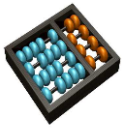
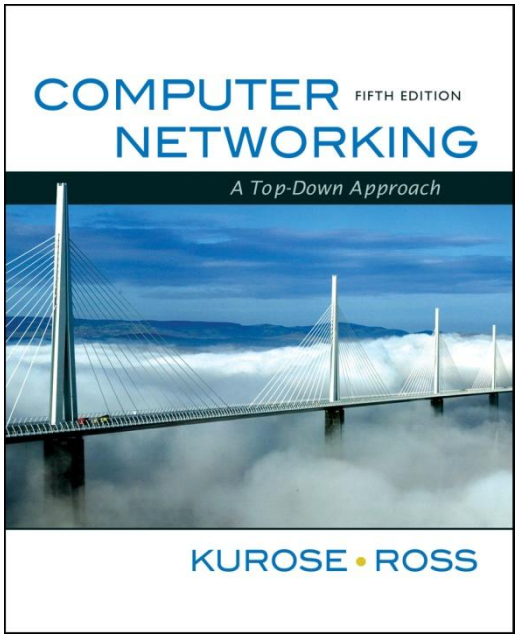
Camada de Enlace de Dados

Prof Nelson Fonseca



Chapter 5
Link Layer and LANs



A note on the use of these ppt slides:

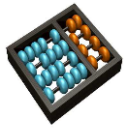
We're making these slides freely available to all (faculty, students, readers). They're in PowerPoint form so you can add, modify, and delete slides (including this one) and slide content to suit your needs. They obviously represent a *lot* of work on our part. In return for use, we only ask the following:

- ❑ If you use these slides (e.g., in a class) in substantially unaltered form, that you mention their source (after all, we'd like people to use our book!)
- ❑ If you post any slides in substantially unaltered form on a www site, that you note that they are adapted from (or perhaps identical to) our slides, and note our copyright of this material.

Thanks and enjoy! JFK/KWR

Computer Networking: A Top Down Approach
5th edition.
Jim Kurose, Keith Ross
Addison-Wesley, April 2009.

All material copyright 1996-2009
J.F Kurose and K.W. Ross, All Rights Reserved



- Alguns slides nesse arquivo foram gentilmente cedidos pelos autores do livro:
- Computer Networks: An Open Source Approach, Ying-Dar Lin, Ren-Hung Hwang, Fred Baker, published by McGraw Hill, Feb 2011

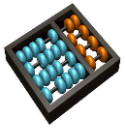
Capítulo 5: A Camada de Enlace

Nossos objetivos:

- entender os princípios por trás dos serviços da camada de enlace:
  - ✓ detecção de erros, correção
  - ✓ compartilhando um canal broadcast: acesso múltiplo
  - ✓ endereçamento da camada de enlace
  - ✓ trasnferência de dados confiável, controle de fluxo: *já visto!*
- instanciação e implementação de várias tecnologias da camada de enlace

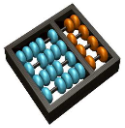
Visão Geral:

- serviços da camada de enlace
- detecção de erros, correção
- protocolos de acesso múltiplo e LANs
- endereçamento da camada de enlace, ARP
- tecnologias específicas da camada de enlace:
  - ✓ Ethernet
  - ✓ hubs, switches
  - ✓ MPLS
  - ✓ Data centers



Camada de Enlace de Dados

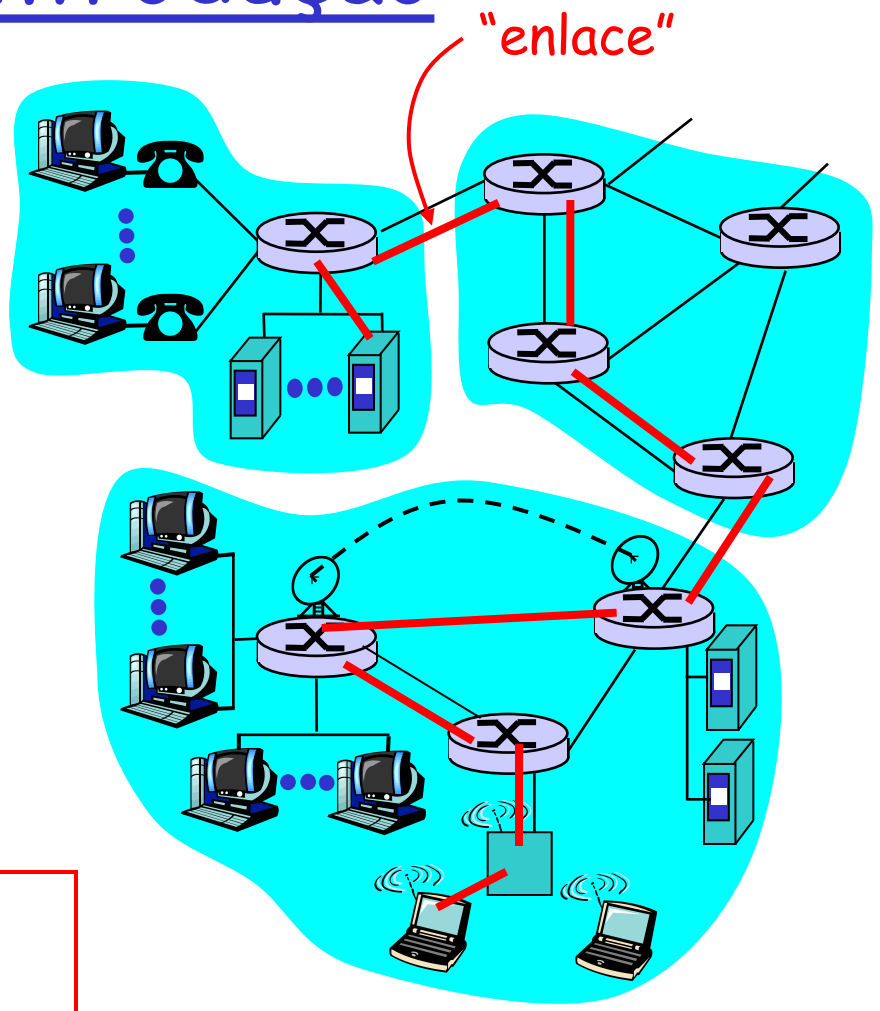
- 5.1 Introdução e Serviços
- 5.2 Correção e detecção de
- 5.3protocolos Múltilpo Acesso
- 5.4 Endereçamento
- 5.5 Ethernet
- 5.6 Switches
- 5.7 PPP
- 5.8 Virtualização: MPLS
- 5.9 Data Center Networks



Camada de Enlace: Introdução

Alguma terminologia:

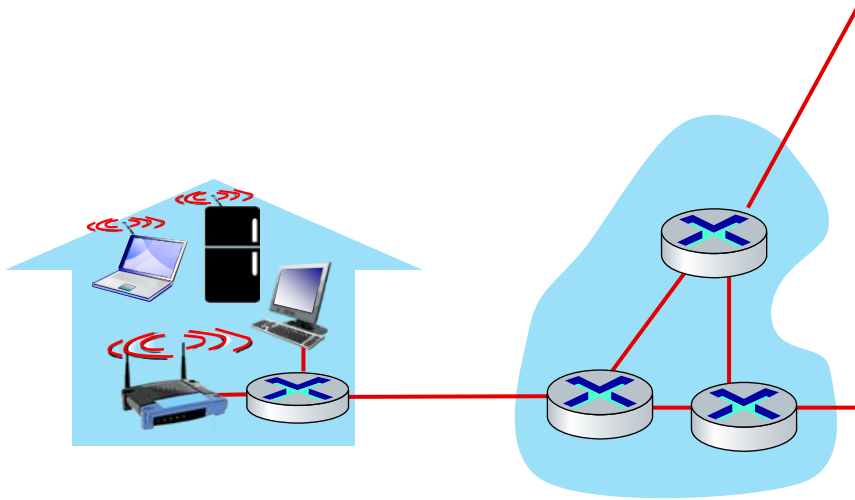
- hosts e roteadores são **nós** (pontes e comutadores também)
- **Enlaces** são canais de comunicação que conectam nós adjacentes ao longo dos caminhos de comunicação
  - ✓ Enlaces cabeados
  - ✓ Enlaces sem fios
  - ✓ LANs
- 2-PDU é um **quadro**, que encapsula um datagrama



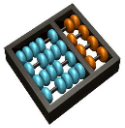
Camada de enlace tem a responsabilidade de transferir datagramas de um nó para o nó ficamente adjacente sobre um enlace

Link layer: context

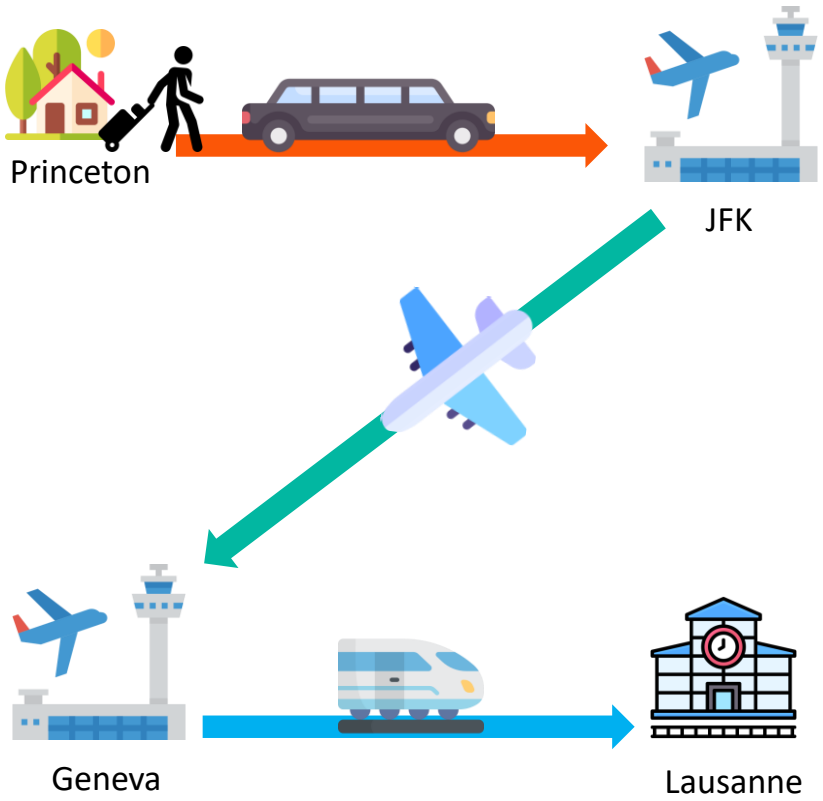
- datagram transferred by different link protocols over different links:
  - e.g., WiFi on first link, Ethernet on next link
- each link protocol provides different services
  - e.g., may or may not provide reliable data transfer over link



UNICAMP



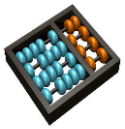
Transportation analogy



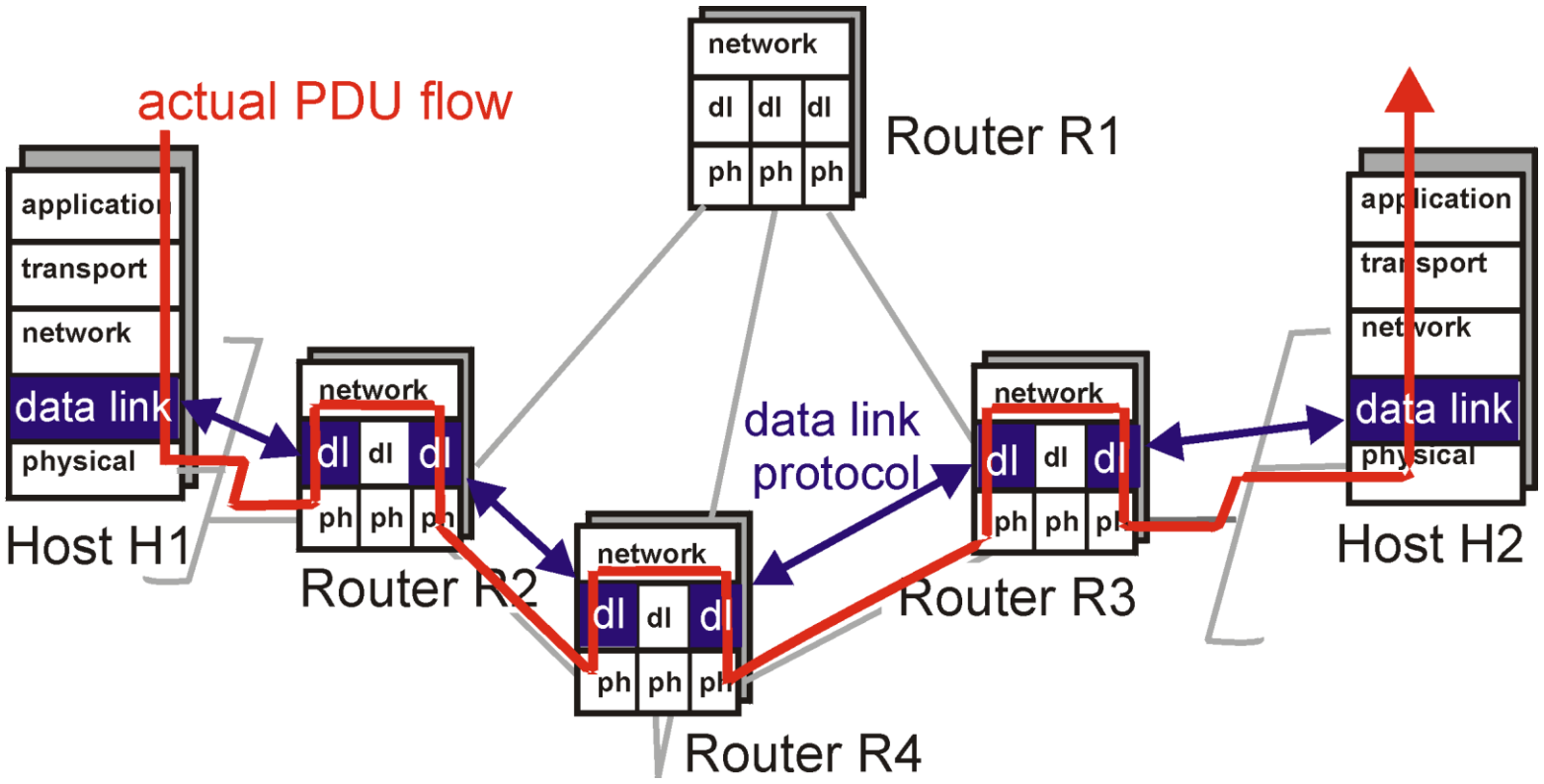
transportation analogy:

- trip from Princeton to Lausanne
  - limo: Princeton to JFK
  - plane: JFK to Geneva
  - train: Geneva to Lausanne
- tourist = datagram
- transport segment = communication link
- transportation mode = link-layer protocol
- travel agent = routing algorithm

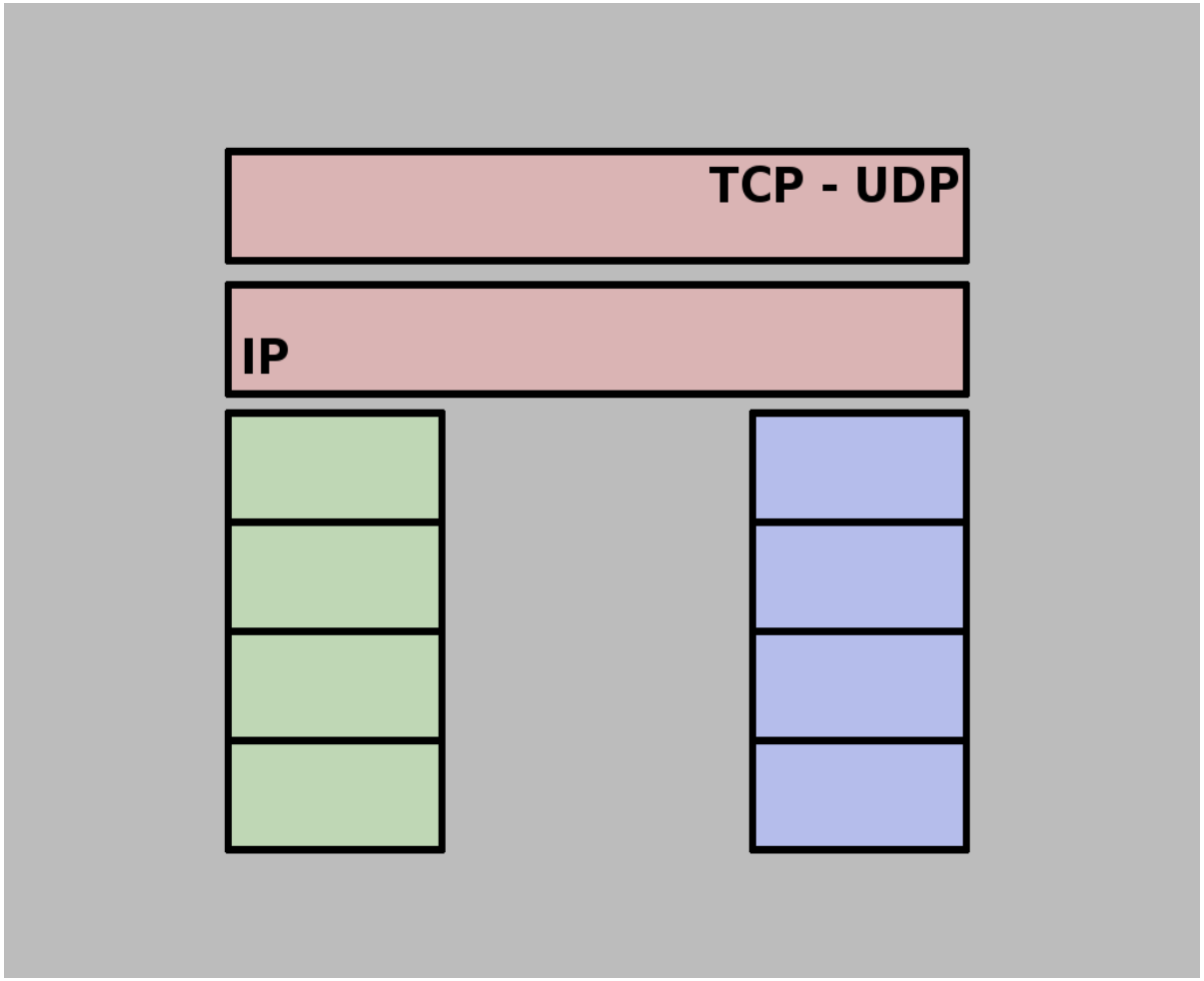
UNICAMP



Protocolos da Camada de Enlace

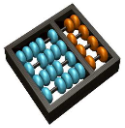


A Internet



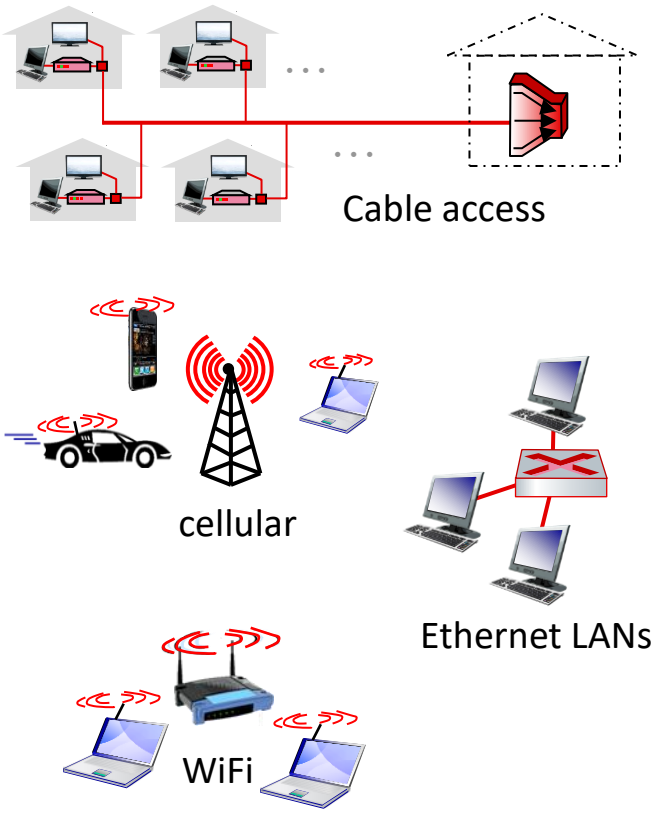
Tecnologia de redes

	PAN/LAN	MAN/WAN
Obsolete or Fading away	Token bus (802.4) Token ring (802.5) HIPPI Fiber Channel Isochronous (802.9) Demand Priority (802.12) ATM FDDI HIPERLAN	DQDB (802.6) B-ISDN HDLC X.25 Frame Relay SMDS ISDN
Mainstream or Still active	Ethernet (802.3) WLAN (802.11) Bluetooth (802.15) Fiber channel HomeRF HomePlug	Ethernet (802.3) Point-to-Point Protocol (PPP) DOCSIS xDSL SONET Cellular(3G, LTE, WiMAX(802.16)) Resilient Packet Ring (802.17) ATM



