)
- Sequence Number (32 bits): Sequence number of the first byte in the segment.
- Ack number (32 bits): Next byte expected
- Data offset (4 bits): # of 32-bit words in the header
- Reserved (6 bits)
- Window (16 bits): Will accept [Ack] to [Ack]+[window]

User Datagram Protocol (UDP)

- **Connectionless** end-to-end service
- **Unreliable**: No flow control.
  No error recovery (No acks. No retransmissions.)
- Used by network management and Audio/Video.
- Provides port addressing
- Error detection (Checksum) optional.





Summary

- IP is the forwarding protocol between networks
- IPv4 uses 32-bit addresses
- IPv6 uses 128 bit addresses
- DNS: Maps names to addresses
- TCP provides reliable full-duplex connections.
- UDP is connectionless and simple. No flow/error control.

Ethernet

Raj Jain
Professor of Computer Science and Engineering
Washington University in Saint Louis
Saint Louis, MO, USA
jain@acm.org
http://www.cse.wustl.edu/~jain/

- CSMA/CD
- IEEE 802 Address Format
- Interconnection Devices
- Distance-B/W Principle
- Gigabit Ethernet
- Spanning Tree
- 10Gbps Ethernet PHYs
- Metro Ethernet Services

CSMA/CD

- Aloha at Univ of Hawaii:
  Transmit whenever you like
  Worst case utilization = 1/(2e) =18%
- Slotted Aloha: Fixed size transmission slots
  Worst case utilization = 1/e = 37%
- CSMA: Carrier Sense Multiple Access
  Listen before you transmit
- p-Persistent CSMA: If idle, transmit with probability p. Delay by one time unit with probability 1-p
- CSMA/CD: CSMA with Collision Detection
  Listen while transmitting. Stop if you hear someone else

IEEE 802.3 CSMA/CD

- If the medium is idle, transmit (1-persistent).
- If the medium is busy, wait until idle and then transmit immediately.
- If a collision is detected while transmitting,
  - Transmit a jam signal for one slot (= 51.2 s = 64 byte times)
  - Wait for a random time and reattempt (up to 16 times)
  - Random time = Uniform[0,$2^{\min(k,10)}$-1] slots
- Collision detected by monitoring the voltage
  High voltage two or more transmitters Collision
  Length of the cable is limited to 2 km

Ethernet Standards

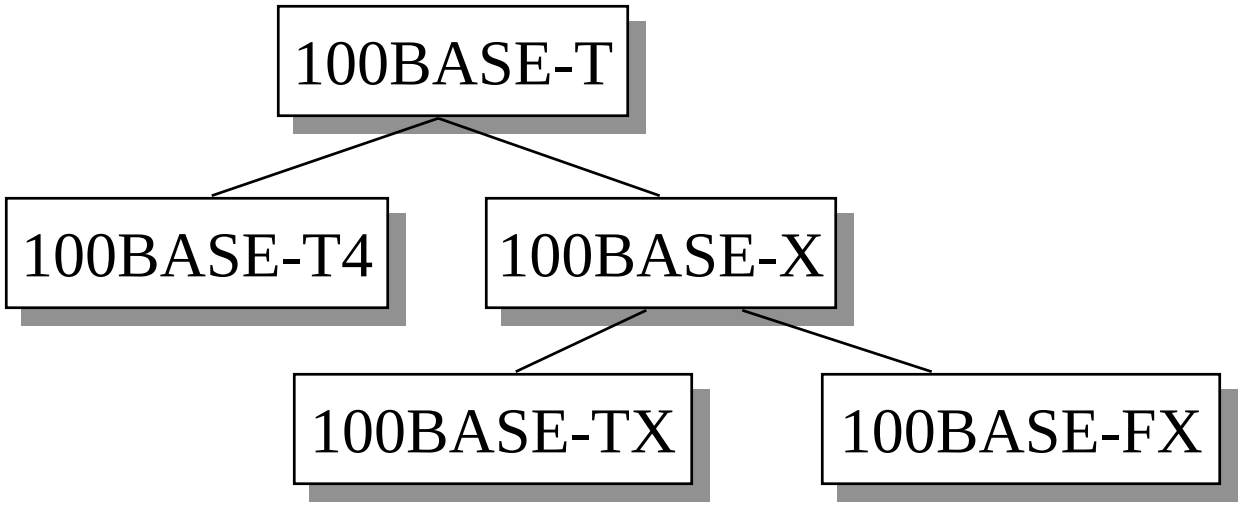
- 10BASE5: 10 Mb/s over coaxial cable (ThickWire)
- 10BROAD36: 10 Mb/s over broadband cable, 3600 m max segments
- 1BASE5: 1 Mb/s over 2 pairs of UTP
- 10BASE2: 10 Mb/s over thin RG58 coaxial cable (ThinWire), 185 m max segments
- 10BASE-T: 10 Mb/s over 2 pairs of UTP
- 10BASE-FL: 10 Mb/s fiber optic point-to-point link
- 10BASE-FB: 10 Mb/s fiber optic backbone (between repeaters). Also, known as synchronous Ethernet.

Ethernet Standards (Cont)

- 10BASE-FP: 10 Mb/s fiber optic passive star + segments
- 10BASE-F: 10BASE-FL, 10BASE-FB, or 10BASE-FP
- 100BASE-T4: 100 Mb/s over 4 pairs of CAT-3, 4, 5 UTP
- 100BASE-TX: 100 Mb/s over 2 pairs of CAT-5 UTP or STP
- 100BASE-FX: 100 Mbps CSMA/CD over 2 optical fiber

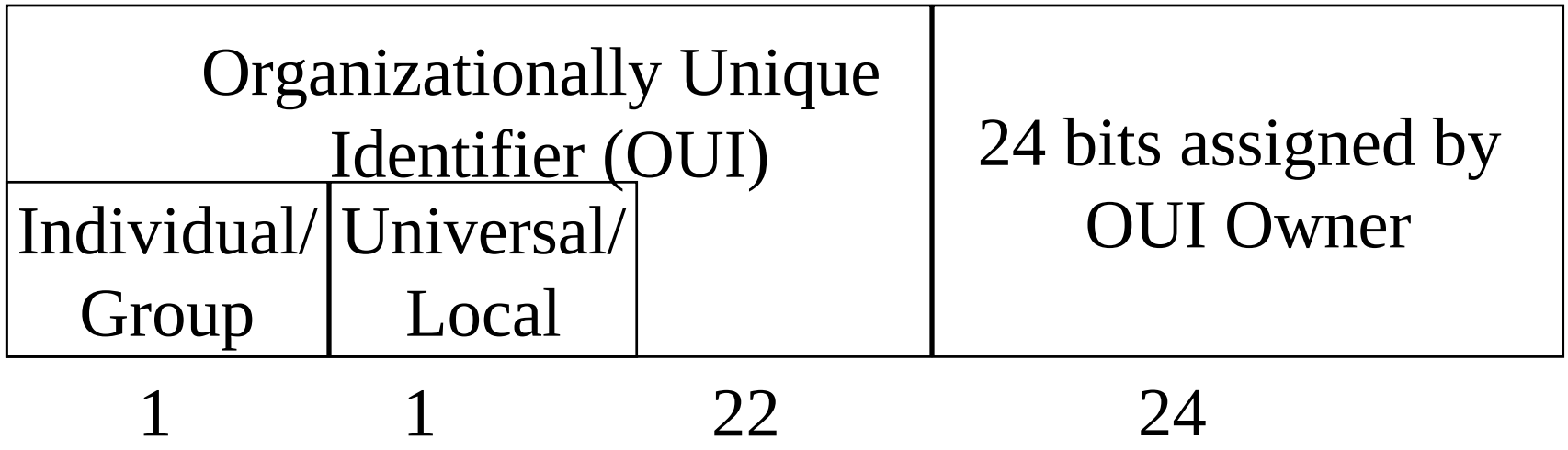
Ethernet Standards (Cont)

- 100BASE-X: 100BASE-TX or 100BASE-FX
- 100BASE-T: 100BASE-T4, 100BASE-TX, or 100BASE-FX
- 1000BASE-T: 1 Gbps (Gigabit Ethernet)



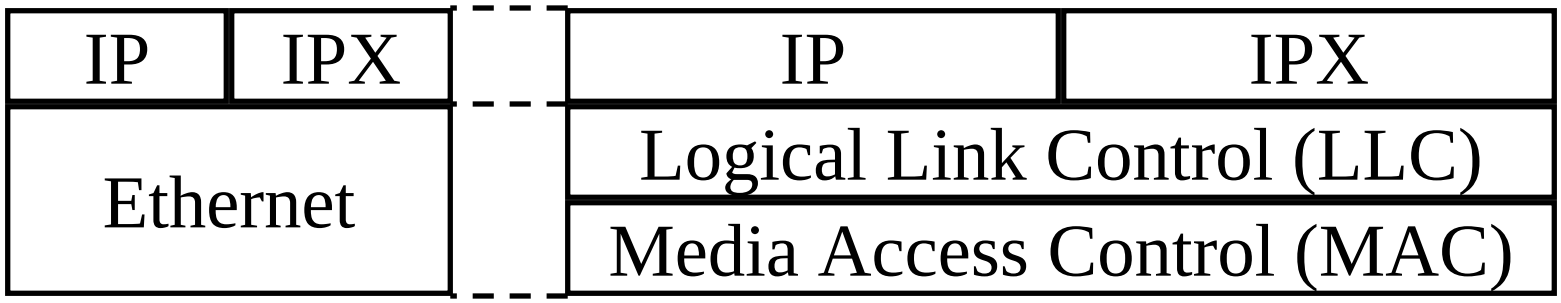
IEEE 802 Address Format

- 48-bit:1000 0000 : 0000 0001 : 0100 0011 : 0000 0000 : 1000 0000 : 0000 1100
  = 80:01:43:00:80:0C



- Multicast = "To all bridges on this LAN"
- Broadcast = "To all stations"
  = 111111....111 = FF:FF:FF:FF:FF:FF

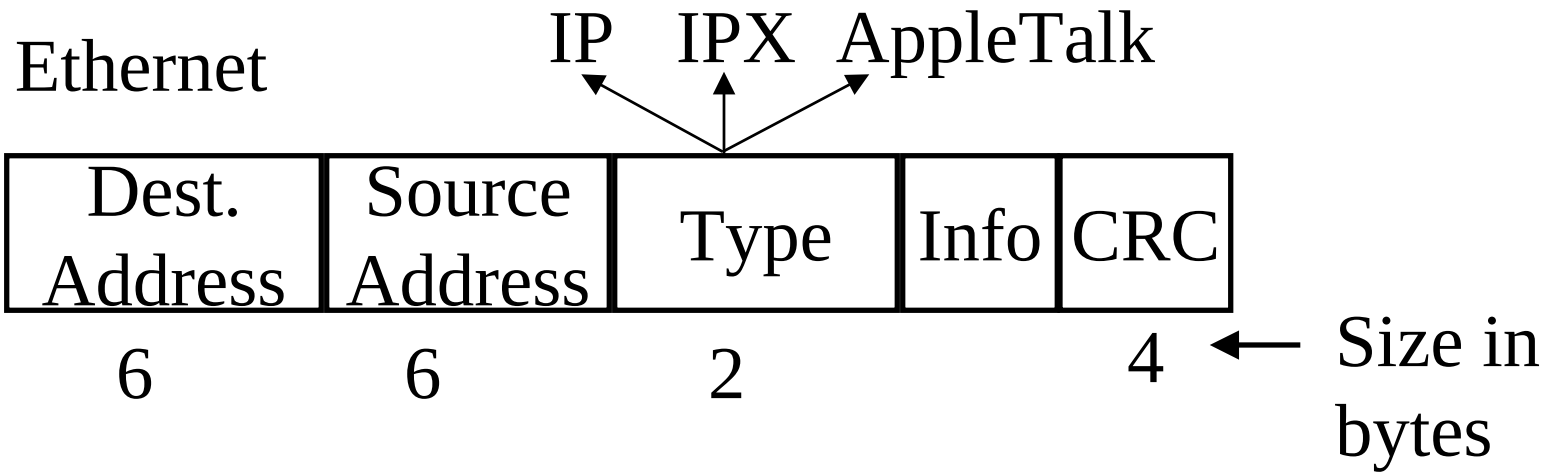
Ethernet vs IEEE 802.3



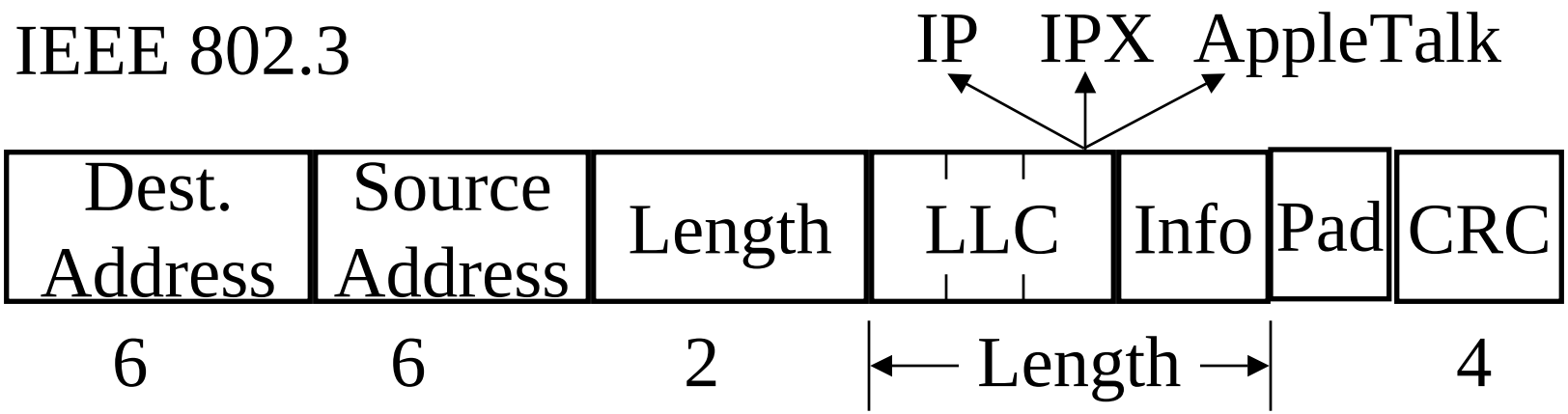
- In 802.3, datalink was divided into two sublayers: LLC and MAC
- LLC provides protocol multiplexing. MAC does not.
- MAC does not need a protocol type field.

Frame Format

- Ethernet



- IEEE 802.3



LLC Type 1

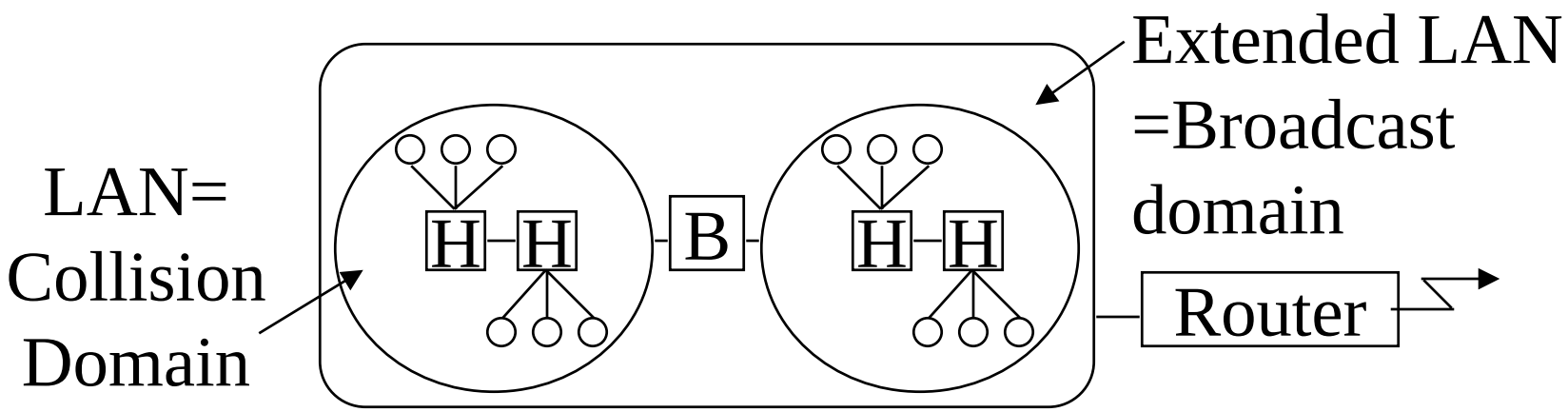
- Unacknowledged connectionless (on 802.3)
  No flow or error control.
  Provides protocol multiplexing.
  Uses 3 types of protocol data units (PDUs):
  UI = Unnumbered informaton
  XID = Exchange ID
  = Types of operation supported, window
  Test = Loop back test

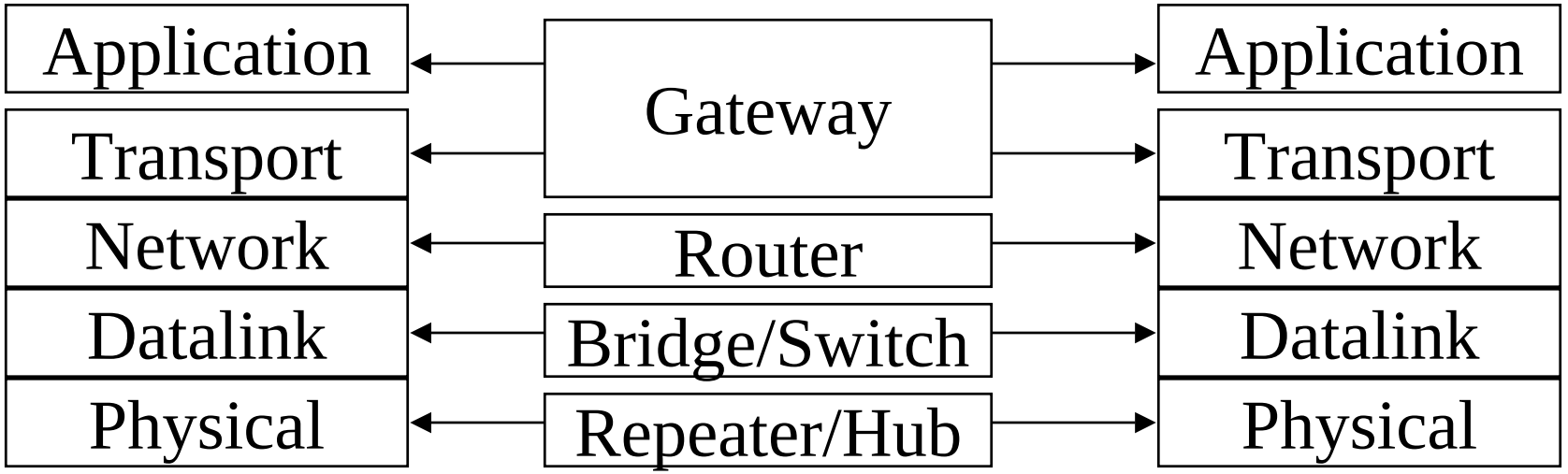
Interconnection Devices

- **Repeater**: PHY device that restores data and collision signals
- **Hub:** Multiport repeater + fault detection and recovery
- **Bridge:** Datalink layer device connecting two or more collision domains. MAC multicasts are propagated throughout "extended LAN."
- **Router:** Network layer device. IP, IPX, AppleTalk. Does not propagate MAC multicasts.
- **Switch**: Multiport bridge with parallel paths

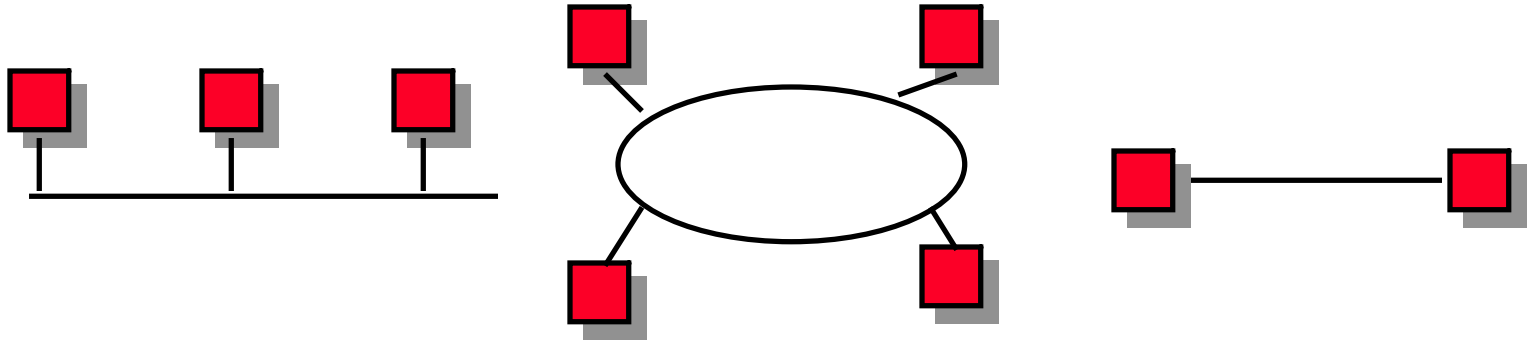
These are functions. Packaging varies.

Interconnection Devices





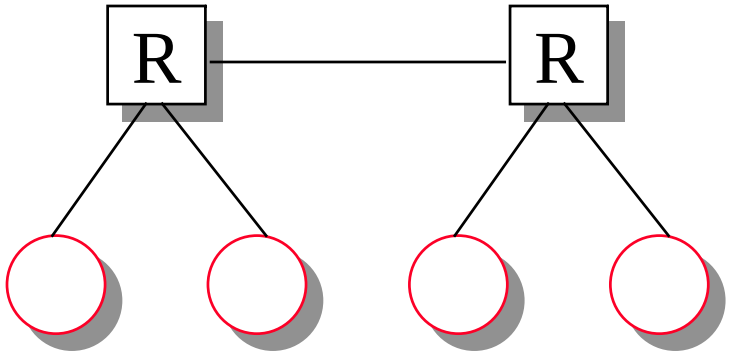
Distance-B/W Principle



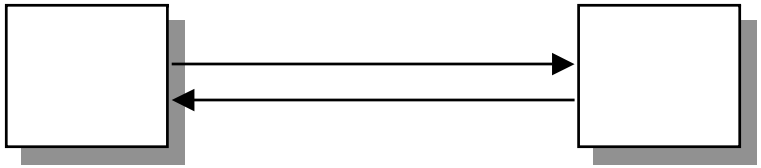
- Efficiency = Max throughput/Media bandwidth
- Efficiency is a nonincreasing function of α
  α = Propagation delay /Transmission time
  = (Distance/Speed of light)/(Transmission size/Bits/sec)
  = Distance×Bits/sec/(Speed of light)(Transmission size)
- Bit rate-distance-transmission size tradeoff.
- 100 Mb/s $\Rightarrow$ Change distance or frame size

Ethernet vs Fast Ethernet

	Ethernet	Fast Ethernet
Speed	10 Mbps	100 Mbps
MAC	CSMA/CD	CSMA/CD
Network diameter	2.5 km	205 m
Topology	Bus, star	Star
Cable	Coax, UTP, Fiber	UTP, Fiber
Standard	802.3	802.3u
Cost	X	2X



Full-Duplex Ethernet



- Uses point-to-point links between TWO nodes
- Full-duplex bi-directional transmission
- Transmit any time
- Many vendors are shipping switch/bridge/NICs with full duplex
- No collisions $\Rightarrow$ 50+ Km on fiber.
- Between servers and switches or between switches

1000Base-X

- 1000Base-LX: 1300-nm laser transceivers
  - 2 to 550 m on 62.5-μm or 50-μm multimode, 2 to 3000 m on 10-μm single-mode
- 1000Base-SX: 850-nm laser transceivers
  - 2 to 300 m on 62.5-μm, 2 to 550 m on 50-μm. Both multimode.
- 1000Base-CX: Short-haul copper jumpers
  - 25 m 2-pair shielded twinax cable in a single room or rack.
    Uses 8b/10b coding ⇒ 1.25 Gbps line rate

1000Base-T

- 100 m on 4-pair Cat-5 UTP
  ⇒ Network diameter of 200 m
- 250 Mbps/pair full duplex DSP based PHY
  ⇒ Requires new 5-level (PAM-5) signaling with 4-D 8-state Trellis code FEC
- Automatically detects and corrects pair-swapping, incorrect polarity, differential delay variations across pairs
- Autonegotiation ⇒ Compatibility with 100Base-T
- 802.3ab task force began March'97, ballot July'98, Final standard by March'99.

2. Spanning Tree

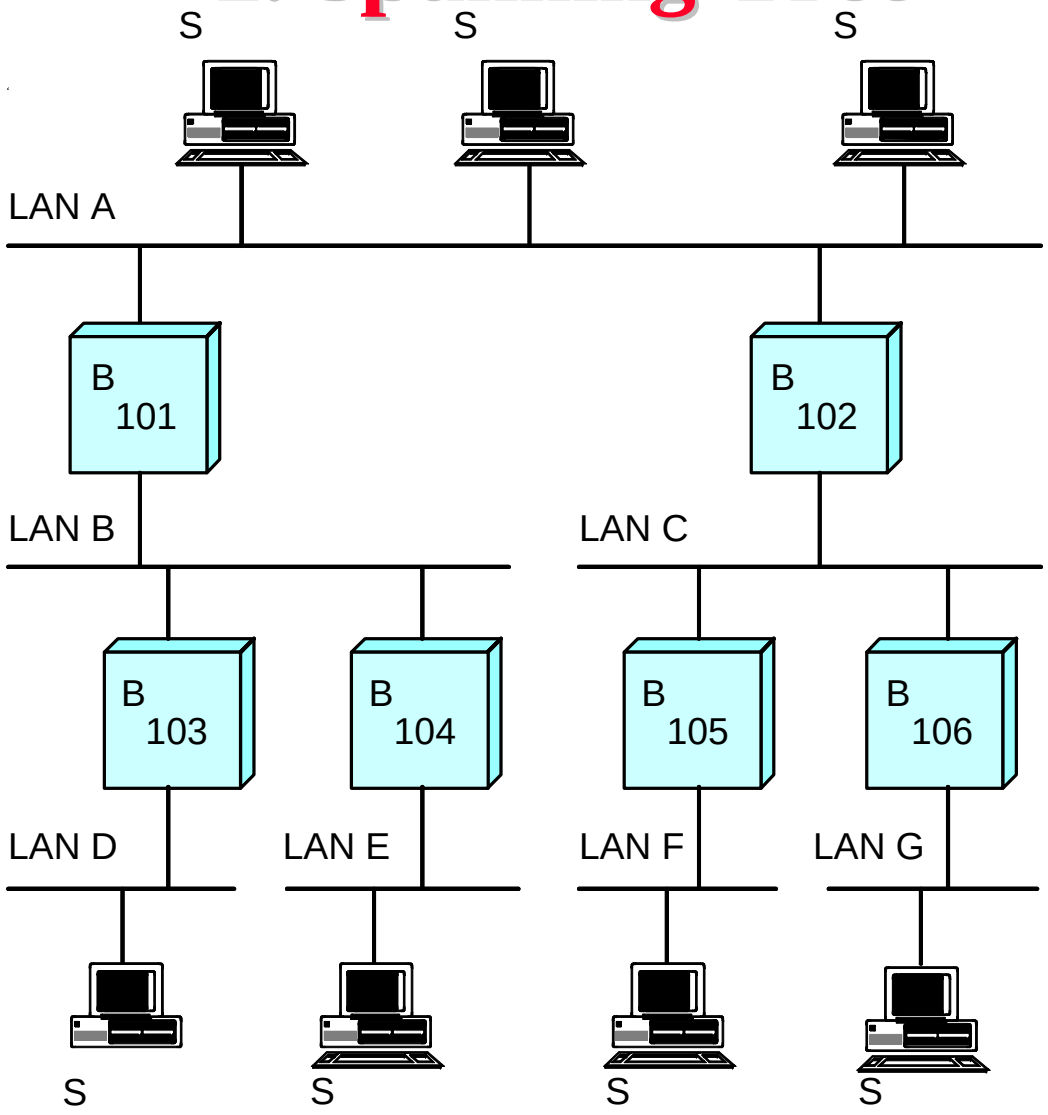


Fig 14.5

Spanning Tree (Cont)

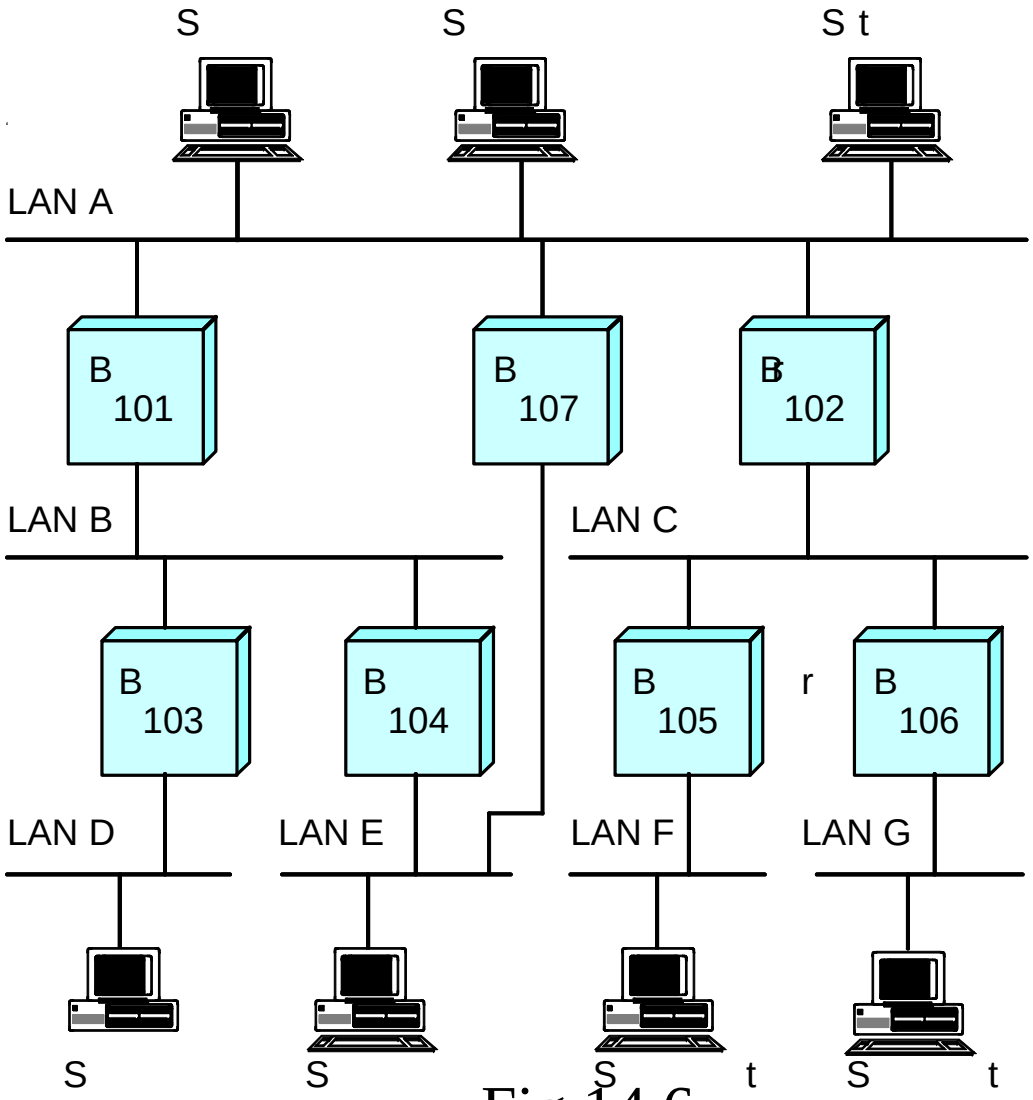
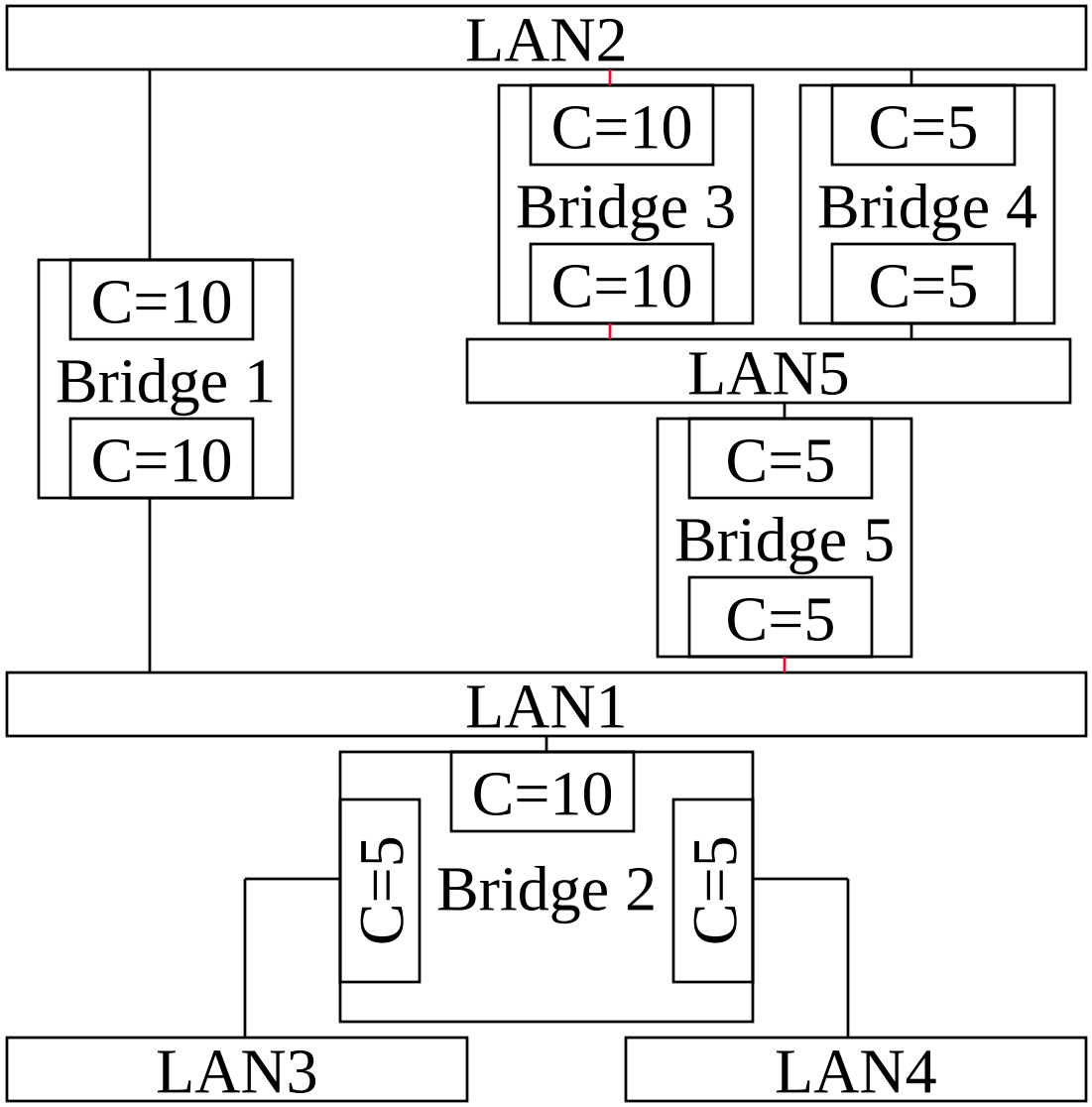


Fig 14.6

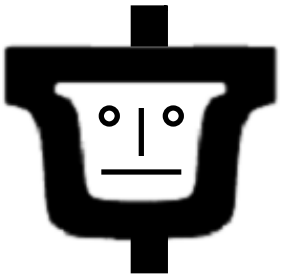
Spanning Tree Algorithm

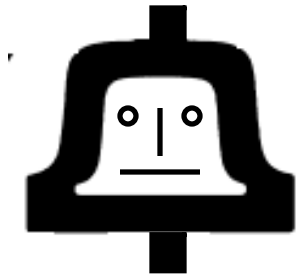
- All bridges multicast to "All bridges"
  - My ID
  - Root ID
  - My cost to root
- The bridges update their info using Dijkstra's algorithm and rebroadcast
- Initially all bridges are roots but eventually converge to one root as they find out the lowest Bridge ID.
- On each LAN, the bridge with minimum cost to the root becomes the Designated bridge
- All ports of all non-designated bridges are blocked.