Link layer: services

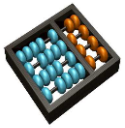
- framing, link access:
  - encapsulate datagram into frame, adding header, trailer
  - channel access if shared medium
  - "MAC" addresses in frame headers identify source, destination (different from IP address!)
- reliable delivery between adjacent nodes
  - we already know how to do this!
  - seldom used on low bit-error links
  - wireless links: high error rates
    - ***Q:*** **why both link-level and end-end reliability?**



UNICAMP

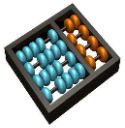
Serviços da Camada de Enlace

- ***Enquadramento e acesso ao enlace:***
  - ✓ encapsula datagrama num quadro incluindo cabeçalho e cauda,
  - ✓ implementa acesso ao canal se meio for compartilhado,
  - ✓ 'endereços físicos' são usados em cabeçalhos de quadros para identificar origem e destino de quadros em enlaces multiponto
    - diferente do endereço IP !
- ***Entrega confiável:***
  - ✓ já aprendemos como isto deve ser feito (capítulo 3)!
  - ✓ raramente usado em enlaces com baixa taxa de erro (fibra, alguns tipos de par trançado)
  - ✓ usada em enlaces sem-fio (wireless): altas taxas de erro
    - Q: porque prover confiabilidade fim-a-fim e na camada de enlace?

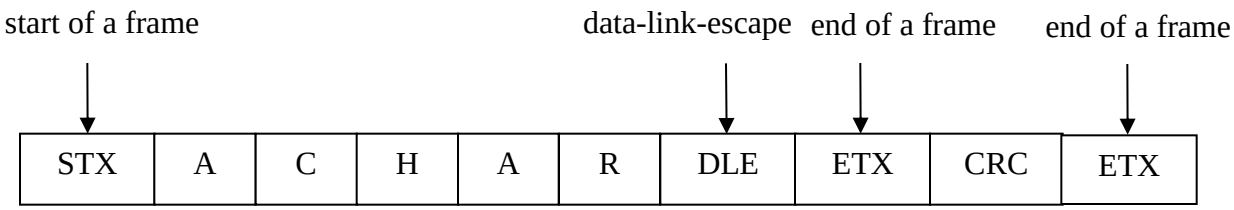


Enquadramento

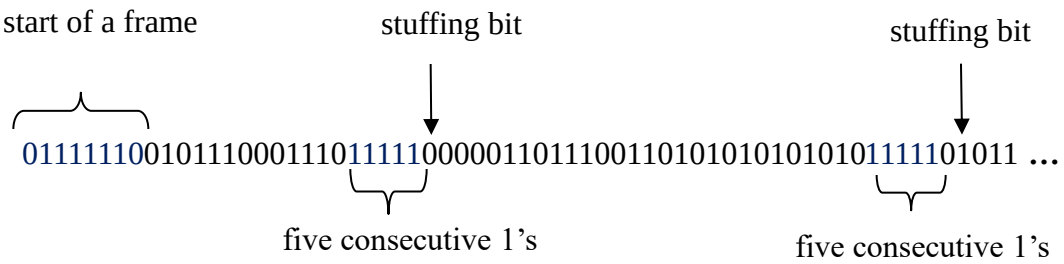
- Métodos para delimitar os quadros:
  - ✓ Caracteres especiais
    e.g. STX (Start of text), ETX (End of text)
  - ✓ Padrão específico de bits
    e.g. a bit pattern 01111110
  - ✓ Padrão específico de símbolo na camada física
    e.g. /J/K/ and /T/R/ code group in 100BASE-X
- Bit (or byte) stuffing para evitar ambiguidade



Bit-Stuffing and Byte-Stuffing



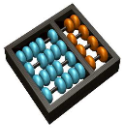
(a) byte-stuffing



(b) bit-stuffing

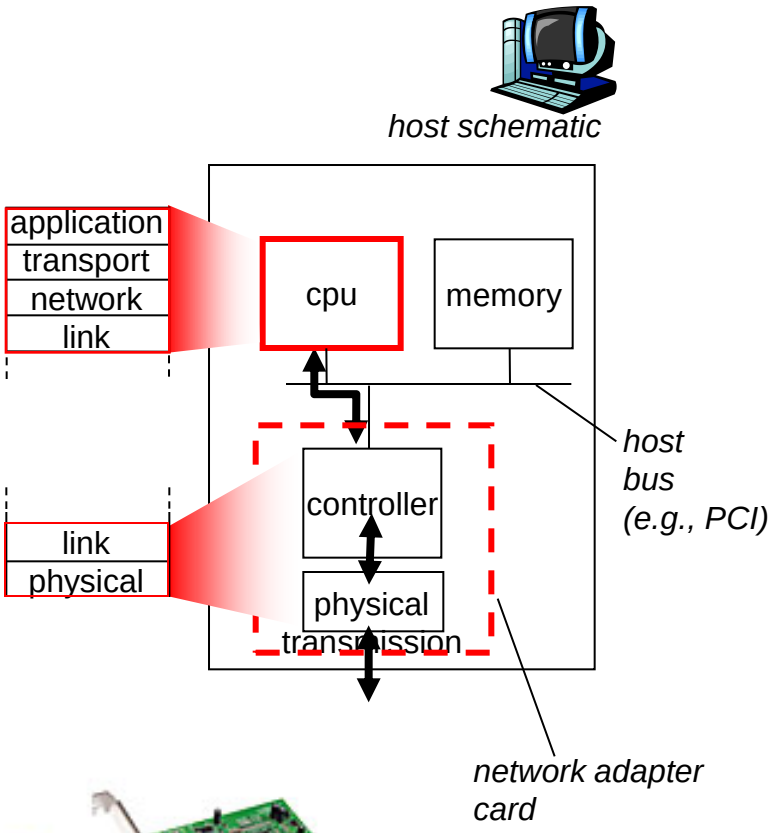
Serviços da Camada de Enlace (mais)

- ***Controle de Fluxo***:
  - ✓ compatibilizar taxas de produção e consumo de quadros entre remetentes e receptores
- ***Detecção de Erros***:
  - ✓ erros são causados por atenuação do sinal e por ruído
  - ✓ receptor detecta presença de erros
    - receptor sinaliza ao remetente para retransmissão, ou simplesmente descarta o quadro em erro
- ***Correção de Erros***:
  - ✓ mecanismo que permite que o receptor localize e corrija o erro sem precisar da retransmissão
- ***Half-duplex e full-duplex***
  - ✓ Com half duplex, os dois nós do enlace podem transmitir, mas não ao mesmo tempo



Aonde é implementado o protocolo de enlace de dados?

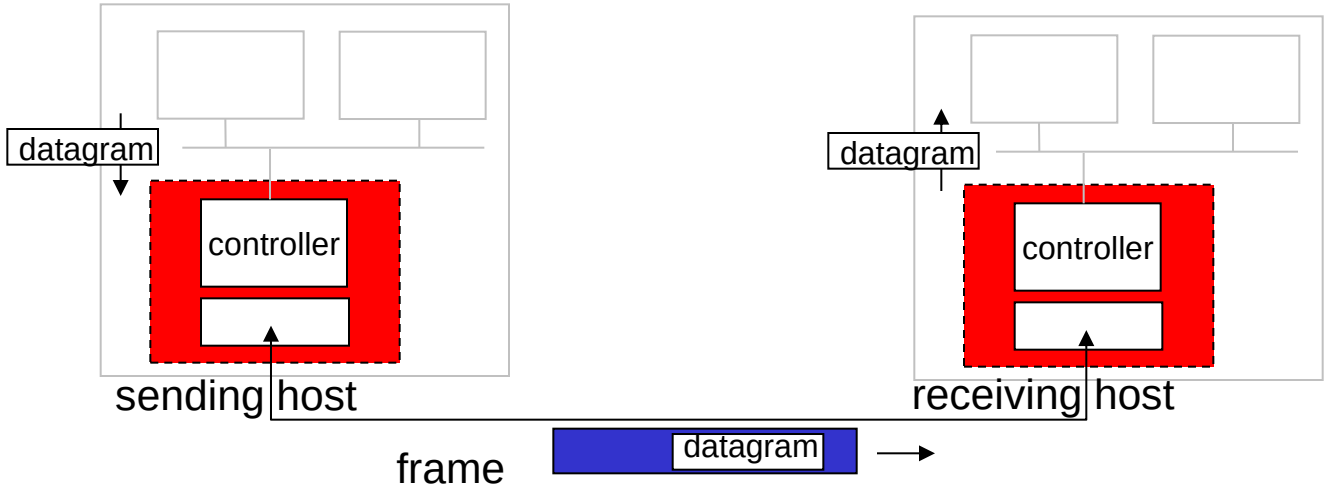
- Implementado em todos os host e interfaces
- Implementado na placa de redes" (*network interface card* NIC)
  - ✓ Ethernet card, PCMCI card, 802.11 card
  - ✓ Implementa camada de redes e física correspondente
- Acoplado ao barramento da rede
- Combinacão de hardware, software, firmware



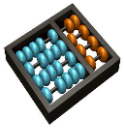
Implementação de Protocolo da Camada de Enlace

- Protocolo da camada de enlace é implementado totalmente no adaptador (p.ex., cartão PCMCIA).
  - ✓ Adaptador tipicamente inclui: RAM, circuitos de processamento digital de sinais, interface do barramento do computador, e interface do enlace
  - ✓ Adaptador é semi-autônomo
- Enlace e camadas físicas

Implementação de Protocolo da Camada de Enlace

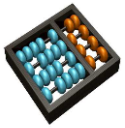


- **transmissão** do adaptador:
  - encapsula (coloca número de seqüência, info de realimentação, etc.)
  - inclui bits de detecção de erros
  - implementa acesso ao canal para meios compartilhados
  - coloca no enlace
- **recepção** do adaptador:
  - verificação e correção de erros
  - interrompe computador para enviar quadro para a camada superior
  - atualiza info de estado a respeito de realimentação para o remetente, número de seqüência, etc.



Camada de Enlace de Dados

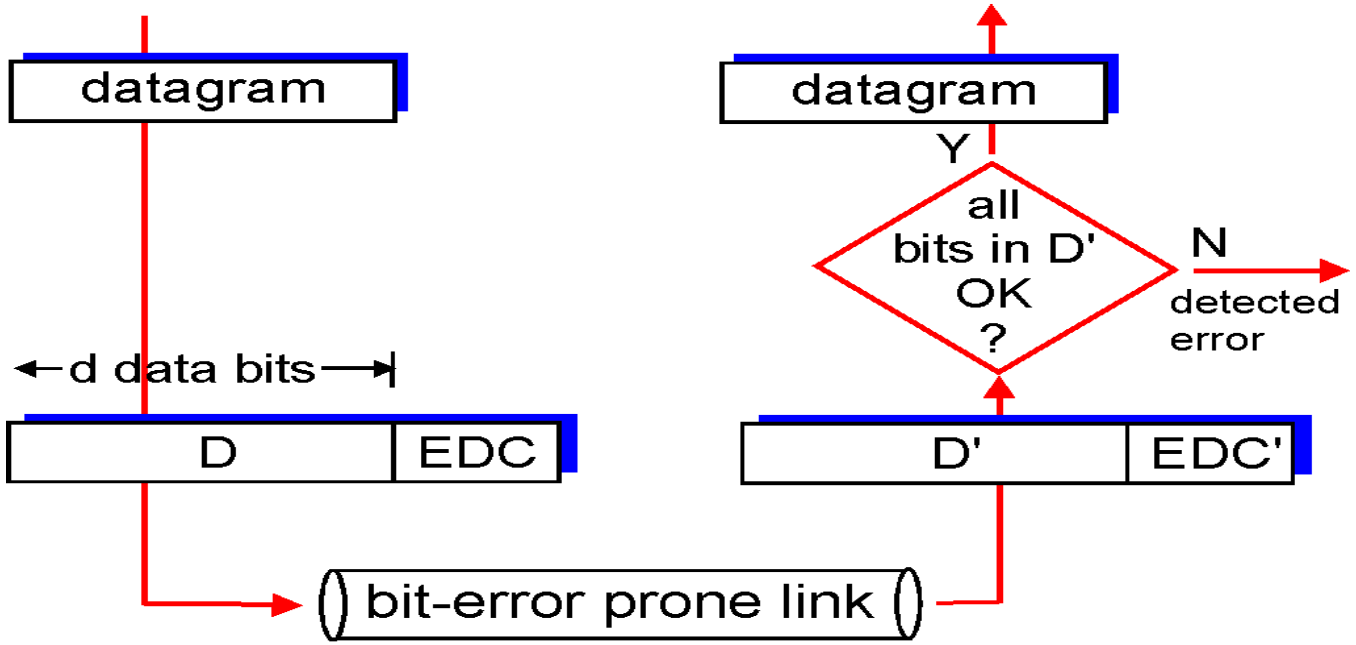
- 5.1 Introdução e Serviços
- 5.2 Correção e detecção de erros
- 5.3 Protocolos Múltilpo Acesso
- 5.4 Endereçamento
- 5.5 Ethernet
- 5.6 Switches
- 5.7 PPP
- 5.8 Virtualização: MPLS
- 5.9 Data Center Networks



Detecção de Erros

EDC= bits de Detecção e Correção de Erros (redundância)
D = Dados protegidos por verificação de erros,
podem incluir alguns campos do cabeçalho

- detecção de erros não é 100% perfeita;
- protocolo pode não identificar alguns erros, mas é raro
- Quanto maior o campo EDC melhor é a capacidade de detecção e correção de erros



Checksum da Internet

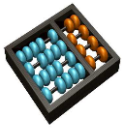
Objetivo: detectar "erros" (ex. bits trocados) num segmento transmitido (nota: usado *apenas* na camada de transporte)

Emissor:

- trata o conteúdo de segmentos como seqüências de números inteiros de 16 bits
- checksum: adição (soma em complemento de um) do conteúdo do segmento
- transmissor coloca o valor do checksum no campo checksum do UDP

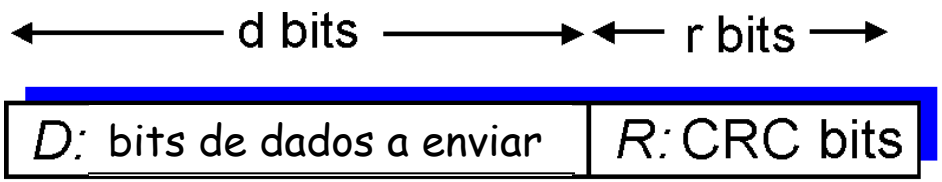
Receptor:

- computa o checksum do segmento recebido
- verifica se o checksum calculado é igual ao valor do campo checksum:
  - ✓ NÃO - erro detectaado
  - ✓ SIM - não detectou erro. *Mas talvez haja erros apesar disso?* Mais depois….



Códigos de Redundância Cíclica (Cyclic Redundancy Codes):

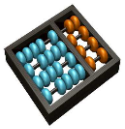
- encara os bits de dados, D, como um número binário
- escolhe um padrão gerador de r+1 bits, G
- objetivo: escolhe r CRC bits, R, tal que
  - ✓ <D,R> é divisível de forma exata por G (módulo 2)
  - ✓ receptor conhece G, divide <D,R> por G. Se o resto é diferente de zero: erro detectado!
  - ✓ pode detectar todos os erros em seqüência (burst errors) com comprimento menor que r+1 bits
- largamente usado na prática (ATM, HDCL)



padrão de bits

$$D * 2^r \quad \text{XOR} \quad R$$

fórmula matemática



CRC

frame content: 11010001110(11 bits)

pattern: 101011 (6 bits)

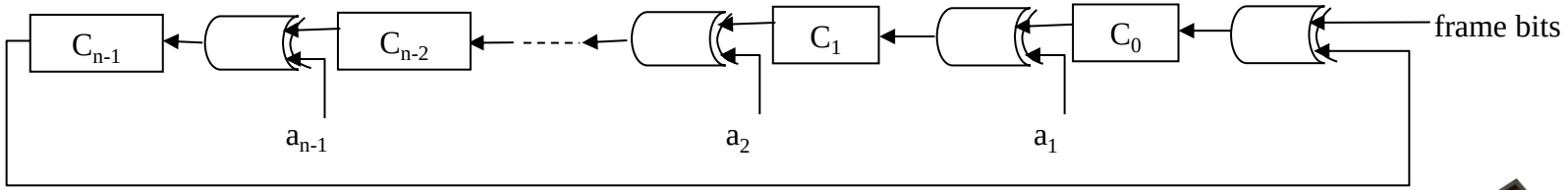
frame check sequence = (5 bits)

```
                11100000111
          ------------------
101011  ) 11010001110000000
          101011
          ------
           111110
           101011
           ------
            101011
            101011
            ------
             110000
             101011
             ------
              110110
              101011
              ------
               111010
               101011
               ------
                10001   <-- the remainder
```

```
                11100000111
          ------------------
101011  ) 11010001110010001
          101011
          ------
           111110
           ...
          ------
                0   <-- correct
```

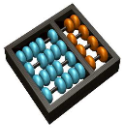
frame check sequence

Hardware implementation



Implementação de CRC (cont)

- Remetente realiza em tempo real por hardware a divisão da seqüência D pelo polinômio G e acrescenta o resto R a D
- O receptor divide <D,R> por G; se o resto for diferente de zero, a transmissão teve erro
- Padrões internacionais de polinômios G de graus 8, 12, 15 e 32 já foram definidos
- ATM utiliza um CRC de 32 bits em AAL5



Enlaces e Protocolos de Múltiplo Acesso

Três tipos de enlace:

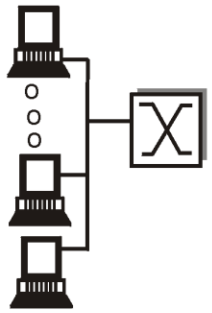
(a) Ponto-a-ponto (um cabo único)

(b) Difusão (cabo ou meio compartilhado: Ethernet, 802.11 wireless LAN.)