Spanning Tree Example



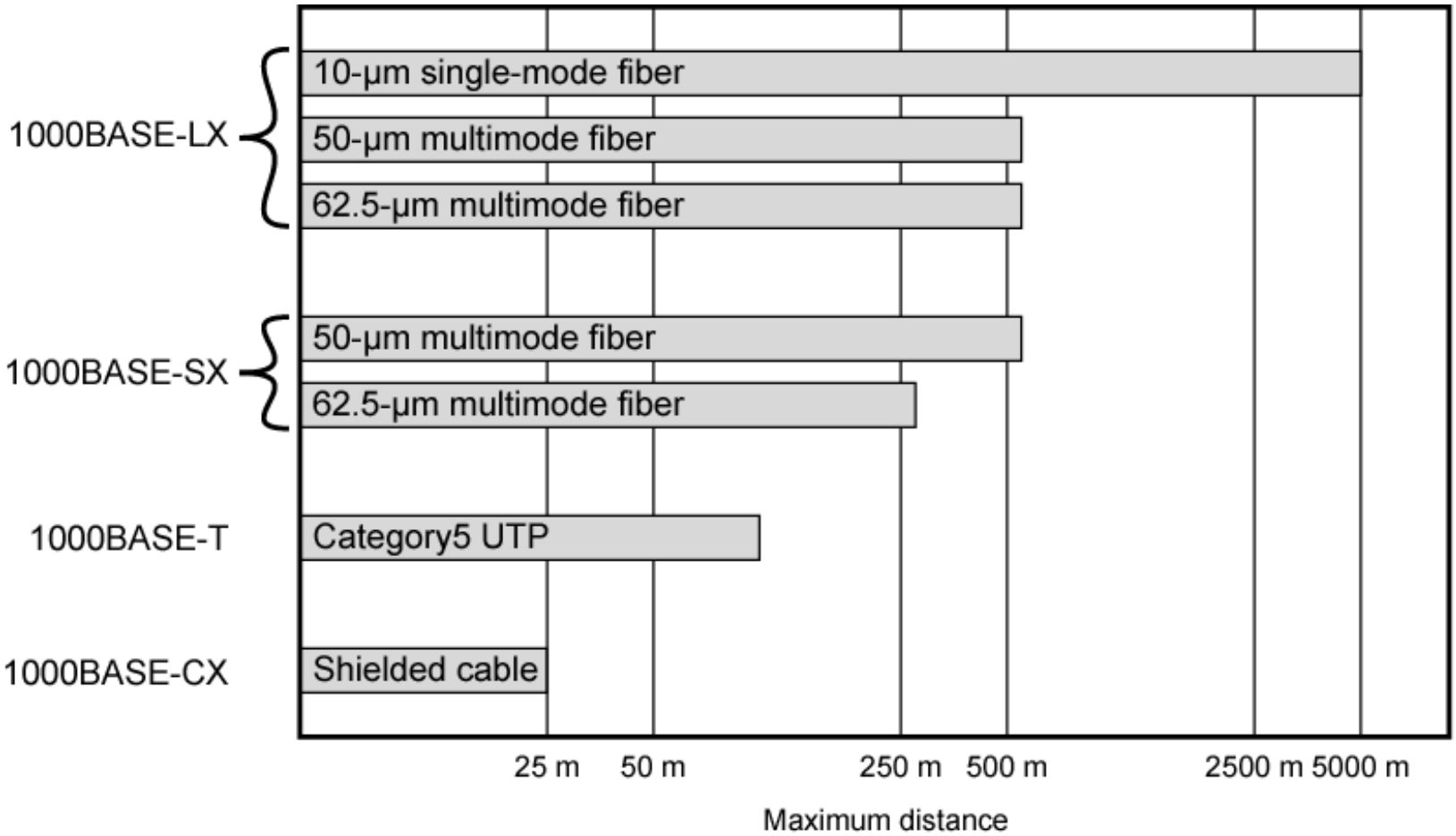
Ethernet: 1G vs 10G Designs





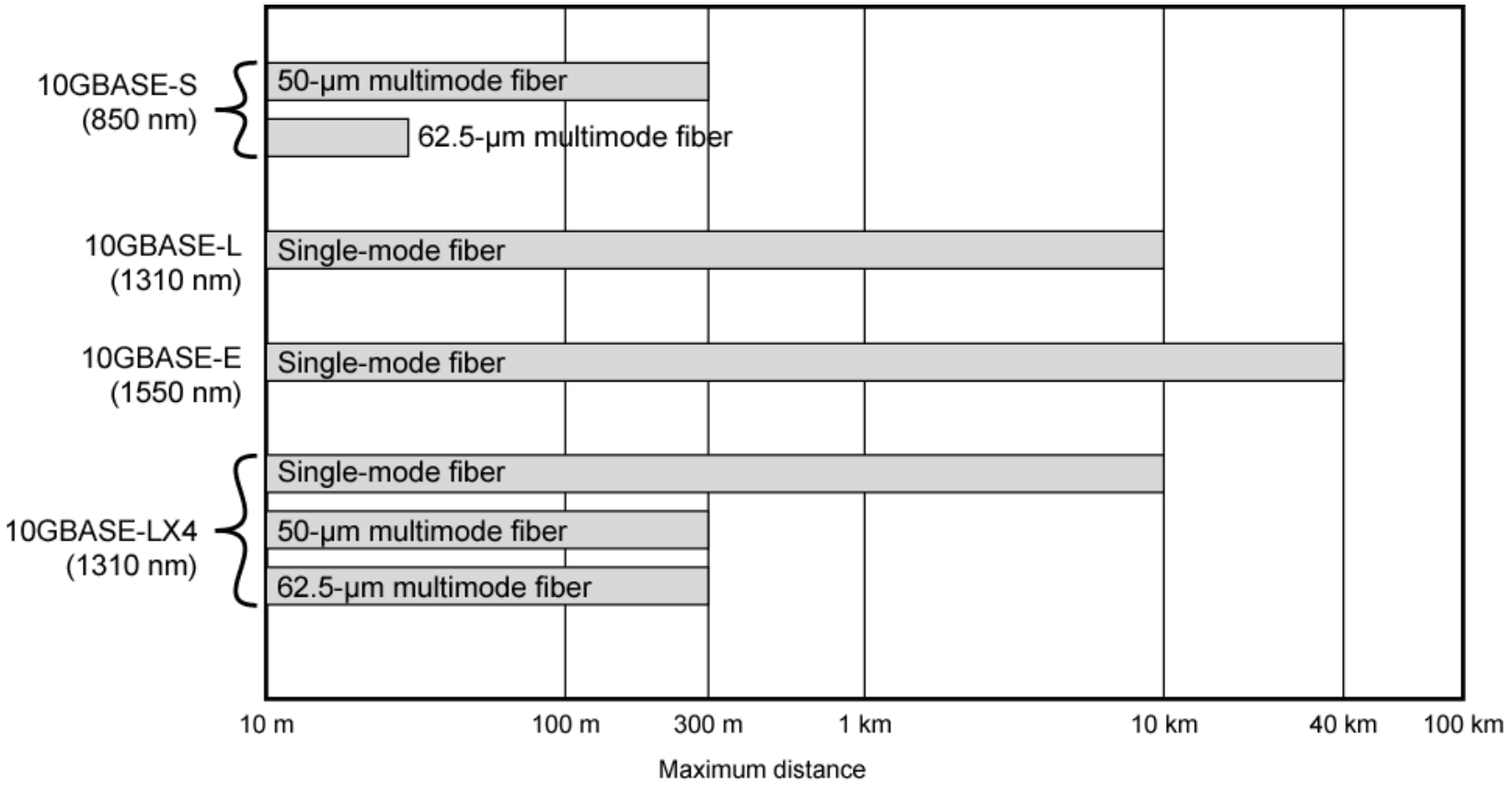
1G Ethernet	10G Ethernet
❑ 1000 / ~~800~~ / ~~622~~ Mbps **Single** data rate	❑ 10.0/9.5 Gbps **Both** rates.
❑ **LAN** distances only	❑ LAN and **MAN** distances
❑ No Full-duplex only ⇒ **Shared** Mode	❑ Full-duplex only ⇒ **No Shared** Mode
❑ Changes to **CSMA/CD**	❑ **No CSMA/CD** protocol ⇒ No distance limit due to MAC ⇒ *Ethernet* End-to-End

Gigabit Ethernet PHYs



- S = Short Wave, L=Long Wave, E=Extra Long Wave
- R = Regular reach (64b/66b), W=WAN (64b/66b + SONET Encapsulation), X = 8b/10b
- 4 = 4 λ's

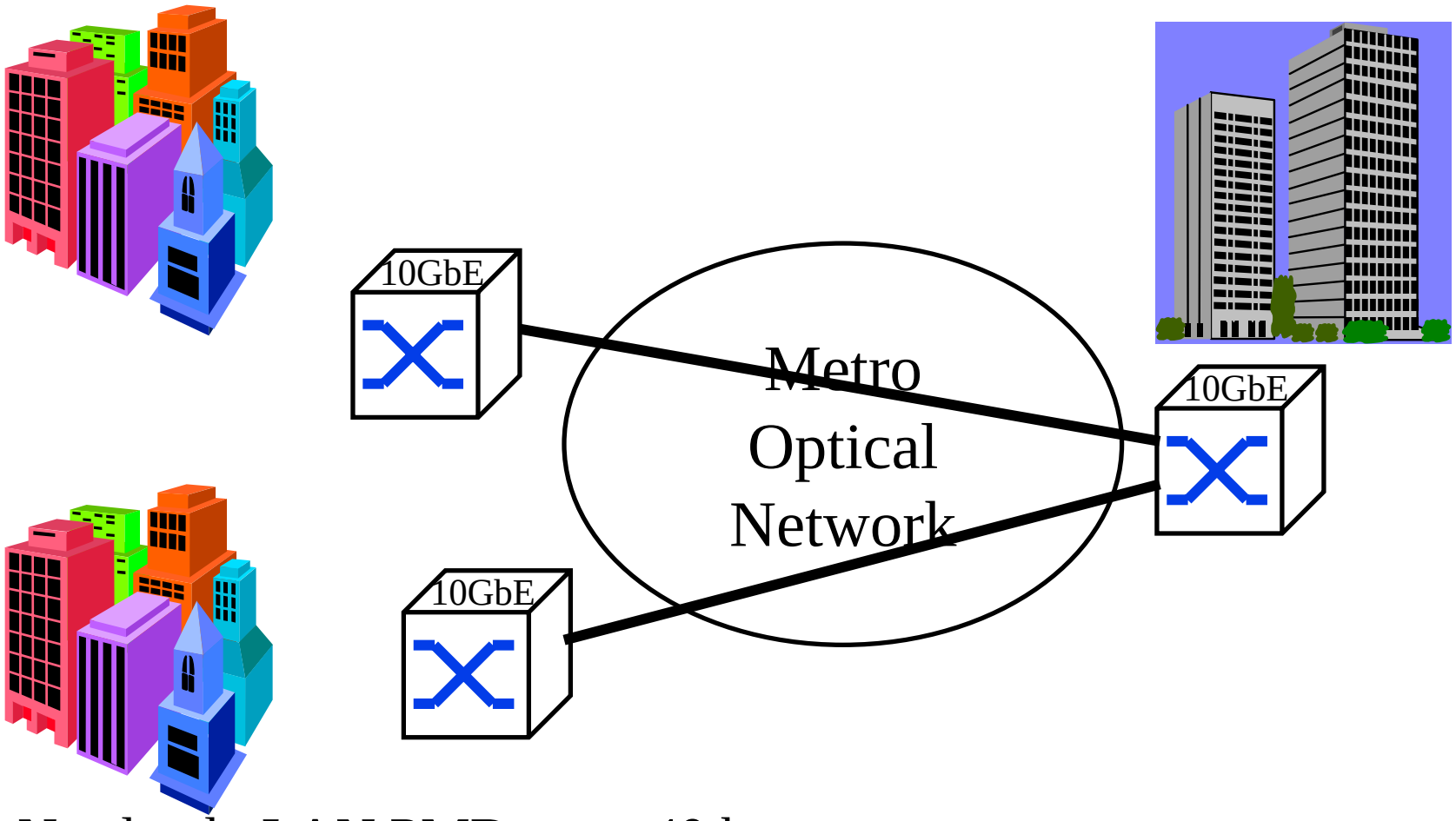
10Gbps Ethernet PHYs



10GBASE-T

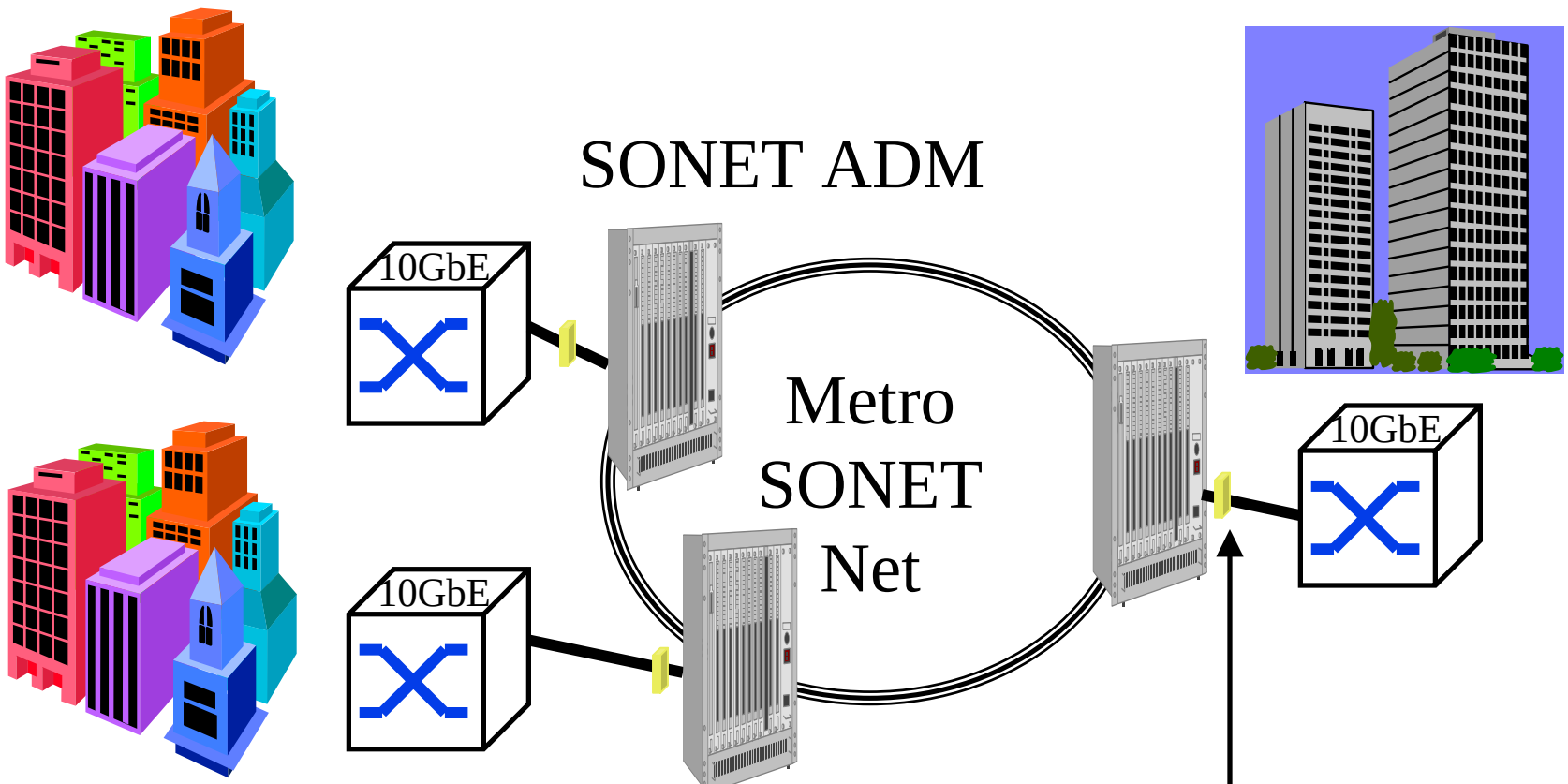
- New PHY for data center and horizontal wiring
- Compatible with existing 802.3ae MAC, XGMII, XAUI
- Standard: Start: Nov 2003 Finish: Jul 2006
- 100 m on Cat-7 and 55+ m on Cat-6
- Cost 0.6 of optical PHY. Greater reach than CX4
- 10-level coded PAM signaling with 3 bits/symbol
  833 MBaud/pair => 450 MHz bandwidth w FEXT cancellation
  (1GBASE-T uses 5-level PAM with 2 bits/symbol, 125 MBaud/pair, 80 MHz w/o FEXT)
- Full-duplex only. 1000BASE-T line code and FEC designed for half-duplex.
- http://www.ieee802.org/3/10GBT

10 GbE over Dark Fiber



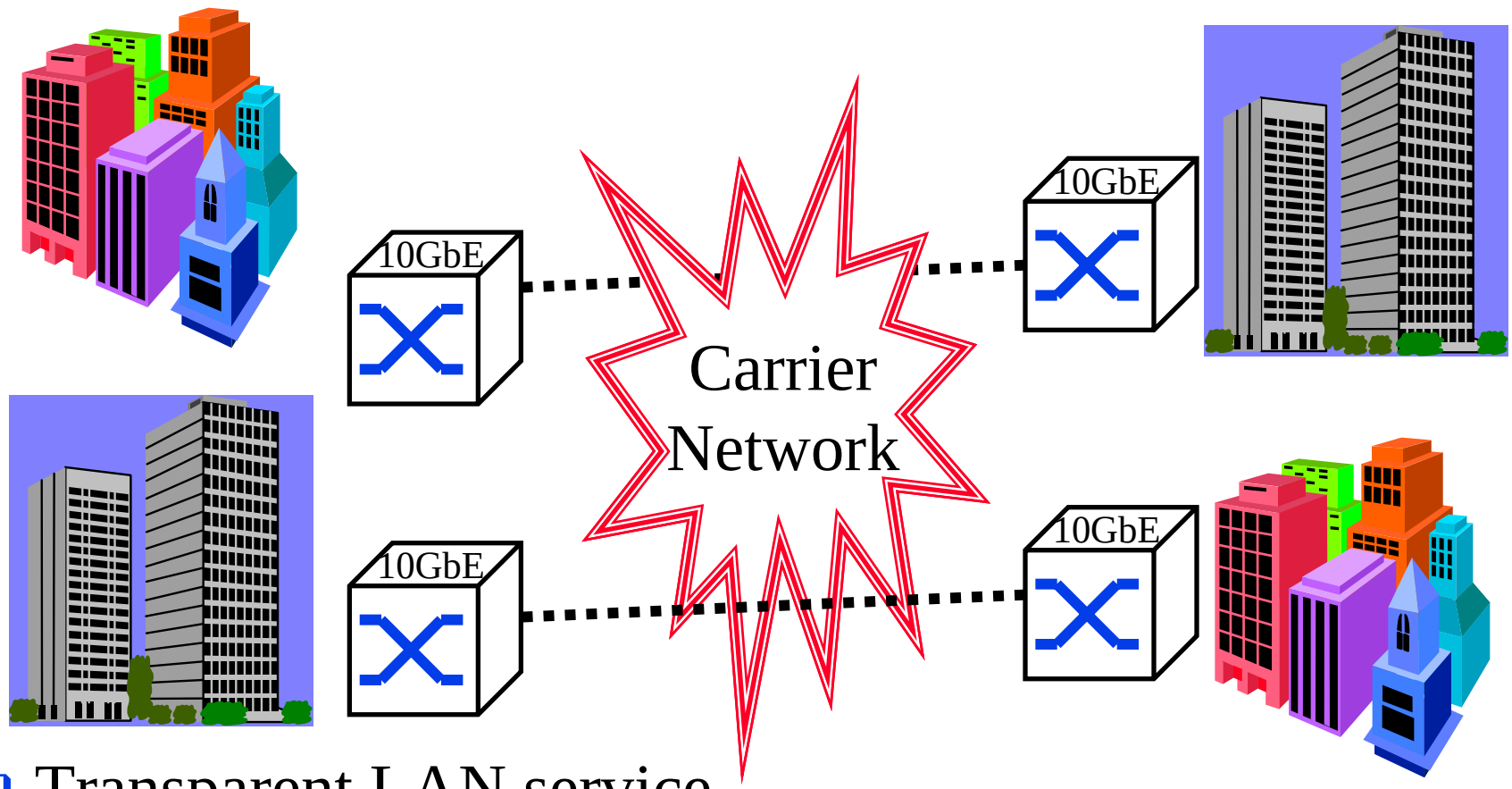
- Need only LAN PMD up to 40 km.
  No SONET overhead. No protection.

10 GbE over SONET/SDH



- Using WAN PMD.
  Legacy SONET. Protection via rings.
  ELTE = Ethernet Line Terminating Equipment

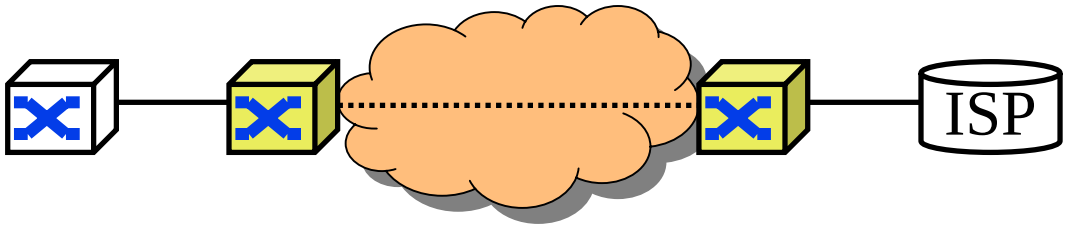
Metro Ethernet Services



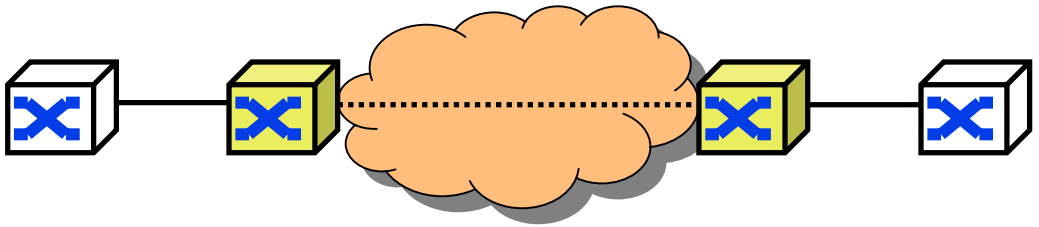
- Transparent LAN service

Virtual Private LAN Services (VPLS)

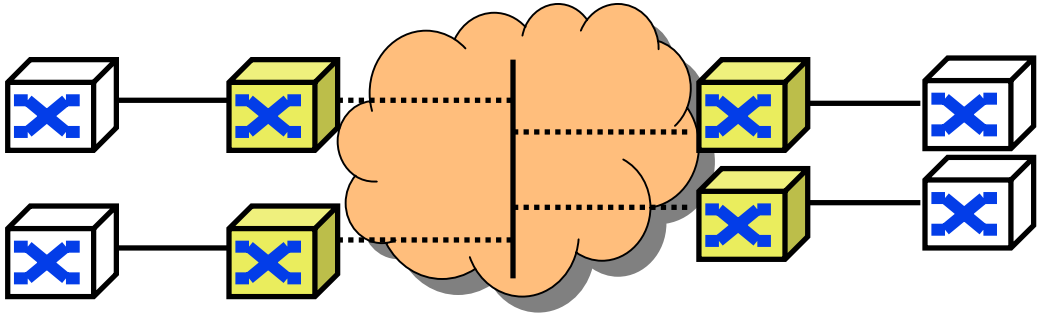
- Ethernet Internet Access



- Ethernet Virtual Private Line

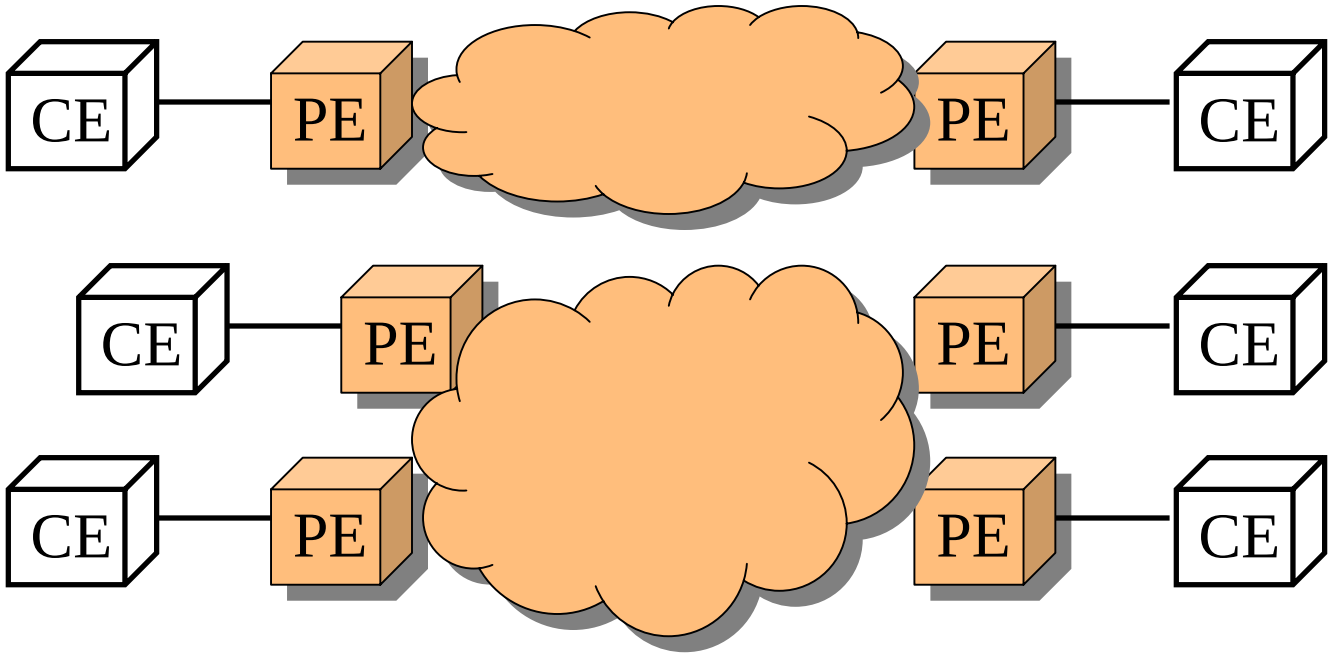


- Ethernet Virtual Private LAN



Metro Ethernet Services

- User-to-network Interface (UNI) = RJ45
- Ethernet Virtual Connection (EVC) = Flows
- Ethernet Line Service (ELS) = Point-to-point
- Ethernet LAN Service (E-LAN) = multipoint-to-multipoint



Enterprise vs Carrier Ethernet

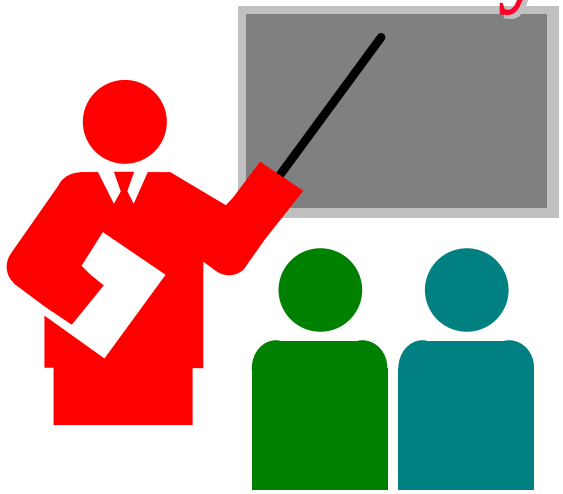
Enterprise

- Distance: up to 2km
- Scale:
  - Few K MAC addresses
  - 4096 VLANs
- Protection: Spanning tree
- Path determined by spanning tree
- Simple service
- Priority ⇒ Aggregate QoS
- No performance/Error monitoring (OAM)

Carrier

- Up to 100 km
- Millions of MAC Addresses
- Millions of VLANs Q-in-Q
- Rapid spanning tree (Gives 1s, need 50ms)
- Traffic engineered path
- SLA
- Need per-flow QoS
- Need performance/BER

Summary



- Ethernet/IEEE 802.3 initially used CSMA/CD.
- Distance bandwidth principle
  ⇒ Bit rate-distance-transmission size tradeoff
- Ethernet Standards:
  10Base5, 10Base2, 10Base-T, 100Base-TX, etc.
- Addresses: Local vs Global, Unicast vs Broadcast
- Spanning Tree
- Metro Ethernet Services

GbE, 10 GbE, RPR: Key References

- For a detailed list of references, see http://www.cse.wustl.edu/~jain/refs/gbe_refs.htm
  Also reproduced at the end of this tutorial book.
- Gigabit Ethernet Overview, http://www.cse.wustl.edu/~jain/cis788-97/gigabit_ethernet/index.htm
- 10 Gigabit Ethernet, http://www.cse.wustl.edu/~jain/cis788-99/10gbe/index.html
- 10 Gigabit Ethernet Alliance, http://www.10gea.org
- 10 GbE Resource Site, http://www.10gigabit-ethernet.com
- RPR Alliance, http://www.rpralliance.org/

References (Cont)

- IEEE 802.3 Higher Speed Study Group, http://grouper.ieee.org/groups/802/3/10G_study/public/index.html
- Email Reflector, http://grouper.ieee.org/groups/802/3/10G_study/email/thrd1.html
- IEEE 802.3ae 10Gb/s Ethernet Task Force, http://grouper.ieee.org/groups/802/3/ae/index.html
- IEEE 802.3ae email list, send a message with "subscribe stds-802-3-hssg <email adr>" in body to majordomo@majordomo.ieee.org

Quality of Service in IP Networks

Raj Jain
Professor of Computer Science and Engineering
Washington University in Saint Louis
Saint Louis, MO, USA
jain@acm.org
http://www.cse.wustl.edu/~jain/

Overview

- ATM QoS and Issues
- Integrated Services and RSVP
- Differentiated Services: Expedited and Assured Forwarding
- Subnet Bandwidth Manager (SBM)
- COPS Protocol for Policy
- IEEE 802.1D Model

ATM Classes of Service

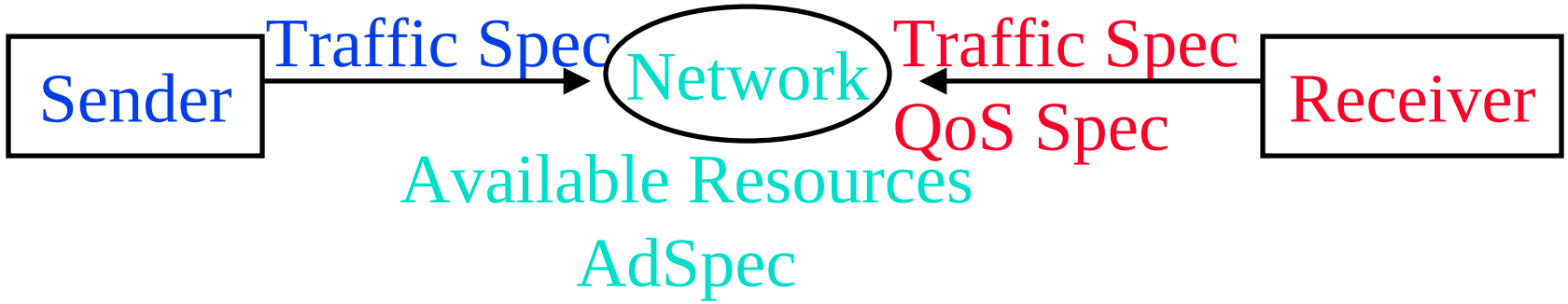
- ABR (Available bit rate): Source follows feedback. Max throughput with minimum loss.
- UBR (Unspecified bit rate): User sends whenever it wants. No feedback. No guarantee. Cells may be dropped during congestion.
- CBR (Constant bit rate): User declares required rate. Throughput, delay and delay variation guaranteed.
- VBR (Variable bit rate): Declare avg and max rate.
  - rt-VBR (Real-time): Conferencing. Max delay guaranteed.
  - nrt-VBR (non-real time): Stored video.
- GFR (Guaranteed Frame Rate): Min Frame Rate

Integrated Services

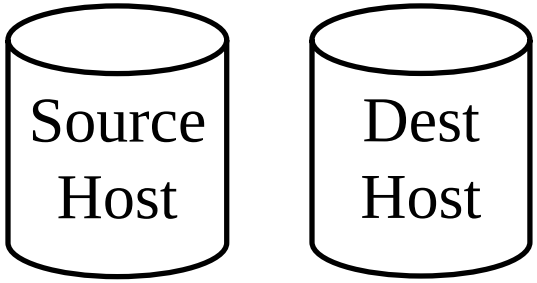
- Best Effort Service: Like UBR.
- Controlled-Load Service: Performance as good as in an unloaded datagram network. No quantitative assurances. Like nrt-VBR or UBR w MCR
- Guaranteed Service: rt-VBR
  - Firm bound on data throughput and delay.
  - Delay jitter or average delay not guaranteed or minimized.
  - Every element along the path must provide delay bound.
  - Is not always implementable, e.g., Shared Ethernet.
  - Like CBR or rt-VBR

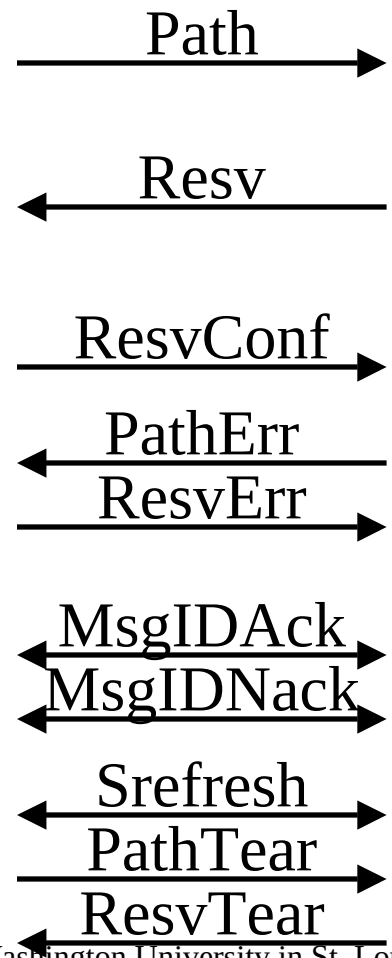
RSVP

- Resource ReSerVation Protocol
- Internet signaling protocol
- Carries resource reservation requests through the network including traffic specs, QoS specs, network resource availability
- Sets up reservations at each hop



RSVP Messages





- Path: Sender's traffic spec
  Path state is created. Regularly refereshed.
- Resv: Resv state is created. Regularly refereshed
- ResvConf: Sent upon request to confirm Resv
- PathErr, ResvErr: Error in path or reservation Installation
- MsgIdAck/Nack: Trigger reporting of an event
- Srefresh: To refresh a group of path/resv states
- PathTear/ResvTear: Remove Path/Resv States

Problems with RSVP and Integrated Services

- Complexity in routers: multi-field packet classification, scheduling
- Per-flow signaling, packet handling, state.
  O(n) $\Rightarrow$ Not scalable with # of flows.
  Number of flows in the backbone may be large.
  $\Rightarrow$ Suitable for small private networks
- Need a concept of "Virtual Paths" or aggregated flow groups for the backbone
- Need policy controls: Who can make reservations? Support for accounting and security.
  $\Rightarrow$ RSVP admission policy (rap) working group.

Problems (Cont)

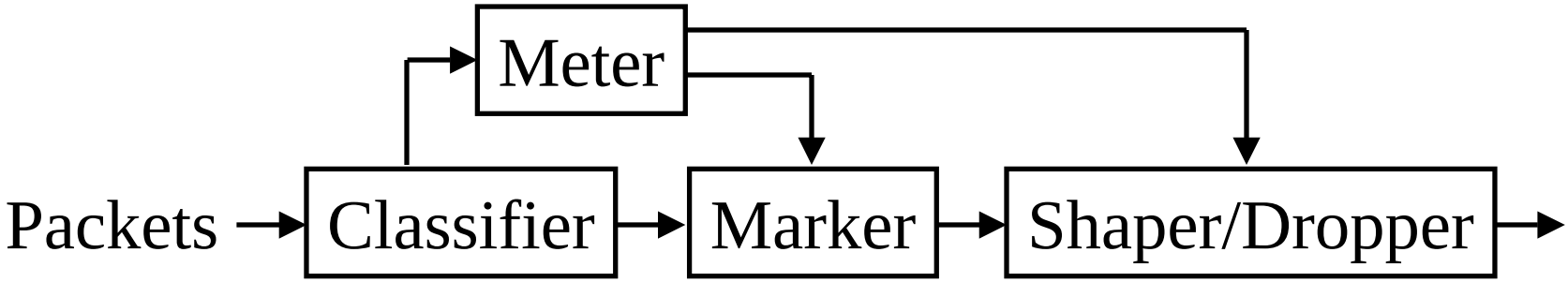
- Receiver Based:
  Need sender control/notifications in some cases.
  Which receiver pays for shared part of the tree?
- Soft State: Need route/path pinning (stability).
  Limit number of changes during a session.
- RSVP does not have negotiation and backtracking
- Throughput and delay guarantees require support of lower layers. Shared Ethernet ⇒ IP can't do GS or CLS. Need switched full-duplex LANs.
- RSVP is being revived to for MPLS and DiffServ signaling. Also, policy, aggregation, security concepts are being developed

Differentiated Services

- IPv4: 3-bit precedence + 4-bit ToS
- OSPF and integrated IS-IS can compute paths for each ToS
- Many vendors use IP precedence bits but the service varies ⇒ Need a standard ⇒ Differentiated Services
- DS working group formed February 1998
- Charter: Define ds byte (IPv4 ToS field)
- Mail Archive: http://www-nrg.ee.lbl.gov/diff-serv-arch/

DiffServ Concepts

- Micro-flow = A single application-to-application flow
- Traffic Conditioners: Meters (token bucket), Markers (tag), Shapers (delay), Droppers (drop)
- Behavior Aggregate (BA) Classifier:
  Based on DS byte only
- Multi-field (MF) Classifiers:
  Based on IP addresses, ports, DS-byte, etc..



Diff-Serv Concepts (Cont)

- Service: Offered by the protocol layer
  - Application: Mail, FTP, WWW, Video,...
  - Transport: Delivery, Express Delivery,...
    Best effort, controlled load, guaranteed service
  - DS group will not develop services
    They will standardize "Per-Hop Behaviors"

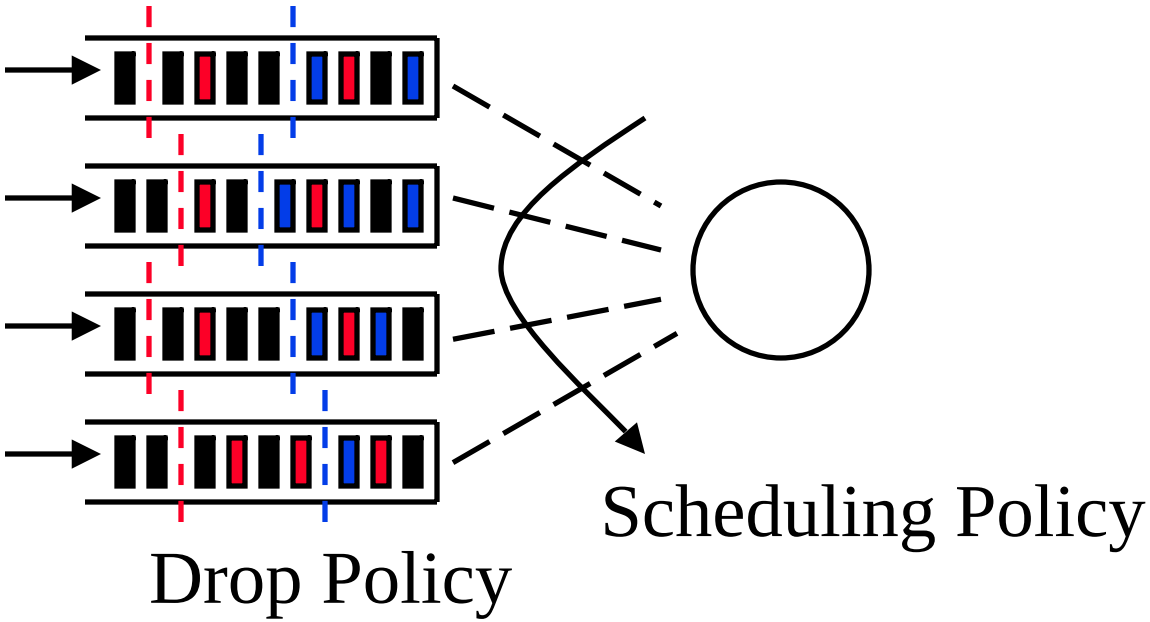
Per-hop Behaviors

- Externally Observable Forwarding Behavior
- x% of link bandwidth
- Minimum x% and fair share of excess bandwidth
- Priority relative to other PHBs
- PHB Groups: Related PHBs. PHBs in the group share common constraints, e.g., loss priority, relative delay

Expedited Forwarding

- Also known as "Premium Service"
- Virtual leased line
- Similar to CBR
- Guaranteed minimum service rate
- Policed: Arrival rate < Minimum Service Rate
- Not affected by other data PHBs
  $\Rightarrow$ Highest data priority (if priority queueing)
- Code point: 101 110

Assured Forwarding



- PHB Group
- Four Classes: No particular ordering.
  ⇒ Creates 4 distinct networks with specified QoS. Share unused capacity.
- Three drop preference per class

Assured Forwarding (Cont)

- DS nodes SHOULD implement all 4 classes and MUST accept all 3 drop preferences. Can implement 2 drop preferences.
- Similar to nrt-VBR/ABR/GFR
- Code Points:

Drop Prec.	Class 1	Class 2	Class 3	Class 4
Low	001 010	010 010	011 010	100 010
Medium	001 100	010 100	011 100	100 100
High	001 110	010 110	011 110	100 110

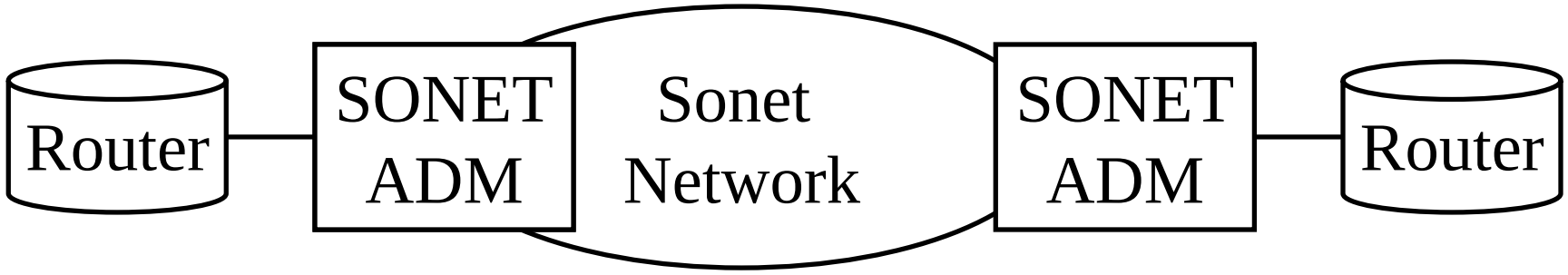
- Avoids xxx000 class selectors. Last bit 0 $\Rightarrow$ Standard

Per-Domain Behavior

- PDBs: Measurable edge to edge behavior across a cloud with same DS policies for all packets of a given PHB
- Existing PHBs have been extended to PDBs:
  - Virtual wire PDB: Based on EF
  - Assured Rate PDB: Based on AF.
    Min Rate. No delay or jitter guarantee
  - Bulk Handling PDB: Less than "Best Effort."
    Dropped if no resources. No need to police.

Problems with DiffServ

- per-hop ⇒ Need at every hop
  One non-DiffServ hop can spoil all QoS
- End-to-end ≠ Σ per-Hop
  Designing end-to-end services with weighted guarantees at individual hops is difficult.
- How to ensure resource availability inside the network?



DiffServ Problems (Cont)

- QoS is for the aggregate not micro-flows.
  - Large number of short flows are better handled by aggregates.
  - High-bandwidth flows (1 Mbps video) need per-flow guarantees.
- Designed for static Service Level Agreements (SLAs) Both the network topology and traffic are highly dynamic.
- Need route pinning or connections.
- Not all DSCPs used by all vendors/providers. DSCPs rewritten at domain boundaries.

QoS Debate Issues

- Massive Bandwidth vs Managed Bandwidth
- Per-Flow vs Aggregate
- Source-Controlled vs Receiver Controlled
- Soft State vs Hard State
- Path based vs Access based
- Quantitative vs Qualitative
- Absolute vs Relative
- End-to-end vs Per-hop
- Static vs Feedback-based
- One-way multicast vs n-way multicast
- Homogeneous multicast vs heterogeneous multicast
- Single vs multiple bottlenecks: Scheduling

Comparison of QoS Approaches

Issue	ATM	IntServ	DiffServ	MPLS	IEEE 802.3D
Massive Bandwidth vs Managed Bandwidth	Managed	Managed	Massive	Managed	Massive
Per-Flow vs Aggregate	Both	Per-flow	Aggregate	Both	Aggregate
Source-Controlled vs Receiver Controlled	Unicast Source, Multicast both	Receiver	Ingress	Both	Source
Soft State vs Hard State	Hard	Soft	None	Hard	Hard
Path based vs Access based	Path	Path	Access	Path	Access
Quantitative vs Qualitative	Quantitative	Quantitative+Qualitative	Mostly qualitative	Both	Qualitative
Absolute vs Relative	Absolute	Absolute	Mostly Relative	Absolute plus relative	Relative

Comparison (Cont)

Issue	ATM	IntServ	DiffServ	MPLS	IEEE 802.3D
End-to-end vs Per-hop	e-e	e-e	Per-hop	e-e	Per-hop
Static vs Feedback-based	Both	Static	Static	Static	Static
One-way multicast vs n-way multicast	Only one-way				
Homogeneous multicast vs heterogeneous multicast	Homogeneous	Heterogeneous	N/A	Homogeneous	N/A
Single vs multiple bottlenecks: Scheduling	Multiple bottleneck	Multiple		Multiple	

Summary

1. ATM: CBR, VBR, ABR, UBR, GFR
2. Integrated Services: GS = rtVBR, CLS = nrt-VBR
3. Signaling protocol: RSVP
4. Differentiated Services will use the DS byte
5. 802.1D allows priority

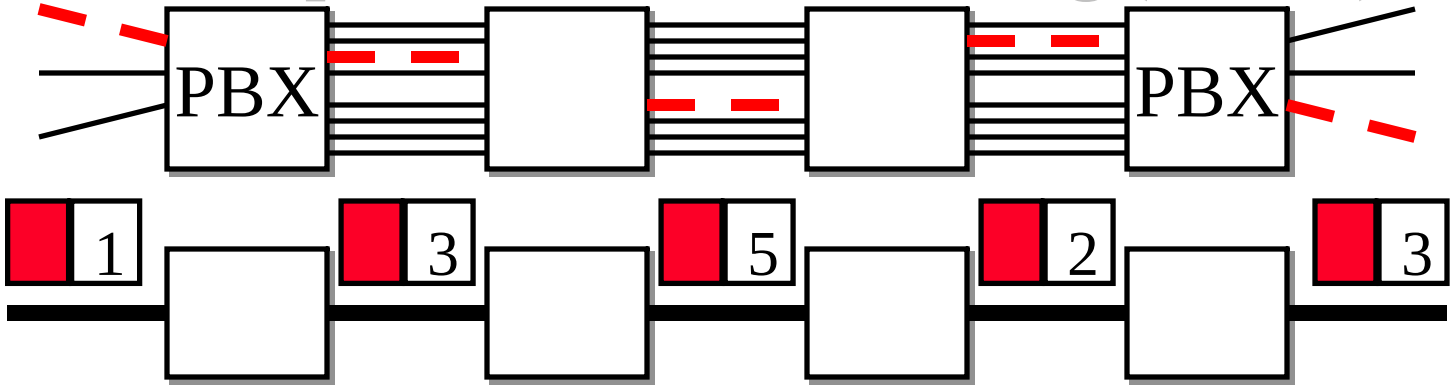
Multi-Protocol Label Switching (MPLS)

Raj Jain
Professor of Computer Science and Engineering
Washington University in Saint Louis
Saint Louis, MO, USA
jain@acm.org
http://www.cse.wustl.edu/~jain/

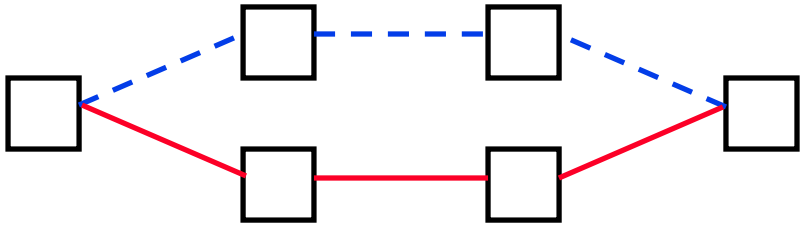
Overview

- Routing vs Switching
- Label Stacks
- Label Distribution Protocol (LDP)
- RSVP Extensions
- Traffic Engineering
- Traffic Trunks
- Traffic Engineering Extensions to OSPF and IS-IS

Multiprotocol Label Switching (MPLS)



- Allows virtual circuits in IP Networks (May 1996)
- Each packet has a virtual circuit number called 'label'
- Label determines the packet's queuing and forwarding
- Circuits are called Label Switched Paths (LSPs)
- LSP's have to be set up before use
- Allows traffic engineering

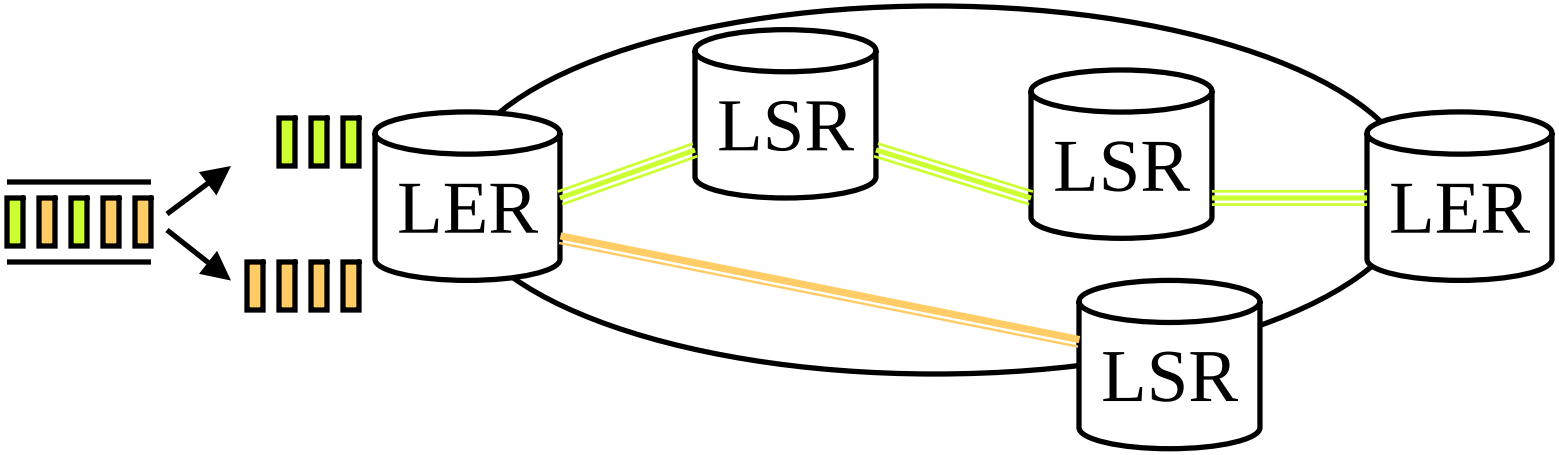


Routing vs Switching

- Routing: Based on address lookup. Max prefix match.
  ⇒ Search Operation
  ⇒ Complexity ≈ $O(\log_2 n)$
- Switching: Based on circuit numbers
  ⇒ Indexing operation
  ⇒ Complexity O(1)
  ⇒ Fast and Scalable for large networks and large address spaces
- These distinctions apply on all datalinks: ATM, Ethernet, SONET

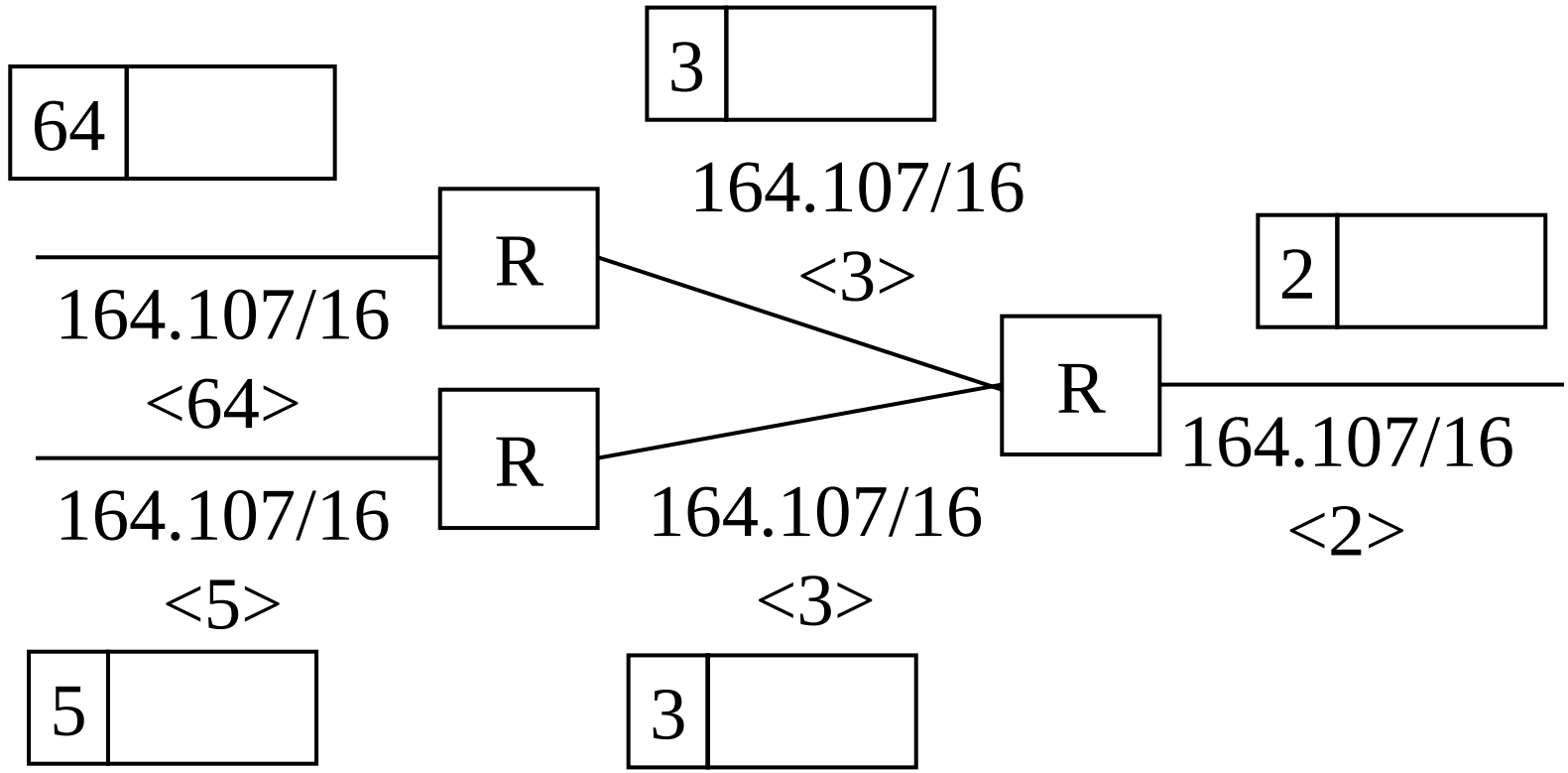
MPLS Terminology

- Label Edge Router (LER)
- Label Switching Router (LSR)
- Label Switched Path (LSP)
- Forwarding Equivalence Class (FEC)



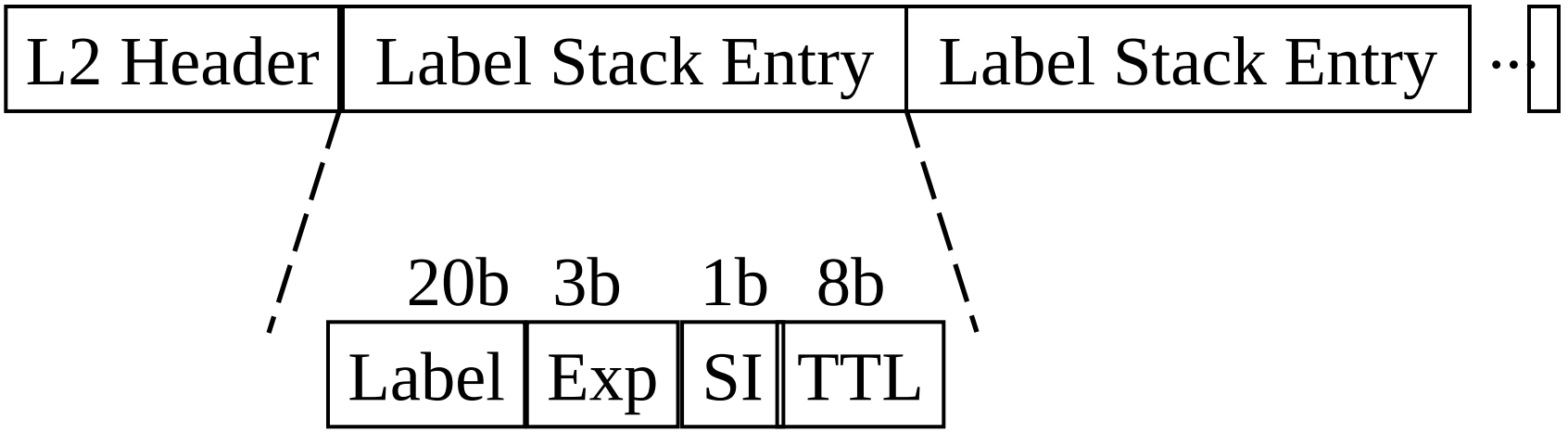
Label Switching Example

- One VC per routing table entry

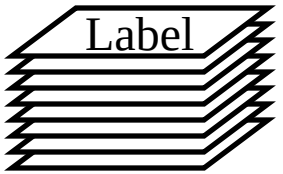


Label Stack Entry Format

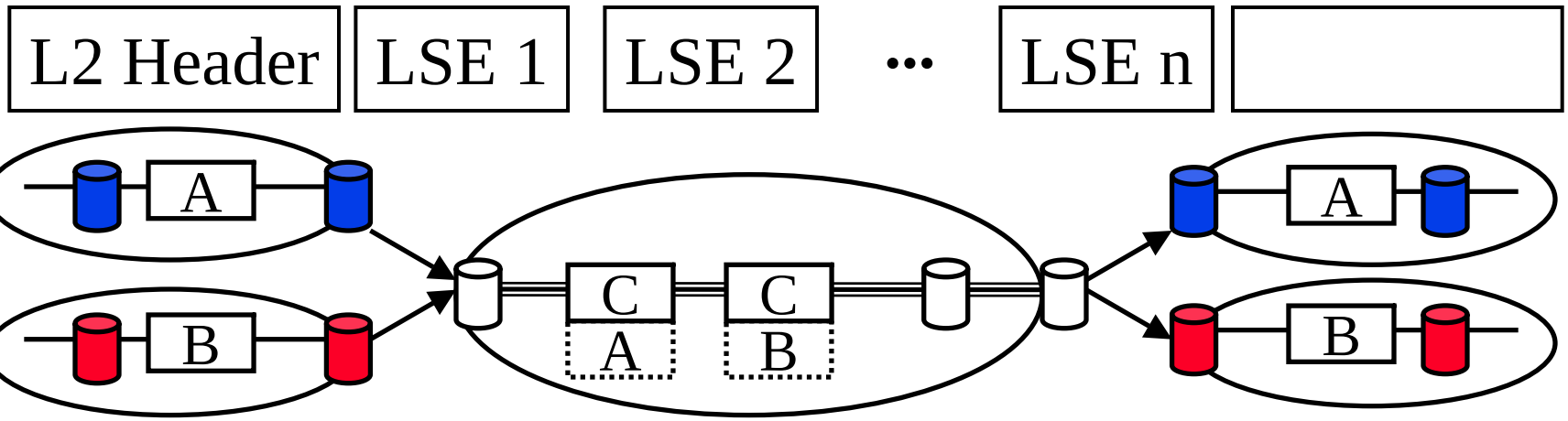
- Labels = Explicit or implicit L2 header
- TTL = Time to live
- Exp = Experimental
- SI = Stack indicator



Label Stacks



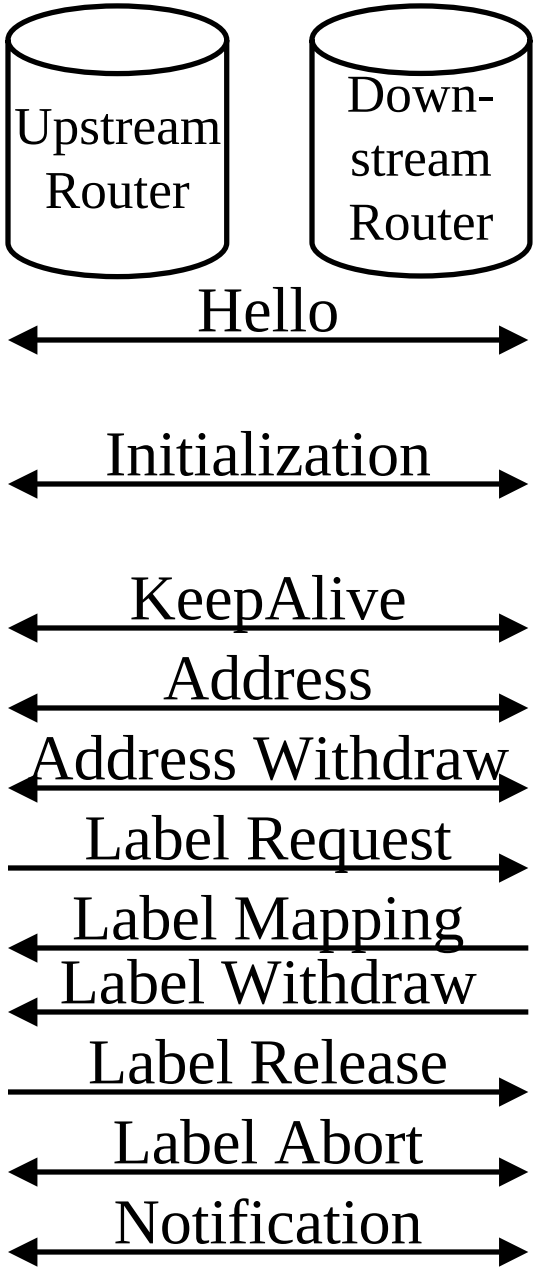
- Labels are pushed/popped as they enter/leave MPLS domain
- Routers in the interior will use Interior Gateway Protocol (IGP) labels. Border gateway protocol (BGP) labels outside.
- Bottom label may indicate protocol (0=IPv4, 2=IPv6)



Label Assignment

- Unsolicited: Topology driven ⇒ Routing protocols exchange labels with routing information.
  Many existing routing protocols are being extended: BGP, OSPF
- On-Demand:
  ⇒ Label assigned when requested,
  e.g., when a packet arrives ⇒ latency
- Label Distribution Protocol called LDP
- RSVP has been extended to allow label request and response

LDP Messages



- Hello: Sent periodically to "all routers on this subnet"
- Initialization: parameters and capability exchange
- KeepAlive: Keep the LDP session alive
- Address (advertise interface addresses)
- Address Withdraw
- Label Request
- Label Mapping (Label Response)
- Label Withdraw (by downstream)
- Label Release (by upstream)
- Label Abort Request
- Notification: Error/advisory information

CR-LDP

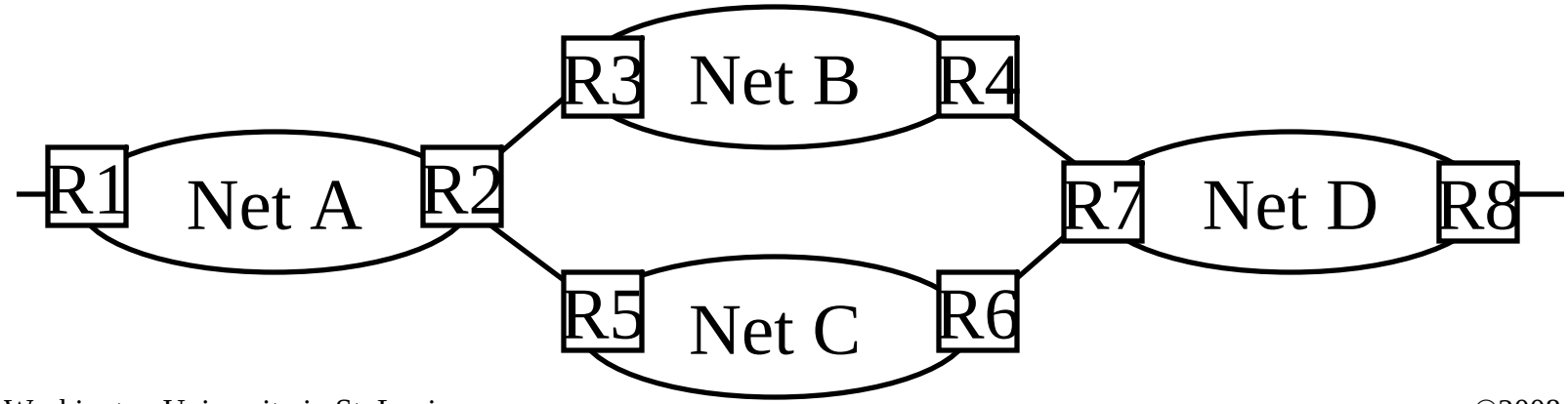
- Extension of LDP for constraint-based routing (CR)
- New Features:
  - Traffic parameters
  - Explicit Routing with Egress Label
  - Preemption of existing route. Based on holding priorities and setup priorities
  - Route pinning: To prevent path changes
  - Label Set: Allows label constraints (wavelengths)
- No new messages
- Enhanced Messages: Label request, Label Mapping, Notification

RSVP Extensions

- Explicit Route Object (ERO): Path messages are forced to go along specified explicit route
- Record Route
- Message Bundling: Multiple messages in one packet
- Refresh Reduction: Srefresh refreshes all reservations related to a given message ID
- Node Failure Detection: Keep alive hello messages
- Quick Fault Notify: Notify msg direct to initiator (and terminator if bidirectional). Multi failures in one msg.
- Aggregation: Resv messages include diffserv marking (DSCP code) or 802.1p tag for the upstream node
- Security: Flow = Dest IP + IPSec Protocol Type + Security Parameter Index (SPI) = Security Association

Explicit Route

- Explicit route specified as a list of Explicit Route Hops (group of nodes)
- Hops can include IPv4 prefix, IPv6 prefix, MPLS tunnels or Autonomous systems
- Example: R1-R2-Net B-R7-R8
- Allows traffic engineering

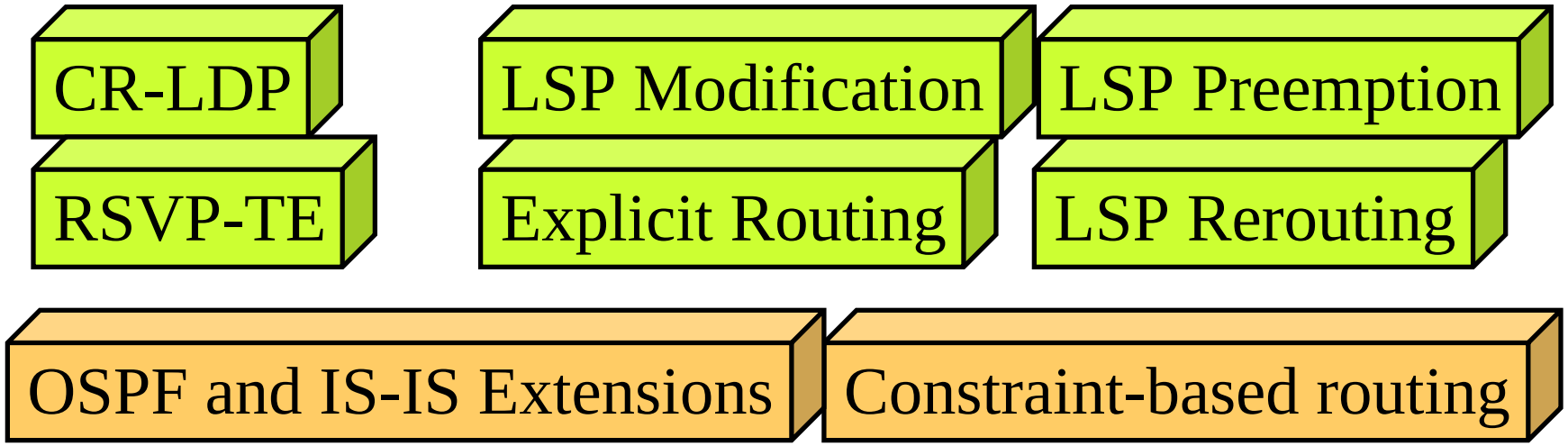


Hop-by-Hop vs Explicit Routing

Issue	Hop-by-hop	Explicit
Topology Awareness	Everywhere	Edge only
Circuit Management	None	LSP setup/ teardown
Signaling	Not required	Requires LDP or RSVP-TE
Recovery Time	Convergence time of routing Protocol	Path switch time
Routing	Fixed	QoS, Policy, or arbitrary
Traffic Engineering	Difficult	Easy

Traffic Engineering Building Blocks

- TE = Directing the traffic to where the capacity exists
- CR-LDP and RSVP-TE allow LSP explicit routing, rerouting, modification, preemption.
- OSPF and IS-IS are being modified to allow constraints



Draft Martini

- 1995-1999: IP over ATM,
  Packet over SONET,
  IP over Ethernet

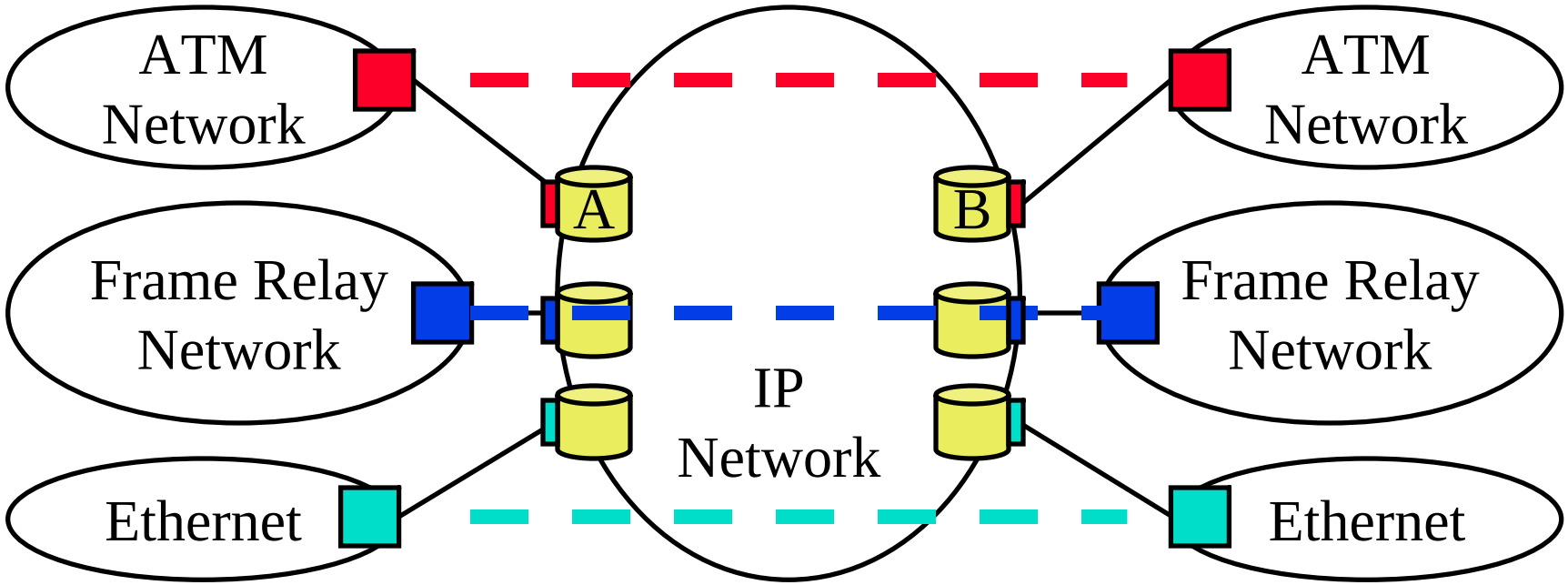
IP		
Ethernet	ATM	PPP

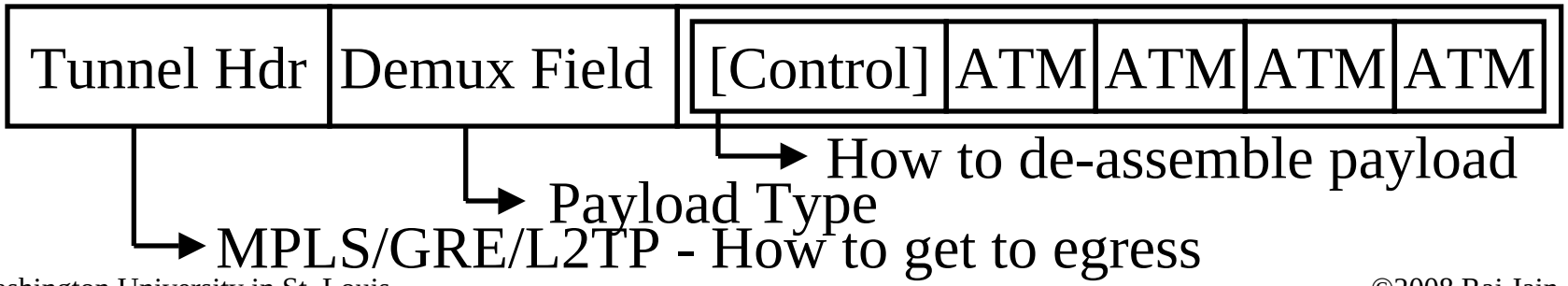
- 2000+: ATM over IP
  Ethernet over IP
  SONET over IP