(c) Comutado (p.ex., E-net comutada, ATM, etc)

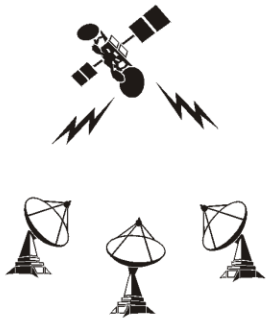
Começamos com enlaces com **Difusão**.

Desafio principal: **Protocolo de Múltiplo Acesso**



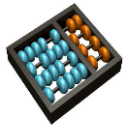
shared wire (e.g. Ethernet)

shared wireless (e.g. Wavelan)



satellite

cocktail party



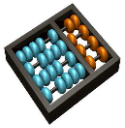
Protocolos de Acesso Múltiplo

- canal de comunicação único e compartilhado
- duas ou mais transmissões pelos nós: interferência
  - ✓ apenas um nó pode transmitir com sucesso num dado instante de tempo
- **protocolo de múltiplo acesso:**
  - ✓ algoritmo distribuído que determina como as estações compartilham o canal, isto é, determinam quando cada estação pode transmitir
  - ✓ comunicação sobre o compartilhamento do canal deve utilizar o própro canal!
  - ✓ o que procurar em protocolos de múltiplo acesso:
    - síncrono ou assíncrono
    - informação necessária sobre as outras estações
    - robustez (ex., em relação a erros do canal) desempenho

Protocolo de Acesso Múltiplo Ideal

Canal Broadcast com taxa de R bps

1. Quando um nó deseja transmitir, ele pode transmitir a taxa R
2. Quando M nós desejam transmitir, cada um transmite a uma taxa igual a R/M
3. Totalmente descentralizado.
   - ✓ Nenhum nó especial coordena as transmissões
   - ✓ Sem sincronização de relógios e de slots;
4. Simples

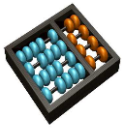


Protocolos MAC: uma taxonomia

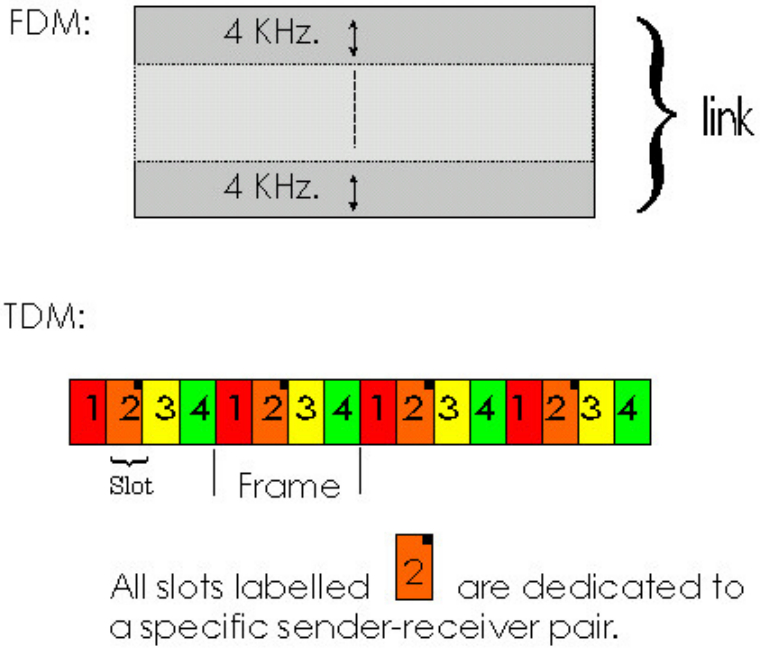
Três grandes classes:

- Particionamento de canal
  - ✓ dividem o canal em pedaços menores (compartimentos de tempo, freqüência)
  - ✓ aloca um pedaço para uso exclusivo de cada nó
- Acesso Aleatório
  - ✓ permite colisões
  - ✓ "recuperação" das colisões
- Passagem de Permissão (revezamento)
  - ✓ compartilhamento estritamente coordenado para evitar colisões

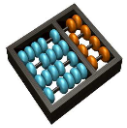
Objetivo: eficiente, justo, simples, descentralizado



Protocolos de Particionamento do Canal



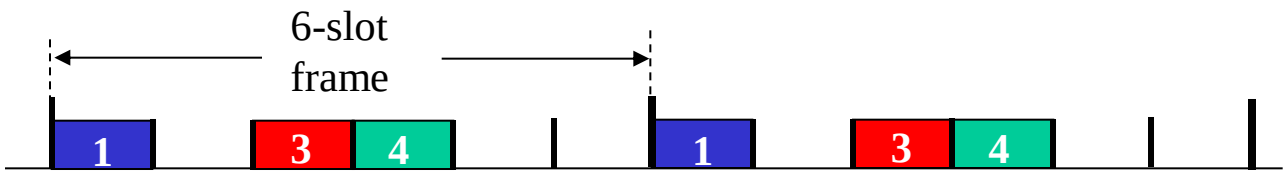
- **TDM** (Multiplexação por Divisão de Tempo): canal dividido em N intervalos de tempo ("slots"), um para cada usuário; ineficiente com usuários de pouco demanda ou quando carga for baixa.
- **FDM** (Multiplexação por Divisão de Freqüência): freqüência subdividida; mesmos problemas de eficiência do TDM.



Multiplexacão por Divisão do Tempo

TDMA: acesso múltiplo por divisão temporal

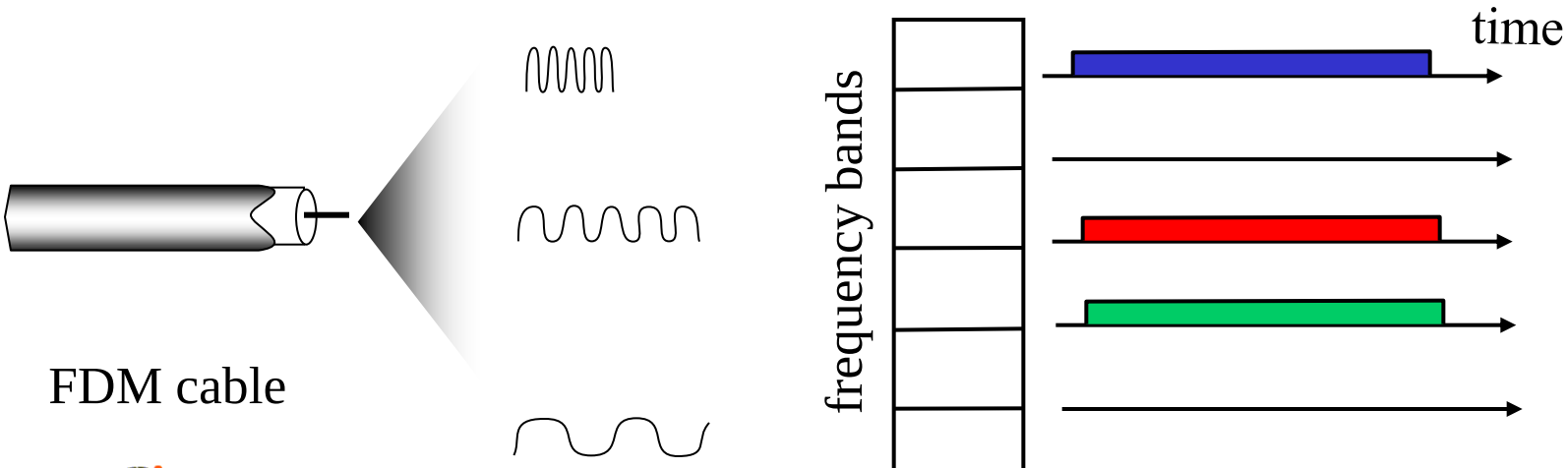
- acesso ao canal é feito por "turnos"
- cada estação controla um compartimento ("slot") de tamanho fixo (tamanho = tempo de transmissão de pacote) em cada turno
- compartimentos não usados são disperdiçados
- exemplo: rede local com 6 estações: 1,3,4 têm pacotes, compartimentos 2,5,6 ficam vazios



Multiplexacão por Divisaão da Freuência

FDMA: acesso múltiplo por divisão de freqüência

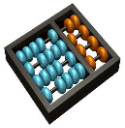
- o espectro do canal é dividido em bandas de freqüência
- cada estação recebe uma banda de freqüência
- tempo de transmissão não usado nas bandas de freqüência é disperdiçado
- exemplo: rede local com 6 estações: 1,3,4 têm pacotes, as bandas de freqüência 2,5,6 ficam vazias



Particionamento de Canal (CDMA)

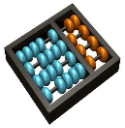
CDMA (Acesso Múltiplo por Divisão de Códigos)

- explora esquema de codificação de espectro espalhado - DS (*Direct Sequence*) ou FH (*Frequency Hopping*)
- "código" único associado a cada canal; ié, particionamento do conjunto de códigos
- muito usado em canais broadcast, sem-fio (celular, satélite,etc)
- todos usuários compartilham a mesma freqüência, mas cada canal tem sua própria seqüência de "chipping" (ié, código)



Protocolos de Acesso Aleatório

- Quando o nó tem um pacote a enviar:
  - ✓ transmite com toda a taxa do canal R.
  - ✓ não há uma regra de coordenação *a priori* entre os nós
- dois ou mais nós transmitindo -> "colisão",
- Protocolo MAC de acesso aleatório especifica:
  - ✓ como detectar colisões
  - ✓ como as estações se recuperam das colisões (ex., via retransmissões atrasadas)
- Exemplos de protocolos MAC de acesso aleatório:
  - ✓ slotted ALOHA
  - ✓ ALOHA
  - ✓ CSMA e CSMA/CD



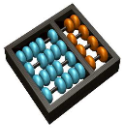
Slotted ALOHA

Considerações

- Todos os quadros tem o mesmo tamanho
- Tempo é dividjdo em intervalos iguais (tempo para transmitir um quadro)
- O nós iniciam a transmissão apenas no início do intervalo;
- Nós são sincronizados
- Se 2 ou mais nós transmitem em um intervalo, todos os nós detectam colisão;

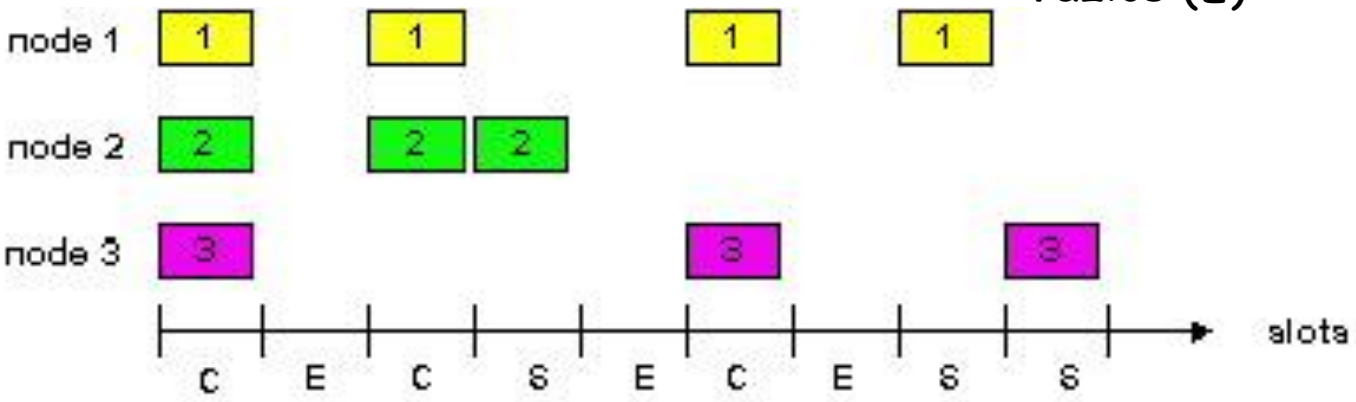
Operação

- Quando um nó tem um quadro novo para transmitir, ele tenta transmiti-lo no próximo intervalo;
- Se não há colisão, o nó pode enviar um novo quadro no próximo intervalo;
- Se tem colisão, o nó retransmite o quadro no intervalo subsequente com probabilidade p, até que o quadro seja transmitido com sucesso;



Slotted ALOHA

Intervalos com Sucesso (S), com Colisão (C), ou Vazios (E)



Pros

- Se há apenas um nó ativo, ele pode transmitir continuamente, utilizando toda a capacidade do canal;
- Altamente descentralizado: apenas os intervalos necessitam ser sincronizados;
- simples

Contras

- Colisões, desperdiça intervalos
- Intervalos vazios
- Nós devem ser capazes de detectar colisões em um intervalo de tempo menor que o tempo para transmitir o pacote;

Eficiência do Slotted Aloha

P: Qual a máxima fração de intervalos com sucesso?

R: Suponha que N estações têm pacotes para enviar cada uma transmite num intervalo com probabilidade p

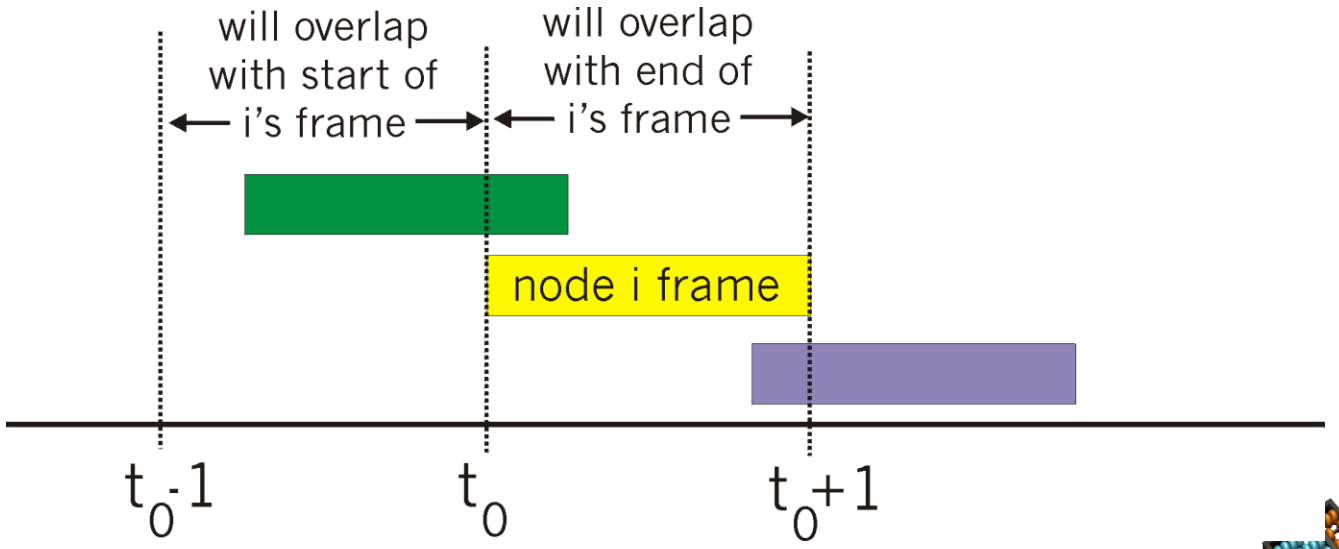
- ✓ prob. sucesso de transmissão, S, é:;
  - ➢ por um único nó: $S = p\,(1-p)^{(N-1)}$;
  - ➢ por qualquer um dos N nó: $S = Np(1-p)^{N-1}$;
  - ➢ Para eficiência máxima com N nós, deve-se escolher p^* que maximize, $Np(1-p)^{N-1}$
  - ➢ Para vários nós, pega-se o limite de $Np^*(1-p^*)^{N-1}$ quando N -> infinito, dá igual a $1/e = .37$

Eficiência é a fração final de intervalos bem-sucedidos no caso em que há um grande número de nós ativos, cada qual com uma grande quantidade de quadros a transmitir;

No máximo: uso do canal para envio de dados úteis: 37% do tempo!

ALOHA Puro (unslotted)

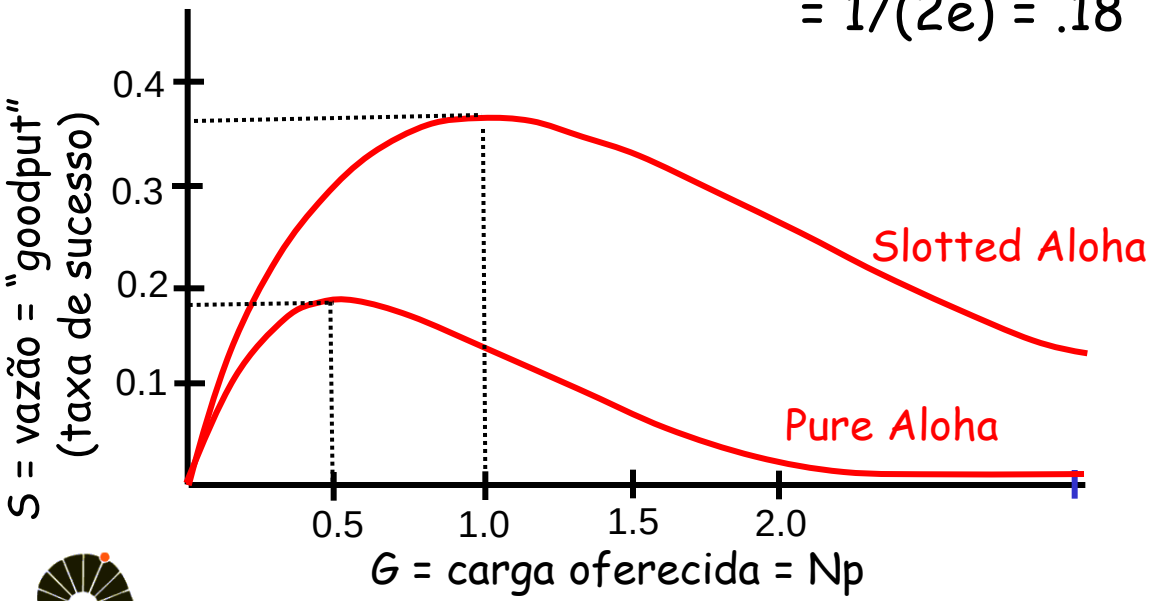
- unslotted Aloha: operação mais simples, não há sincronização
- pacote necessita transmissão:
  - ✓ enviar sem esperar pelo início de um intervalo
- a probabilidade de colisão aumenta:
  - ✓ pacote enviado em t_0 colide com outros pacotes enviados em $[t_0-1, t_0+1]$



Aloha Puro (cont.)

P(sucesso por um dado nó) = P(nó transmite) ·
P(outro nó não transmite em $[t_0-1,t_0]$) ·
P(outro nó não transmite em $[t_0,t_0+1]$)
= p · (1-p) · (1-p)

P(sucesso por um qualquer dos N nós) = N p · (1-p) · (1-p)
… escolhendo p ótimo quando n -> infinito …
= 1/(2e) = .18

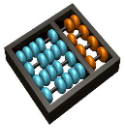


protocolo limita a vazão efetiva do canal!

CSMA: Carrier Sense Multiple Access

CSMA: escuta antes de transmitir:

- Se o canal parece vazio: transmite o pacote
- Se o canal está ocupado, adia a transmissão
  - ✓ CSMA Persistente: tenta outra vez imediatamente com probabilidade p quando o canal se torna livre (pode provocar instabilidade)
  - ✓ CSMA Não-persistente: tenta novamente após um intervalo aleatório
- analogia humana: não interrompa os outros!