Ethernet	ATM	PPP
IP		

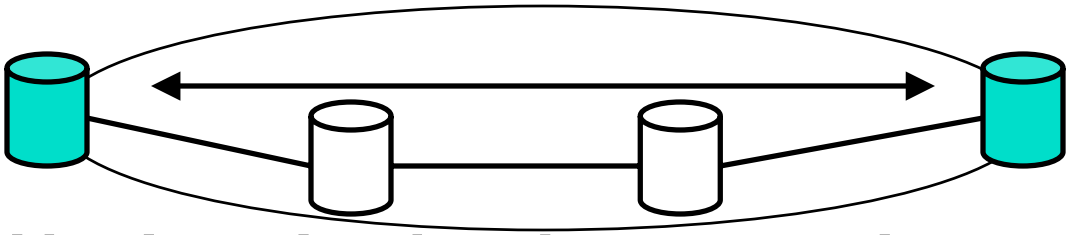
- Ref: draft-martini-*.txt

L2 Circuits over IP





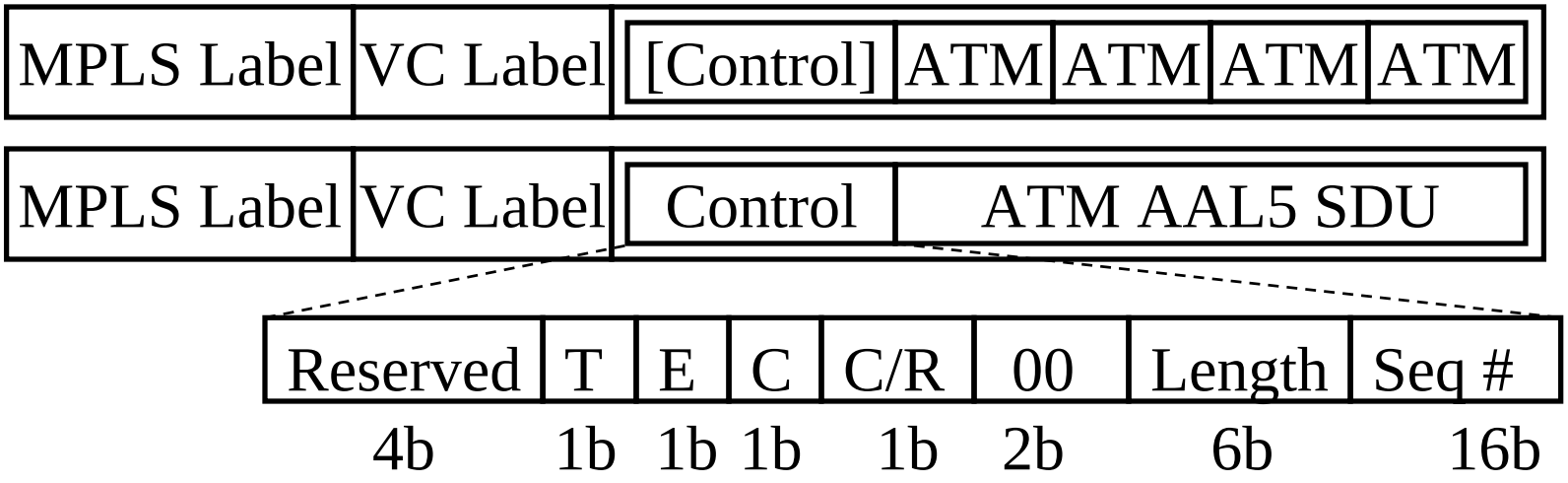
VC Label



- VC Label bindings distributed using LDP downstream unsolicited mode between ingress and egress LSRs
- Circuit specific parameters such as MTU, options are exchanged at the time VC Label exchange
- VC Label: S=1 ⇒ Bottom of stack, TTL=2
- VC Type:

1 Frame Relay DLCI	6 HDLC
2 ATM AAL5 VCC Transport	7 PPP
3 ATM Transparent Cell Transport	8 Circuit Emulation
4 Ethernet VLAN	9 ATM VCC Cell Transport
5 Ethernet	10 ATM VPC Cell Transport

ATM over MPLS

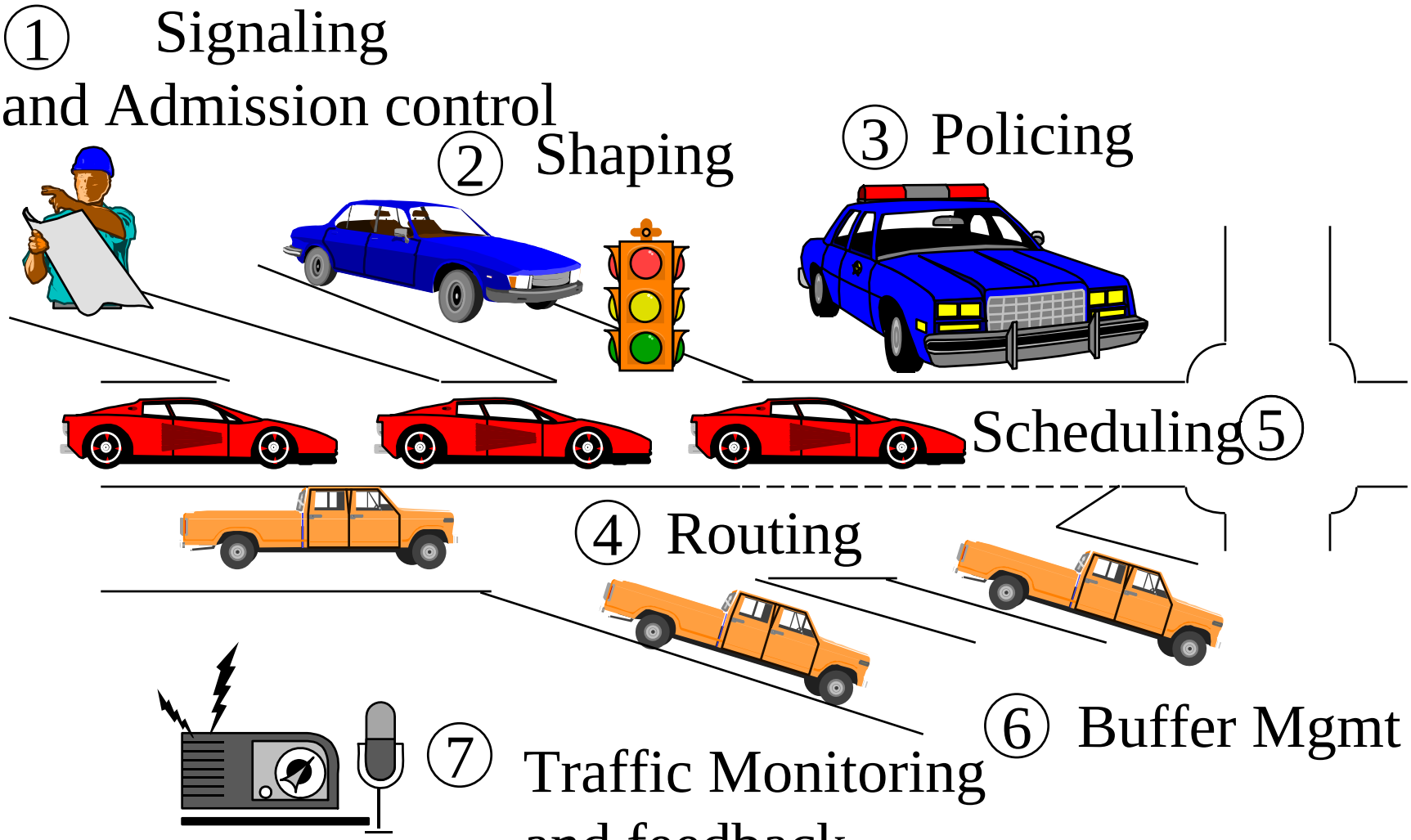


- T = Transport Type: 0=> Cells, 1=> SDU
- E = EFCI
- C = CLP
- C/R = Command/Response
- Length of payload + Control word
  0 => Greater than or equal to 64 bytes
- Ref: draft-martini-atm-encap-mpls-00.txt

Traffic Engineering Objectives

- User's Performance Optimization
  ⇒ Maximum throughput, Min delay, min loss, min delay variation
- Efficient resource allocation for the provider
  ⇒ Efficient Utilization of all links
  ⇒ Load Balancing on parallel paths
  ⇒ Minimize buffer utilization
  - Current routing protocols (e.g., RIP and OSPF) find the shortest path (may be over-utilized).
- QoS Guarantee: Selecting paths that can meet QoS
- Enforce Service Level agreements
- Enforce policies: Constraint based routing ⊇ QoSR

Traffic Engineering Components



Traffic Engineering Components

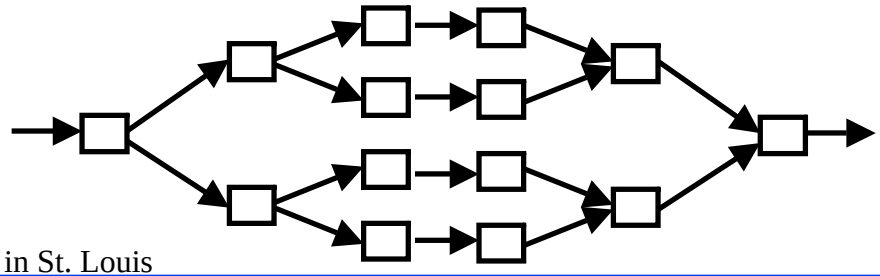
1. Signaling: Tell the network about traffic and QoS.
   Admission Control: Network may deny the request.
2. Shaping: Smoothen the bursts
3. Policing: Ensure that users are following rules
4. Routing: Path Selection, Request Prioritization, Preemption, Re-optimization/Pinning, Fault Recovery
5. Scheduling: Weight, Prioritization, Preemption
6. Buffer Management: Drop Thresholds, Drop Priority
7. Feedback: Implicit, Explicit

Accounting/Billing

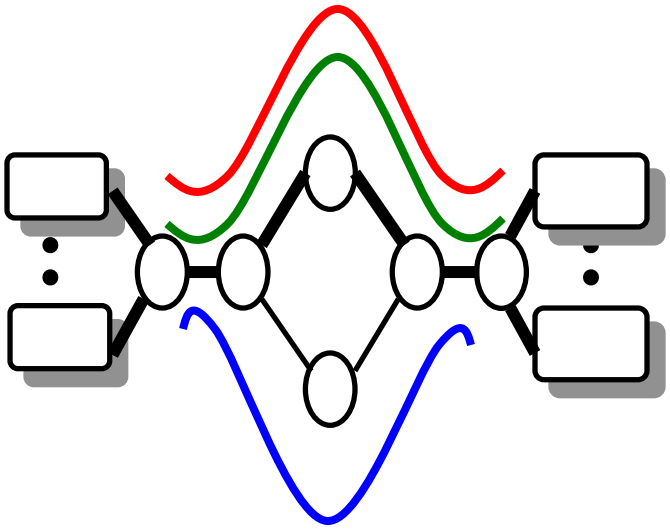
Performance Monitoring/Capacity Planning

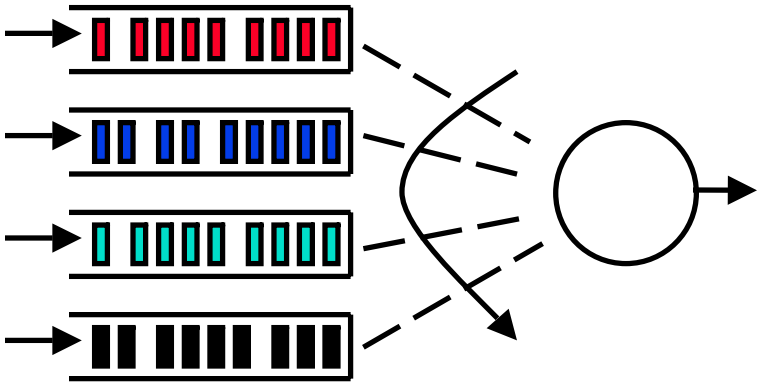
MPLS Mechanisms for TE

- Signaling, Admission Control, Routing
- Explicit routing of LSPs
- Constrained based routing of LSPs
  Allows both Traffic constraints and Resource Constraints (Resource Attributes)
- Hierarchical division of the problem (Label Stacks)
- Traffic trunks allow aggregation and disaggregation (Shortest path routing allows only aggregation)



Traffic Trunks

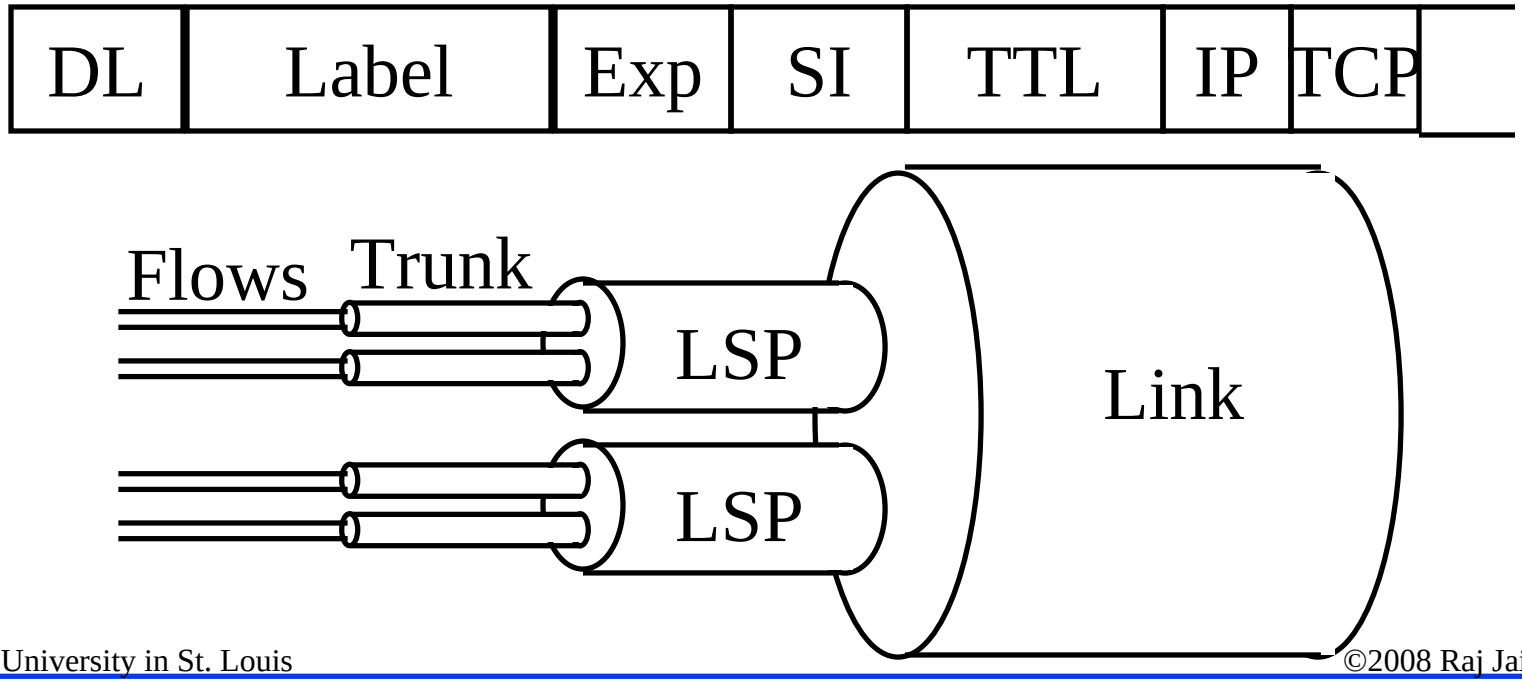




- Trunk: Aggregation of flows of same class on same LSP
- Trunks are routable
  ⇒ LSP through which trunk passes can be changed
- Class ⇒ Queue, LSP ⇒ Next hop
  Class can be coded in Exp or Label field. Assume Exp.

Flows, Trunks, LSPs, and Links

- Label Switched Path (LSP):
  Path for all packets with the same label
- Trunk: Same Label+Exp
- Flow: Same MPLS+IP+TCP headers



Traffic Trunks

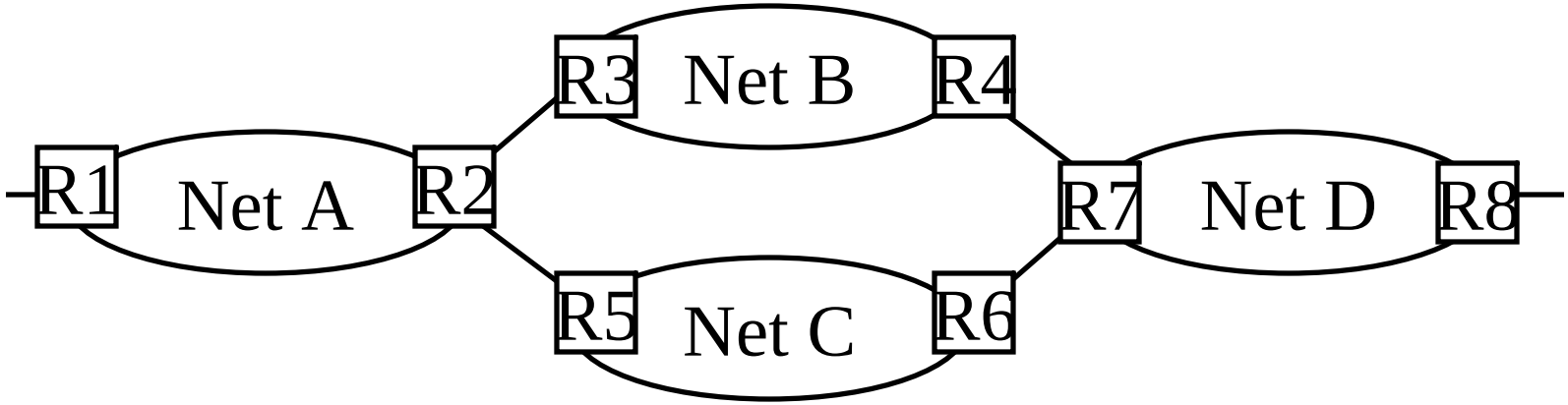
- Each traffic trunk can have a set of associated characteristics, e.g., priority, preemption, policing
- Some trunks may preempt other trunks. A trunk can be preemptor, non-preemptor, preemptable, or non-preemptable.
- Trunk paths are setup based on policies or specified resource availability.
- A traffic trunk can have alternate sets of paths in case of failure of the main path. Trunks can be rerouted.
- Multiple LSPs can be used in parallel to the same egress.

Trunk Attributes

- **Signaling**: Routing Protocols, RSVP, CR-LDP
- **Admission Control**: Network may deny the request.
- **Policing**: Token Bucket
- Shaping: Smoothen the bursts
- **Routing: Path Selection, Request Prioritization, Preemption, Re-optimization/Pinning, Fault Recovery**
- Scheduling: Class Weight, Prioritization, Preemption
- Buffer Management: Class drop thresholds/priority
- Feedback: Implicit, Explicit (ICMP being discussed)
- Accounting/Billing
- Performance Monitoring/Capacity Planning

Explicit Route

- Explicit route specified as a list of Explicit Route Hops (group of nodes)
- Hops can include IPv4 prefix, IPv6 prefix, MPLS tunnels or Autonomous systems
- Example: R1-R2-Net B-R7-R8



Explicit Route (Cont)

- All or a subset may be traversed
- The list is specified by edge router based on imperfect info (Strict/loose)
  - Strict ⇒ Path must include only nodes from the previous and this abstract node
  - Loose ⇒ path between two nodes may include other nodes
- Managed like ATM PNNI Designated Transit Lists (DTLs)

Path Selection

- Manual/Administrative
- Dynamically computed
- Explicitly specified: Partially/fully, strict/loose, Mandatory/non-mandatory, Single/Set
- Non-Mandatory
  $\Rightarrow$ Use any available path if specified not available
- Set $\Rightarrow$ Preference ordered list
- Resource class affinity

Resource Attributes

- Capacity
- Overbooking Factor: Maximum Allocation Multiplier
- Class: Allows policy enforcement
- Class Examples: secure/non-secure, transit/local-only
- A resource can be member of multiple classes

Resource Class Affinity

- Each resource has a class
- Affinity = Desirability
- Binary Affinity: 0 ⇒ Must Exclude,
  1 ⇒ Must Include, Not-specified ⇒ Don't care
- <Class, affinity> pair can be used to implement policies

Adaptivity and Resilience

- Stability: Route pinning
- Resource availability is dynamic
- Trunks can live for long time
- Adaptivity: Re-optimization when availability changes
- Resilience: Reroute if path breaks
- Adaptivity $\Rightarrow$ Resilience. Resilience /$\Rightarrow$ Adaptivity
- Idea: Adaptivity is not binary $\Rightarrow$ Rerouting period

Priority and Preemption

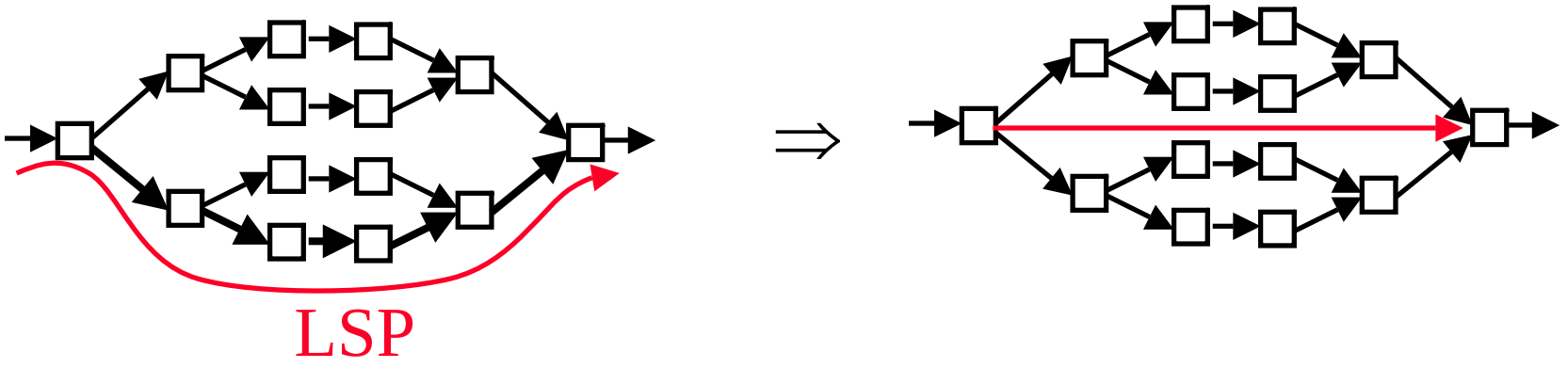
- Preemptor-enabled: Can preempt other trunks
- Non-Preemptor: Can't preempt other trunks
- Preemptable: Can be preempted by other trunks
- Non-Preemptable: Can't be preempted by other trunks
- These attributes and priority are used to decide preemption

Traffic Engineering Extensions to OSPF

- Add to Link State Advertisements:
- TE Metric: May be different from standard OSPF link metric
- Maximum bandwidth
- Maximum Reservable Bandwidth:
  May be more than maximum bandwidth
- Unreserved Bandwidth
- Resource Class/color
- Ref: draft-katz-yeung-ospf-traffic-00.txt

TE Extensions to OSPF (Cont)

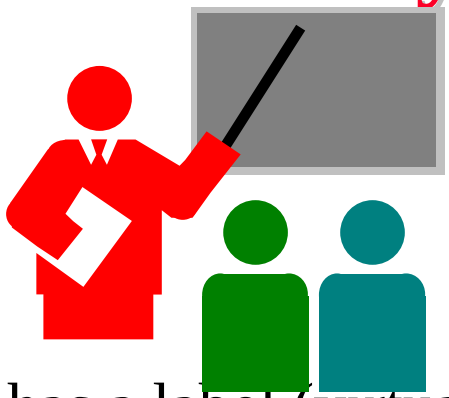
- Link Delay and Link Loss rate also proposed in draft-wimer-ospf-traffic-00.txt
- In path calculations, TE tunnels are used as links to tunnel egress



Traffic Engineering Extensions to IS-IS

- Add to Link State Protocol Data Units:
- TE Metric
- Maximum bandwidth
- Maximum Reservable Bandwidth: May be more than maximum bandwidth
- Unreserved Bandwidth
- Resource Class/color
- Ref: draft-ietf-isis-traffic-01.txt

Summary



- MPLS: Each packet has a label (virtual circuit number)
- Label Distribution Protocol (LDP) and RSVP are used for label distribution.
- MPLS traffic trunks are like ATM VCs that can be routed based on explicit route or policies
- CR-LDP allows explicit routing, constraint-based routing, traffic parameters, and QoS
- OSPF and IS-IS is being modified for traffic engg

Label Switching: Key References

- See http://www.cse.wustl.edu/~jain/refs/ipsw_ref.htm
  Also reproduced at the end of this tutorial book.
- Multiprotocol Label Switching (mpls) working group at IETF. Email: mpls-request@cisco.com
- IP Switching, http://www.cse.wustl.edu/~jain/cis788-97/ip_switching/index.htm
- IP Switching and MPLS, http://www.cse.wustl.edu/~jain/cis777-00/g_fipsw.htm
- MPLS Resource Center, http://www.mplsrc.com

Routing Protocols

Raj Jain
Professor of Computer Science and Engineering
Washington University in Saint Louis
Saint Louis, MO, USA
jain@acm.org
http://www.cse.wustl.edu/~jain/

- Building Routing Tables
- Routing Information Protocol Version 1 (RIP V1)
- RIP V2
- OSPF
- BGP and IDRP.

Autonomous Systems

- An internet connected by homogeneous routers under the administrative control of a single entity

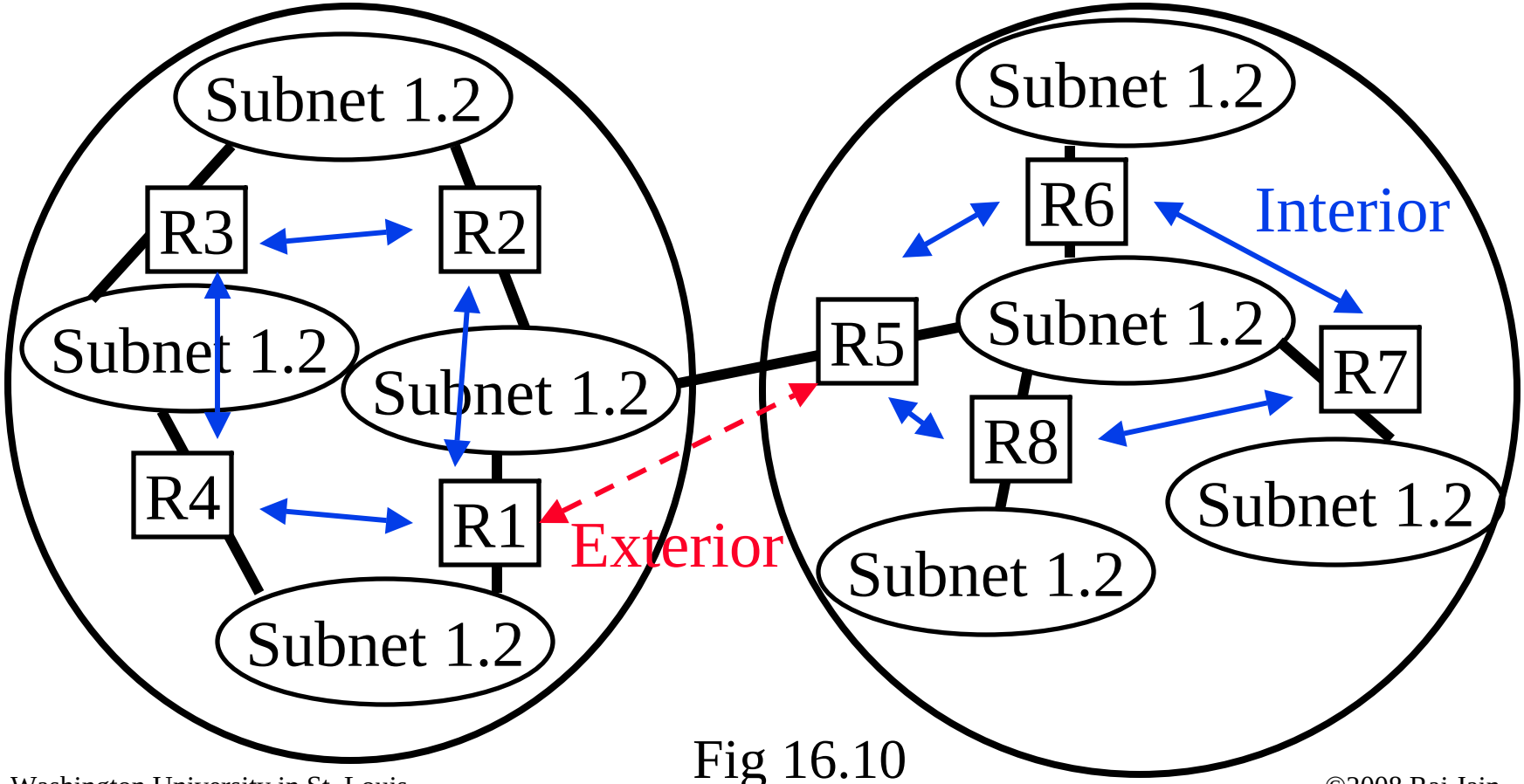


Fig 16.10

Routing Protocols

- Interior Router Protocol (IRP): Used for passing routing information among routers internal to an autonomous system. Also known as IGP.
  - Examples: RIP, OSPF
- Exterior Router Protocol (ERP): Used for passing routing information among routers between autonomous systems. Also known as EGP.
  - Examples: EGP, BGP, IDRP
    Note: EGP is a class as well as an instance in that class.

Routing Information Protocol

- RIP uses distance vector $\Rightarrow$ A vector of distances to all nodes is sent to neighbors
- Each router computes new distances:
  - Replace entries with new lower hop counts
  - Insert new entries
  - Replace entries that have the same next hop but higher cost
  - Each entry is aged.
    Remove entries that have aged out
- Send out updates every 30 seconds.

Distance-Vector Example

Destination	Delay	Next node
1	0	Ñ
2	2	2
3	5	3
4	1	4
5	6	3
6	8	3

D^1 S^1

(a) Node 1ås routing table before update

D^2	D^3	D^4
2	3	1
0	3	2
3	0	2
2	2	0
3	1	1
5	3	3

(b) Delay vectors sent t neighbor nodes

Destination	Delay	Next node
1	0	Ñ
2	2	2
3	3	4
4	1	4
5	2	4
6	4	4

$1^{1,2}$ = 2

$1^{1,3}$ = 5

$1^{1,4}$ = 1

(c) Node 1ås routing table after update and link c

Fig 9.9 Stallings

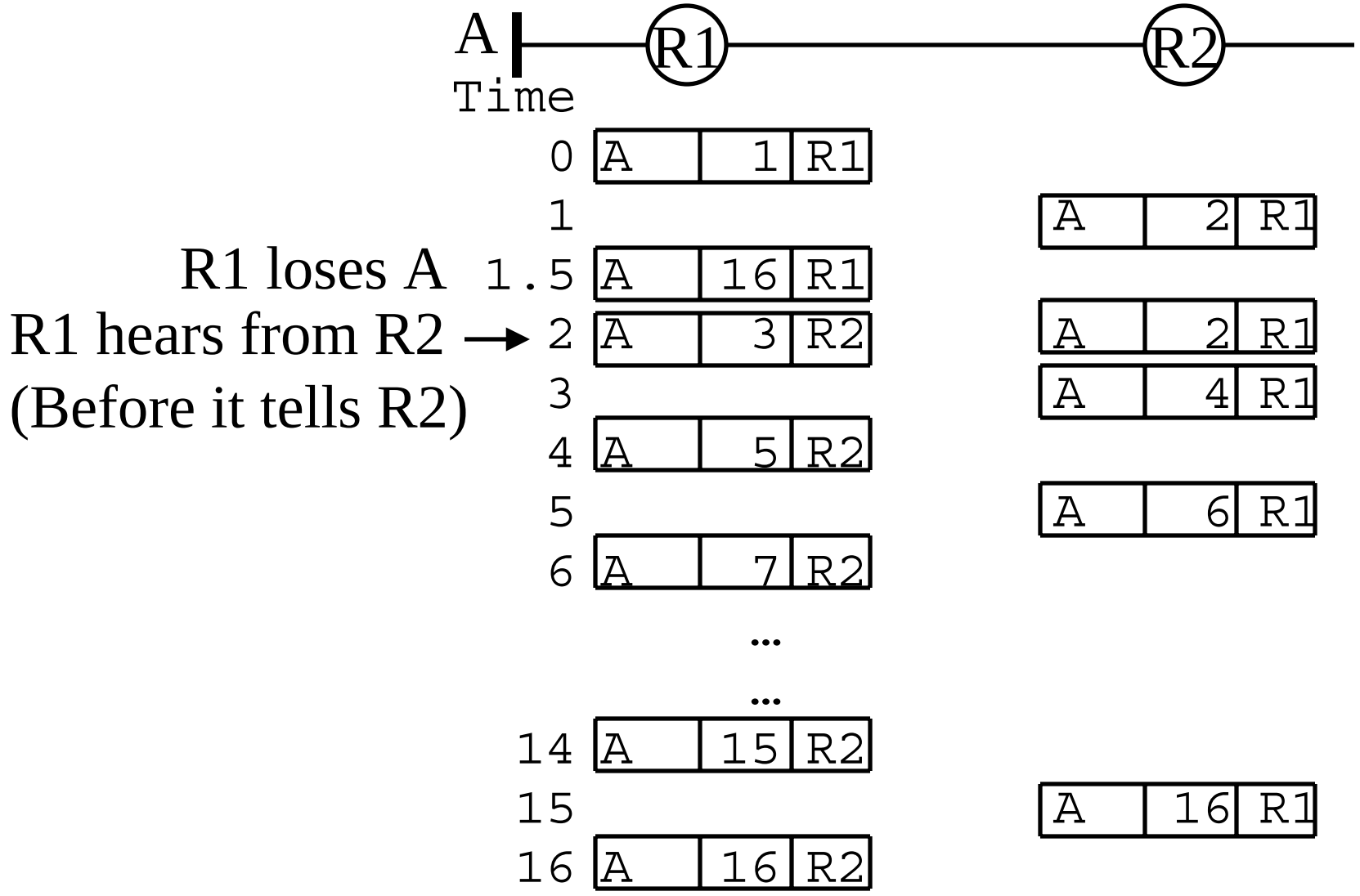
RIP V1

- RFC 1058 adopted in 1988
- Implemented in Berkeley UNIX as “routed” (pronounced route d)
- Both hosts and routers can implement RIP
- Hosts use passive mode Þ Do not send out updates
- Runs on UDP
- RIP packets do not leave local network

Shortcomings of RIP

- Maximum network diameter = 15 hops
- Cost is measured in hops
  Only shortest routes. May not be the fastest route.
- Entire tables are broadcast every 30 seconds.
  Bandwidth intensive.
- Uses UDP with 576-byte datagrams.
  Need multiple datagrams.
  300-entry table needs 12 datagrams.
- An error in one routing table is propagated to all routers
- Slow convergence

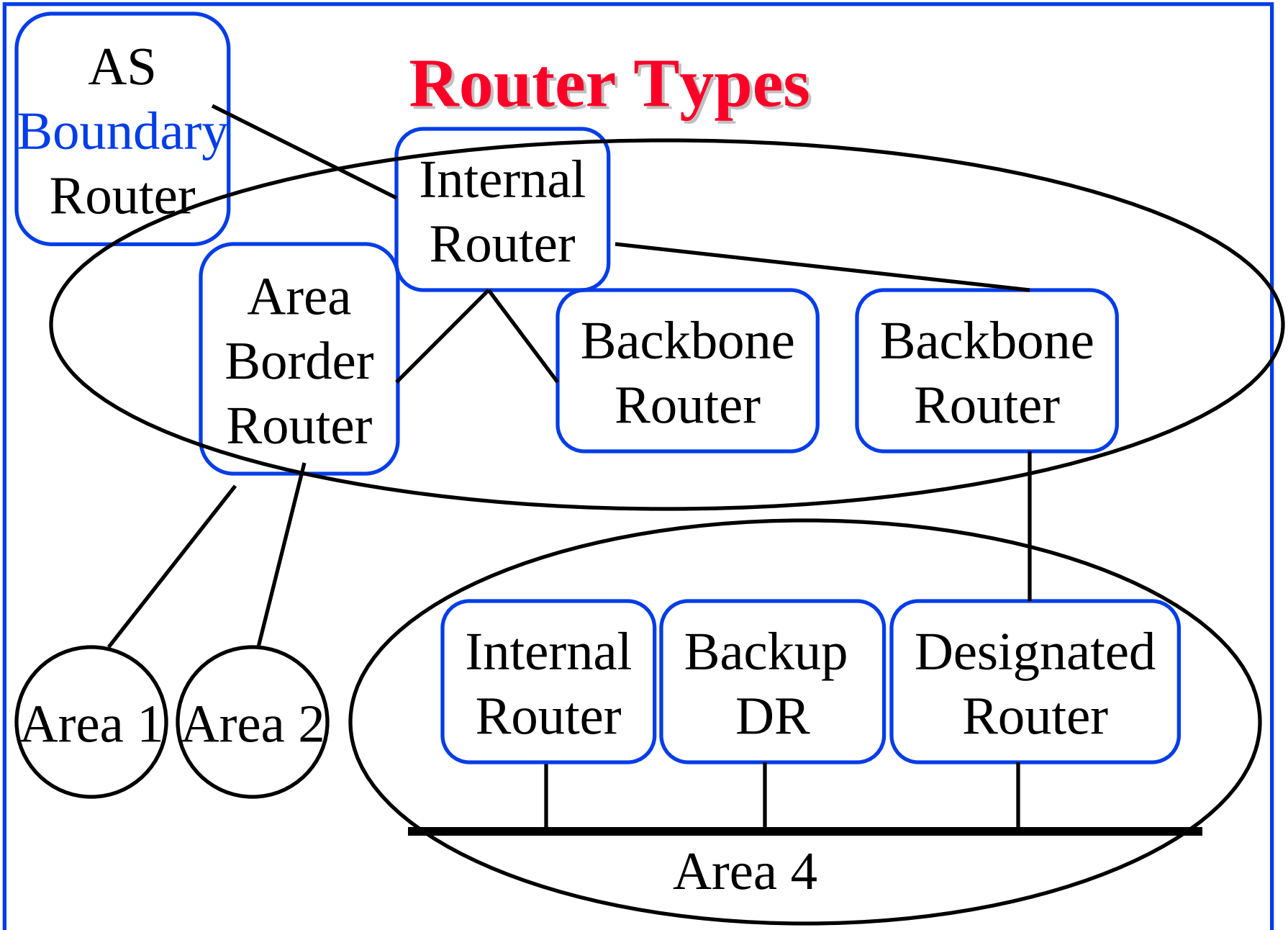
Counting to Infinity Problem



Open Shortest Path First (OSPF)

- Uses true metrics (not just hop count)
- Uses subnet masks
- Allows load balancing across equal-cost paths
- Supports type of service (ToS)
- Allows external routes (routes learnt from other autonomous systems)
- Authenticates route exchanges
- Quick convergence
- Direct support for multicast
- Link state routing $\Rightarrow$ Each router broadcasts its connectivity with neighbors to entire network

Router Types



Router Types (Cont)

- **Internal Router (IR)**: All interfaces belong to the same area
- **Area Border Router (ABR):** Interfaces to multiple areas
- **Backbone Router (BR):** Interfaces to the backbone
- **Autonomous System Boundary Router (ASBR):** Exchanges routing info with other autonomous systems
- **Designated Router (DR):** Generates link-state info about the subnet
- **Backup Designated Router (BDR):** Becomes DR if DR fails.