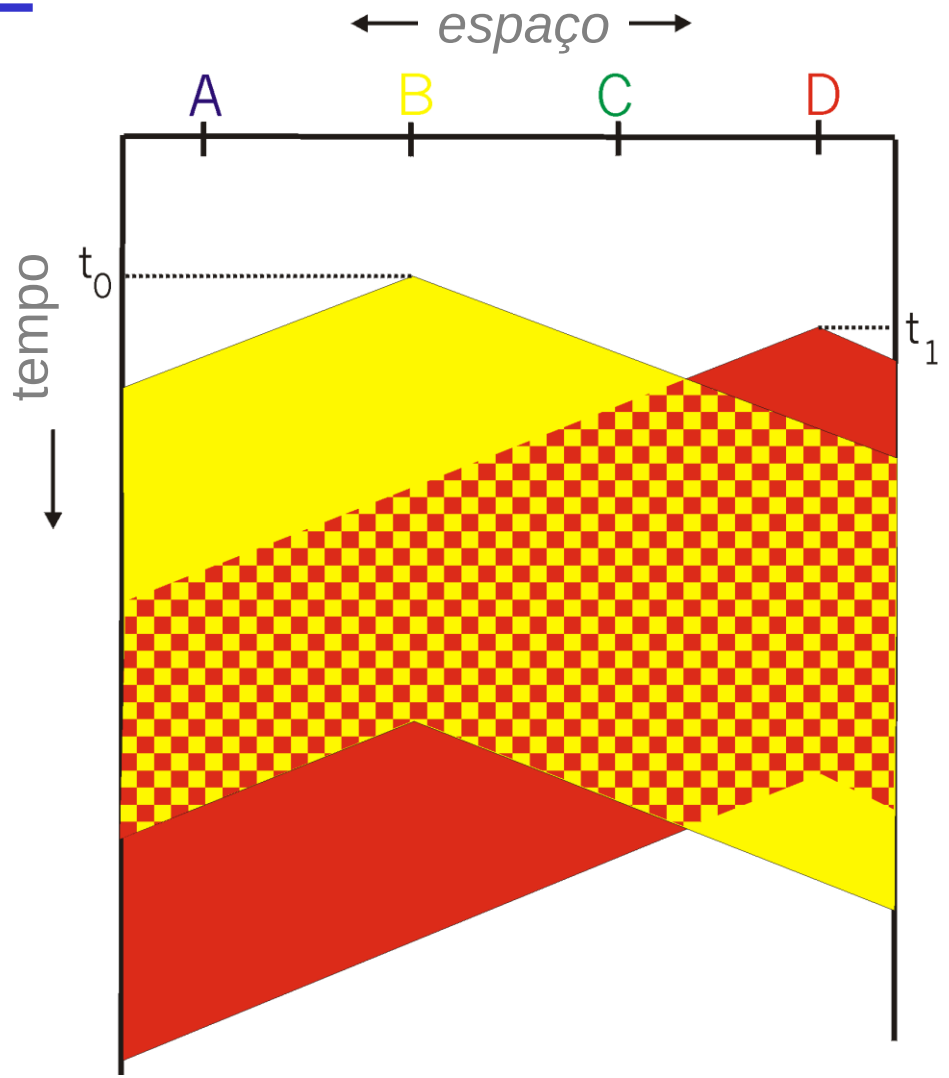


Colisões no CSMA

colisões *podem* ocorrer:
o atraso de propagação implica que dois nós podem não ouvir as transmissões de cada outro

colisão:
todo o tempo de transmissão do pacote é desperdiçado

nota:
papel da distância e do atraso de propagação na determinação da probabilidade de colisão.



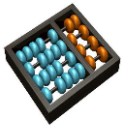
arranjo espacial dos nós na rede

CSMA/CD (Detecção de Colisão)

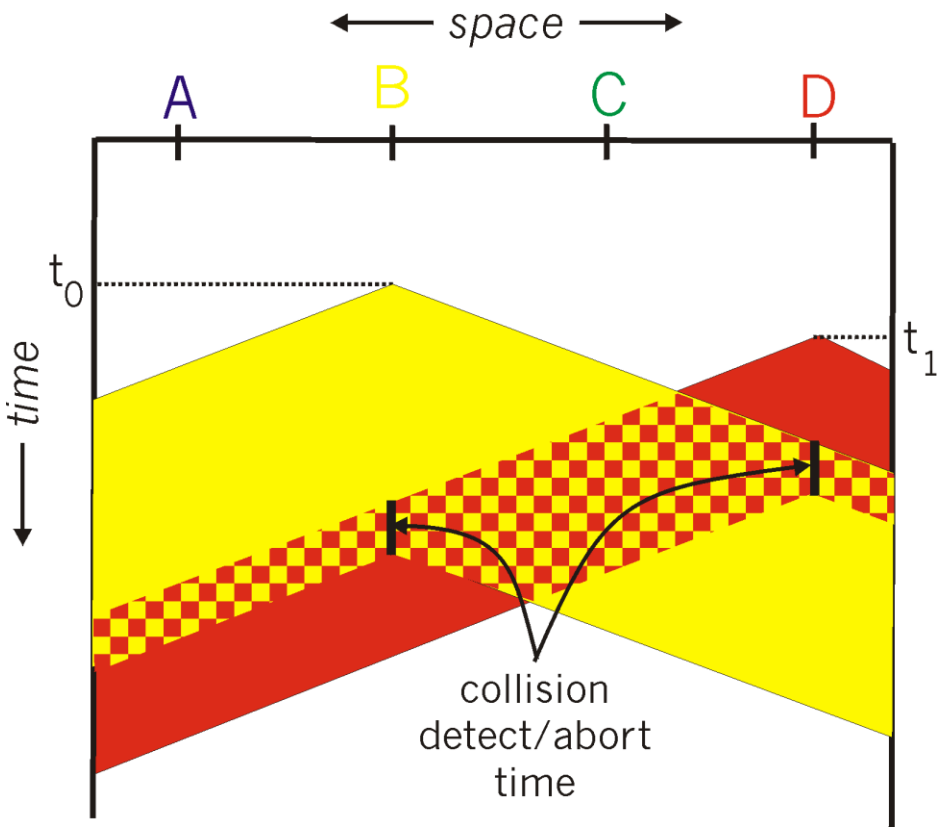
CSMA/CD: detecção de portadora, deferência como no CSMA

- ✓ colisões *detectadas* num tempo mais curto
- ✓ transmissões com colisões são interrompidas, reduzindo o desperdício do canal
- ✓ retransmissões **persistentes** ou não-persistentes

- detecção de colisão:
  - ✓ **fácil em LANs cabeadas**: (p.ex., E-net): pode-se medir a intensidade do sinal na linha, detectar violações do código, ou comparar sinais Tx e Rx
  - ✓ **difícil em LANs sem fio**: o receptor é desligado durante transmissão, para evitar danificá-lo com excesso de potência
- CSMA/CD pode conseguir utilização do canal perto de 100% em redes locais (se tiver baixa razão de tempo de propagação para tempo de transmissão do pacote)
- analogia humana: o "bom-de-papo" educado



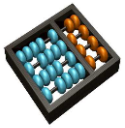
Detecção de colisões em CSMA/CD



Ethernet CSMA/CD algorithm

1. Ethernet receives datagram from network layer, creates frame
2. If Ethernet senses channel:
   if idle: start frame transmission.
   if busy: wait until channel idle, then transmit
3. If entire frame transmitted without collision - done!
4. If another transmission detected while sending: abort, send jam signal
5. After aborting, enter *binary (exponential) backoff:*
   - after *m*th collision, chooses *K* at random from $\{0,1,2, \ldots, 2^m-1\}$. Ethernet waits $K \cdot 512$ bit times, returns to Step 2
   - more collisions: longer backoff interval

UNICAMP



Protocolos MAC com Passagem de Permissão (revezamento)

Protocolos MAC com particionamento de canais:

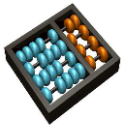
- ✓ compartilham o canal eficientemente quando a carga é alta e bem distribuída
- ✓ ineficiente nas cargas baixas: atraso no acesso ao canal. A estação consegue uma banda de 1/N da capacidade do canal, mesmo que haja apenas 1 nó ativo!

Protocolos MAC de acesso aleatório

- ✓ eficiente nas cargas baixas: um único nó pode usar todo o canal
- ✓ cargas altas: excesso de colisões

Protocolos de passagem de permissão

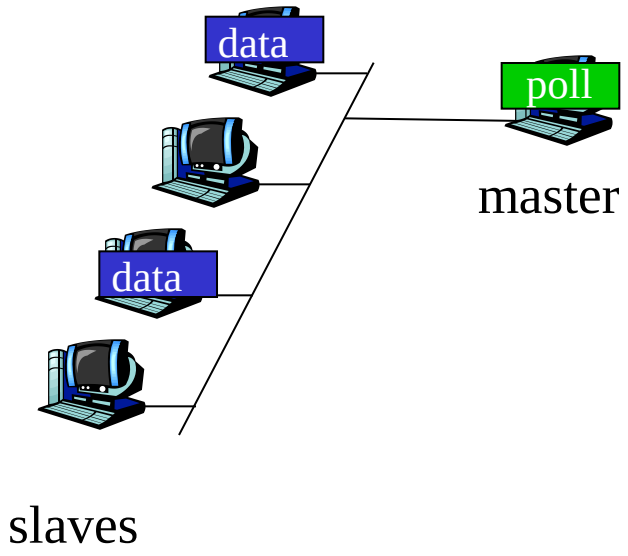
buscam o melhor dos dois mundos!



Protocolos de polling

Polling:

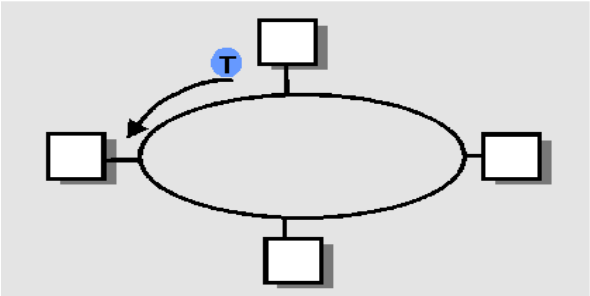
- nó mestre "convida" em ordem as estações escravas a transmitir seus pacotes (até algum Máximo).
- Mensagens Request to Send e Clear to Send
- problemas:
  - ✓ custo de Request to Send/Clear to Send
  - ✓ latência
  - ✓ ponto único de falha (mestre)



Protocolos MAC com Passagem de Permissão (revezamento)

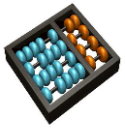
Passagem de ficha de permissão:

- a **ficha de permissão** é passada seqüencialmente de estação a estação;
- É possível aliviar a latência e melhorar tolerância a falhas (numa configuração de barramento de fichas).
- problemas:
  - ✓ token overhead
  - ✓ latência
  - ✓ ponto único de falha (token): procedimentos complexos para recuperar de **perda de ficha**, etc



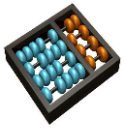
Resumo dos Protocolos MAC

- O que se pode fazer com um meio compartilhado ?
  - ✓ Particionamento do canal, por tempo, freqüência ou código
    - TDM, FDM, CDMA, WDMA (wave division)
  - ✓ Particionamento aleatório (dinâmico),
    - ALOHA, S-ALOHA, CSMA, CSMA/CD
    - Detecção de portadora: fácil em alguns meios físicos (cabeados), difícil em outros (sem fio)
    - CSMA/CD usado na rede Ethernet
  - ✓ Passagem de Permissão (revezamento)
    - polling a partir de um ponto central, passagem da ficha de permissão



Camada de Enlace de Dados

- 5.1 Introdução e Serviços
- 5.2 Correção e detecção de erros
- 5.3 Protocolos Múltilpo Acesso
- 5.4 Endereçamento
- 5.5 Ethernet
- 5.6 Switches
- 5.7 PPP
- 5.8 Virtualização: MPLS
- 5.9 Data Center Networks



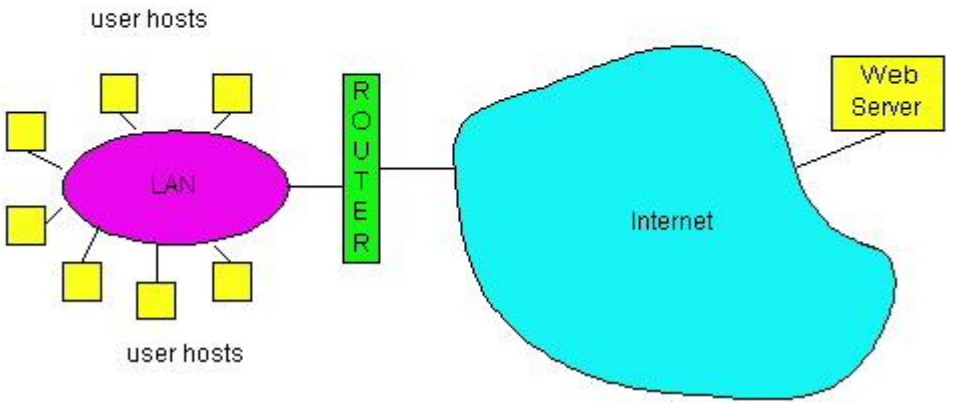
Tecnologias de Rede Local

Camada de enlace até agora:

- ✓ serviços, detecção de erros/correção, acesso múltiplo;
- ✓ Protocolos MAC usados em redes locais, para controlar acesso ao canal

A seguir: tecnologias de redes locais (LAN)

- ✓ endereçamento
- ✓ Ethernet
- ✓ hubs, switches
- ✓ PPP



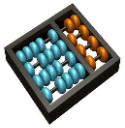
Endereços físicos e ARP

Endereços IP de 32-bit:

- endereços da *camada de rede*
- usados para levar o datagrama até a rede de destino (lembre da definição de rede IP)

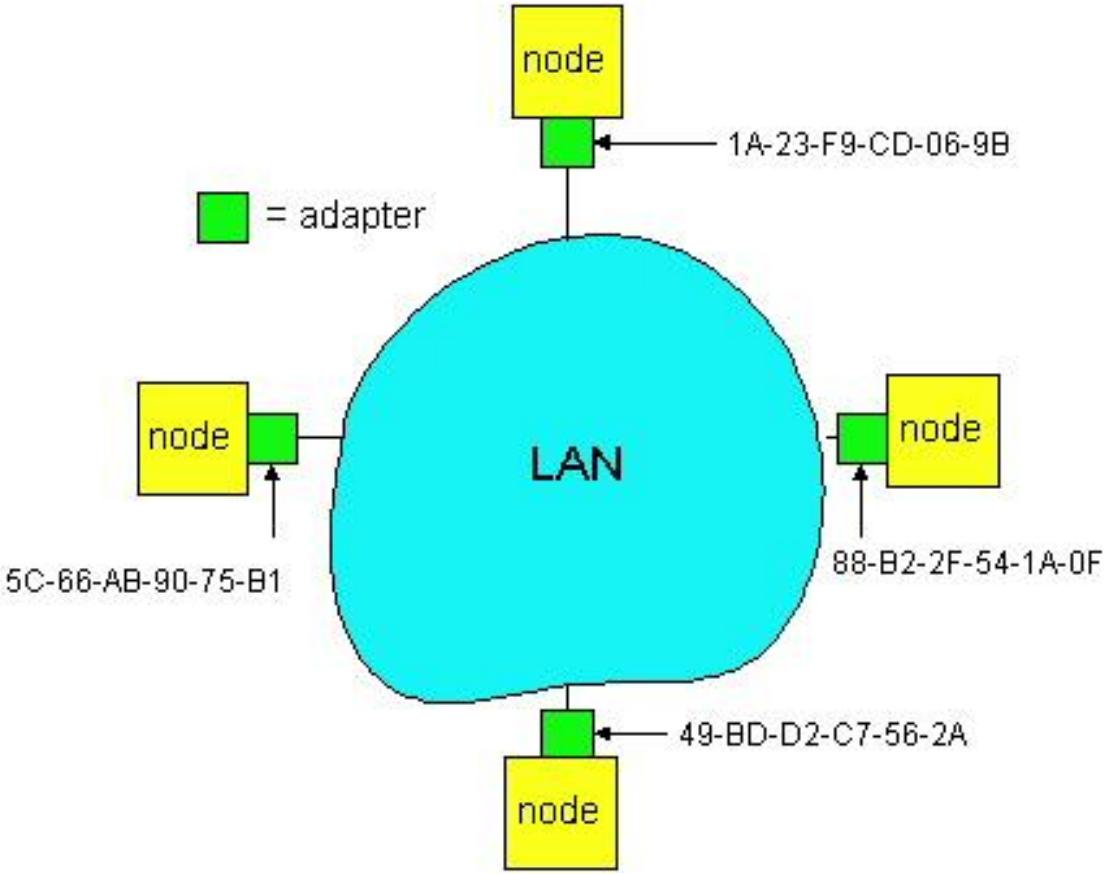
Endereço de LAN (ou MAC ou físico):

- usado para levar o datagrama de uma interface física a outra fisicamente conectada com a primeira (isto é, na mesma rede)
- Endereços MAC com 48 bits (na maioria das LANs) gravado na memória fixa (ROM) do adaptador de rede



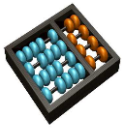
Endereços físicos e ARP

Cada adaptador numa LAN tem um único endereço de LAN



Endereços de LAN (mais)

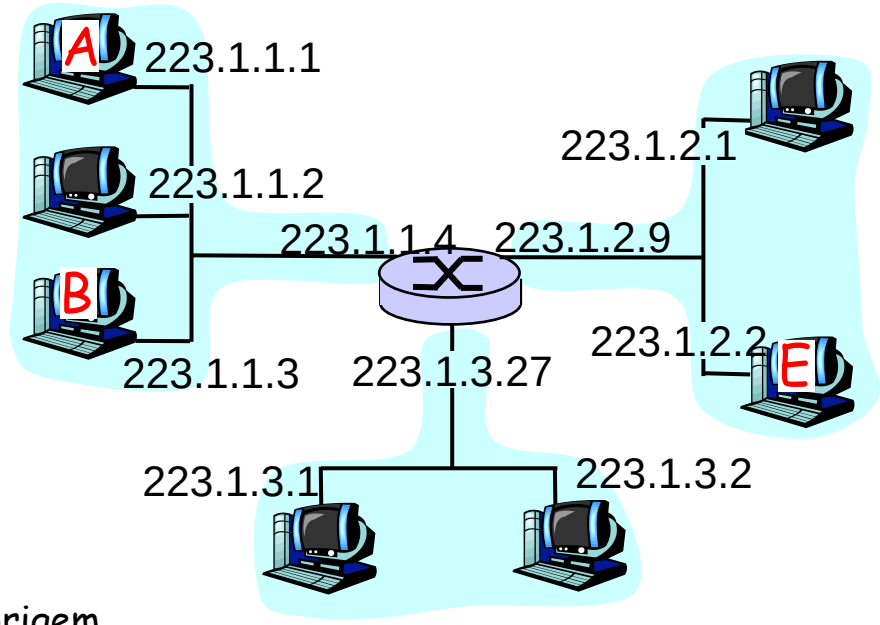
- A alocação de endereços MAC é administrada pelo IEEE
- O fabricante compra porções do espaço de endereço MAC (para assegurar a unicidade)
- Analogia:
  - (a) endereço MAC: senelhante ao número do CPF
  - (b) endereço IP: semelhante a um endereço postal
- endereçamento MAC é "flat" => portabilidade
  - ✓ é possível mover uma placa de LAN de uma rede para outra sem reconfiguração de endereço MAC
  - ✓ endereço MAC de difusão (*broadcast*): 1111……1111
- endereçamento IP "hierárquico" => NÃO portável
  - ✓ depende da rede na qual se está ligado