OSPF Message Types

- Type 1 - Router Link-State Advertisements (LSAs):
  Neighbor's address and cost
  Flooded within the area by all routers.
- Type 2 - Network LSAs:
  Addresses of all routers on the LAN and cost
  Flooded within the area by Designated Router
- Type 3 - Summary LSAs: Flooded into area by ABR.
  Describes reachable networks in other areas.
- Type 4 - AS Boundary Router Summary LSAs:
  Describes cost from the router to ASBR.
  Flooded into the area by ABR.

Message Types (Cont)

- Type 5 - AS External LSAs:
  Flooded to all areas by ASBR.
  Describes external network reachable via the ASBR.
- Type 6 - Multicast Group Membership LSAs:
- Type 7 - Multicast OSPF
- All LSAs contain 32-bit sequence numbers.
  Used to detect duplicate and old LSAs.
- All database entries have an expiration timer (age field)

Metrics (Cost)

- RFC 1253: Metric = 10^8/Speed

Bit Rate	Metric
9.6 kbps	10,416
19.2 kbps	5208
56 kbps	1785
64 kbps	1562
T1 (1.544 Mbps)	65
E1 (2.048 Mbps)	48
Ethernet/802.3 (10 Mbps)	10
100 Mbps or more	1

Hello Protocol

- Routers periodically transmit hello packet
  Multicast to "All-SPF-Routers" (224.0.0.5)
- Used to find neighours and elect DR and BDR
- Packets stay on local subnet.
  Not forwarded by routers.
- Packet contains:
  - Router's selection of DR and BDR
  - Priority of DR and BDR
  - Timers: Hello interval and dead interval (time before a router is declared down)
  - List of neighbor routers from which hellos have been received

Adjacency

- Adjacency is formed between:
  - Two routers on a point-to-point link
  - DR or BDR and routers on LANs
  - Other routers on the LAN do not form adjacency between them
- Adjacent routers should have "synchronized databases"
- Routers send to adjacent routers a summary list of LSAs using database description packets
- Routers then compares the databases and request missing information.
- Database is synchronized $\Rightarrow$ Fully adjacent. Dykstra algorithm is then run to find OSPF routes.

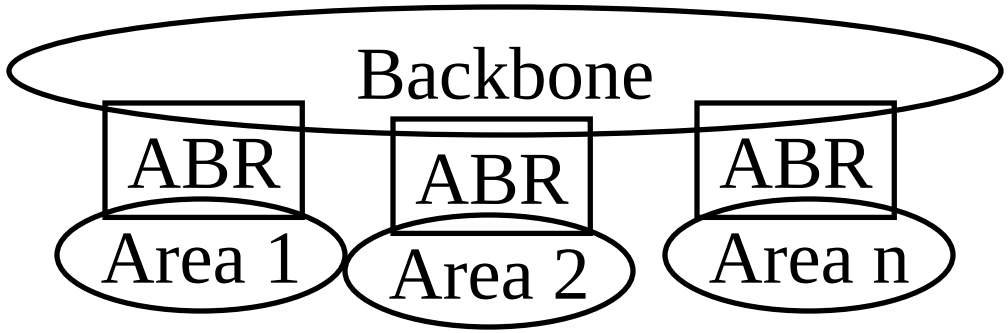
Maintaing the Database

- Databases are continually checked for synchronization by flooding LSAs
- All flooded LSAs are acked. Unacked LSAs are flooded again.
- Database information is checked. If new info, it is forwarded to other adjacencies using LSAs.
- When an entry is aged out, the info is flooded.
- Dykstra algorithm is run on every new info, to build new routing tables.

OSPF Areas

- LSAs are flooded throughout the area
- Area = domain
- Large networks are divided into areas to reduce routing traffic.
- Each area has a 32-bit area ID.
- Although areas are written using dot-decimal notation, they are locally assigned.
- The backbone area is area 0 or 0.0.0.0
  Other areas may be 0.0.0.1, 0.0.0.2, …
- Each router has a router ID. Typically formed from IP address of one of its interfaces.

Backbone Area

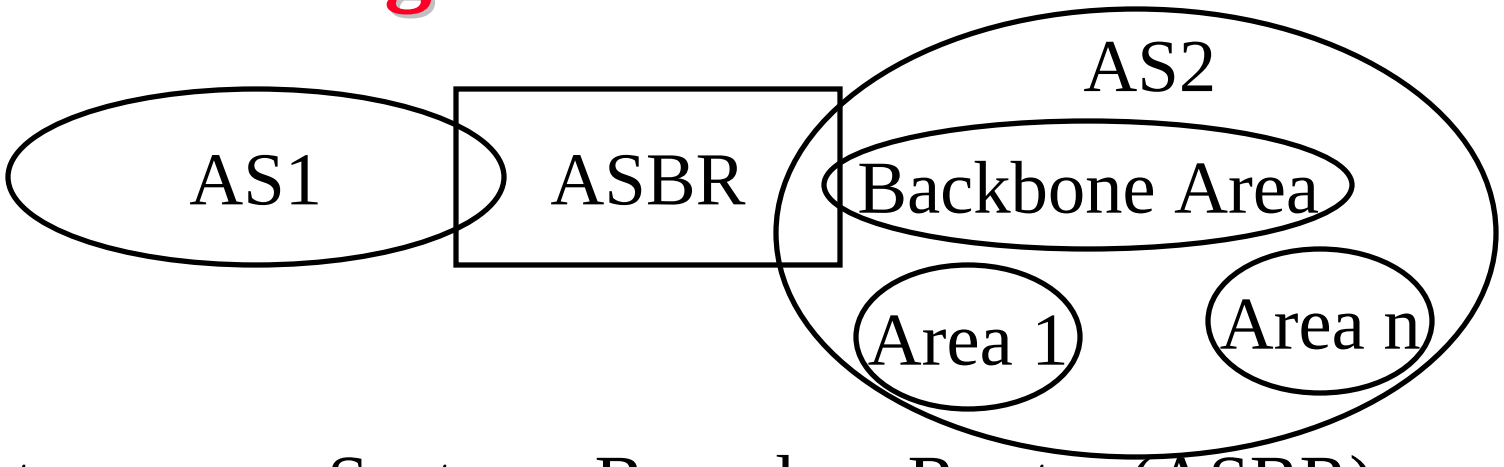


- Area border routers (ABRs) summarize the topology and transmit it to the backbone area
- Backbone routers forward it to other areas
- ABRs connect an area with the backbone area.
  ABRs contain OSPF data for two areas.
  ABRs run OSPF algorithms for the two areas.
- If there is only one area in the AS, there is no backbone area and there are no ABRs.

Inter-Area Routing

- Packets for other areas are sent to ABR
- ABR transmits the packet on the backbone
- Backbone routers send it to the destination area ABR
- Destination ABR forwards it in the destination area.

Routing Info from Other ASs



- Autonomous Systems Boundary Router (ASBR) exchanges "exterior gateway protocol (EGP)" messages with other autonomous systems
- ASBRs generate "external link advertisements." These are flooded to all areas of the AS. There is one entry for every external route.

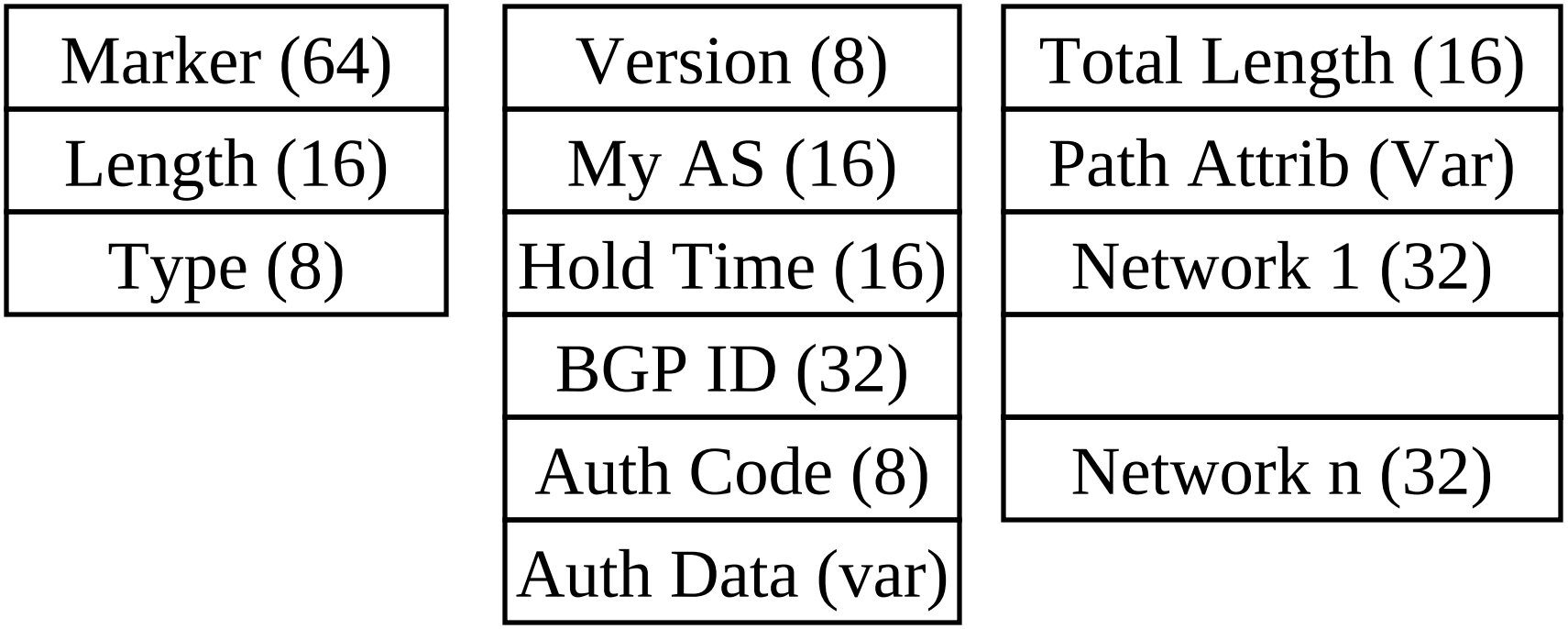
Border Gateway Protocol

- Inter-autonomous system protocol [RFC 1267]
- Used since 1989 but not extensively until recently
- Runs on TCP (segmentation, reliable transmission)
- Advertises all transit ASs on the path to a destination address
- A router may receive multiple paths to a destination
  $\Rightarrow$ Can choose the best path
- No loops and no count-to-infinity problems

BGP Operations

- BGP systems initially exchange entire routing tables. Afterwards, only updates are exchanged.
- BGP messages have the following information:
  - Origin of path information: RIP, OSPF, …
  - AS_Path: List of ASs on the path to reach the dest
  - Next_Hop: IP address of the border router to be used as the next hop to reach the dest
  - Unreachable: If a previously advertised route has become unreachable
- BGP speakers generate update messages to all peers when it selects a new route or some route becomes unreachable.

BGP Messages



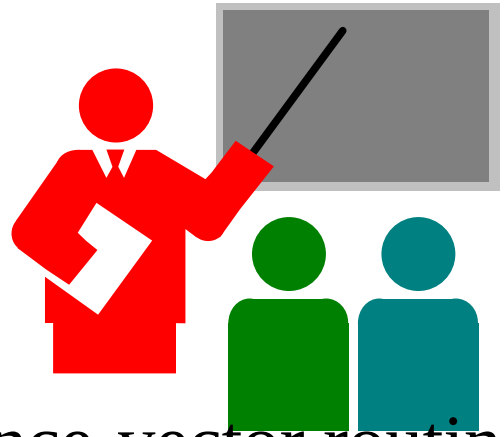
BGP Messages (Cont)

- Marker field is used for authentication or to detect a lost of synch
- Types of messages: Open, update, notification, keep-alive
- Open messages are used to establish peer relationship
  - Hold time: max time between successive keep-alive, update, or notification messages
  - BGP ID: IP address of one of the sender interfaces. Same value is used for all interfaces.
- Update messages are used to exchange routing info.
  - Path attributes = bit mask indicating optional/required, partial/full, etc.

IDRP

- Interdomain Routing Protocol (an EGP)
- Recent extension of BGP concepts
- Distributes path vectors
- Allows multiple routes to a destination
- Allows an additional hierarchy entity: Routing domain confederation $\Rightarrow$ A domain can belong to several RDCs
- Each domain has a Routing Domain Identifier (RDI)
- Each RDC has a RDC identifier (RDCI)
- Uses link attributes, such as, throughput, delay, security
- IDRP has its own reliability mechanism
  $\Rightarrow$ Does not need TCP

Summary



- RIP uses distance-vector routing
- RIP v2 fixes the slow convergence problem
- OSPF uses link-state routing and divides the autonomous systems into multiple areas.
  Area border router, AS boundary router, designated router
- BGP and IDRP are exterior gateway protocols

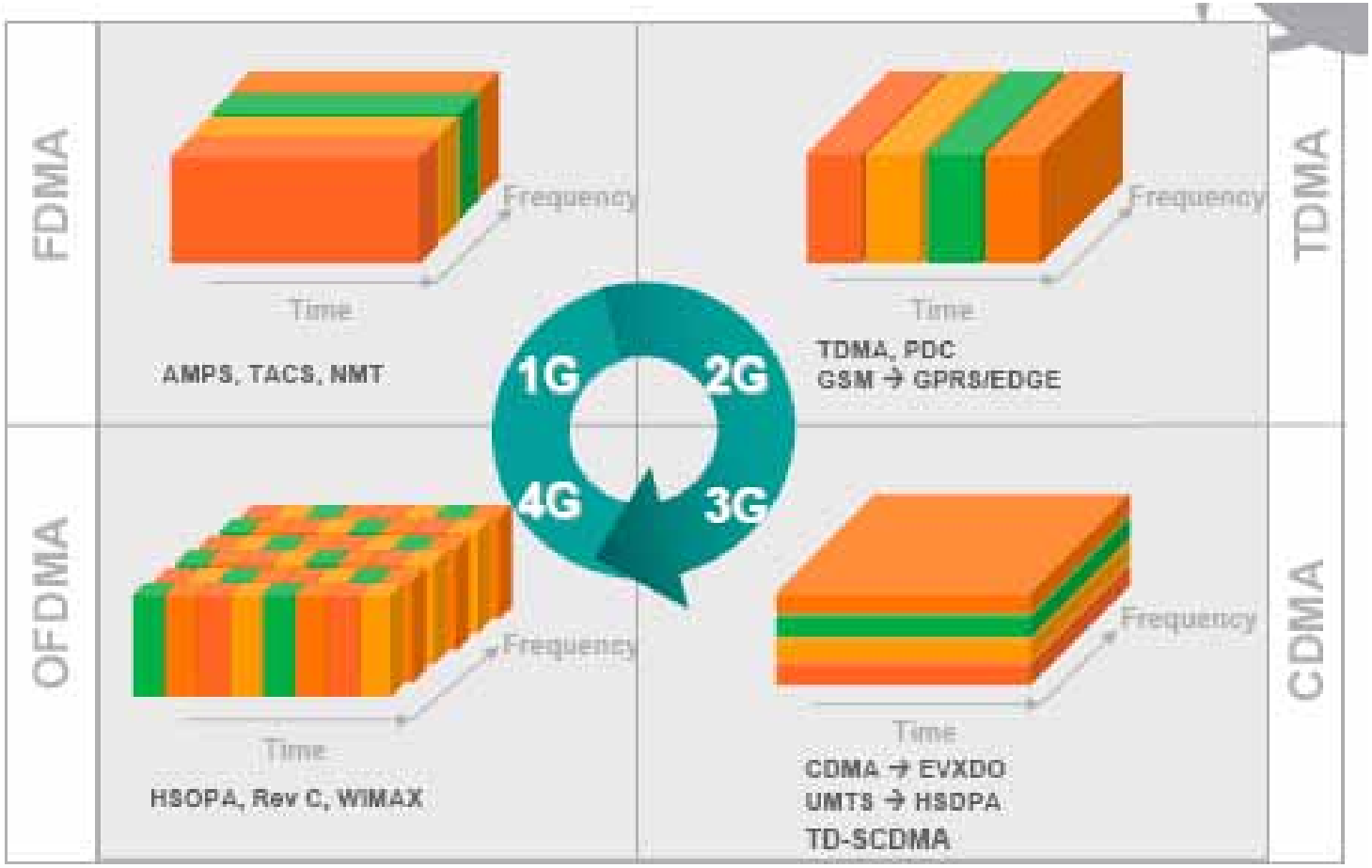
Wireless Networks

Raj Jain
Department of Computer Science and Engineering
Washington University in Saint Louis
Saint Louis, MO 63130
Jain@cse.wustl.edu
http://www.cse.wustl.edu/~jain/

Overview

1. Recent advances in wireless PHY
2. WiMAX Broadband Wireless Access
3. Cellular Telephony Generations
4. WiMAX vs LTE
5. 4G: IMT-Advanced
6. 700 MHz

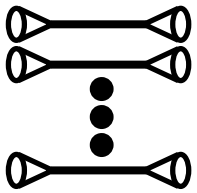
Multiple Access Methods



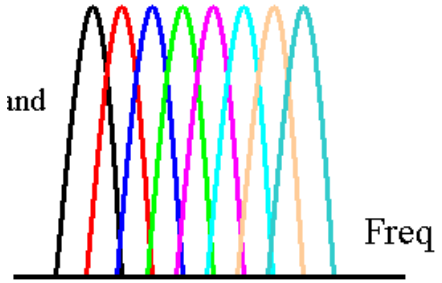
Source: Nortel

1. OFDM

- Orthogonal Frequency Division Multiplexing
- Ten 100 kHz channels are better than one 1 MHz Channel
  ⇒ Multi-carrier modulation



- Frequency band is divided into 256 or more sub-bands.
  Orthogonal ⇒ Peak of one at null of others
- Each carrier is modulated with a BPSK, QPSK, 16-QAM, 64-QAM etc depending on the noise (Frequency selective fading)
- Used in 802.11a/g, 802.16,
  Digital Video Broadcast handheld (DVB-H)
- Easy to implement using FFT/IFFT



Advantages of OFDM

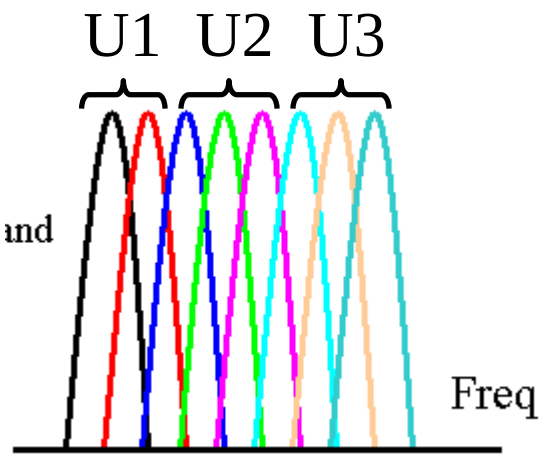
- Easy to implement using FFT/IFFT
- Computational complexity = O(B log BT) compared to previous $O(B^2T)$ for Equalization. Here B is the bandwidth and T is the delay spread.
- Graceful degradation if excess delay
- Robustness against frequency selective burst errors
- Allows adaptive modulation and coding of subcarriers
- Robust against narrowband interference (affecting only some subcarriers)
- Allows pilot subcarriers for channel estimation

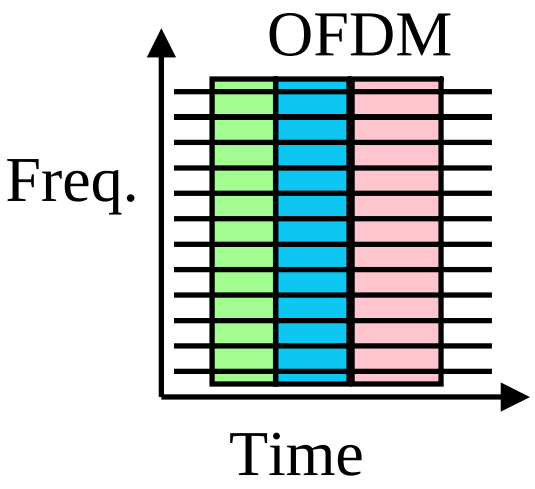
OFDM: Design considerations

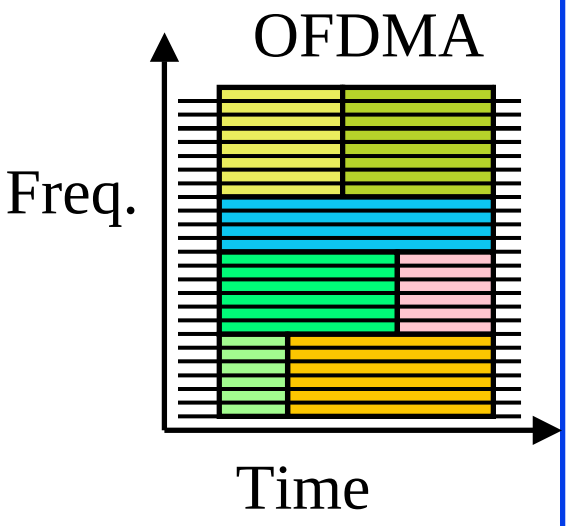
- Large number of carriers ⇒ Larger symbol duration ⇒ Less inter-symbol interference
- Reduced subcarrier spacing ⇒ Increased inter-carrier interference due to Doppler spread in mobile applications
- Easily implemented as Inverse Discrete Fourier Transform (IDFT) of data symbol block
- Fast Fourier Transform (FFT) is a computationally efficient way of computing DFT

OFDMA

- Orthogonal Frequency Division Multiple Access
- Each user has a subset of subcarriers for a few slots
- OFDM systems use TDMA
- OFDMA allows Time+Freq DMA ⇒ 2D Scheduling



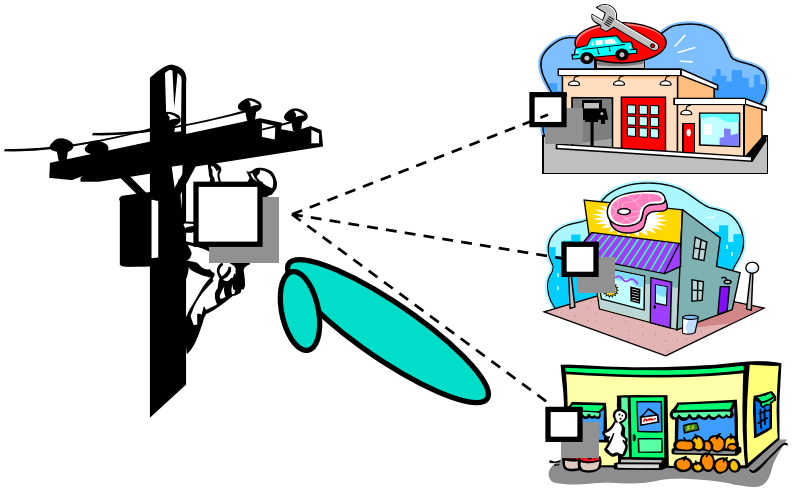




Scalable OFDMA (SOFDMA)

- OFDM symbol duration = f(subcarrier spacing)
- Subcarrier spacing = Frequency bandwidth/Number of subcarriers
- Frequency bandwidth=1.25 MHz, 3.5 MHz, 5 MHz, 10 MHz, 20 MHz, etc.
- Symbol duration affects higher layer operation
  $\Rightarrow$ Keep symbol duration constant at 102.9 us
  $\Rightarrow$ Keep subcarrier spacing 10.94 kHz
  $\Rightarrow$ Number of subcarriers $\propto$ Frequency bandwidth
  This is known as scalable OFDMA

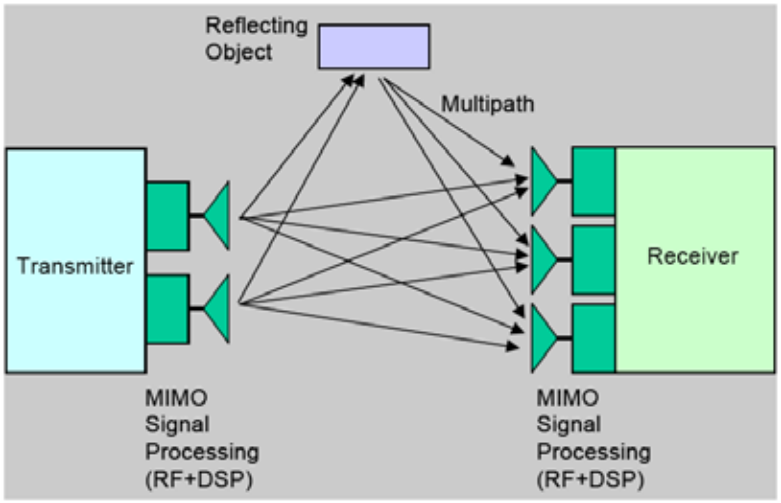
2. Beamforming



- Phased Antenna Arrays:
  Receive the same signal using multiple antennas
- By phase-shifting various received signals and then summing ⇒ Focus on a narrow directional beam
- Digital Signal Processing (DSP) is used for signal processing ⇒ Self-aligning

3. MIMO

- Multiple Input Multiple Output
- RF chain for each antenna
  $\Rightarrow$ Simultaneous reception or transmission of multiple streams



2x3

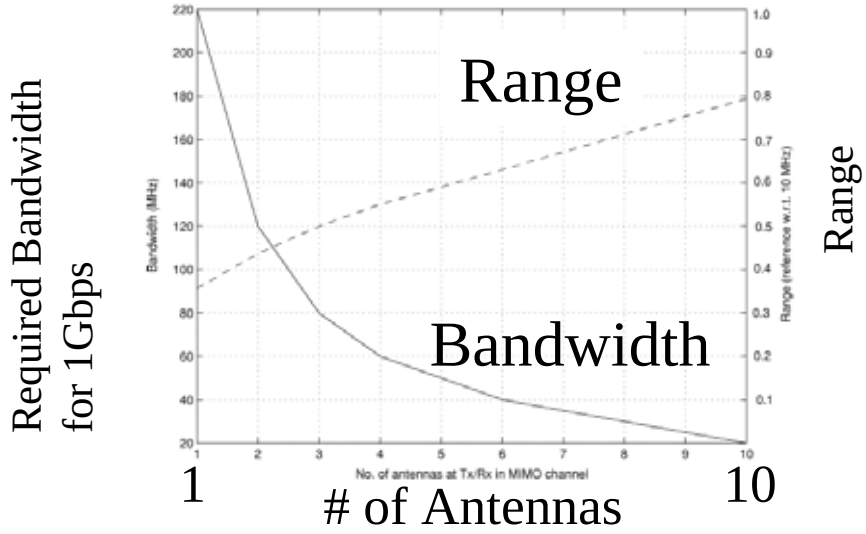
802.16e at 2.5 GHz, 10 MHz TDD, D:U=2:1

T:R	1x1	1x2	2x2	2x4	4x2	4x4
b/Hz	1.2	1.8	2.8	4.4	3.7	5.1

MIMO

- **Antenna Diversity**: Multiple transmit or receive antenna but a single transmit/receive chain
- **MIMO**: RF chain for each antenna ⇒ Simultaneous reception or transmission of multiple streams
  1. **Array Gain**: Improved SNR. Requires channel knowledge (available at receiver, difficult at transmitter)
  2. **Diversity Gain**: Multiple independently fading paths. Get $N_T \times N_R$th order diversity. Transmitter can code the signal suitably ⇒ Space time coding.
  3. **Spatial Multiplexing Gain**: Transmitting independent streams from antennas. Min (N_T, N_R) gain
  4. **Interference Reduction**: Co-channel interference reduced by differentiating desired signals from interfering signals

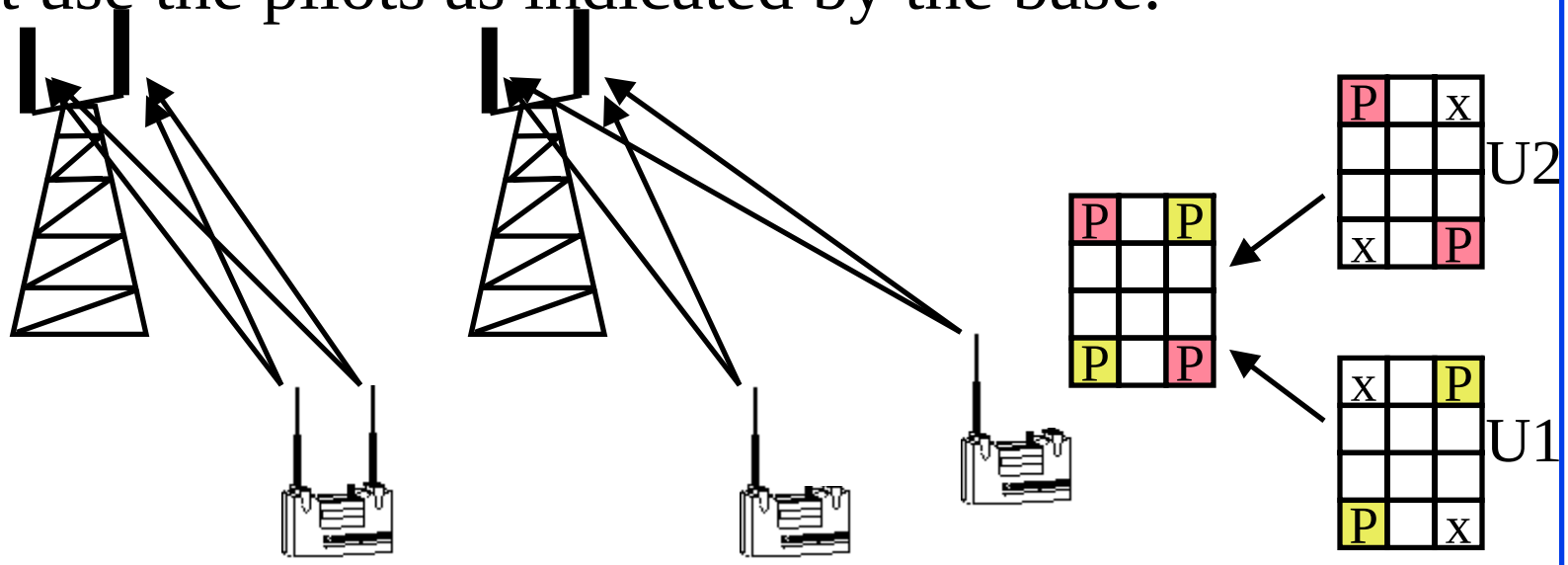
Gigabit Wireless



- Max 9 b/Hz in fixed and 2-4 b/Hz in mobile networks ⇒ Need too much bandwidth ⇒ High frequency ⇒ Line of sight
- Single antenna will require too much power ⇒ high cost amplifiers
- MIMO improves the range as well as reduces the required bandwidth
- Ref: Paulraj et al, Proc of IEEE, Feb 2004.

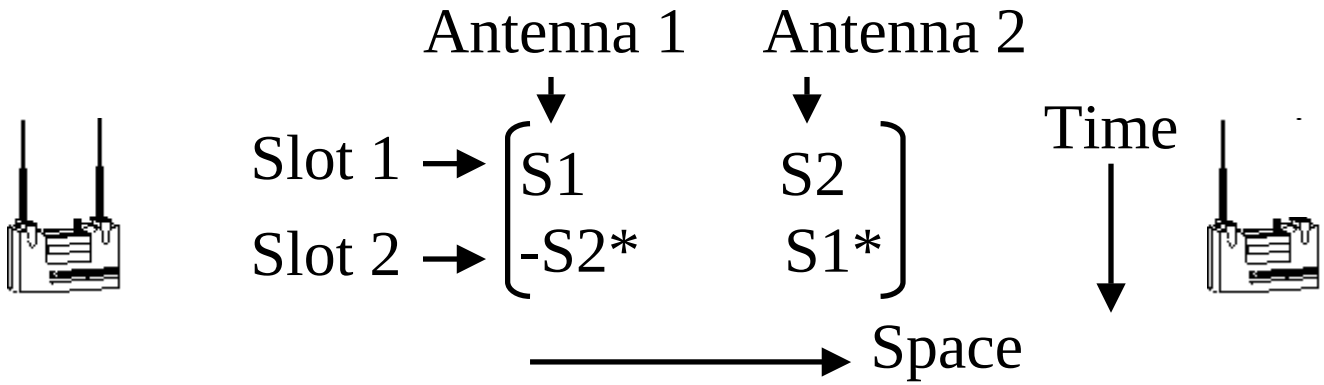
Cooperative MIMO

- Two subscribers with one antenna each can transmit at the same frequency at the same time
- The users do not really need to know each other. They just use the pilots as indicated by the base.