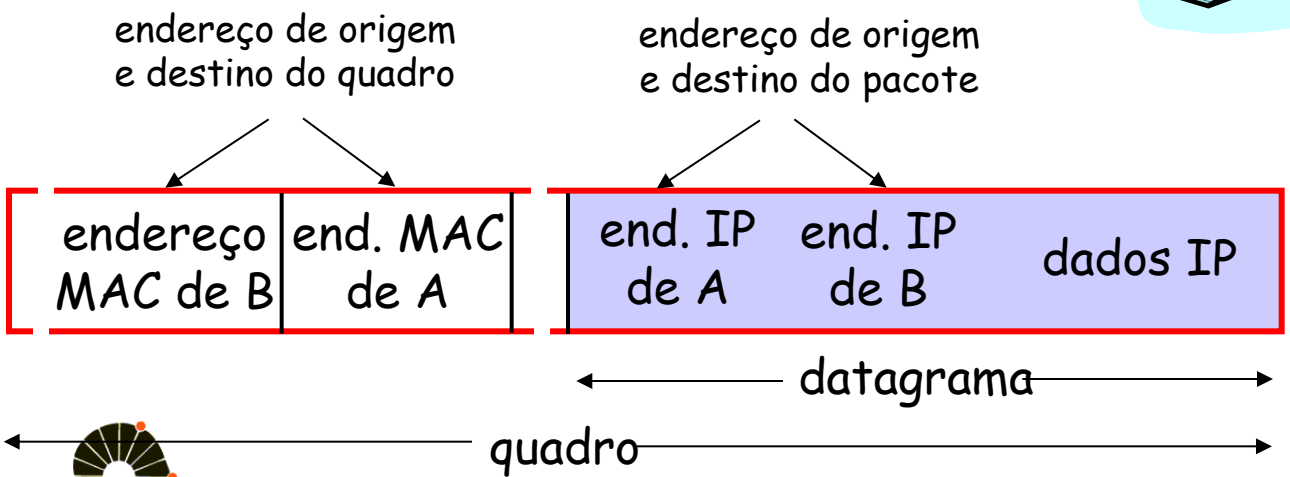


Lembrando a discussão anterior sobre roteamento

Começando em A, dado que o datagrama está endereçado para B (endereço IP):

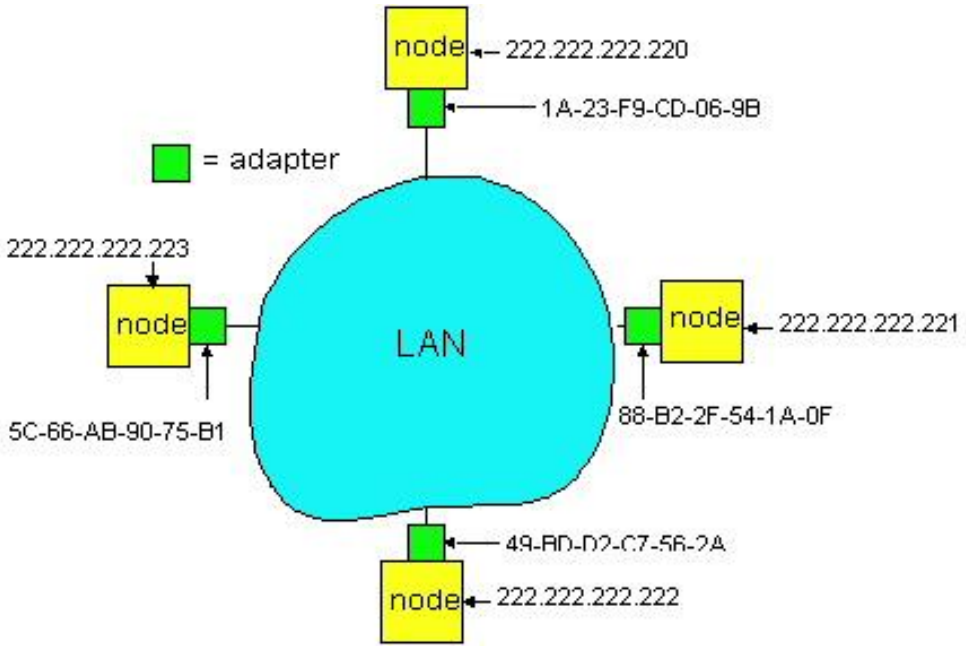
- procure rede.endereço de B, encontre B em alguma rede, no caso igual à rede de A
- camada de enlace envia datagrama para B dentro de um quadro da camada de enlace





ARP: Address Resolution Protocol

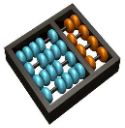
Questão: como determinar o endereço MAC de B dado o endereço IP de B?



- Cada nó IP (Host, Roteador) numa LAN tem um módulo e uma tabela ARP
- Tabela ARP: mapeamento de endereços IP/MAC para alguns nós da LAN

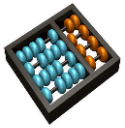
  < endereço IP; endereço MAC; TTL>

  - ✓ TTL (Time To Live): tempo depois do qual o mapeamento de endereços será esquecido (tipicamente 20 min)



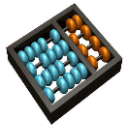
Protocolo ARP

- A conhece o endereço IP de B, quer aprender o endereço físico de B
- A envia em broadcast um pacote ARP de consulta contendo o endereço IP de B
  - ✓ todas as máquinas na LAN recebem a consulta ARP
- B recebe o pacote ARP, responde a A com o seu (de B) endereço de camada física
- A armazena os pares de endereço IP-físico até que a informação se torne obsoleta (esgota a temporização)
  - ✓ soft state: informação que desaparece com o tempo se não for re-atualizada



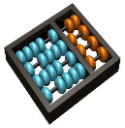
Protocolo ARP

- A deseja enviar um datagrama para B, e conhece o seu endereço IP;
- Suponha que o endereço MAC de B não esteja na tabela ARP de A;
- A envia em broadcast um pacote ARP de consulta com o endereço IP de B
  - ✓ todas as máquinas na LAN recebem a consulta
- B recebe o pacote ARP, responde a A com o seu endereço de camada física
  - ✓ Quadro enviado para o endereço MAC de A;
- A armazena os pares de endereço IP-físico até que a informação se torne obsoleta (esgota a temporização)
  - ✓ soft state: informação que desaparece com o tempo se não for re-atualizada
- ARP é "plug-and-play":
  - ✓ Nós criam suas tabelas ARP sem a intervenção do administrador da rede;



Camada de Enlace de Dados

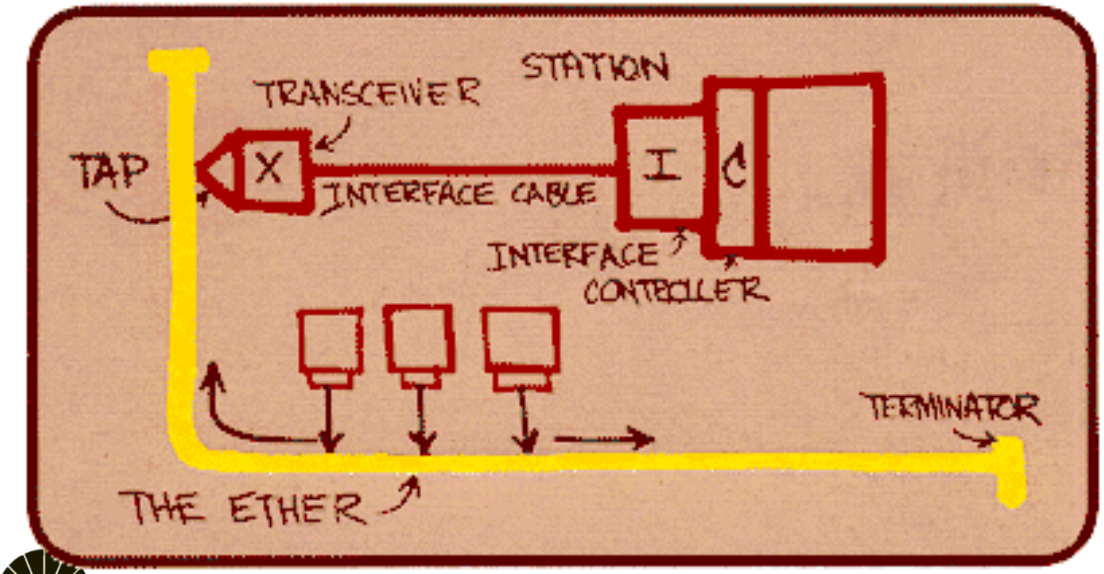
- 5.1 Introdução e Serviços
- 5.2 Correção e detecção de
- 5.3protocolos Múltilpo Acesso
- 5.4 Endereçamento
- 5.5 Ethernet
- 5.6 Switches
- 5.7 PPP
- 5.8 Virtualização: MPLS
- 5.9 Data Center Networks



Ethernet

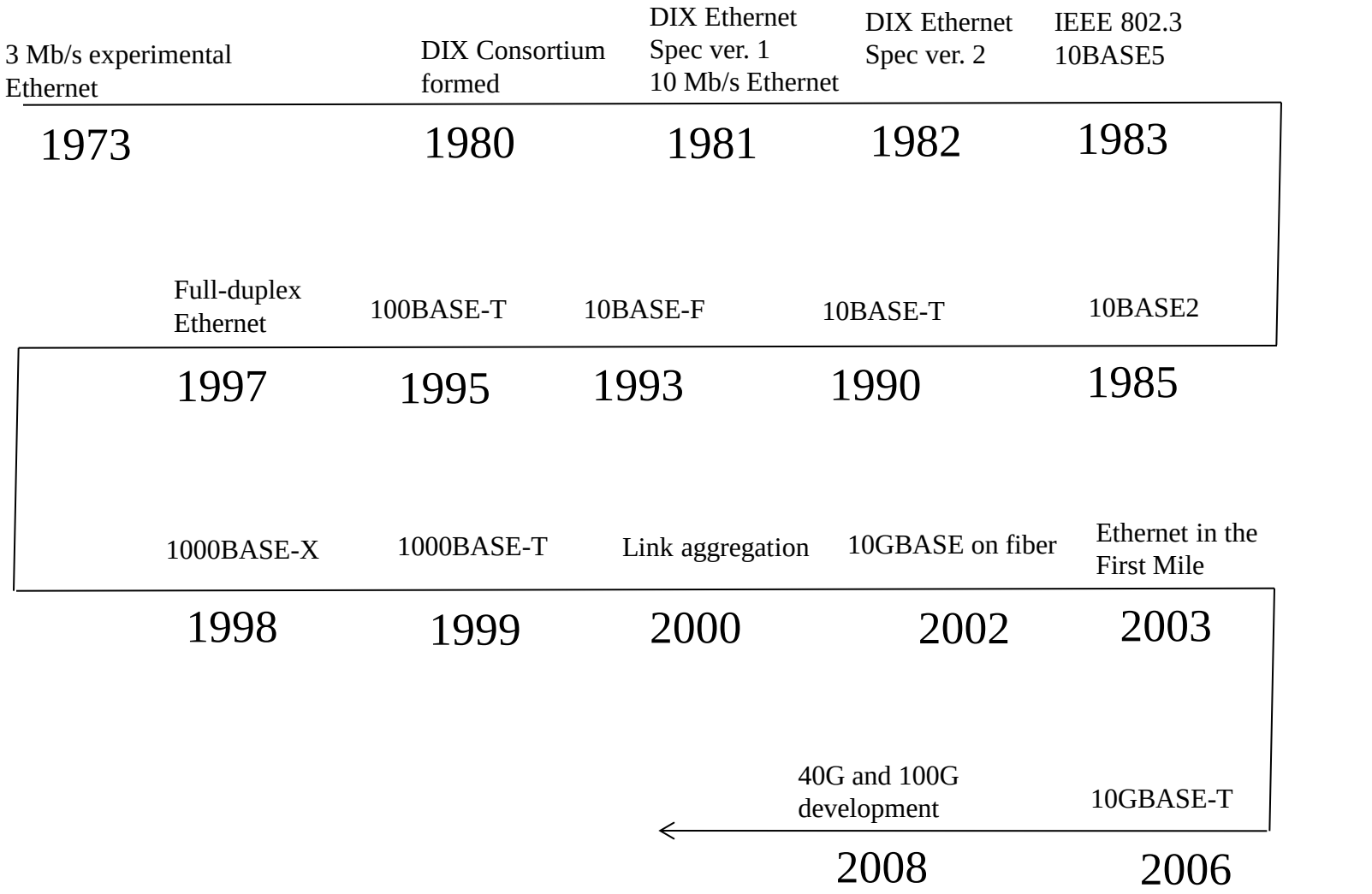
Tecnologia de rede local "dominante" :

- barato R$20 por 100Mbs!
- primeira tecnologia de LAN largamente usada
- Mais simples, e mais barata que redes usando ficha e ATM
- Velocidade crescente: 10, 100, 1000, 10000 Mbps
- Muitas tecnologias E-net (cabo, fibra, etc). Mas todas compartilham características comuns



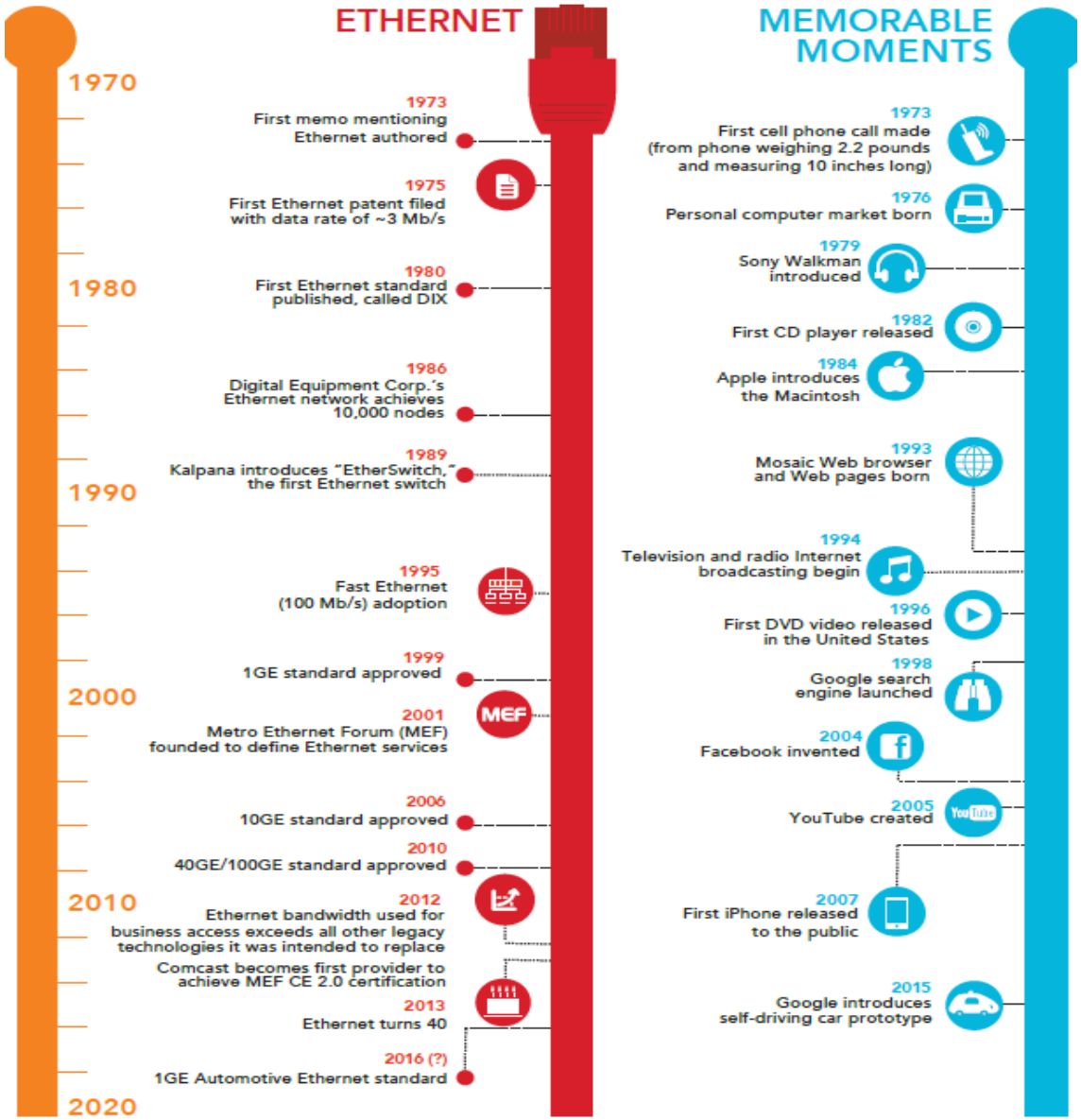
Esboço da Ethernet por Bob Metcalf

Evolução da Tecnologia



THE HISTORY OF ETHERNET

A timeline of innovation in Ethernet and otheemorable tech moments

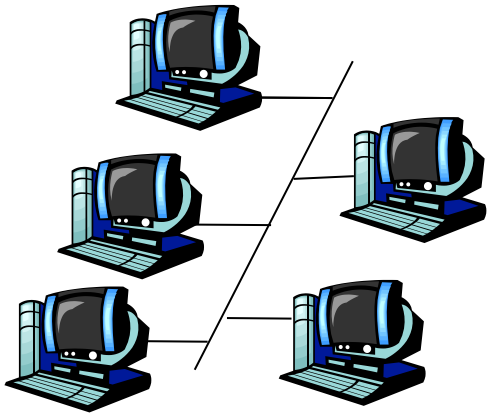


SOURCES: *MEF, Wikipedia, Heavy Reading, TIME, Business Insider, CBS News, University of North Carolina, Bright Hub, Tech Hive, Vertical Systems Group*