4. Space Time Block Codes (STBC)

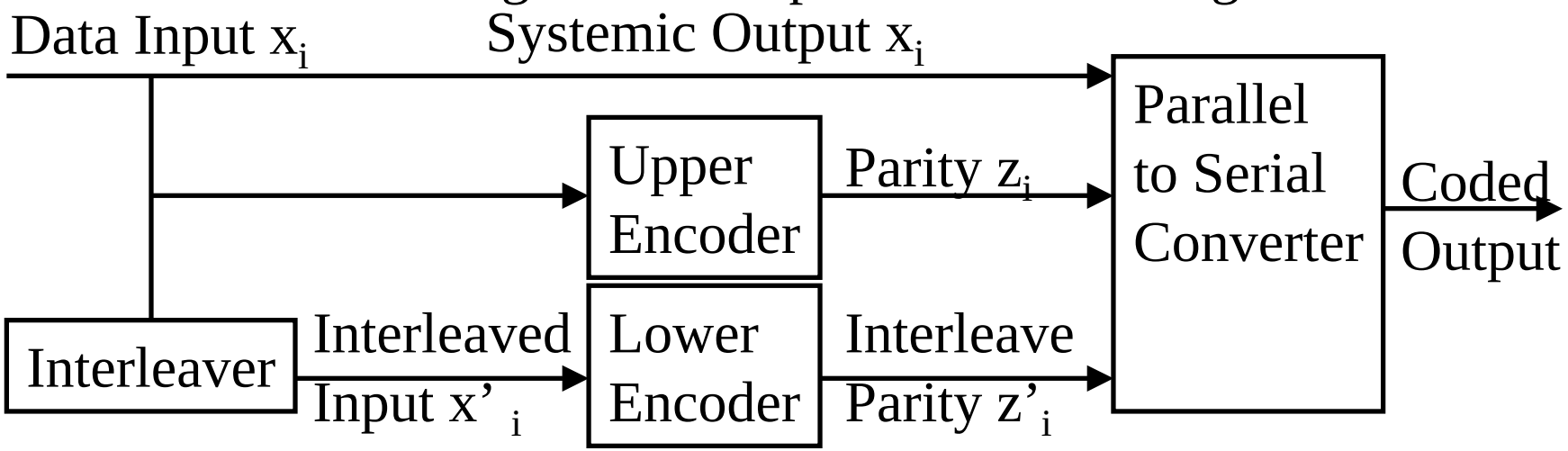
- Invented 1998 by Vahid Tarokh.
- Transmit multiple redundant copies from multiple antennas
- Precisely coordinate distribution of symbols in space and time.
- Receiver combines multiple copies of the received signals optimally to overcome multipath.
- Example: Two antennas:



S1* is complex conjugate of S1 ⇒ columns are orthogonal

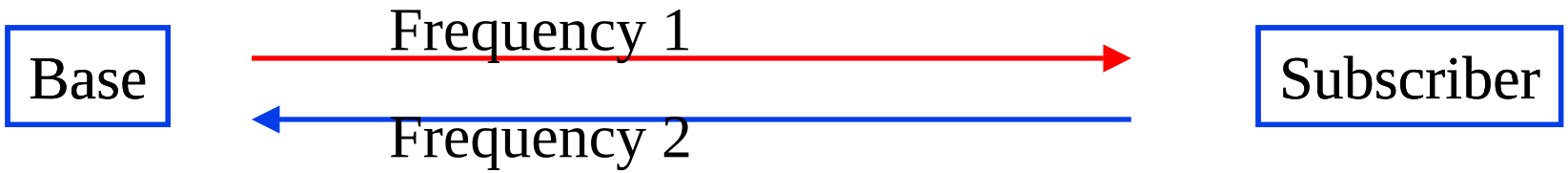
5. Turbo Codes

- Shannon Limit:= B $\log_2$ (1+S/N)
- Normal FEC codes: 3dB below the Shannon limit
- Turbo Codes: 0.5dB below Shannon limit
  Developed by French coding theorists in 1993
- Use two coders with an interleaver
- Interleaver rearranges bits in a prescribed but irregular manner

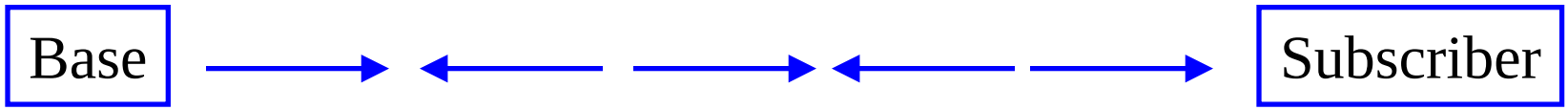


6. Time Division Duplexing (TDD)

- Duplex = Bi-Directional Communication
- Frequency division duplexing (FDD) (Full-Duplex)



- Time division duplex (TDD): Half-duplex

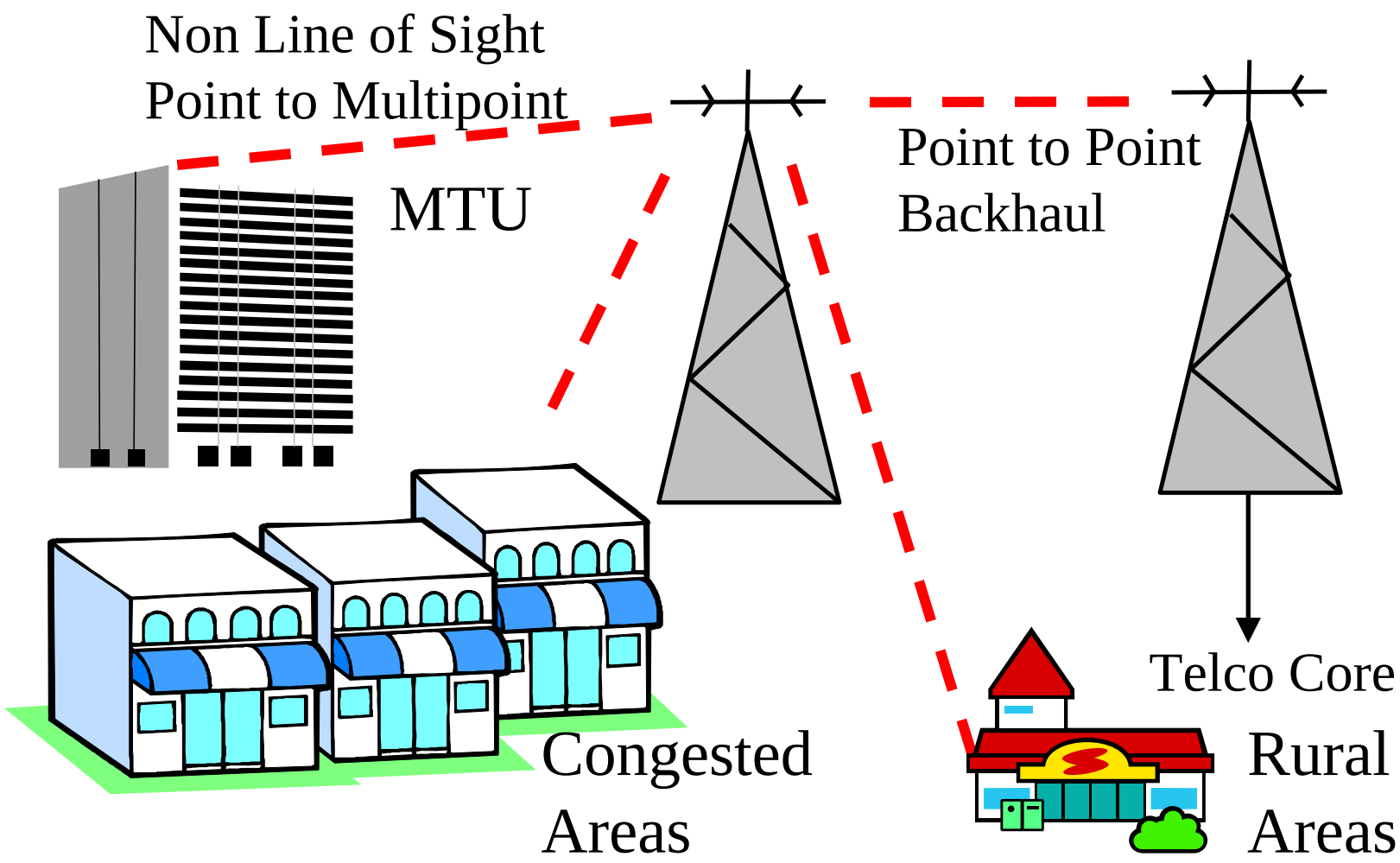


- Most WiMAX deployments will use TDD.
  - Allows more flexible sharing of DL/UL data rate
  - Does not require paired spectrum
  - Easy channel estimation ⇒ Simpler transceiver design
  - Con: All neighboring BS should time synchronize

7. Software Defined Radio

- GSM and CDMA incompatibility ⇒ Need multimode radios
- Military needs to intercept signals of different characteristics
- Radio characteristics (Channel bandwidth, Data rate, Modulation type) can be changed by software
- Multiband, multi-channel, multi-carrier, multi-mode (AM, FM, CDMA), Multi-rate (samples per second)
- Generally using Digital Signal Processing (DSP) or field programmable gate arrays (FPGAs)
- Signal is digitized as close to the antenna as possible
- Speakeasy from Hazeltine and Motorola in mid 80's was one the first SDRs. Could handle 2 MHz to 2 GHz.

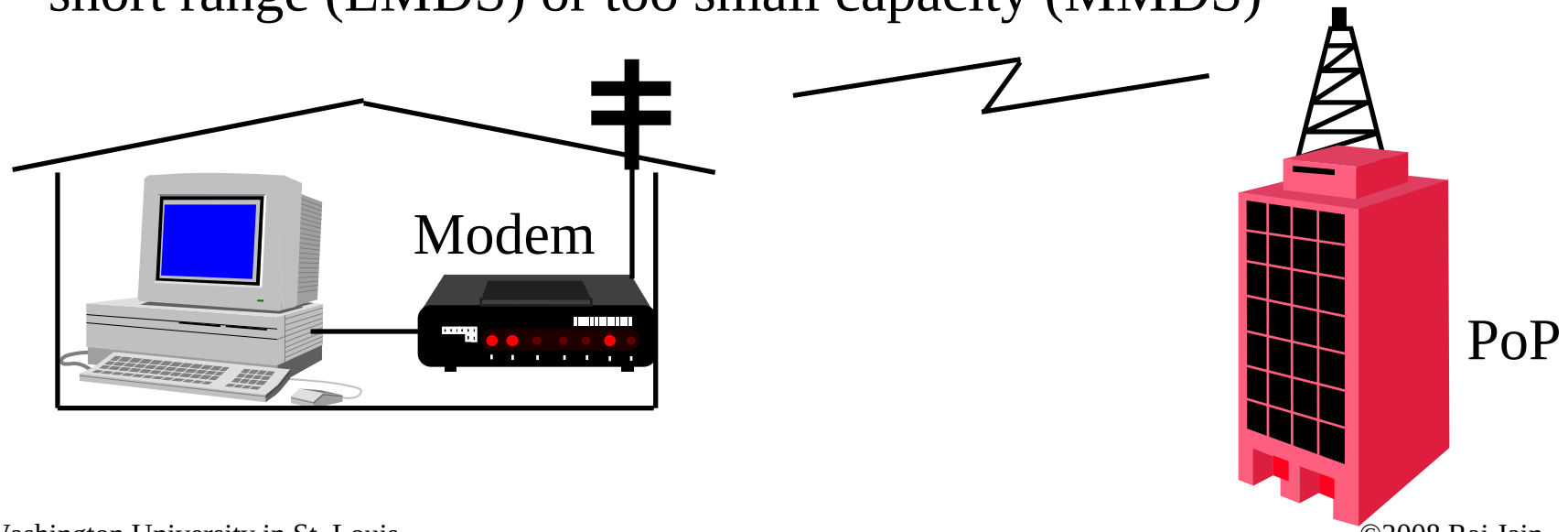
Broadband Wireless Access



Prior Attempts: LMDS & MMDS

- Local Multipoint Distribution Service (1998)
- 1.3 GHz around 28 GHz band (Ka Band)
  28 GHz ⇒ Rain effects
- Multi-channel Multipoint Distribution Services (1999-2001)
- 2.1, 2.5-2.7 GHz Band ⇒ Not affected by rain

Issues: Equipment too expensive, Roof top **LoS** antennas, short range (LMDS) or too small capacity (MMDS)



IEEE 802.16: Key Features

- Broadband Wireless Access
- Up to 50 km or Up to 70 Mbps.
- Data rate vs Distance trade off w adaptive modulation. 64QAM to BPSK
- Offers non-line of site (NLOS) operation
- 1.5 to 28 MHz channels
- Hundreds of simultaneous sessions per channel
- Both Licensed and license-exempt spectrum
- Centralized scheduler
- QoS for voice, video, T1/E1, and bursty traffic
- Robust Security

WiMAX

- WiMAX ≠ IEEE 802.16
- Worldwide Interoperability for Microwave Access
- 420+ members including Semiconductor companies, equipment vendors, integrators, service providers. Like Wi-Fi Alliance
- Narrows down the list of options in IEEE 802.16
- Plugfests started November 2005
- WiMAX forum lists certified base stations and subscriber stations from many vendors
- http://www.wimaxforum.org

Spectrum Options

Designation	Frequency GHz	Bandwidth MHz	Notes
3.5 GHz	3.4-3.6; 3.3-3.4; 3.6-3.8	200 Total. 2×(5 to 56)	Not in US. Considering 3.65-3.70 for unlicensed
2.5 GHz	2.495-2.690	194 Total. 16.5+6 paired.	In USA.
2.3 GHz	2.305-2.320; 2.345-2.360	2×5 paired. 2×5 unpaired.	US, Kr, Au, Nz
2.4 GHz	2.405-2.4835	80 Total	Lic exempt. World-wide.
5 GHz	5.250-5.350; 5.725-5.825	200 MHz	Worldwide.
700 MHz	0.698-0.746; 0.747-0.792	30+48	US
Adv W. Serv.	1.710-1.755; 2.110-2.155	2×45 paired	Used for 3G

Status of WiMAX

- WiBro service started in Korea in June 2006
- More than 200 operators have announced plans for WiMAX
- About half are already trialing or have launched pre-WiMAX
- Two dozen networks in trial or deployed in APAC
- 15 in Western Europe
- Sprint-Nextel in 2.3/2.5 GHz with equipment supplied by Intel, Motorola, Samsung, Nokia, and LG
- Initial deployment in Washington DC and Chicago
- Intel will sample a multi-band WiMAX/WiFi chipset in late 2007
- M-Taiwan

Sample WiMAX Subscriber Stations

Alvarion

Airspan

Axxcelera

Siemens

Aperto

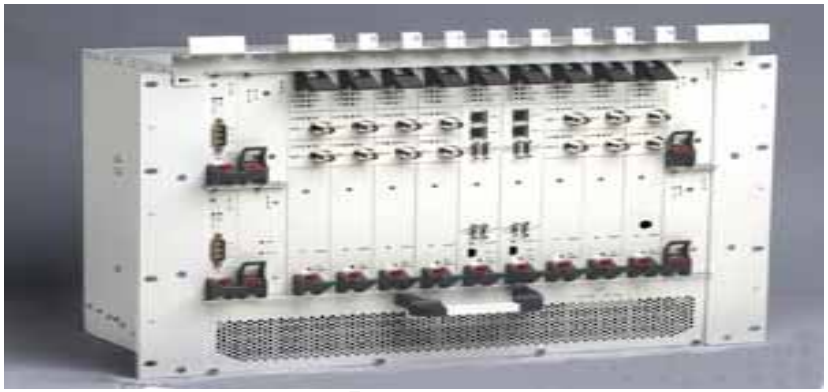
Redline

SR Telecom

Telsima

Sample WiMAX Base Stations

Axxcelara



Alverian

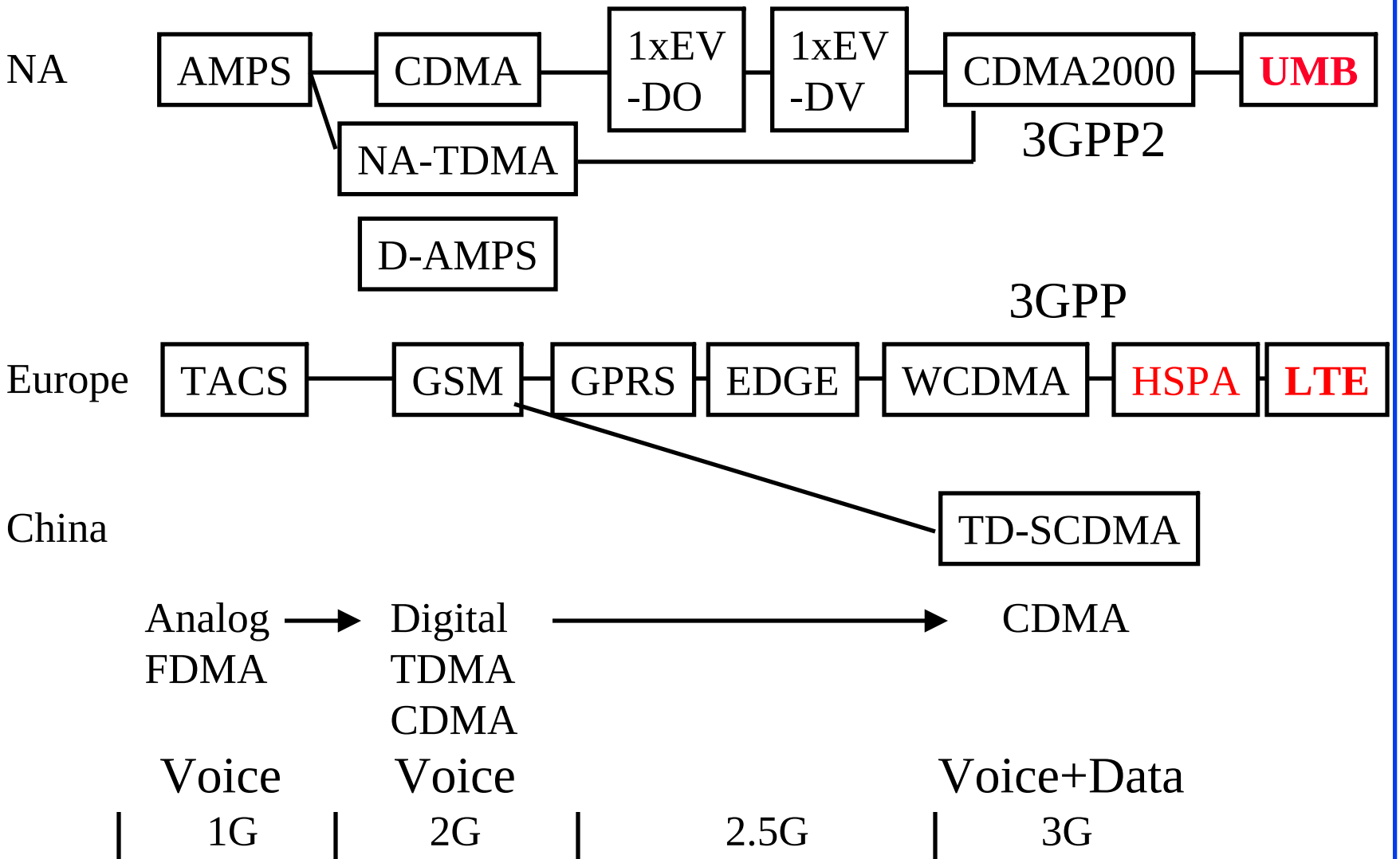
Redline

Airspan

Aperto

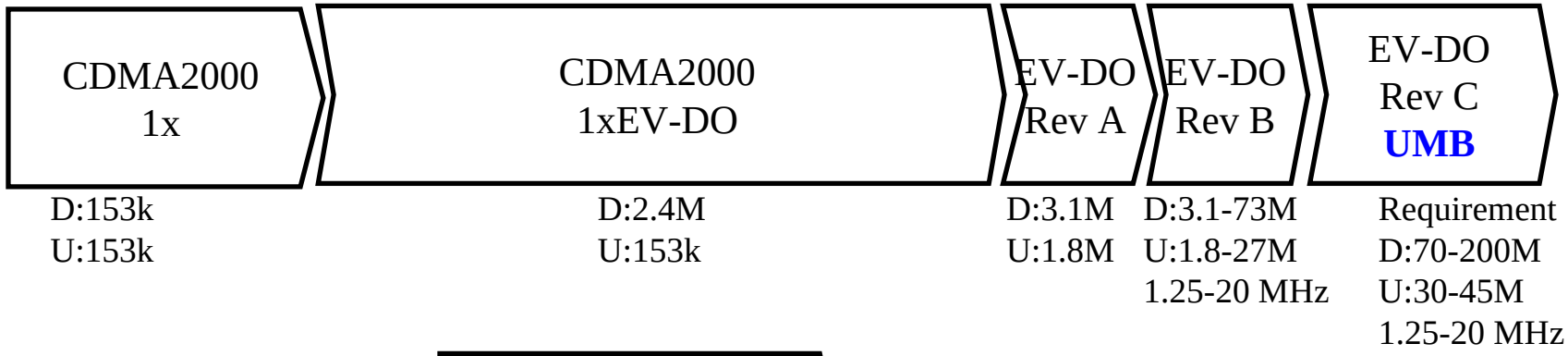
SR Telecom

Cellular Telephony Generations

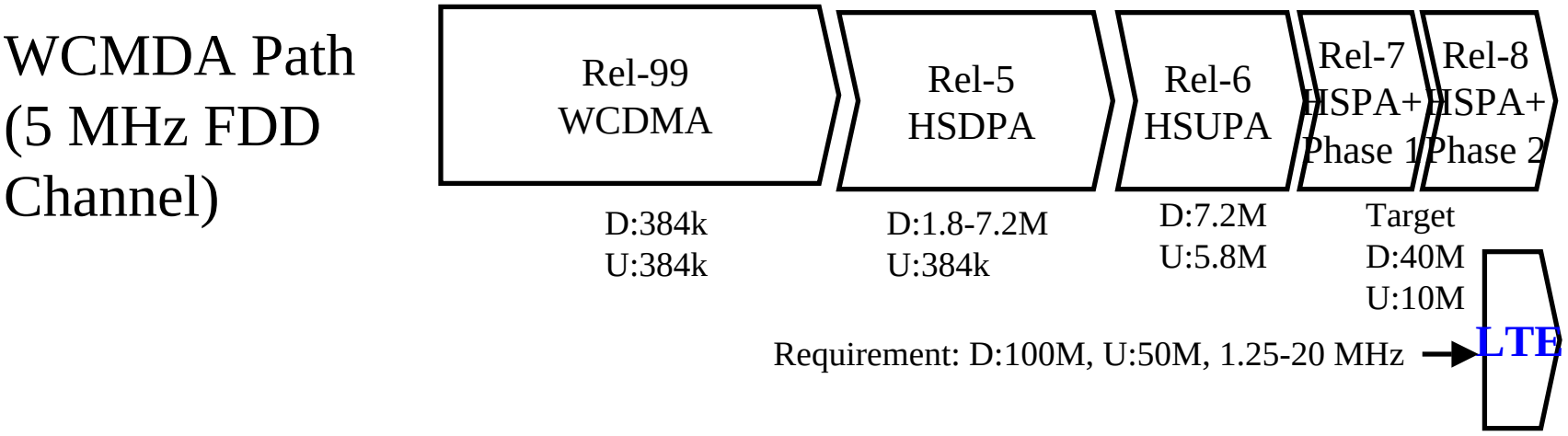


3G Technologies: Bit Rates

CDMA2000 Path (1.25 MH FDD Channel)



WCMDA Path (5 MHz FDD Channel)

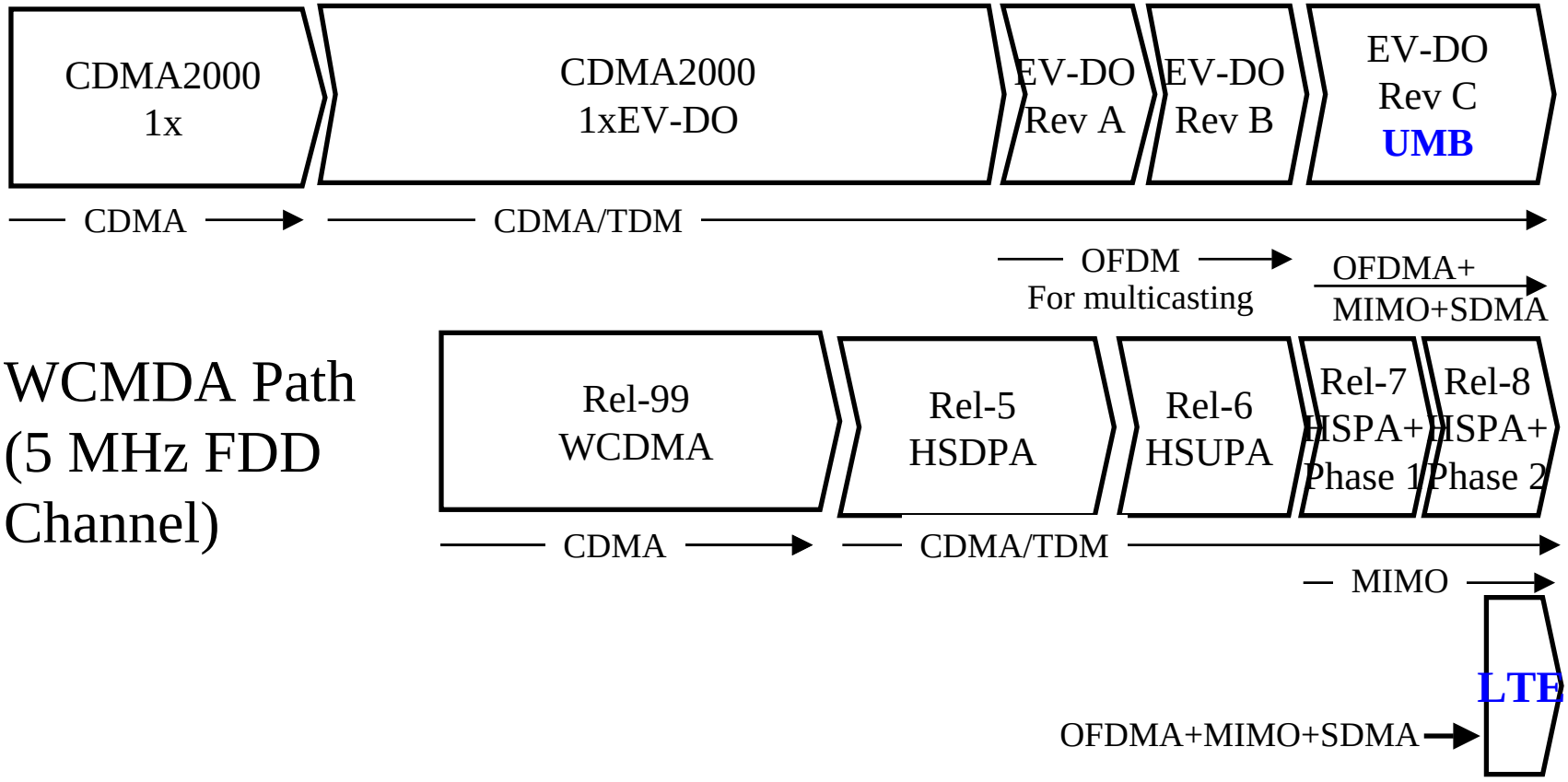


Note: WiMAX is also a 3G Technology now.

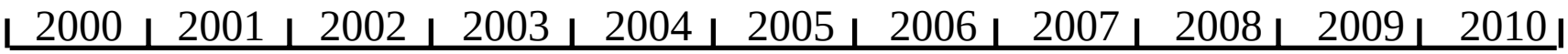
Source: www.cdg.org

3G Technologies: PHY

CDMA2000 Path (1.25 MH FDD Channel)



WCMDA Path (5 MHz FDD Channel)



Source: www.cdg.org

3G Technologies (Cont)

- All data rates are for FDD
  $\Rightarrow$ 20MHz = 2×20 MHz
- On the downlink, LTE uses a modified version of OFDMA called DFT-Spread OFDMA, also known as single-carrier FDMA.
- UMB may utilize a combination of OFDMA and CDMA or OFDM and CDMA
- Data rates depend upon level of mobility

HSDPA

- High-Speed Downlink Packet Access for W-CDMA
- Improved spectral efficiency for downlink ⇒ Asymmetric
- Up to 10 Mbps in theory, 2Mbps+ in practice
- Announced by Siemens, then by Ericsson, Alcatel, Fujitsu
- Adaptive modulation and coding (AMC)
- Multi-code (multiple CDMA channels) transmission
- Fast physical layer (L1) hybrid ARQ (H-ARQ)
- Packet scheduler moved from the radio network controller (RNC) to the Node-B (base station)
  ⇒ advanced packet scheduling techniques
  ⇒ user data rate can be adjusted to match the instantaneous radio channel conditions.

4G: IMT-Advanced

- International Mobile Telecommunications – Advanced or 4G
- Wireless broadband access to be standardized around 2010 and deployed around 2015
- 1 Gbps for nomadic/fixed and 100 Mbps for high mobility (150 km/h)
- Requirements will be set in 2008
- Set of 4G technologies will be selected by 2010

Ref: ITU-R M.1645, "Framework and overall objectives of the future development of IMT-2000 and systems beyond IMT-2000" (2003)

IEEE 802.16m

- Peak data rate:
  - Downlink (BS->MS) > 6.5 bps/Hz,
    Uplink (MS->BS) > 2.8 bps/Hz
    After PHY overhead
  - 20 MHz => 130 Mbps
- Mobility: Optimized for 0-15 km/h, marginal degradation 15-120 km/h, maintain connection 120-350 km/h
- 3 dB improvement in link budget over 16e
- Optimized for cell sizes of up to 5km. Graceful degradation in spectral efficiency for 5-30km. Functional for 30-100 km.

Ref: Draft IEEE 802.16m requirements, June 8, 2007,
http://ieee802.org/16/tgm/docs/80216m-07_002r2.pdf

700 MHz

- February 19, 2009: TV vacates 700-MHz
- FCC just approved 700 MHz for broadband access
- 108 MHz total available
  - 60 MHz available by Auction in January 16, 2008
  - 24 MHz for Public Safety
  - 24 MHz already owned by Access Spectrum, Aloa Partners, Pegasus Comm, Qualcomm,Verizon, DirecTV, Echostar, Google, Intel, Skype, and Yahoo!
- **Open Access**: Open applications, Open devices, Open services, and open networks
- **White spaces**: Unused spectrum between 54 and 698 MHz. (Channel 2 through 51)

Effect of Frequency

- Higher Frequencies have higher attenuation,
  e.g., 18 GHz has 20 dB/m more than 1.8 GHz
- Higher frequencies need smaller antenna
  Antenna $\geq$ Wavelength/2, 800 MHz $\Rightarrow$ 6”
- Higher frequencies are affected more by weather
  Higher than 10 GHz affected by rainfall
  60 GHz affected by absorption of oxygen molecules
- Higher frequencies have more bandwidth and higher data rate
- Higher frequencies allow more frequency reuse
  They attenuate close to cell boundaries. Low frequencies propagate far.
- Mobility $\Rightarrow$ Below 10 GHz

Summary

1. Key developments in wireless are: OFDMA,
2. WiMAX Broadband Wireless Access
3. Cellular Telephony 3G is all CDMA based
4. All 4G technologies will be OFDMA based

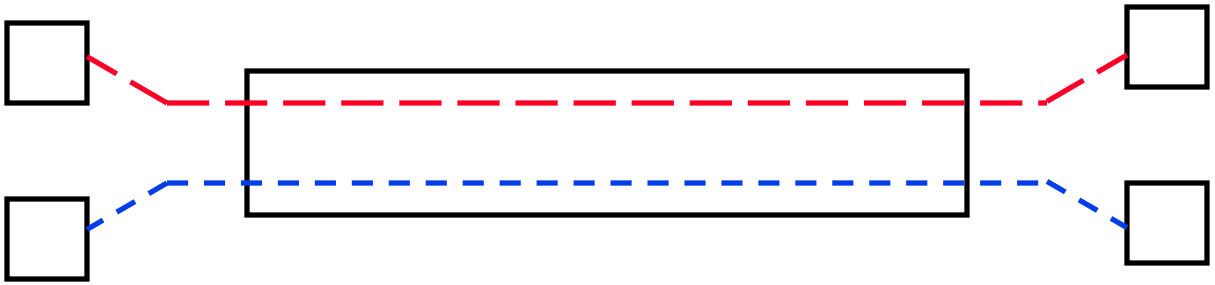
Optical Networking

Raj Jain
Professor of Computer Science and Engineering
Washington University in Saint Louis
Saint Louis, MO, USA
jain@acm.org
http://www.cse.wustl.edu/~jain/

Overview

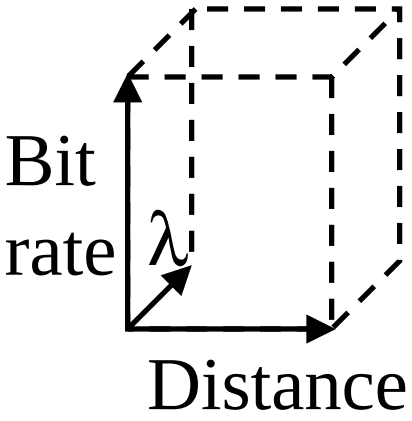
1. Recent DWDM Records
2. OEO vs OOO Switches
3. More Wavelengths
4. Ultra-Long Haul Transmission
5. Passive Optical Networks
6. IP over DWDM: MPλS, GMPLS
7. Free Space Optical Comm
8. Optical Packet Switching

Sparse and Dense WDM



- 10Mbps Ethernet (10Base-F) uses 850 nm
- 100 Mbps Ethernet (100Base-FX) + FDDI use 1310 nm
- Some telecommunication lines use 1550 nm
- WDM: 850nm + 1310nm or 1310nm + 1550nm
- Dense ⇒ Closely spaced ≈ 0.1 - 2 nm separation
- Coarse = 2 to 25 nm = 4 to 12 λ's
- Wide = Different Wavebands

Recent DWDM Records

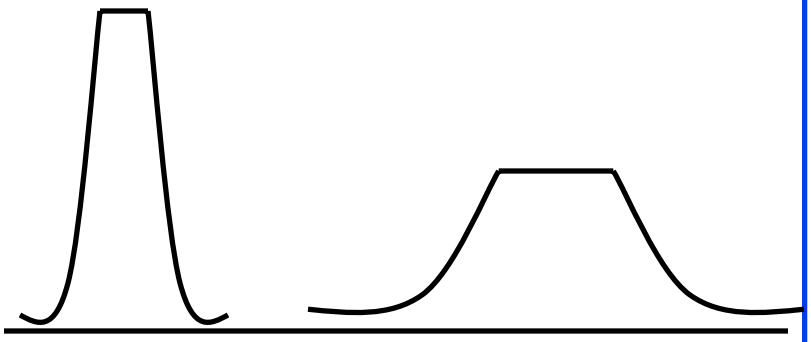


- 32λ× 5 Gbps to 9300 km (1998)
- 16λ× 10 Gbps to 6000 km (NTT'96)
- 160λ× 20 Gbps (NEC'00)
- 128λ× 40 Gbps to 300 km (Alcatel'00)
- 64λ× 40 Gbps to 4000 km (Lucent'02)
- 19λ× 160 Gbps (NTT'99)
- 7λ× 200 Gbps (NTT'97)
- 1λ×1200 Gbps to 70 km using TDM (NTT'00)
- 1022 Wavelengths on one fiber (Lucent'99)

Potential: 58 THz = 50 Tbps on 10,000 λ's

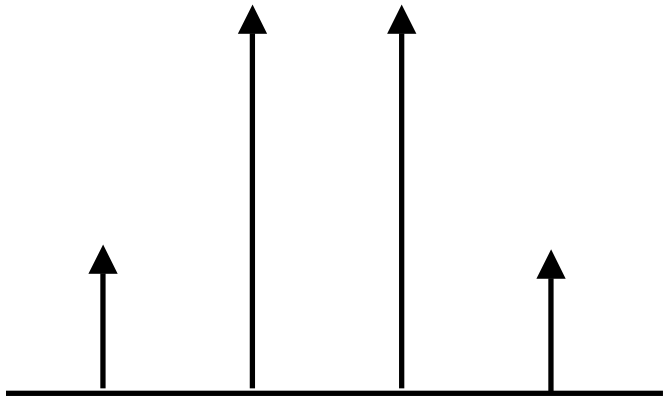
Ref: IEEE J. on Selected Topics in Quantum Electronics, 11/2000.

Attenuation and Dispersion



- Pulses become shorter and wider as they travel through the fiber

Four-Wave Mixing



- If two signals travel in the same phase for a long time, new signals are generated.

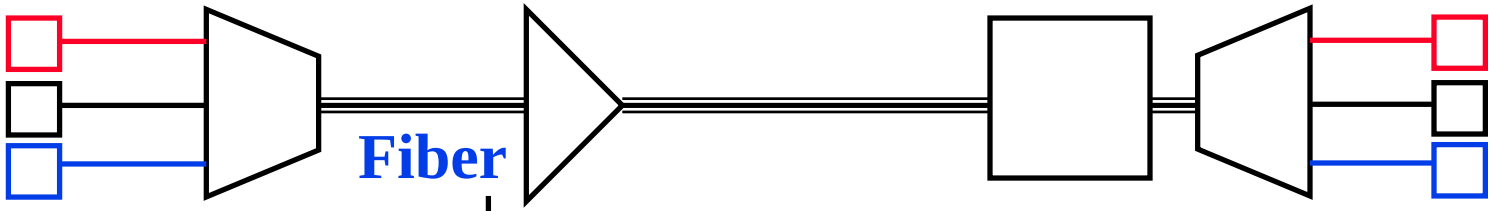
Core Optical Networks

- Higher Speed: 10 Gbps to 40 Gbps to 160 Gbps
- Longer Distances: 600 km to 6000 km
- More Wavelengths: 16 λ's to 160 λ's
- All-optical Switching: OOO vs OEO Switching

OEO vs OOO Switches

- OEO:
  - Requires knowing data rate and format, e.g., 10 Gbps SONET
  - Can multiplex lower rate signals
  - Cost/space/power increases linearly with data rate
- OOO:
  - Data rate and format independent
    $\Rightarrow$ Data rate easily upgraded
  - Sub-wavelength mux/demux difficult
  - Cost/space/power relatively independent of rate
  - Can switch multiple ckts per port (waveband)
  - Issues: Wavelength conversion, monitoring

40 Gbps



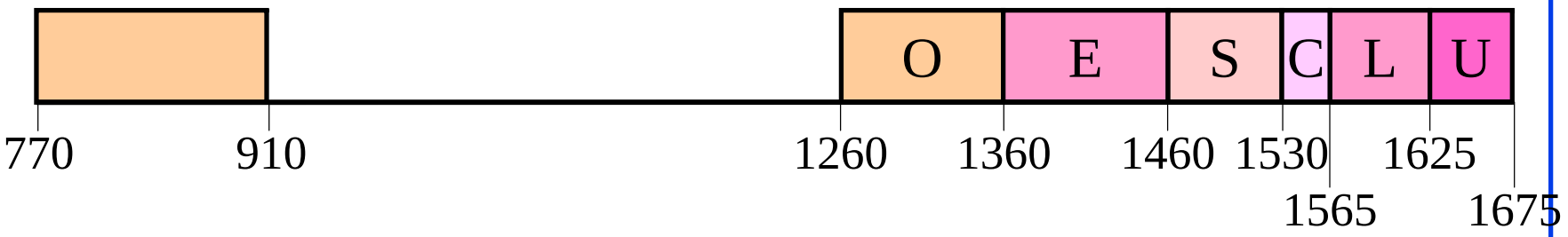
Transmitter	Mux/Demux	Amplifier	Switching	Receivers
Sources	Filters	Gain Equalizers	ADM	Detectors
Modulators	Interleavers	Performance Monitors		
Wavelockers		Dispersion compensators		
		PMD compensators		

- Need all new optical and electronic components
- Non-linearity's reduce distance by square of rate.
- Deployment may be 2-3 years away
- Development is underway. To avoid 10 Gbps mistake.
- Cost goal: 2.5×10 Gbps

More Wavelengths

- C-Band (1535-1560nm), 1.6 nm (200 GHz) ⇒ 16 λ's
- Three ways to increase # of wavelengths:

1. **Narrower Spacing**: 100, 50, 25, 12.5 GHz
   Spacing limited by data rate. Cross-talk (FWM)
   Tight frequency management: Wavelength monitors, lockers, adaptive filters
2. **Multi-band**: C+L+S Band
3. **Polarization Muxing**



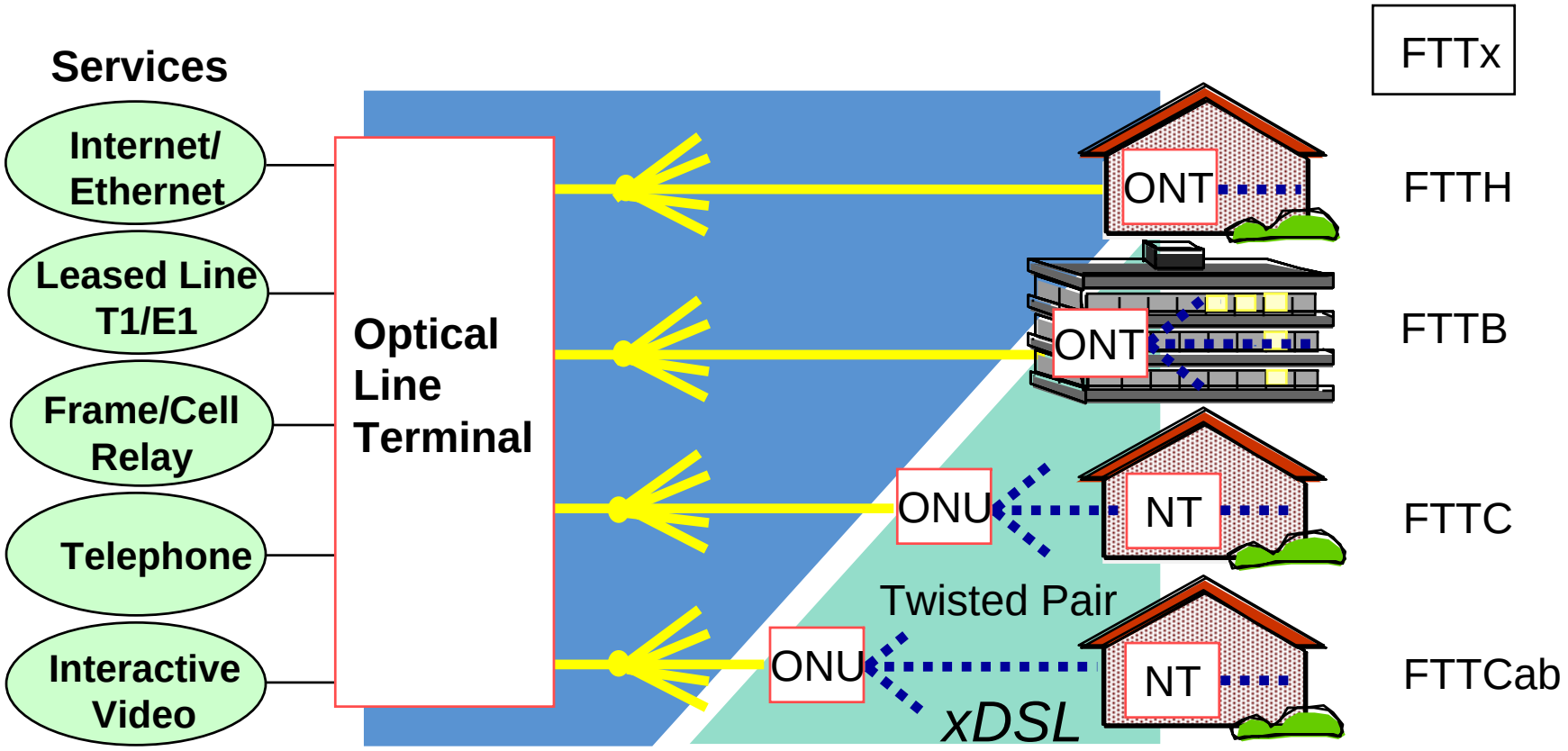
More Wavelengths (Cont)

- More wavelengths ⇒ More Power
  ⇒ Fibers with large effective area
  ⇒ Tighter control of non-linearity's
  ⇒ Adaptive tracking and reduction of polarization mode dispersion (PMD)

Ultra-Long Haul Transmission

1. Strong out-of-band Forward Error Correction (FEC)
   Changes regeneration interval from 80 km to 300km
   Increases bit rate from 40 to 43 Gbps
2. Dispersion Management: Adaptive compensation
3. More Power: Non-linearity's $\Rightarrow$ RZ coding
   Fiber with large effective area
   Adaptive PMD compensation
4. Distributed Raman Amplification:
   Less Noise than EDFA
5. Noise resistant coding: 3 Hz/bit by Optimight

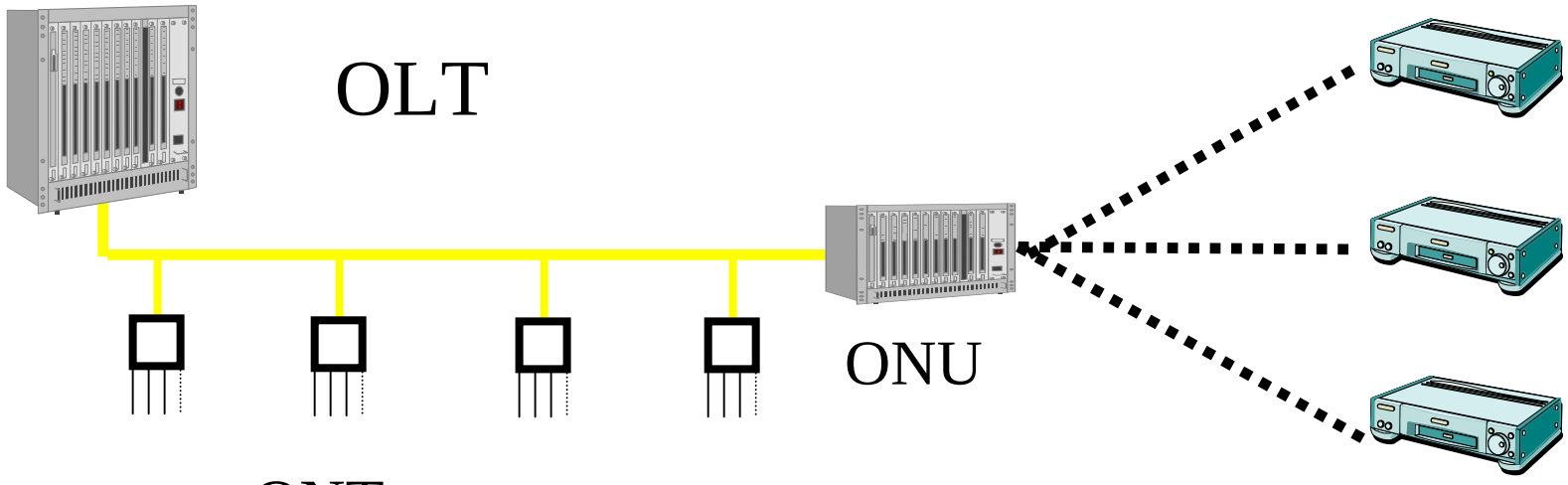
Access: Fiber To The X(FTTx)



FTTH :Fiber To The Home
FTTB :Fiber To The Building
FTTC:Fiber To The Curb
FTTCab :Fiber To The Cabinet

Passive Optical Networks

- A single fiber is used to support multiple customers
- No active equipment in the path ⇒ Highly reliable
- Both upstream and downstream traffic on ONE fiber (1490nm down, 1310nm up). OLT assigned time slots upstream.
- Optical Line Terminal (OLT) in central office
- Optical Network Terminal (ONT) on customer premises
  Optical Network Unit (ONU) at intermediate points w xDSL



Why PONs?

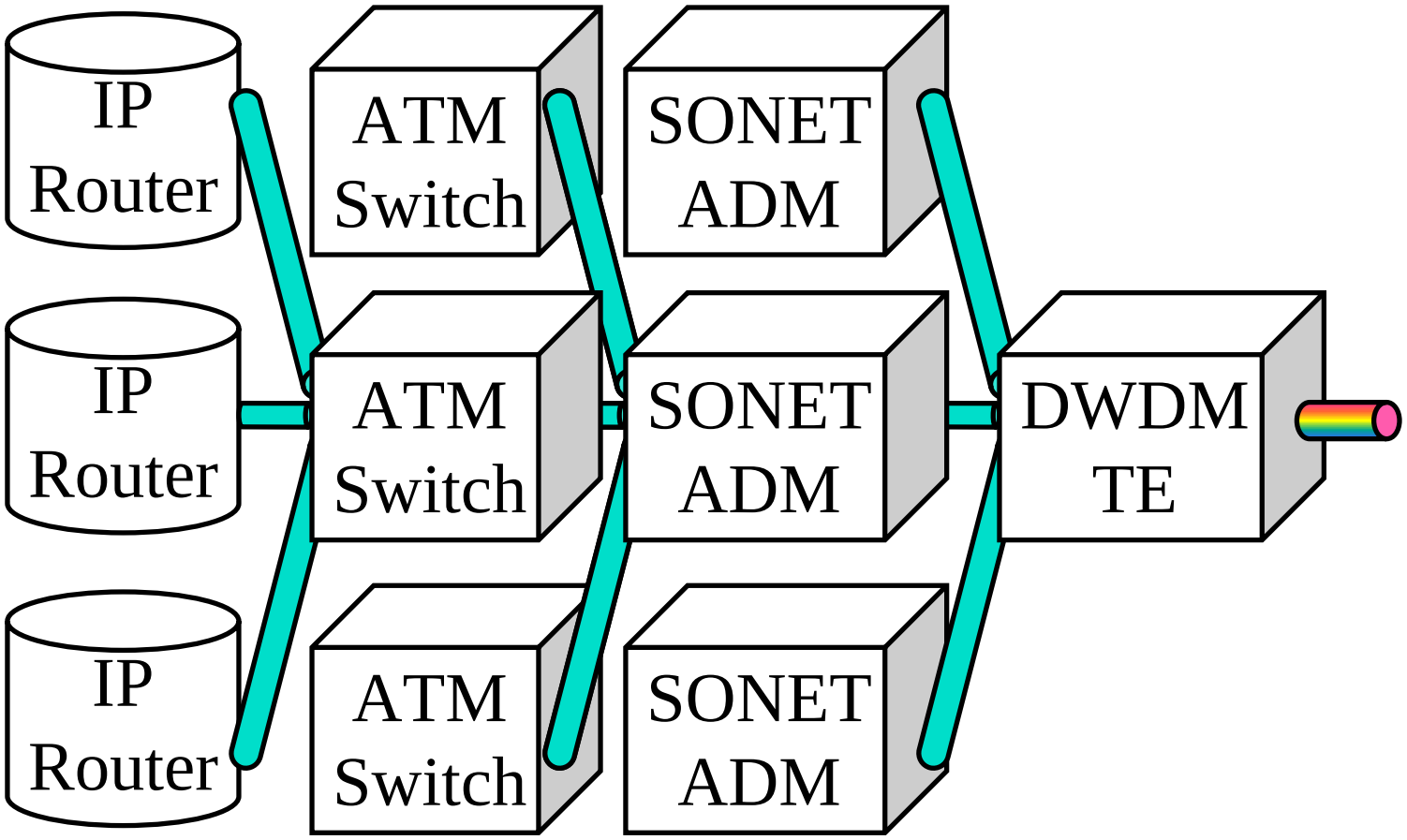
1. Passive ⇒ No active electronics or regenerators in distribution network ⇒ Very reliable. Easy to maintain. Reduced truck rolls. Shorter installation times. Reduced power expences. ⇒ Lower OpEx.
2. Single fiber for bi-directional communication ⇒ Reduced cabling and plant cost ⇒ Lower CapEx
3. A single fiber is shared among 16 to 64 customers ⇒ Relieves fiber congestion
4. Single CO equipment is shared among 16 to 64 customers 2N fibers + 2N transceivers vs 1 fiber + (N+1) transceivers ⇒ Significantly lower CapEx.
5. Scalable ⇒ New customers can be added. Exisiting Customer bandwidth can be changed
6. Multi-service: Voice, T1/E1, SONET/SDH, ATM, Video, Ethernet. Most pt-pt networks are single service.

Useful if customers are clustered ⇒ Asia (Korea, China)

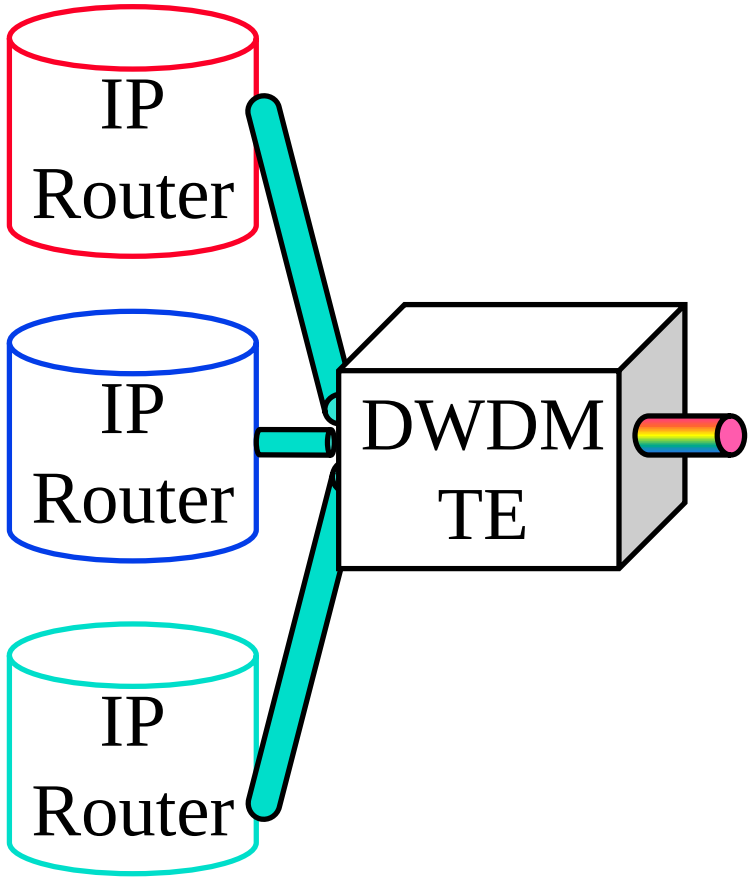
Types of PONs

- **APON**: Initial name for ATM based PON spec. Designed by Full Service Access Network (FSAN) group
- **BPON**: Broadband PON standard specified in ITU G.983.1 thru G.893.7 = APON renamed
  - 155 or 622 Mbps downstream, 155 upstream
- **EPON**: Ethernet based PON draft being designed by IEEE 802.3ah.
  - 1000 Mbps down and 1000 Mbps up.
- **GPON**: Gigabit PON standard specified in ITU G.984.1 and G.984.2
  - 1244 and 2488 Mbps Down, 155/622/1244/2488 up

IP over DWDM (Past)

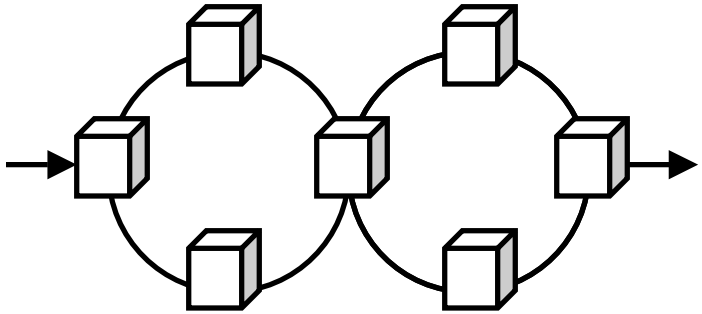


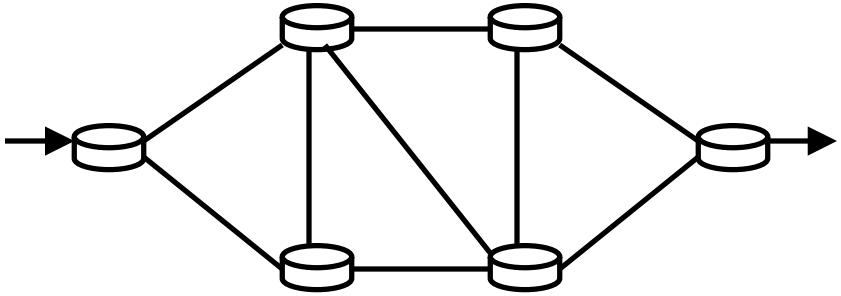
IP over DWDM (Future)



Telecom vs Data Networks

	Telecom Networks	Data Networks
Topology Discovery	Manual	Automatic
Path Determination	Manual	Automatic
Circuit Provisioning	Manual	No Circuits
Transport & Control Planes	Separate	Mixed
User and Provider Trust	No	Yes
Protection	Static using Rings	No Protection



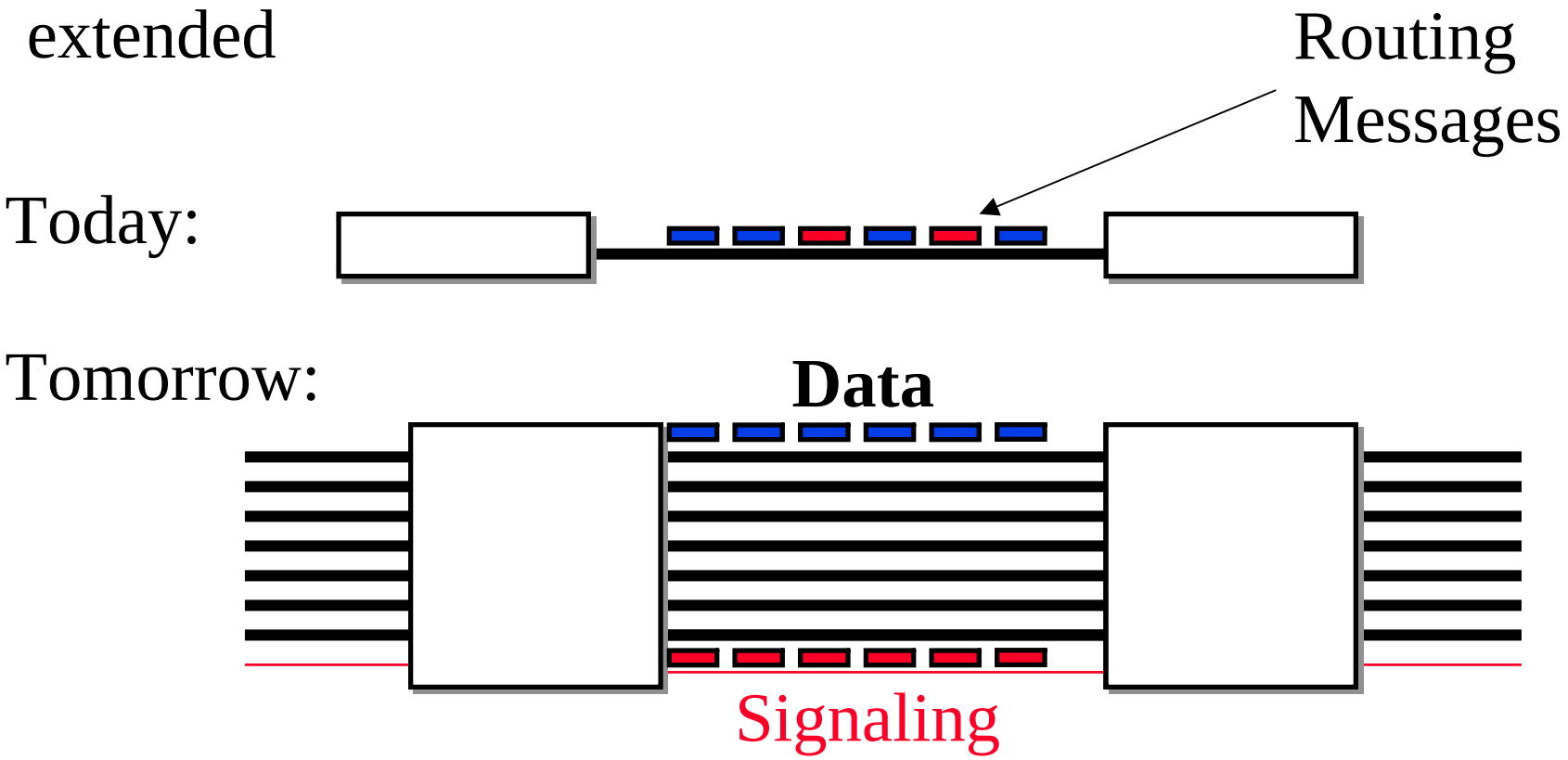


IP over DWDM Issues

1. Data and Control plane separation
2. Circuits
3. Signaling
4. Addressing
5. Protection and Restoration

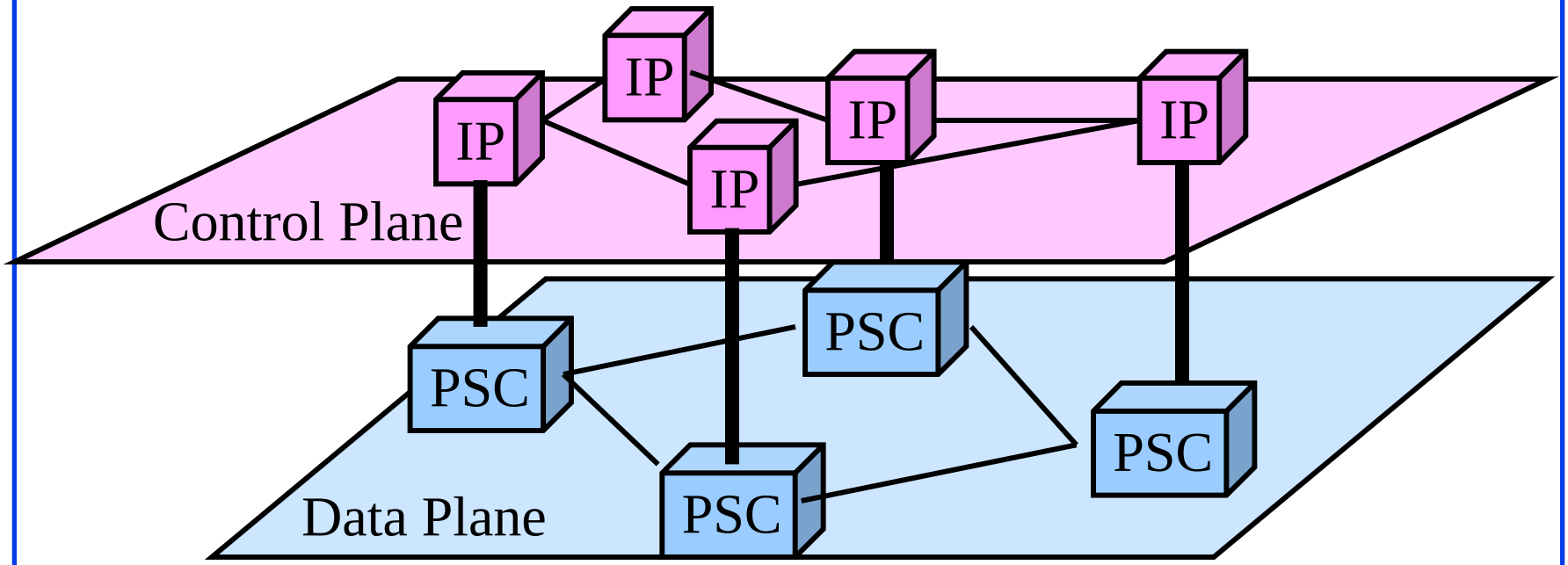
Control and Data Plane Separation

- Separate control and data channels
- IP routing protocols (OSPF and IS-IS) are being extended



IP-Based Control Plane

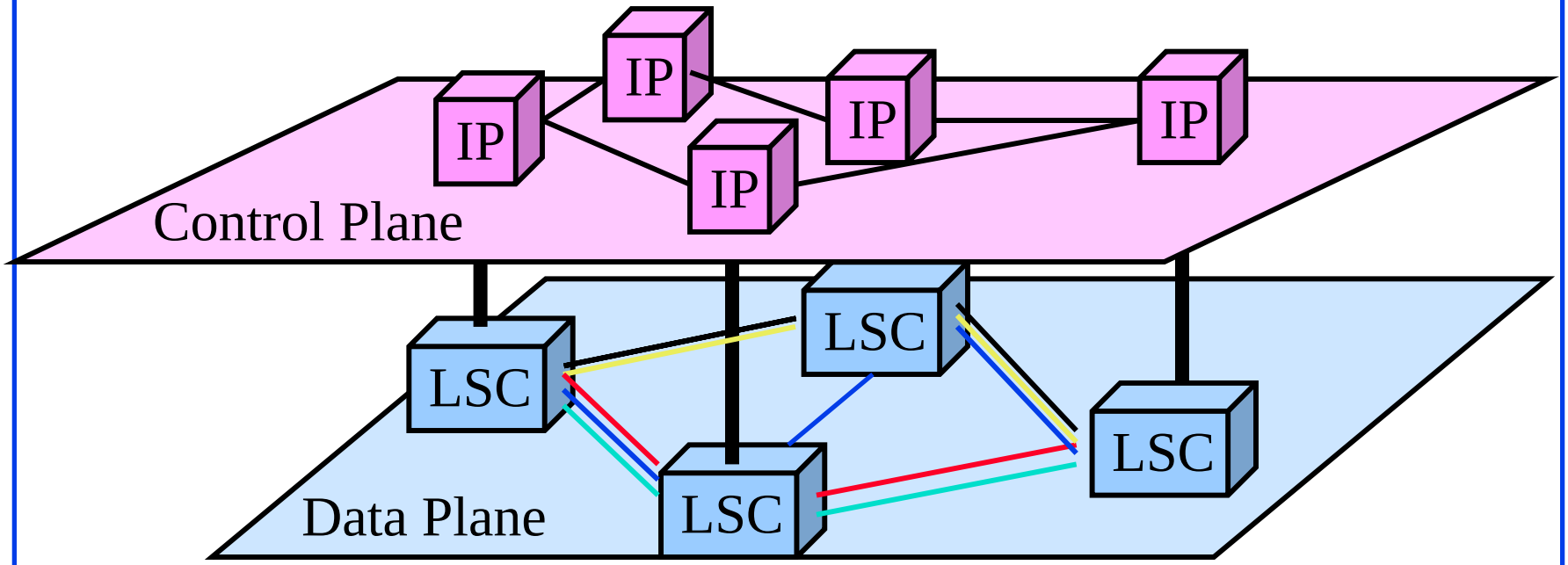
- Control is by IP packets (electronic).
  Data can be any kind of packets (IPX, ATM cells).
  ⇒ MPLS



PSC = Packet Switch Capable Nodes

MPλS

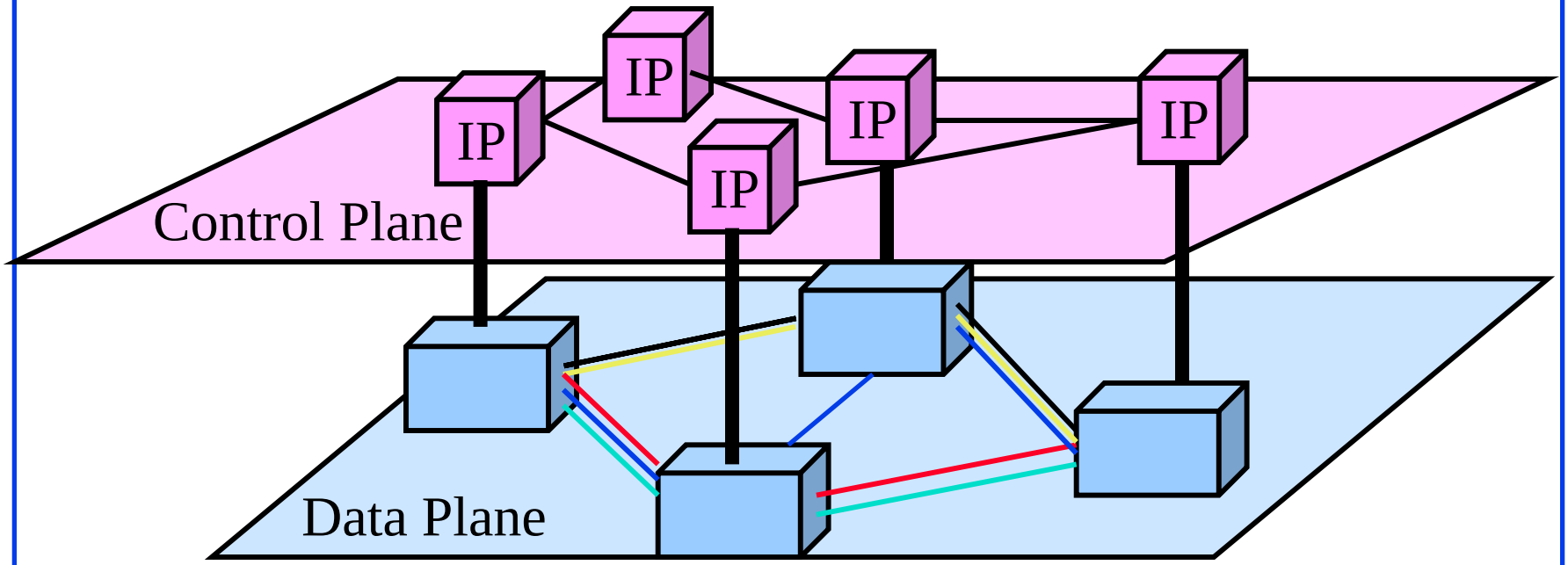
- Control is by IP packets (electronic).
  Data plane consists of wavelength circuits
  ⇒ Multiprotocol Lambda Switching (October 1999)



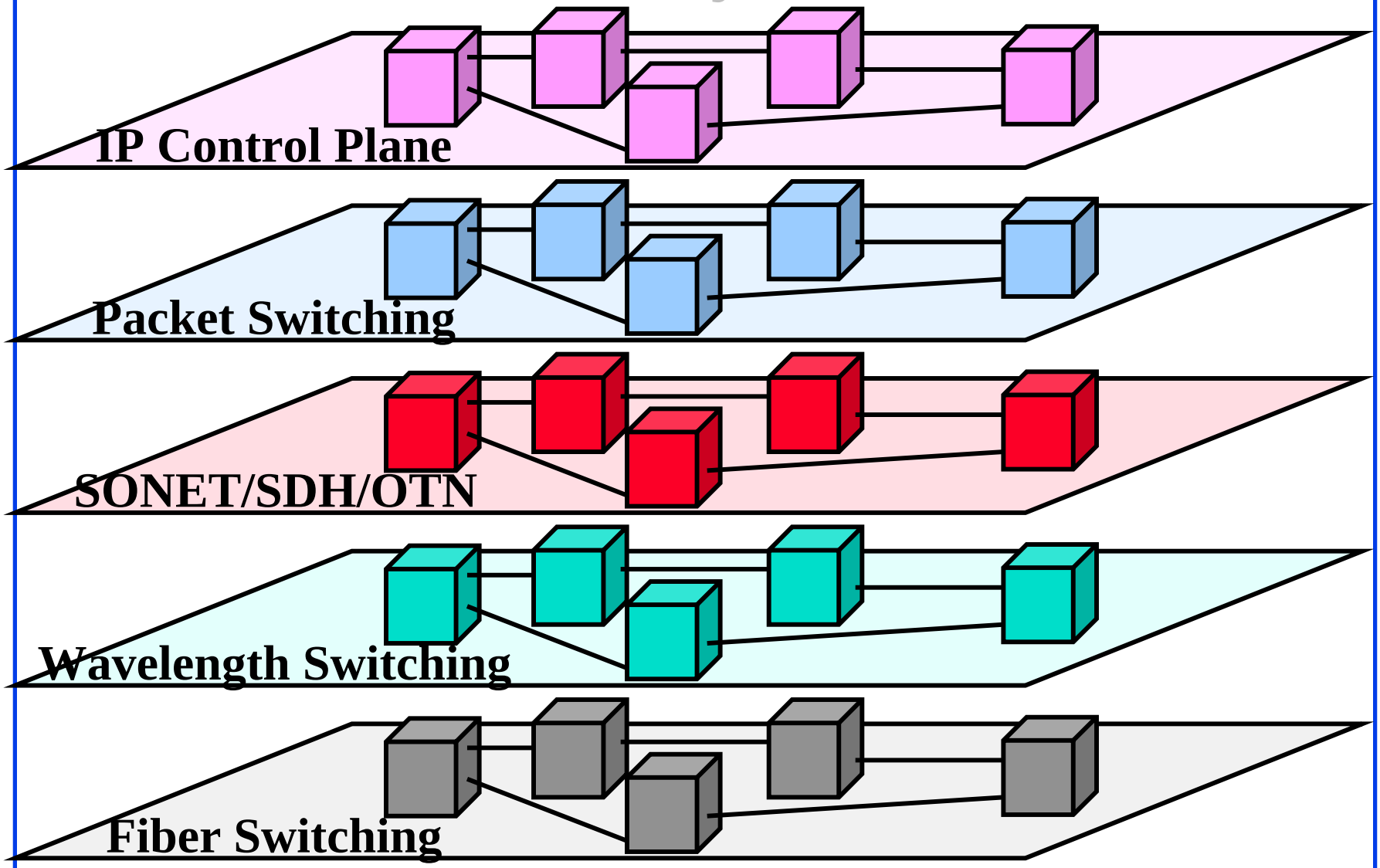
LSC = Lambda Switch Capable Nodes
= Optical Cross Connects = OXC