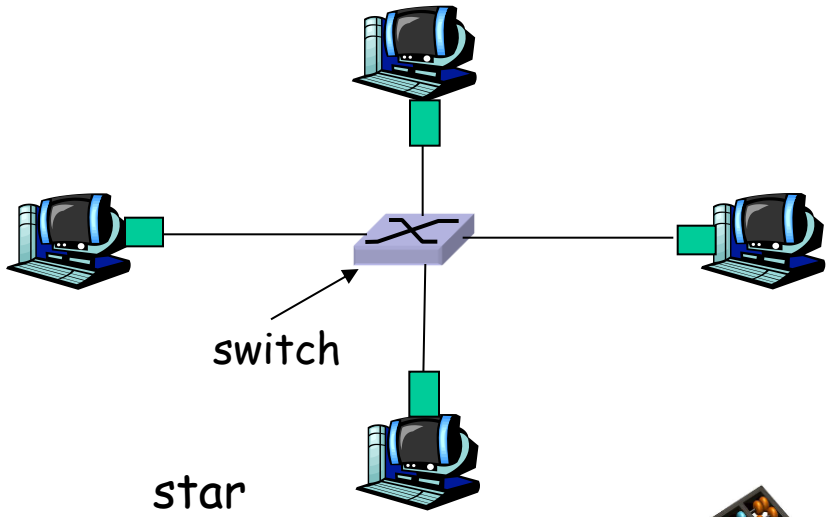
UNICAMP

Topologia

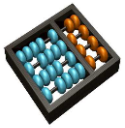
- Topologia barramento popular até meados dos ans 90s
  - ✓ Todos os nós no mesmo domínio de colisão
- Topologia estrela atualmente
  - ✓ Comutador central
  - ✓ Cada perna executa o protocolo



bus: coaxial cable

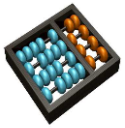


star

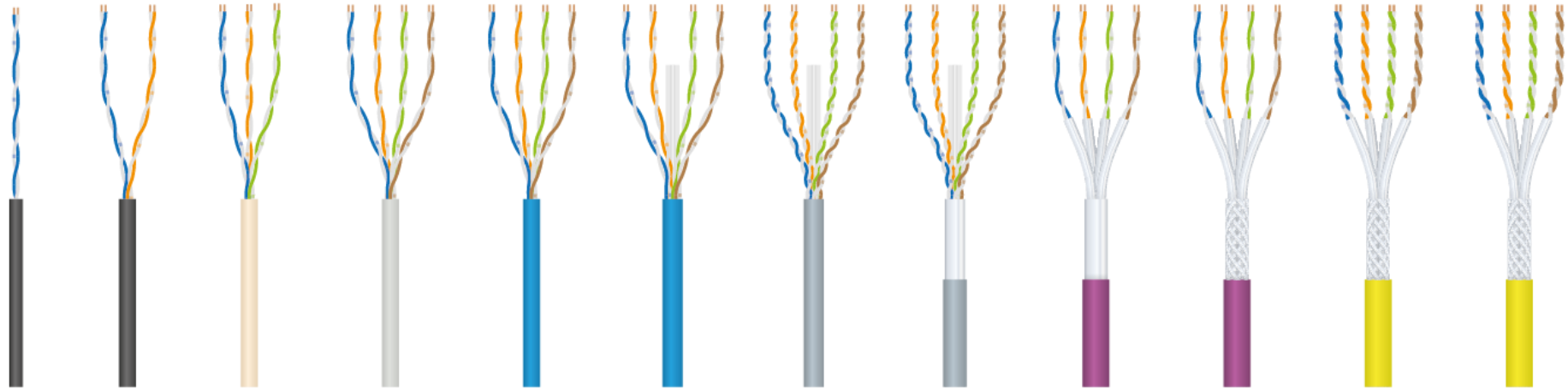


Meios de transmissão

speed \ medium	Coaxial cable	Twisted pairs	Fiber
under 10 Mb/s		1BASE5 (1987) 2BASE-TL (2003)	
10 Mb/s	10BASE5 (1983) 10BASE2 (1985) 10BROAD36 (1985)	10BASE-T (1990) 10BASE-TS (2003)	10BASE-FL (1993) 10BASE-FP (1993) 10BASE-FB (1993)
100 Mb/s		100BASE-TX (1995) 100BASE-T4 (1995) 100BASE-T2 (1997)	100BASE-FX (1995) 100BASE-LX/BX10 (2003)
1 Gb/s		1000BASE-CX (1998) 1000BASE-T (1999)	1000BASE-SX (1998) 1000BASE-LX (1998) 1000BASE-LX/BX10 (2003) 1000BASE-PX10/20 (2003)
10 Gb/s		10GBASE-T (2006)	10GBASE-R (2002) 10GBASE-W (2002) 10GBASE-X (2002)



History of Ethernet LAN Cables' Categories

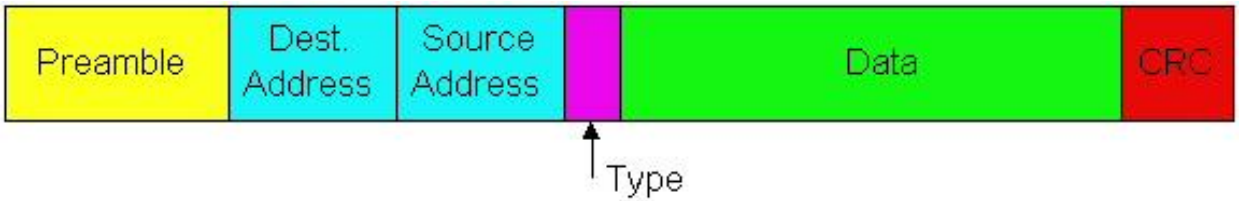


CAT 1	CAT 2	CAT 3	CAT 4	CAT 5	CAT 5e	CAT 6	CAT 6A	CAT 7	CAT 7A	CAT 8.1	CAT 8.2
1 Mbps	4 Mbps	10 Mbps	16 Mbps	100 Mbps	1 Gbps	1 Gbps	10 Gbps	10 Gbps	10 Gbps	25 Gbps	40 Gbps
400 KHz	4 MHz	16 MHz	20 MHz	100 MHz	100 MHz	250 MHz	500 MHz	600 MHz	1000 MHz	2000 MHz	2000 MHz
1983	1987	1991	1993	1995	2001	2002	2008	2010	2013	2016	2018

All Rights Reserved, Samm Teknoloji / telecom.samm.com / telecom@samm.com

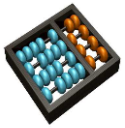
Estrutura do Quadro Ethernet

Adaptador do transmissor encapsula o datagrama IP (ou outro pacote de protocolo da camada de rede) num quadro Ethernet



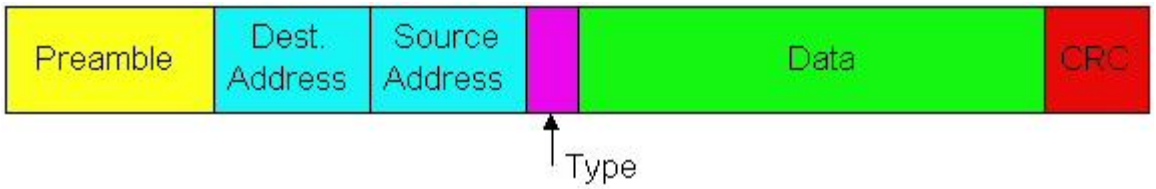
Preâmbulo:

- 7 bytes com padrão 10101010 seguido por um byte com padrão 10101011
- usado para sincronizar as taxas de relógio do transmissor e do receptor



Estrutura de Quadro Ethernet (cont)

- **Cabeçalho** contém Endereços de Destino e Origem e um campo Tipo
- **Endereços**: 6 bytes, o quadro é recebido por todos adaptadores numa rede local e descartado se não casar o endereço de destino com o do receptor
- **Tipo**: indica o protocolo da camada superior, usualmente IP, mas existe suporte para outros (tais como IPX da Novell e AppleTalk)
- **CRC**: verificado pelo receptor: se for detectado um erro, o quadro será descartado



Taxa Máxima de Quadros

Quadro com tamanho mínimo tem:

- 7 bytes Preamble + 1 byte SFD
- 64 bytes tamanho mínimo quadro
- 12 bytes Inter-frame gap (IFG)

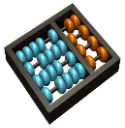
A 10 Mb/s:

Taxa máxima de quadros = $10*10^6$ / ((7+1+64+12)*8)

= 14,880 quadros / s

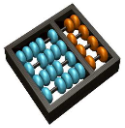
100 Mb/s → 148,809 quadros / s

1 Gb/s → 1,488,095 quadros / s



Ethernet utiliza CSMA/CD

- Sem intervalos (slots)
- Detecção de portadora: o adaptador não transmite se verifica que algum outro adaptador esteja transmitindo
- Detecção de colisão: o adaptador transmissor aborta a transmisão quando verifica que um outro adaptador está transmitindo
- Acesso aleatório: antes de tentar retransmitir um pacote, o adaptador transmissor espera um intervalo de tempo aleatório



Algoritmo Ethernet CSMA/CD

Adaptador recebe o datagrama e monta o quadro

A: escuta canal, **se** ocioso

então {

transmite e monitora o canal;

se detectou outra transmissão

então {

aborta e envia sinal de "jam";

atualiza número de colisões "m";

retarda de acordo com o algoritmo de retardamento exponencial (o adaptador escolhe um valor K aleatório entre {0,1,2,…,2^m-1} e espera um intervalo de K*512)

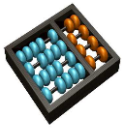
vai para *A*

}

senão {terminado este quadro; zera número de colisões}

}

senão {espera o final da transmissão atual e **vai para** *A*}



Ethernet CSMA/CD (mais)

Sinal "Jam":

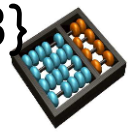
- garante que todos os outros transmissores estão cientes da colisão; 48 bits;

Bit time:

.1 microsec for 10 Mbps Ethernet ;
for K=1023, wait time is about 50 μsec

Retardamento Exponencial (Exponential Backoff)

- *Objetivo*: adaptar tentativas de retransmissão para carga atual da rede
  - ✓ carga pesada: espera aleatória será mais longa
- primeira colisão: escolha K entre {0,1}; espera é K x 512 tempos de transmissão de bit
- após a segunda colisão: escolha K entre {0,1,2,3}...
- após 10 ou mais colisões, escolha K entre {0,1,2,3,4,...,1023}

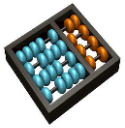


Eficiência CSMA/CD

- T_{prop} = tempo máximo de propagação entre 2 nós na rede;
- t_{trans} = tempo para se transmitir um quadro de tamanho máximo;

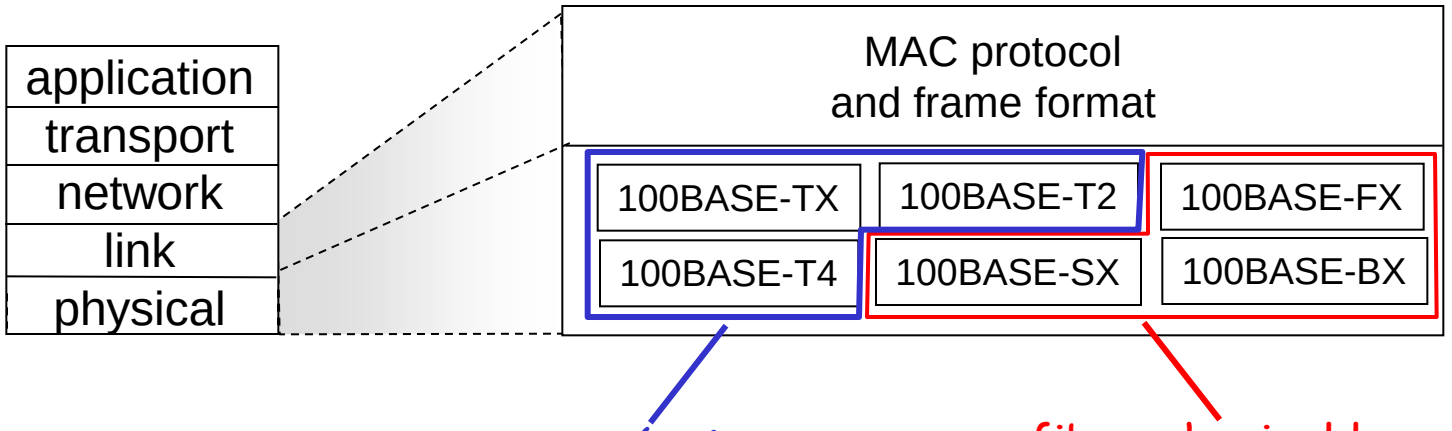
$$\text{efficiency} = \frac{1}{1 + 5t_{prop} / t_{trans}}$$

- Eficiência se aproxima de 1 quando t_{prop} se aproxima de 0;
- Eficiência se aproxima de 1 quando t_{trans} vai para infinito;
- Bem melhor que ALOHA, mais ainda descentralizado, simples e barato;
- Nota-se que neste esquema um quadro novo tem uma chance de sucesso na primeira tentativa, mesmo com tráfego pesado



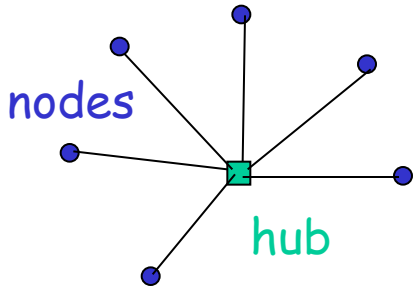
Standard 802.3

- *vários* padrões
  - ✓ Protocolo MAC e quadro padronizados
  - ✓ Taxas diferentes: 2 Mbps, 10 Mbps, 100 Mbps, 1Gbps, 10Gbps, 100Gbps, 400 Gbps (next… 800Gbps)
  - ✓ Diferentes medias: fibras, cabos, par

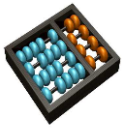


10BaseT e 100BaseT

- T significa "Twisted Pair" (par trançado)
- Os nós se conectam a um concentrador (hub) por um meio físico em "par trançado", portanto trata-se de uma "topologia em estrela";

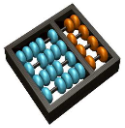


- Os hubs são essencialmente repetidos da camada física:
  - ✓ Bits que chegam em um enlace vão para todos os outros enlaces;
  - ✓ Não existe armazenamento de quadros;
  - ✓ Não se tem CSMA/CD no hub: adaptadores detectam colisões;



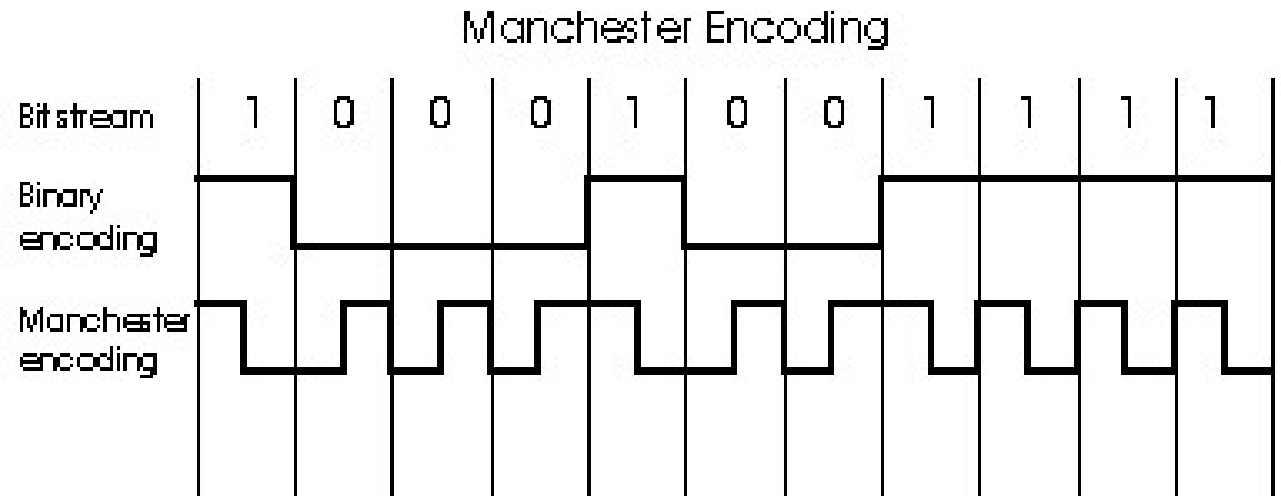
10BaseT e 100BaseT (cont)

- Distância máxima do nó ao hub é de 100 metros
- Hub pode desligar da rede um adaptador falho ("jabbering"); 10Base2 não funcionaria se um adaptador não pára de transmitir no cabo
- Hub pode coletar informação e estatísticas de monitoramento para administradores da rede
- 100BaseT não usa codificação Manchester; usa 4B5B para maior eficiência



Codificação Manchester

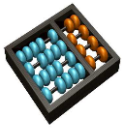
- Banda básica significa que não se usa modulação de portador; ao invés disto, bits são codificados usando codificação Manchester e transmitidos diretamente, modificando a voltagem de sinal de corrente contínuo
- Codificação Manchester garante que ocorra uma transição de voltagem a cada intervalo de bit, ajudando sincronização entre relógios do remetente e receptor
  - Não é necessário a existência de um relógio global centralizado entre os nós;
- Usado em 10BaseT



Codificação 4B/5B

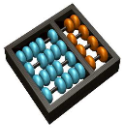
Name	4B	5B	description
0	0000	11110	hex data 0
1	0001	01001	hex data 1
2	0010	10100	hex data 2
3	0011	10101	hex data 3
4	0100	01010	hex data 4
5	0101	01011	hex data 5
6	0110	01110	hex data 6
7	0111	01111	hex data 7
8	1000	10010	hex data 8
9	1001	10011	hex data 9
A	1010	10110	hex data A
B	1011	10111	hex data B
C	1100	11010	hex data C
D	1101	11011	hex data D
E	1110	11100	hex data E
F	1111	11101	hex data F
Q	n/a	00000	Quiet (signal lost)
I	n/a	11111	Idle
J	n/a	11000	Start #1
K	n/a	10001	Start #2
T	n/a	01101	End
R	n/a	00111	Reset
S	n/a	11001	Set
H	n/a	00100	Halt

UNICAMP



Gbit Ethernet

- Usa formato do quadro Ethernet padrão
- Admite enlaces ponto-a-ponto e canais de difusão compartilhados
- Em modo compartilhado, usa-se CSMA/CD; para ser eficiente, as distâncias entre os nós devem ser curtas (poucos metros)
- Os Hubs usados são chamados de Distribuidores com Buffers ("Buffered Distributors")
- Full-Duplex em 1 Gbps para enlaces ponto-a-ponto
- Nota: o uso de enlaces ponto-a-ponto também foi estendido a 10Base-T e 100Base-T.