GMPLS

- Data Plane = Wavelengths, Fibers, SONET Frames, Packets (October 2000)
- Two separate routes: Data route and control route

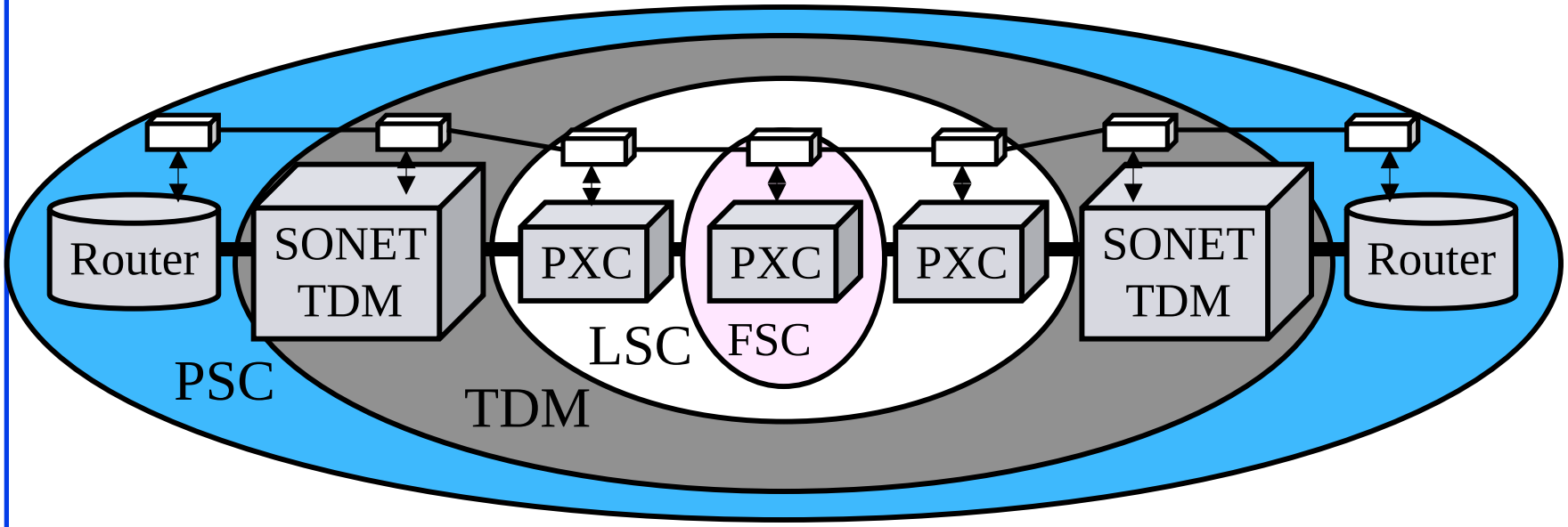


GMPLS: Layered View



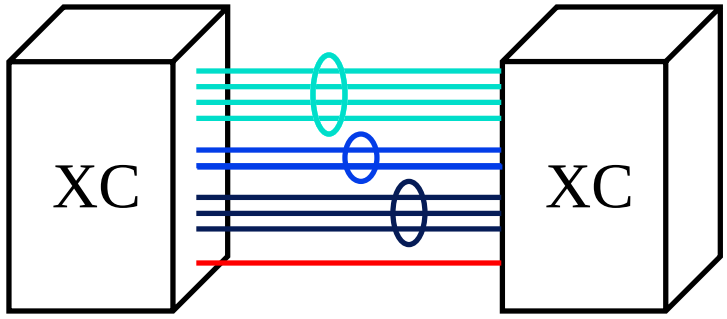
GMPLS: Hierarchical View

- Packets over SONET over Wavelengths over Fibers
- Packet switching regions, TDM regions, Wavelength switching regions, fiber switching regions
- Allows data plane connections between SONET ADMs, PXCs. FSCs, in addition to routers

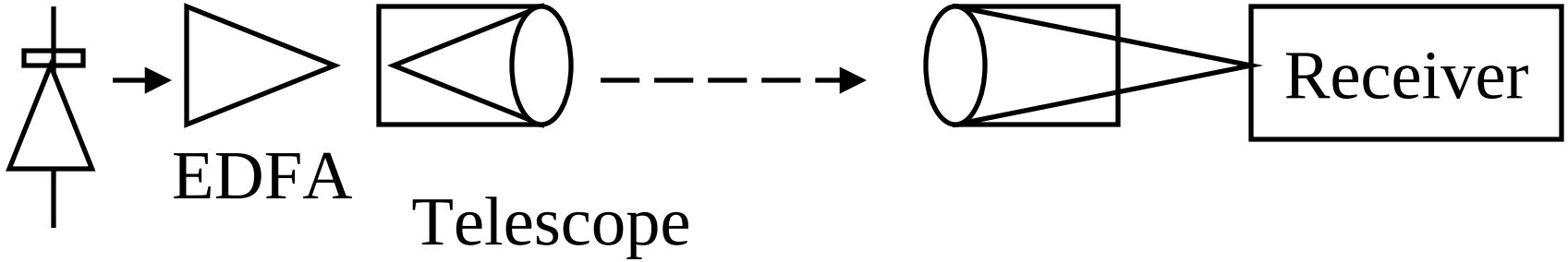


MPLS vs GMPLS

Issue	MPLS	GMPLS
Data & Control Plane	Same channel	Separate
Types of Nodes and labels	Packet Switching	PSC, TDM, LSC, FSC, …
Bandwidth	Continuous	Discrete: OC-n, λ's, ..
# of Parallel Links	Small	100-1000's
Port IP Address	One per port	Unnumbered
Fault Detection	In-band	Out-of-band or In-Band



Free Space Optical Comm



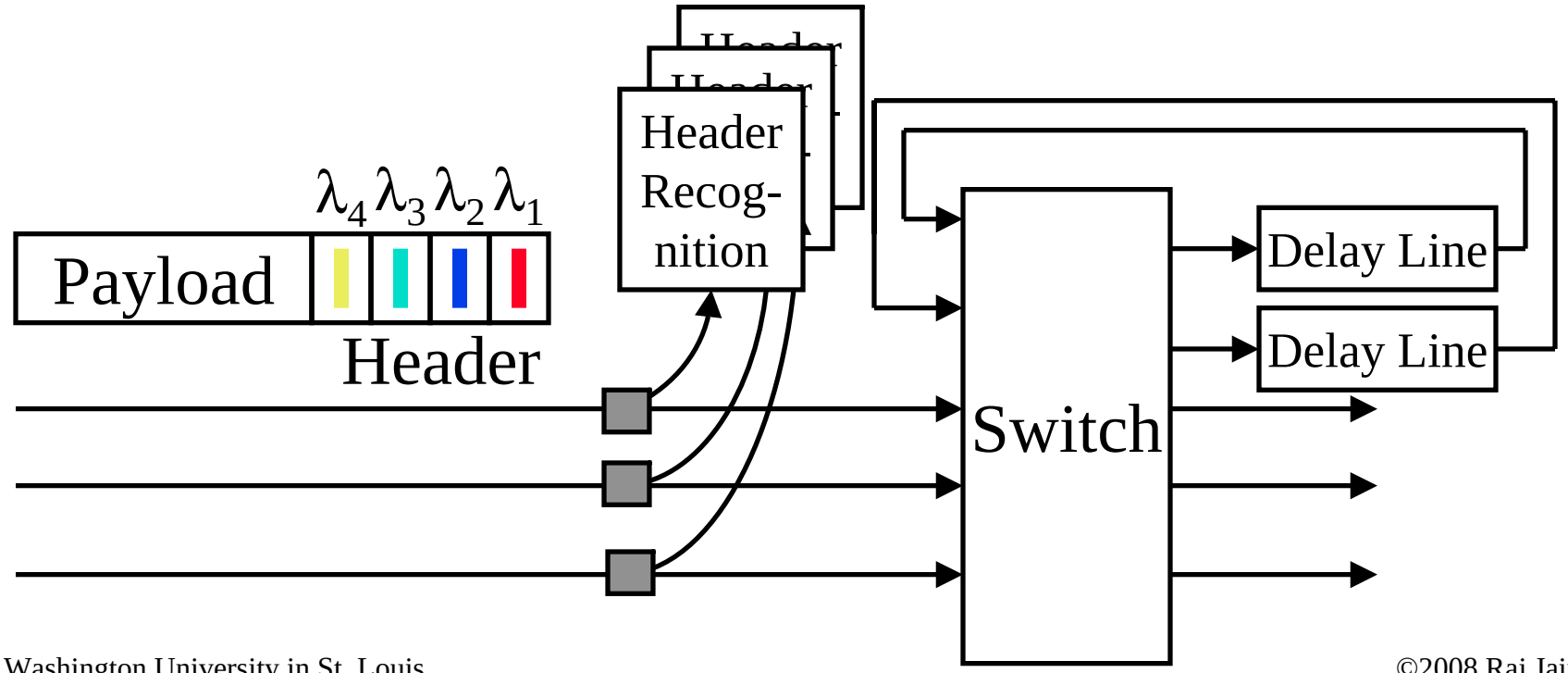
- Uses WDM in open air
- Sample Product:
  Lucent WaveStar OpticAir: 4×2.5Gbps to 5 km
  Available March'00.
- EDFA = Erbium Doped Fiber Amplifier

Free Space Optical Comm

- No FCC Licensing required
- Immunity from interference
- Easy installation
  $\Rightarrow$ Unlimited bandwidth, Easy Upgrade
- Transportable upon service termination or move
- Affected by weather (fog, rain)
  $\Rightarrow$ Need lower speed Microwave backup
- Example Products: Optical Crossing Optibridge 2500 2.5Gbps to 2km, Texas Instruments TALP1135 Chipset for 10/100 Mbps up to 50m

Optical Packet Switching

- Header Recognition: Lower bit rate or different λ
- Switching
- Buffering: Delay lines, Dispersive fiber



Summary

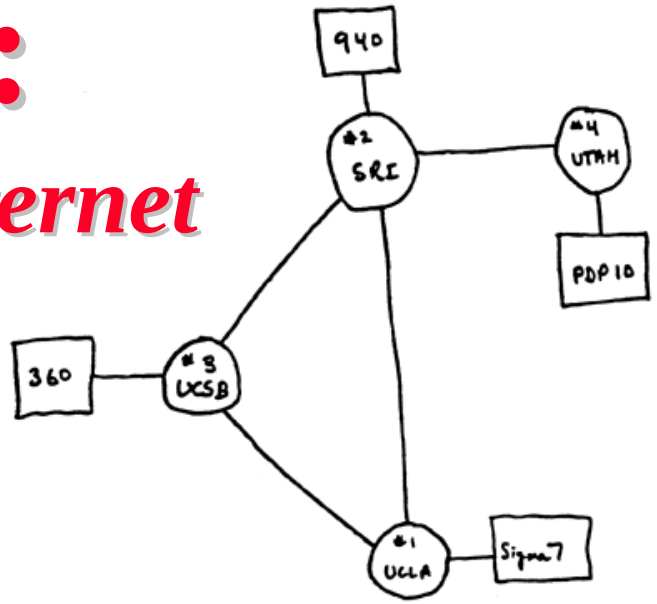
- O/O/O switches are bit rate and data format independent
- PONs provide a scalable, upgradeable, cost effective solution.
- High speed routers
  ⇒ IP directly over DWDM
- Separation of control and data plane
  ⇒ IP-Based control plane
- Transport Plane = Packets ⇒ MPLS
  Transport Plane = Wavelengths
  ⇒ MPλS
  Transport Plane = λ, SONET, Packets ⇒ GMPLS

References

- Detailed references in http://www.cse.wustl.edu/~jain/refs/opt_refs.htm
- Recommended books on optical networking, http://www.cse.wustl.edu/~jain/refs/opt_book.htm
- Optical Networking and DWDM, http://www.cse.wustl.edu/~jain/cis788-99/dwdm/index.html
- IP over Optical: A summary of issues, (internet draft) http://www.cse.wustl.edu/~jain/ietf/issues.html
- Lightreading, http://www.lightreading.com

Internet 3.0:

The Next Generation Internet



Raj Jain
Washington University in Saint Louis
Saint Louis, MO 63130
Jain@wustl.edu

1. What is Internet 3.0?
2. Why should you keep on the top of Internet 3.0?
3. What are we missing in the current Internet?
4. Our Proposed Architecture for Internet 3.0

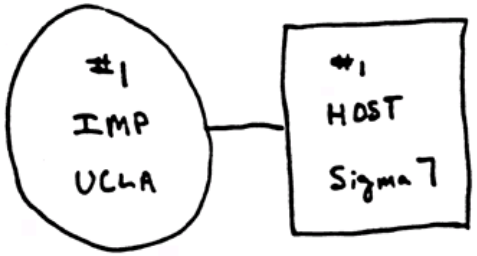
Acknowledgement: This research is sponsored by a grant from Intel Research Council.

Internet 3.0

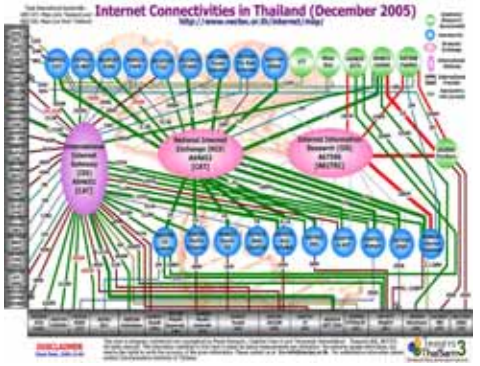
- US National Science Foundation started a large research and infrastructure program on next generation Internet
  - Testbed: "Global Environment for Networking Innovations" (GENI)
  - Architecture: "Future Internet Design" (FIND).
- Q: How would you design Internet today? Clean slate design.
- Ref: http://www.nsf.gov/cise/cns/geni/
- Most of the networking researchers will be working on GENI/FIND for the coming years
- Internet 3.0 is the name of the Washington University project on the next generation Internet
- Named by me along the lines of "Web 2.0"
- Internet 3.0 is more intuitive then GENI/FIND

Internet Generations

- **Internet 1.0** (1969 – 1989) – Research project
  - RFC1 is dated April 1969.
  - ARPA project started a few years earlier
  - IP, TCP, UDP
  - Mostly researchers
  - Industry was busy with proprietary protocols: SNA, DECnet, AppleTalk, XNS

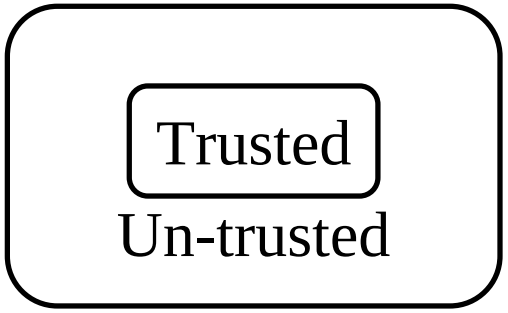


- **Internet 2.0** (1989 – Present) – Commerce ⇒ new requirements
  - Security RFC1108 in 1989
  - NSFnet became commercial
  - Inter-domain routing: OSPF, BGP,
  - IP Multicasting
  - Address Shortage IPv6
  - Congestion Control, Quality of Service,…



Ten Problems with Current Internet

1. Designed for research
   ⇒ Trusted systems
   Used for Commerce
   ⇒ Untrusted systems
2. Control, management, and Data path are intermixed ⇒ security issues
3. Difficult to represent organizational, administrative hierarchies and relationships. Perimeter based.



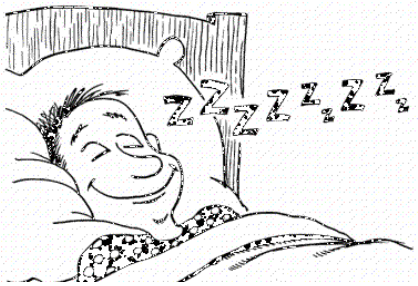
Problems (cont)

4. Identity and location in one (IP Address)
   Makes mobility complex.

5. Location independent addressing
   ⇒ Most services require nearest server.
   ⇒ Also, Mobility requires location

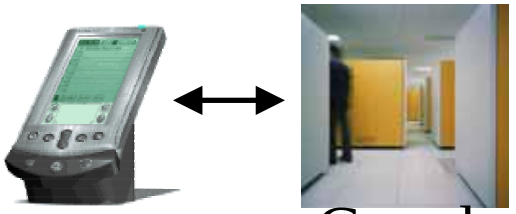
6. No representation for real end system: the human.

Problems (cont)



7. Assumes live and awake end-systems
   Does not allow communication while sleeping.
   Many energy conscious systems today sleep.
8. Single-Computer to single-computer communication ⇒ Numerous patches needed for communication with globally distributed systems and services.

9. Symmetric Protocols
   ⇒ No difference between a PDA and a Google server.



Google

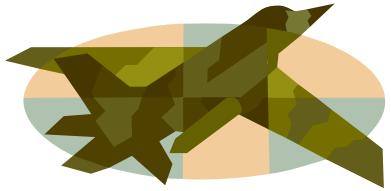
Problems (Cont)

10. Stateless ⇒ Can't remember a flow
    ⇒ QoS difficult.
    QoS is generally for a flow and not for one packet

Our Proposed Solution: Internet 3.0

- Take the best of what is already known
  - Wireless Networks, Optical networks, …
  - Transport systems: Airplane, automobile, …
  - Communication: Wired Phone, Cellular nets,…
- Develop a consistent general purpose, evolvable architecture that can be customized by implementers, service providers, and users



Names, IDs, Addresses

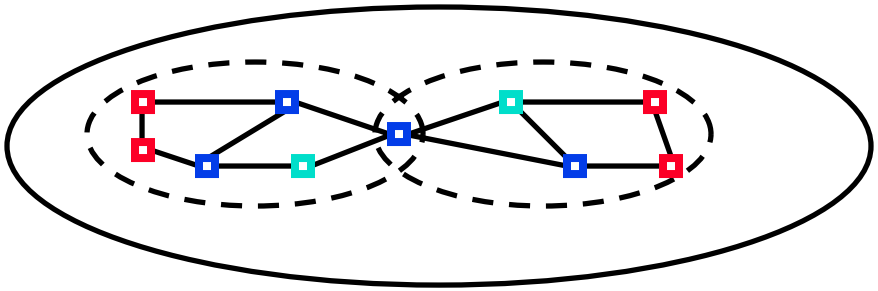
Name: John Smith

ID: 012-34-5678

Address:
1234 Main Street
Big City, MO 12345
USA

- Address changes as you move, ID and Names remain the same.
- **Examples**:
  - Names: Company names, DNS names (microsoft.com)
  - IDs: Cell phone numbers, 800-numbers, Ethernet addresses, Skype ID, VOIP Phone number
  - Addresses: Wired phone numbers, IP addresses

Realms



- Object names and Ids are defined within a realm
- A realm is a **logical** grouping of objects under an administrative domain
- The Administrative domain may be based on Trust Relationships
- A realm represents an organization
  - Realm managers set policies for communications
  - Realm members can share services.
  - Objects are generally members of multiple realms
- Realm Boundaries: Organizational, Governmental, ISP, P2P,…

Realm = Administrative Group

Physical vs Logical Connectivity

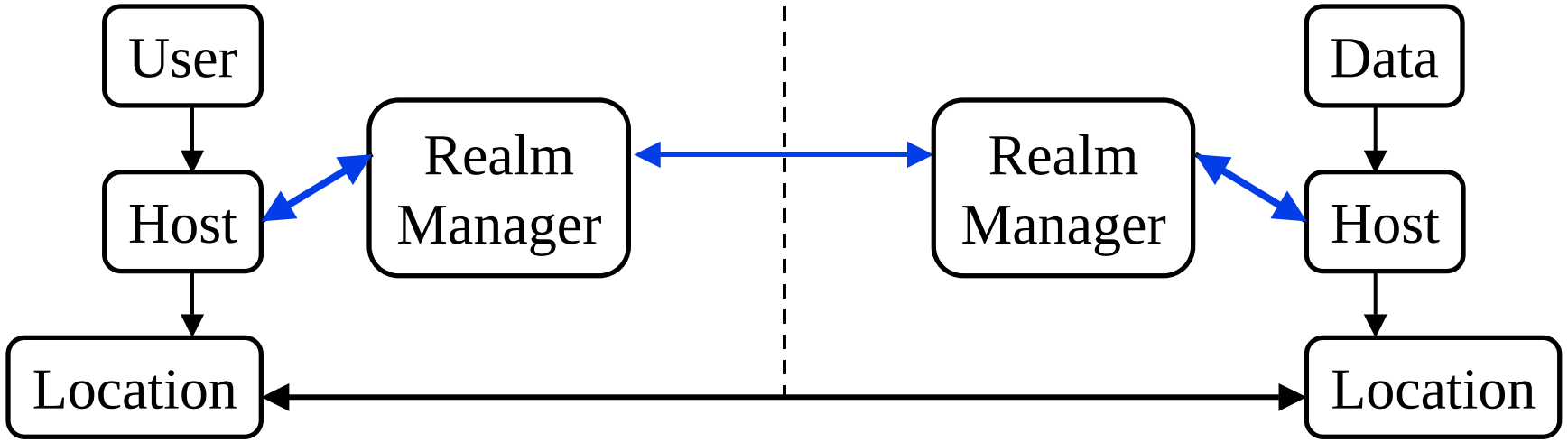
- Physically and logically connected:
  All computers in my lab
  = Private Network,
  Firewalled Network

- Physically disconnected but logically connected:
  My home and office computers

- Physically connected but logically disconnected: Passengers on a plane, Neighbors, Conference attendees sharing a wireless network, A visitor

Physical connectivity ≠ Trust

Id-Locator Split Architecture (MILSA)



- Realm managers:
  - Resolve current location for a given host-ID
  - Enforce policies related to authentication, authorization, privacy
  - Allow mobility, multi-homing, location privacy
  - Similar to several other proposals
- Ref: Our Globecom 2008 paper [2]

Server and Gatekeeper Objects

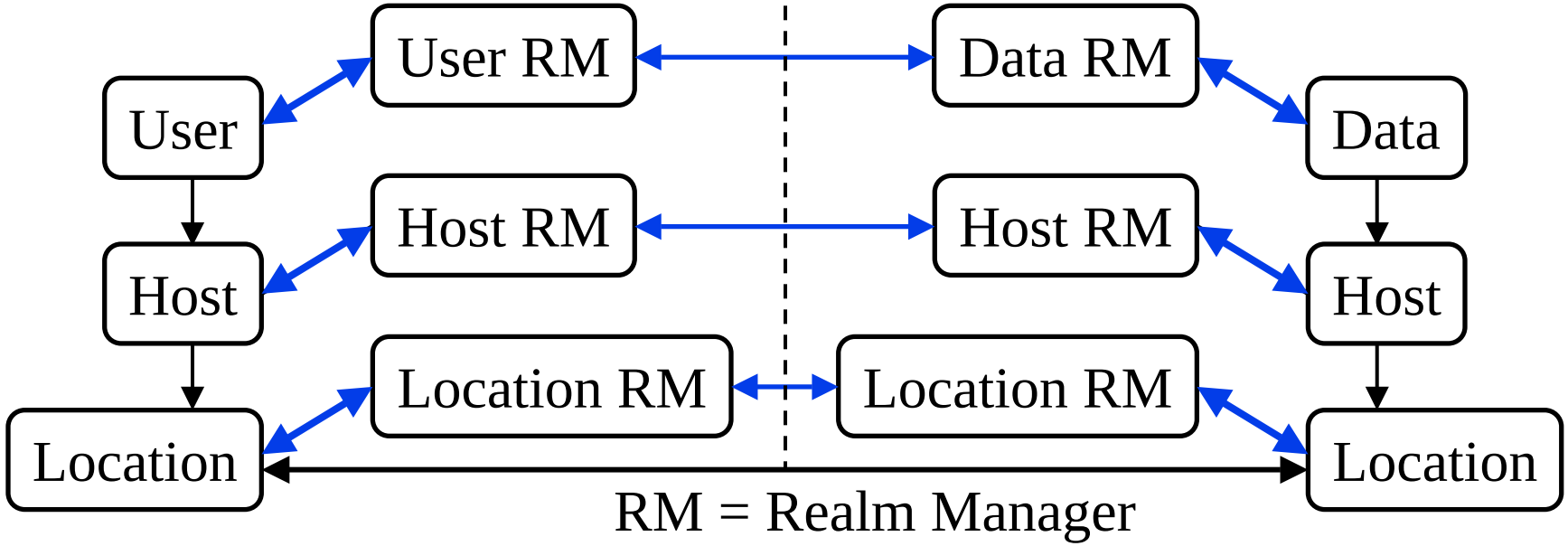
- Each realm has a set of server objects, e.g., forwarding, authentication, encryption, storage, transformation, …
- Some objects have built-in servers, e.g., an "enterprise router" may have forwarding, encryption, authentication services.
- Other objects rely on the servers in their realm
- Authentication servers (AS) add their signatures to packets and verify signatures of received packets..
- Storage servers store packets while the object may be sleeping and may optionally aggregate/compress/transform data. Could wake up objects.
- Objects can appoint proxies for any function(s)
- Gatekeepers enforce policies: Security, traffic, QoS

Servers allow simple energy efficient end devices

User- Host- and Data Centric Models

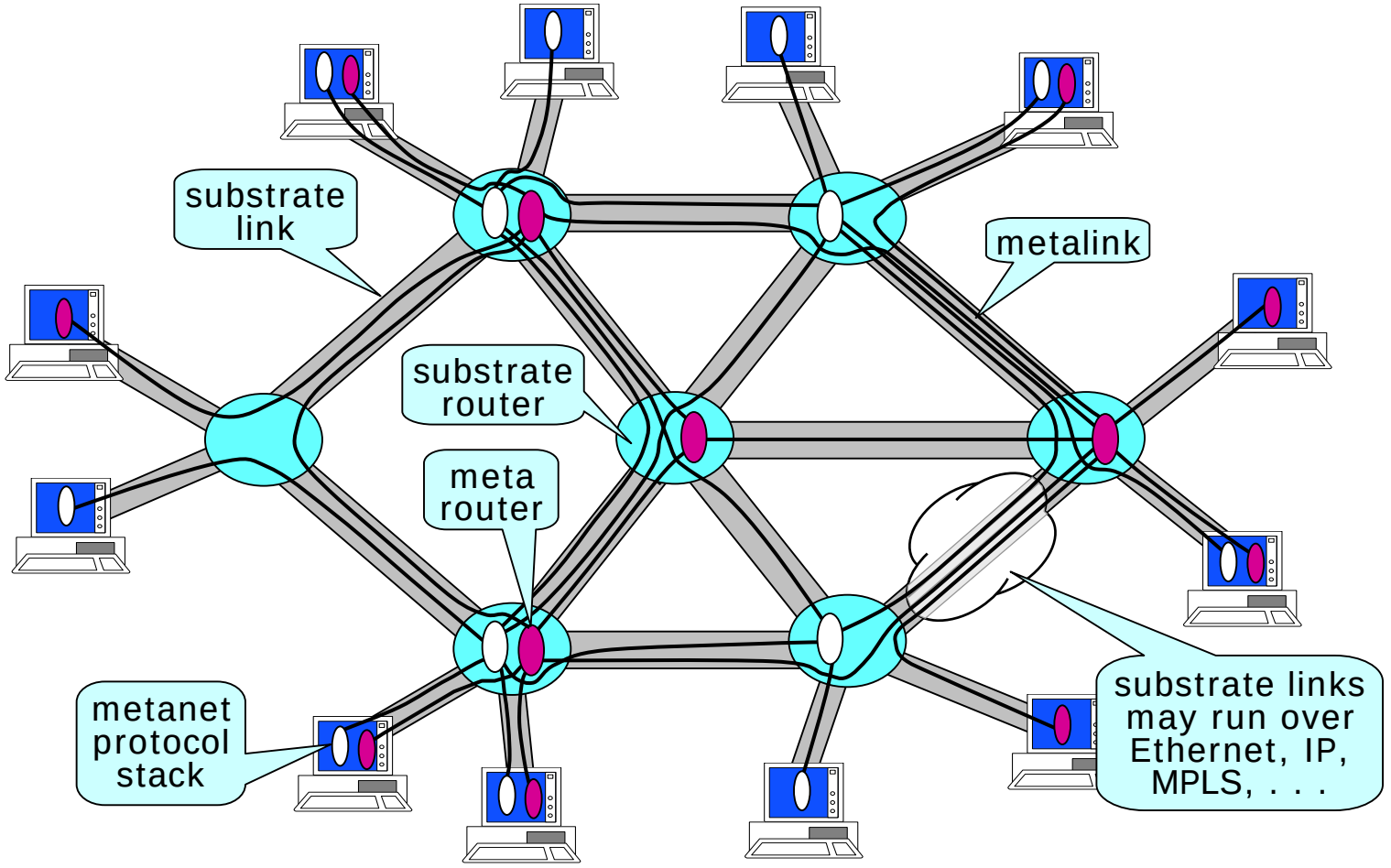
- All discussion so far assumed host-centric communication
  - Host mobility and multihoming
  - Policies, services, and trust are related to hosts
- User Centric View:
  - Bob wants to watch a movie
  - Starts it on his media server
  - Continues on his iPod during commute to work
  - Movie exists on many servers
  - Bob may get it from different servers at different times or multiple servers at the same time
- Can we just give addresses to users and treat them as hosts? No! ⇒ Policy Oriented Naming Architecture (PONA)

Policy Oriented Naming Architecture



- Both Users and data need hosts for communication
- Data is easily replicable. All copies are equally good.
- Users, Hosts, Infrastructure, Data belong to different realms (organizations).
- Each object has to follow its organizational policies.

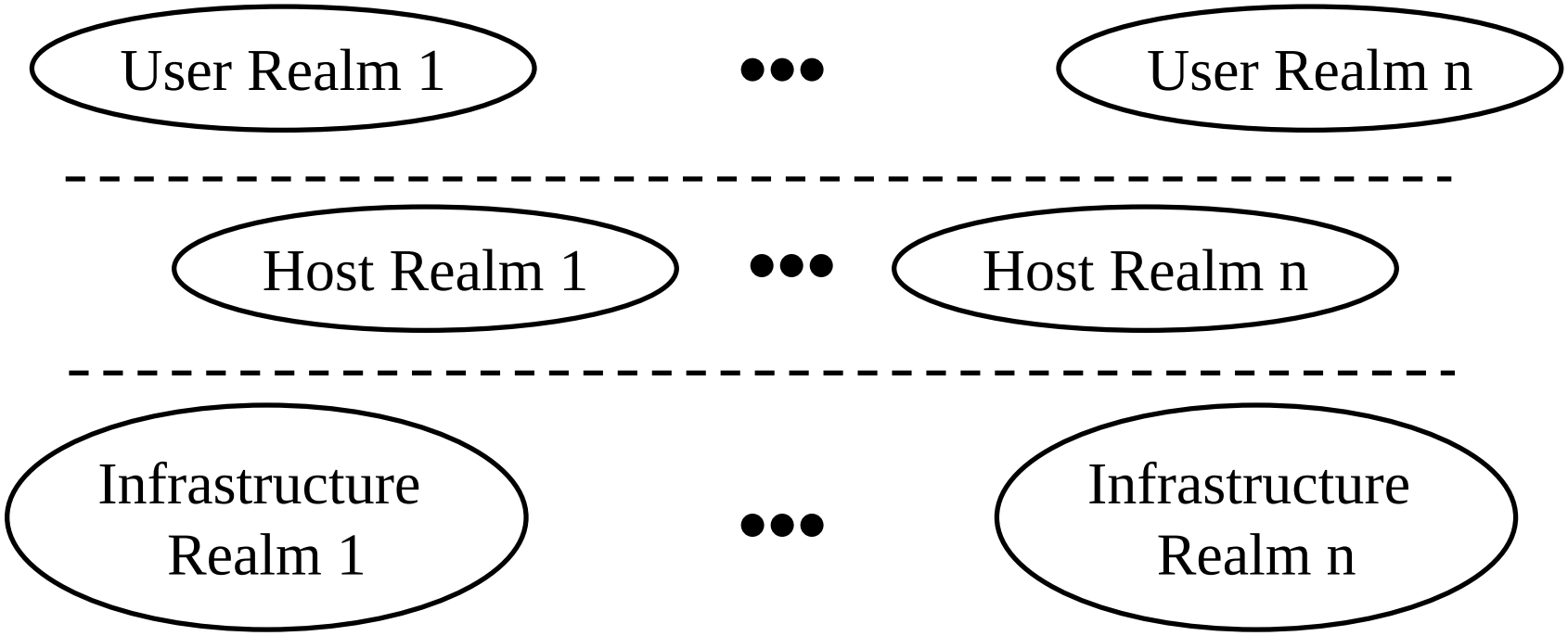
Virtualizable Network Concept



Ref: T. Anderson, L. Peterson, S. Shenker, J. Turner, "Overcoming the Internet Impasse through Virtualization," Computer, April 2005, pp. 34 – 41.

Slide taken from Jon Turner's presentation at Cisco Routing Research Symposium

Realm Virtualization

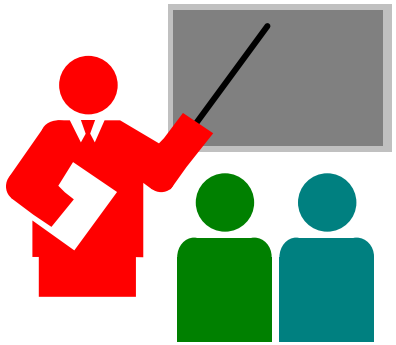


- Old: Virtual networks on a common infrastructure
- New: Virtual user realms on virtual host realms on a group of infrastructure realms. 3-level hierarchy not 2-level. Multiple organizations at each level.

Internet 1.0 vs. Internet 3.0

	Feature	Internet 1.0	Internet 3.0
1.	Energy Efficiency	Always-on	Green ⇒ Mostly Off
2.	Mobility	Mostly stationary computers	Mostly mobile *objects*
3.	Computer-Human Relationship	Multi-user systems ⇒ Machine to machine comm.	Multi-systems user ⇒ Personal comm. systems
4.	End Systems	Single computers	Globally distributed systems
5.	Protocol Symmetry	Communication between equals ⇒ Symmetric	Unequal: PDA vs. big server ⇒ Asymmetric
6.	Design Goal	Research ⇒ Trusted Systems	Commerce ⇒ No Trust Map to organizational structure
7.	Ownership	No concept of ownership	Hierarchy of ownerships, administrations, communities
8.	Sharing	Sharing ⇒ Interference, QoS Issues	Sharing *and* Isolation ⇒ Critical infrastructure
9.	Switching units	Packets	Packets, Circuits, Wavelengths, Electrical Power Lines, …
10.	Applications	Email and Telnet	Information Retrieval, Distributed Computing, Distributed Storage, Data diffusion

Summary



1. Internet 3.0 is the next generation of Internet.
2. It must be secure, allow mobility, and be energy efficient.
3. Must be designed for commerce
   $\Rightarrow$ Must represent multi-organizational structure and policies
4. Moving from host centric view to user-data centric view
   $\Rightarrow$ Important to represent users and data objects
5. Users, Hosts, and infrastructures belong to different realms (organizations). Users/data/hosts should be able to move freely without interrupting a network connection.

References

1. Jain, R., "Internet 3.0: Ten Problems with Current Internet Architecture and Solutions for the Next Generation," in Proceedings of Military Communications Conference (MILCOM 2006), Washington, DC, October 23-25, 2006, http://www.cse.wustl.edu/~jain/papers/gina.htm
2. Subharthi Paul, Raj Jain, Jianli Pan, and Mic Bowman, "A Vision of the Next Generation Internet: A Policy Oriented View," British Computer Society Conference on Visions of Computer Science, Sep 2008, http://www.cse.wustl.edu/~jain/papers/pona.htm
3. Jianli Pan, Subharthi Paul, Raj Jain, and Mic Bowman, "MILSA: A Mobility and Multihoming Supporting Identifier-Locator Split Architecture for Naming in the Next Generation Internet,," Globecom 2008, Nov 2008, http://www.cse.wustl.edu/~jain/papers/milsa.htm