Half-Duplex vs. Full-Duplex

Half-duplex

Somente uma estação transmite *(necessário CSMA/CD)*

Full-duplex (IEEE 802.3x, 1997)

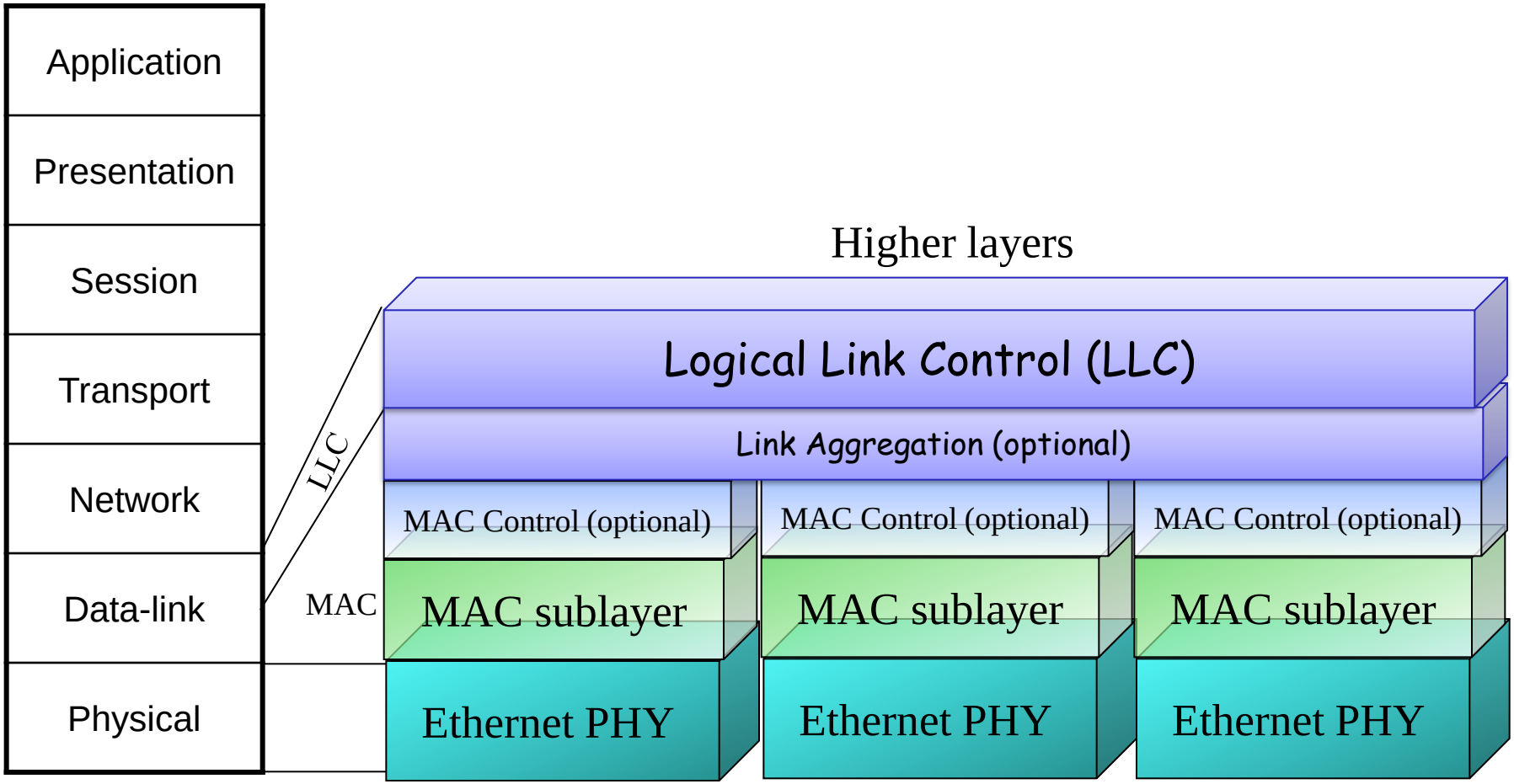
Transmissõa simultânea entre pares de estações em enlace ponto-a-ponto *(elimina CS, MA e CD)*

Três condições necessárias para full-duplex

1. Transmissões simultâneas sem interferência
2. Enlace ponto-a-ponto dedicado com exatamente duas estações
3. Capacidade de configuração das estações a operarem em full-duplex

UNICAMP

Gigabit Ehtenet

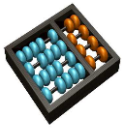


UNICAMP

Controle de Fluxo naEthernet

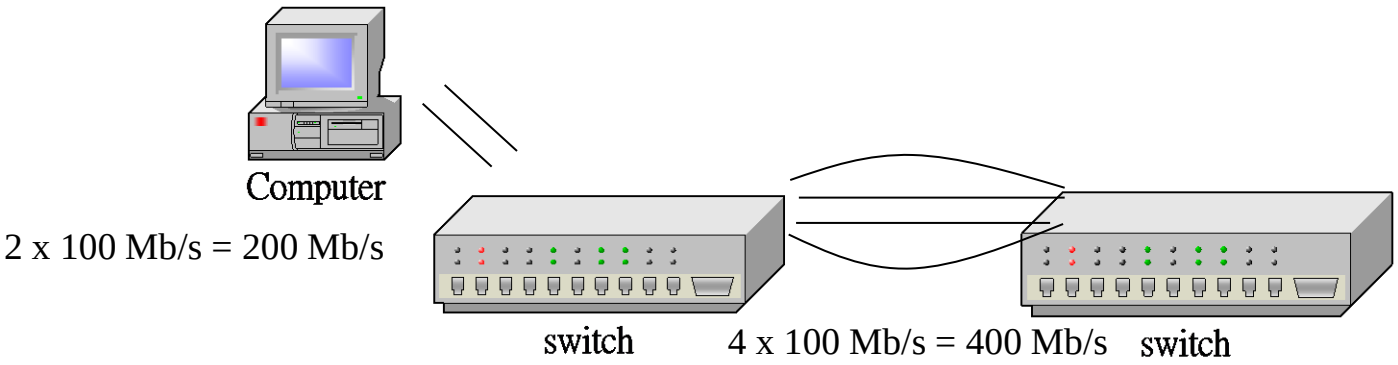
- Back pressure – half-duplex
  - ✓ Força colisão

- PAUSE frame – full-duplex Ethernet
  - ✓ Quadro PAUSE (IEEE 802.3x) enviado do receptor para o transmissor

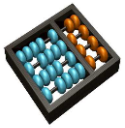


Agregação de Enlace

- Definido no padrão IEEE 802.3ad (2000)
- Permite obter enlaces de maior capacidade
- Balanceamento de carga
- Transparente às camadas superiores



UNICAMP

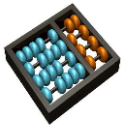


10 Gigabit Ethernet

- Especificado no padrão IEEE 802.3ae (2002)
- Características
  1. Somente Full-duplex
  2. Retro-compatível
  3. Entrando na escala de WANs
     (Longas distâncias, interface com OC-192)

Code name	Wave length	Transmission distance (m)
10GBASE-LX4	1310 nm	300
10GBASE-SR	850 nm	300
10GBASE-LR	1310 nm	10,000
10GBASE-ER	1550 nm	10,000
10GBASE-SW	850 nm	300
10GBASE-LW	1310 nm	10,000
10GBASE-EW	1550 nm	40,000

UNICAMP

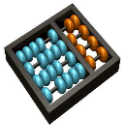


Ethernet na Primeira (Última) Milha

- IEEE 802.3ah finalizado em 2003.
- Orientado a redes de Acesso
- **Novas topologias**: fibras ponto-a-ponto, fibras ponto-a-multiponto, cobre ponto-a-ponto
  - **Novos PHYs**: 1000BASE-X extensão, Ethernet PON, voice-grade copper
  - **OAM**: detecção de falha remota, monitoramento de

Code name	Description
100BASE-LX10	100 Mbps on a pair of optical fibers up to 10 km
100BASE-BX10	100 Mbps on a optical fiber up to 10 km
1000BASE-LX10	1000 Mbps on a pair of optical fibers up to 10 km
1000BASE-BX10	1000 Mbps on a optical fiber up to 10 km
1000BASE-PX10	1000 Mbps on passive optical network up to 10 km
1000BASE-PX20	1000 Mbps on passive optical network up to 20 km
2BASE-TL	At least 2 Mbps over SHDSL up to 2700 m
10PASS-TS	At least 10 Mbps over VDSL up to 750 m

UNICAMP



40G e 100G Ethernet

40GBASE-T

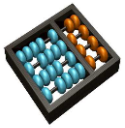
40GBASE-T is a port type for 4-pair balanced twisted-pair Cat.8 copper cabling up to 30 m defined in IEEE 802.3bq.[119] IEEE 802.3bq-2016 standard was approved by The IEEE-SA Standards Board on June 30, 2016.[120] It uses 16-level PAM signaling over four lanes at 3,200 MBaud each, scaled up from 10GBASE-T.

Comparison of twisted-pair-based Ethernet physical transport layers (TP-PHYs)[100]

Name	Standard	Status	Speed (Mbit/s)	Pairs required	Lanes per direction	Bits per hertz	Line code	Symbol rate per lane (MBd)	Bandwidth	Max distance (m)	Cable	Cable rating (MHz)	Usage
40GBASE-T	802.3bq-2016 (CL113)	current	40000	4	4	6.25	PAM-16 RS-FEC (192, 186) LDPC	3200	1600	30	Cat 8	2000	LAN, Data centres

100 Gigabit Ethernet (100 GbE) (2nd Generation: 25GbE-based) - (Data rate: 100 Gbit/s - Line code: 256b/257b × RS-FEC(528,514) × NRZ - Line rate: 4x 25.78125 GBd = 103.125 GBd - Full-Duplex) [101][102][103][105]										
100GBASE-KR4	802.3bj-2014 (CL93)	current	Cu-Backplane	N/A	N/A	1	8	N/A	4	PCBs; total insertion loss of up to 35 dB at 12.9 GHz
100GBASE-KP4	802.3bj-2014 (CL94)	current	Cu-Backplane	N/A	N/A	1	8	N/A	4	PCBs; **Line code: RS-FEC(544,514) × PAM4** **× 92/90 framing and 31320/31280 lane identification** **Line rate: 4x 13.59375 GBd = 54.375 GBd** total insertion loss of up to 33 dB at 7 GHz
100GBASE-CR4 *Direct Attach*	802.3bj-2010 (CL92)	current	twinaxial balanced	**QSFP28** (SFF-8665) CFP2 CFP4	N/A	5	8	N/A	4	Data centres (inter-rack)
100GBASE-SR4	802.3bm-2015 (CL95)	current	Fibre 850 nm	MPO/MTP (MPO-12)	**QSFP28** CFP2 CFP4 CPAK	OM3: 70 OM4: 100	8	1	4	
100GBASE-SR2-BiDi *(BiDirectional)*	*proprietary* *(non IEEE)*	current	Fibre 850 nm 900 nm	LC	**QSFP28**	OM3: 70 OM4: 100	2	2	2	**WDM** **Line rate: 2x (2x 26.5625 GBd with PAM4)** duplex fiber with both being used to transmit and receive; The major selling point of this variant is its ability to run over existing 25G multi-mode fiber (i.e. allowing easy migration from 25G to 100G).
100GBASE-SWDM4	*proprietary* *(MSA, Nov 2017)*	current	Fibre 844 – 858 nm 874 – 888 nm 904 – 918 nm 934 – 948 nm	LC	**QSFP28**	OM3: 75 OM4: 100 OM5: 150	2	4	4	**SWDM**[106]
100GBASE-LR4	802.3ba-2010 (CL88)	current	Fibre 1295.56 nm 1300.05 nm 1304.59 nm	LC	**QSFP28** CFP CFP2 CFP4 CPAK	OSx: 10k	2	4	4	**WDM** **Line code: 64b/66b × NRZ**

UNICAMP



200G Ethernet

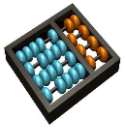
200G port types [edit]

Legend for fibre-based TP-PHYs[35]

MMF FDDI 62.5/125 µm (1987)	MMF OM1 62.5/125 µm (1989)	MMF OM2 50/125 µm (1998)	MMF OM3 50/125 µm (2003)	MMF OM4 50/125 µm (2008)	MMF OM5 50/125 µm (2016)	SMF OS1 9/125 µm (1998)	SMF OS2 9/125 µm (2000)
160 MHz·km @ 850 nm	200 MHz·km @ 850 nm	500 MHz·km @ 850 nm	1500 MHz·km @ 850 nm	3500 MHz·km @ 850 nm	3500 MHz·km @ 850 nm & 1850 MHz·km @ 950 nm	1 dB/km @ 1300/ 1550 nm	0.4 dB/km @ 1300/ 1550 nm

Name	Standard	Status	Media	Connector	Transceiver Module	Reach in m	# Media (⇆)	# Lambdas (→)	# Lanes (→)	Notes
200 Gigabit Ethernet (200 GbE) (1st Generation: 25GbE-based) - (Data rate: 200 Gbit/s - Line code: 256b/257b × RS-FEC(544,514) × NRZ - Line rate: 8x 26.5625 GBd = 212.5 GBd - Full-Duplex) [36][37][38]										
200GAUI-8	802.3bs-2017 (CL120B/C)	current	Chip-to-chip/ Chip-to-module interface	N/A	N/A	0.25	16	N/A	8	PCBs
200 Gigabit Ethernet (200 GbE) (2nd Generation: 50GbE-based) - (Data rate: 200 Gbit/s - Line code: 256b/257b × RS-FEC(544,514) × PAM4 - Line rate: 4x 26.5625 GBd x2 = 212.5 GBd - Full-Duplex) [36][37][38]										
200GAUI-4	802.3bs-2017 (CL120D/E)	current	Chip-to-chip/ Chip-to-module interface	N/A	N/A	0.25	8	N/A	4	PCBs
200GBASE-KR4	802.3cd-2018 (CL137)	current	Cu-Backplane	N/A	N/A	1	8	N/A	4	PCBs; total insertion loss of ≤ 30 dB at 13.28125 GHz
200GBASE-CR4	802.3cd-2018 (CL136)	current	twinaxial copper cable	QSFP56, microQSFP, QSFP-DD, OSFP (SFF-8665)	N/A	3	8	N/A	4	Data centres (in-rack)
200GBASE-SR4	802.3cd-2018 (CL138)	current	Fibre 850 nm	MPO/MTP (MPO-12)	QSFP56	OM3: 70 OM4: 100	8	1	4	uses four fibers in each direction
200GBASE-DR4	802.3bs-2017 (CL121)	current	Fibre 1304.5 – 1317.5 nm	MPO/MTP (MPO-12)	QSFP56	OS2: 500	8	1	4	uses four fibers in each direction
200GBASE-FR4	802.3bs-2017 (CL122)	current	Fibre 1271 – 1331 nm	LC	QSFP56	OS2: 2k	2	4	4	WDM
200GBASE-LR4	802.3bs-2017 (CL122)	current	Fibre 1295.56 – 1309.14 nm	LC	QSFP56	OS2: 10k	2	4	4	WDM
200GBASE-ER4	802.3cn-2019 (CL122)	current	Fibre 1295.56 – 1309.14 nm	LC	QSFP56	OS2: 40k	2	4	4	WDM
200 Gigabit Ethernet (200 GbE) (3rd Generation: 100GbE-based) - (Data rate: 200 Gbit/s - Line code: 256b/257b × RS-FEC(544,514) × PAM4 - Line rate: 2x 53.1250 GBd x2 = 212.5 GBd - Full-Duplex) [36][37][38]										
802.3ck			Chip-to-chip/							

UNICAMP



400G Ethernet

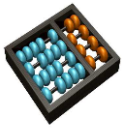
400G port types [edit]

Legend for fibre-based TP-PHYs[35]

MMF FDDI 62.5/125 µm (1987)	MMF OM1 62.5/125 µm (1989)	MMF OM2 50/125 µm (1998)	MMF OM3 50/125 µm (2003)	MMF OM4 50/125 µm (2008)	MMF OM5 50/125 µm (2016)	SMF OS1 9/125 µm (1998)	SMF OS2 9/125 µm (2000)
160 MHz·km @ 850 nm	200 MHz·km @ 850 nm	500 MHz·km @ 850 nm	1500 MHz·km @ 850 nm	3500 MHz·km @ 850 nm	3500 MHz·km @ 850 nm & 1850 MHz·km @ 950 nm	1 dB/km @ 1300/ 1550 nm	0.4 dB/km @ 1300/ 1550 nm

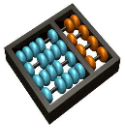
Name	Standard	Status	Media	Connector	Transceiver Module	Reach in m	# Media (⇆)	# Lambdas (→)	# Lanes (→)	Notes
400 Gigabit Ethernet (400 GbE) (1st Generation: 25GbE-based) - (Data rate: 400 Gbit/s - Line code: 256b/257b × RS-FEC(544,514) × NRZ - Line rate: 16x 26.5625 GBd = 425 GBd - Full-Duplex) [36]										
400GAUI-16	802.3bs-2017 (CL120B/C)	current	Chip-to-chip/ Chip-to-module interface	N/A	N/A	0.25	32	N/A	16	PCBs
400GBASE-SR16	802.3bs-2017 (CL123)	current	Fibre 850 nm	MPO/MTP (MPO-32)	CFP8	OM3: 70 OM4: 100 OM5: 100	32	1	16	
400 Gigabit Ethernet (400 GbE) (2nd Generation: 50GbE-based) - (Data rate: 400 Gbit/s - Line code: 256b/257b × RS-FEC(544,514) × PAM4 - Line rate: 8x 26.5625 GBd x2 = 425.0 GBd - Full-Duplex) [36]										
400GAUI-8	802.3bs-2017 (CL 120D/E)	current	Chip-to-chip/ Chip-to-module interface	N/A	N/A	0.25	16	N/A	8	PCBs
400GBASE-KR8	*proprietary (ETC) (CL120)*	current	Cu-Backplane	N/A	N/A	1	8	N/A	8	**WDM**
400GBASE-SR8	802.3cm-2020 (CL138)	current	Fiber 850 nm	MPO/MTP (MPO-16)	QSFP-DD	OM3: 70 OM4: 100 OM5: 100	16	1	8	
400GBASE-SR4.2 *(Bidirectional)*	802.3cm-2020 (CL150)	current	Fiber 850 nm 912 nm	MPO/MTP (MPO-12)	QSFP-DD	OM3: 70 OM4: 100 OM5: 150	8	2	8	Bidirectional WDM
400GBASE-FR8	802.3bs-2017 (CL122)	current	Fibre 1273.54 – 1309.14 nm	LC	QSFP-DD	OS2: 2k	2	8	8	**WDM**
400GBASE-LR8	802.3bs-2017 (CL122)	current	Fibre 1273.54 – 1309.14 nm	LC	QSFP-DD	OS2: 10k	2	8	8	**WDM**
400GBASE-ER8	802.3cn-2019 (CL122)	current	Fibre 1273.54 – 1309.14 nm	LC	QSFP-DD	OS2: 40k	2	8	8	**WDM**
400 Gigabit Ethernet (400 GbE) (3rd Generation: 100GbE-based) - (Data rate: 400 Gbit/s - Line code: 256b/257b × RS-FEC(544,514) × PAM4 - Line rate: 4x 53.1250 GBd x2 = 425.0 GBd - Full-Duplex) [36]										
	802.3ck		Chip-to-chip/							

UNICAMP



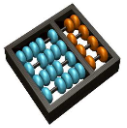
Camada de Enlace de Dados

- 5.1 Introdução e Serviços
- 5.2 Correção e detecção de
- 5.3protocolos Múltilpo Acesso
- 5.4 Endereçamento
- 5.5 Ethernet
- 5.6 Switches
- 5.7 PPP
- 5.8 Virtualização: MPLS
- 5.9 Data Center Networks



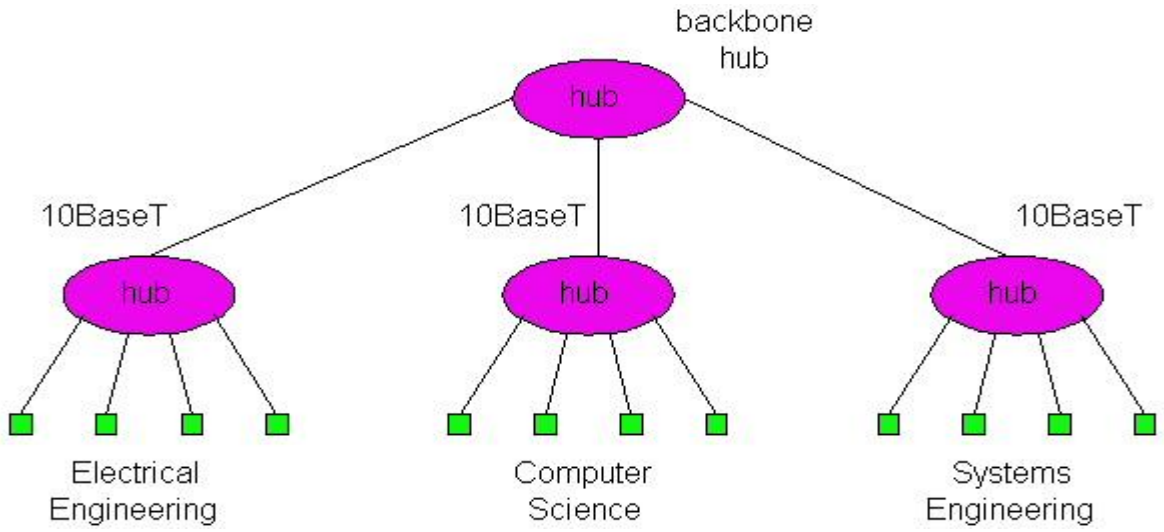
Interconectando segmentos de Redes

- Usados para estender as características das redes locais: cobertura geográfica, número de nós, funcionalidade administrativa, etc.
- Diferem entre si em respeito a:
  - ✓ isolamento de domínios de colisão
  - ✓ camada em que operam
- Diferentes de roteadores
  - ✓ "plug and play"
  - ✓ não provêem roteamento ótimo de pacotes IP
- Concentradores (Hubs), Pontes (Bridges), Comutadores (Switches)
  - ✓ Nota: comutadores são essencialmente pontes com múltiplas portas;
  - ✓ O que se fala para pontes, também é válido para comutadores;



Interconexão utilizando Hubs

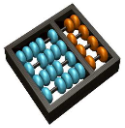
- Dispositivos da camada física: basicamente são repetidores operando ao nível de bit: repete os bits recebidos numa interface para as demais interfaces
- Hubs podem ser dispostos numa hierarquia (ou **projeto de múltiplos níveis**), com um hub **backbone** na raíz;
- Domínios de colisões individuais tornam-se grandes domínios de colisões
  - ✓ Se um nó em CS e um outro nó em EE transmitem ao mesmo tempo: colisão;



Interconexão utilizando Hubs (cont)

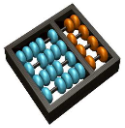
- Cada rede local ligado é chamada de **segmento** de rede local
- Hubs **não isolam** domínios de colisão: um nó pode colidir com qualquer outro nó residindo em qualquer segmento da rede local

- Vantagens de Hubs:
  - ✓ Dispositivos simples, baratos
  - ✓ Configuração em múltiplos níveis provê degradação paulatina: porções da rede local continuam a operar se um dos hubs parar de funcionar
  - ✓ Estende a distância máxima entre pares de nós (100m por Hub)



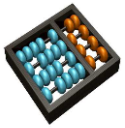
Interconexão utilizando Hubs (cont)

- Limitações de Hubs:
  - ✓ Domínio de colisão único resulta em nenhum aumento na vazão máxima; a vazão no caso de múltiplos níveis é igual à do segmento único
  - ✓ Restrições em redes locais individuais põe limites no número de nós no mesmo domínio de colisão (portanto, por Hub ou coleção de Hubs); e na cobertura geográfica total permitida
  - ✓ Não se pode misturar tipos diferentes de Ethernet (p.ex., 10BaseT and 100BaseT)



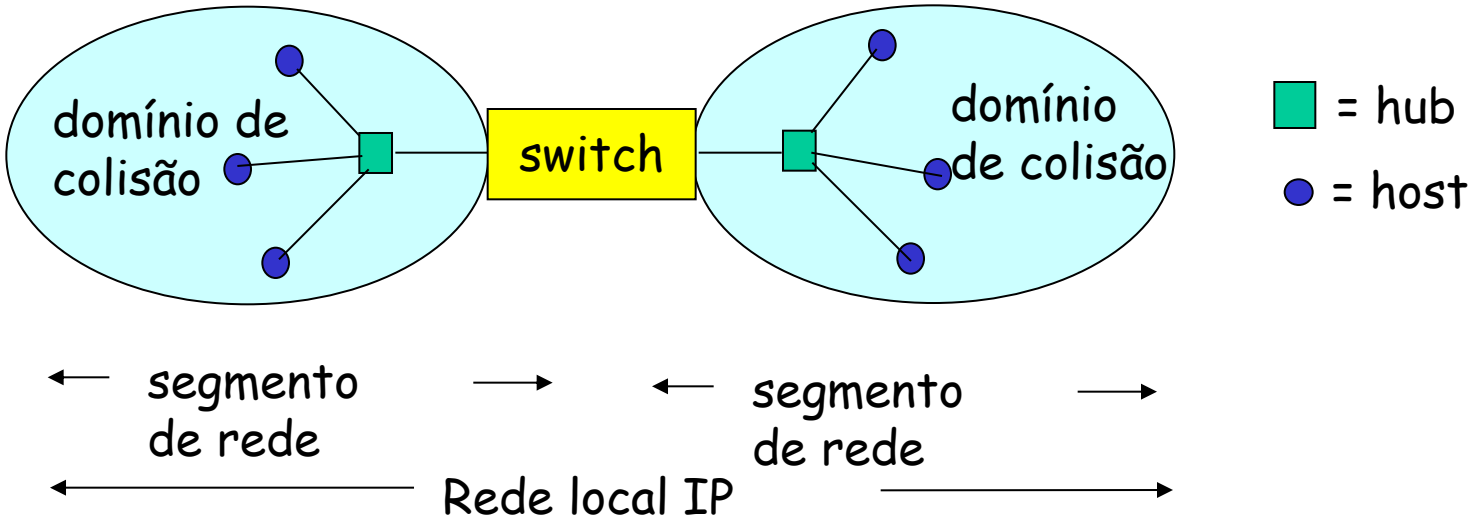
Comutadores ("Switches")

- **Dispositivos da camada de enlace**:
  - ✓ operam em quadros Ethernet
  - ✓ examinam o cabeçalho do quadro, e reencaminham selectivamente um quadro com base no seu endereço de destino
  - ✓ Quando se quer re-encaminhar um quadro num segmento, usa CSMA/CD para fazer acesso ao segmento e transmitir;
- Transparente: hosts desconhecem a existência dos switches;
- plug-and-play, auto aprendizagem
  - ✓ Switches não necessitam ser configuradas

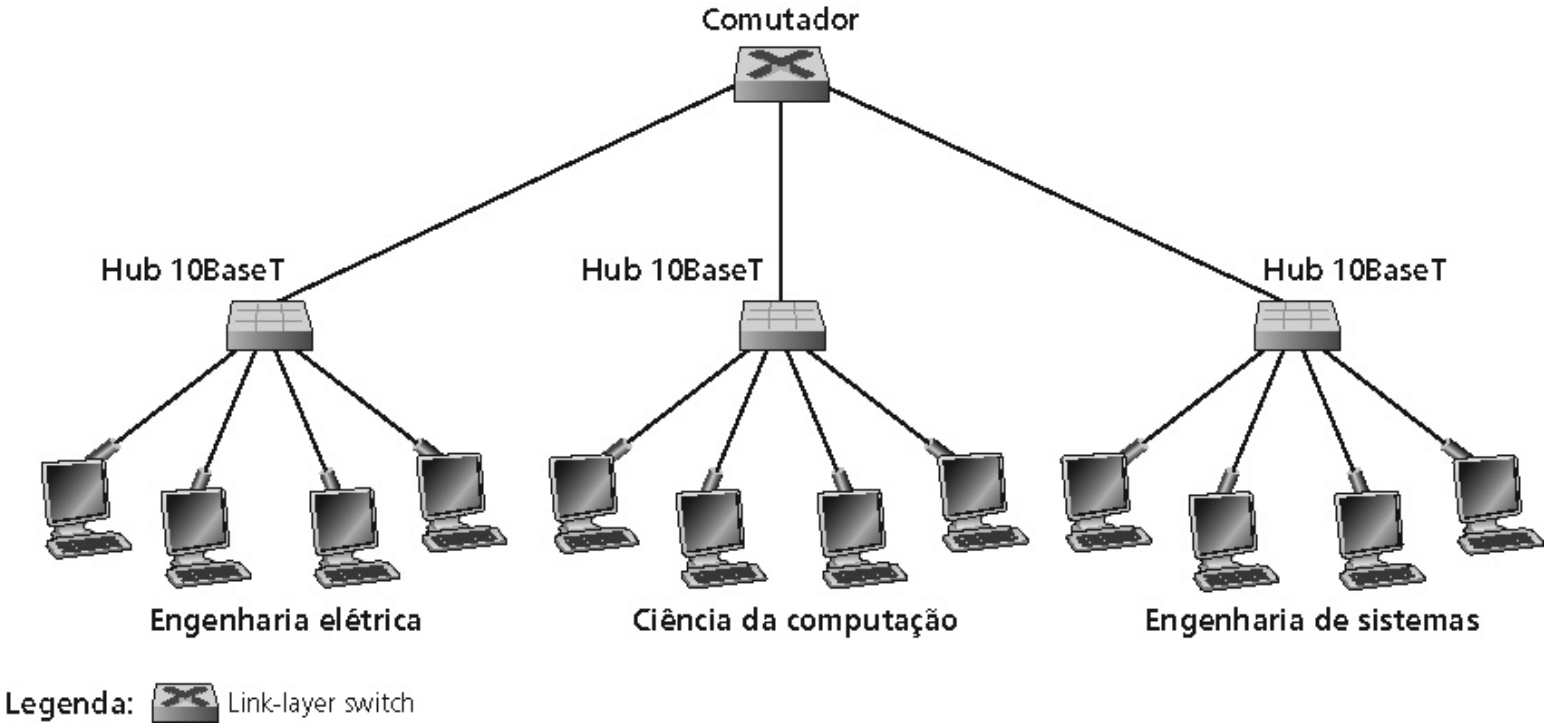


Switches: isolamento de tráfego

- A instalação do switch particiona a rede em segmentos de LAN
- switch filtra pacotes:
  - ✓ Quadros de um segmento de rede não são geralmente reencaminhados para outro segmento de rede;
  - ✓ Segmentos separam os domínios de colisão

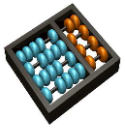


Encaminhamento dos quadros



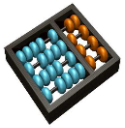
Como determinar para qual segmento de rede um quadro deve ser encaminhado?

- Parece um problema de roteamento ....



Auto Aprendizagem

- Um switch tem uma tabela de switch;
- Entradas da tabela do switch:
  - (endereço do no na rede, interface do switch, tempo corrente)
  - Entradas expiradas na tabela de filtragem são descartadas (TTL geralmente é de 60 min)
- switches aprendem quais hosts podem ser acessados através de quais interfaces
  - Quando um quadro é recebido, o switch "aprende" a localização do emisor: qual segmento de rede ele pertence;
  - Armazena o par emissor/localização na tabela;



Filtragem/Encaminhamento de quadros

Quando um switch recebe um quadro:

se destino estiver na rede local pela qual o quadro foi recebido

então descarta o quadro

senão { faz pesquisa na tabela de filtragem

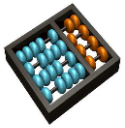
se foi encontrada a entrada para o destino

então re-encaminha o quadro na interface indicada;

senão faz inundação;

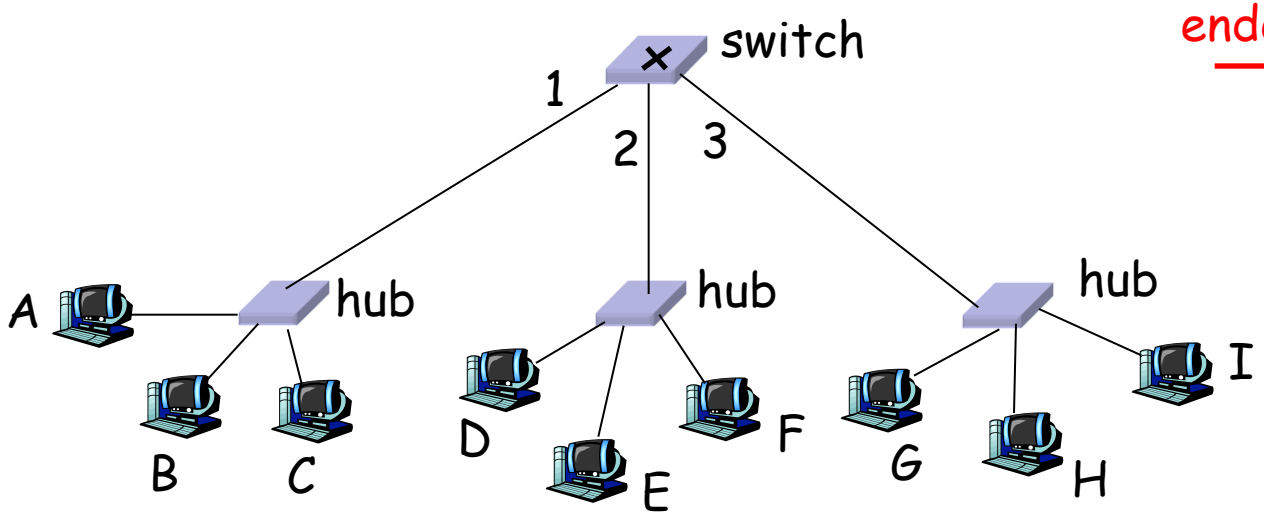
}

re-encaminha em todas as interfaces exceto naquela por onde chegou



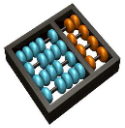
Switch: exemplo

Suponha que C envia um quadro para D



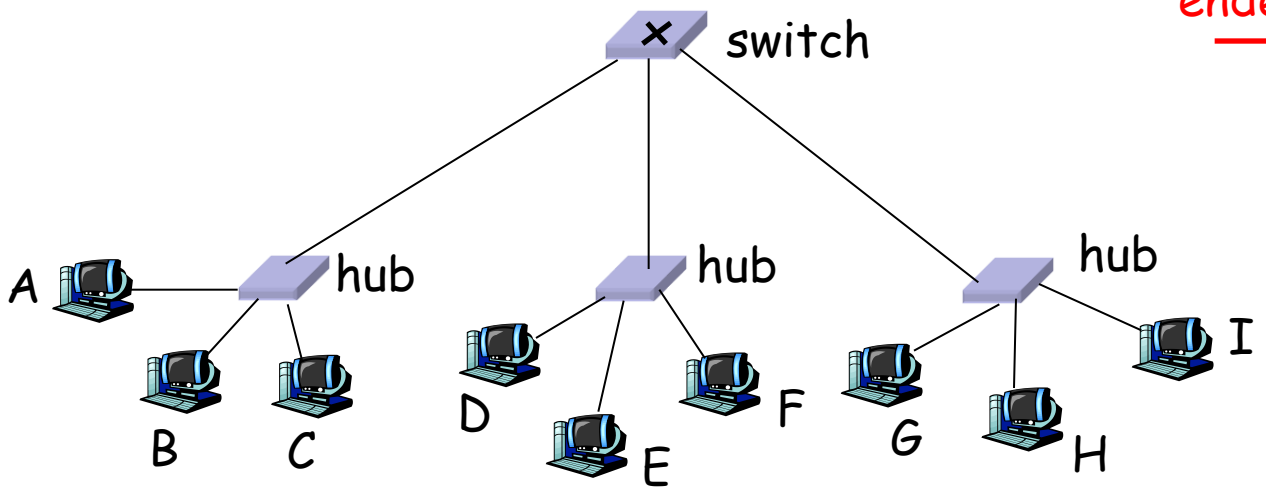
endereço	interface
A	1
B	1
E	2
G	3

- Switch recebe o quadro de C
  - Anota na tabela que C está na interface 1
  - Como D não está na tabela, o switch encaminha o quadro para as interfaces 2 e 3
- Quadro recebido por D



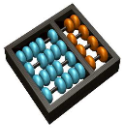
Switch: exemplo

Suponha que D responde com um quadro para C.

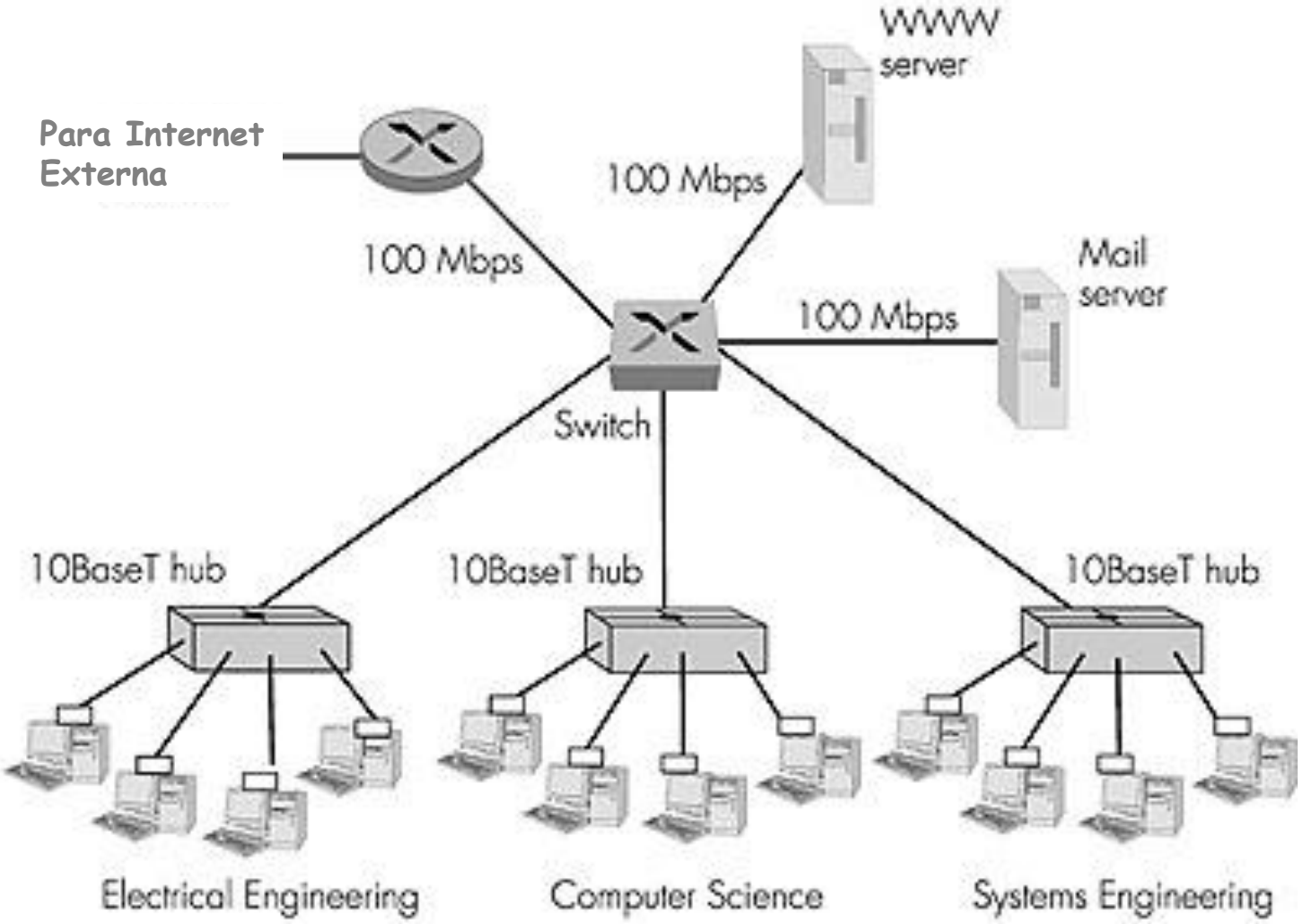


endereço	interface
A	1
B	1
E	2
G	3
C	1

- Switch recebe quadro de D
  - Anota na tabela que D está na interface 2
  - Como C está na tabela, o switch encaminha o quadro apenas para a interface 1
- Quadro recebido por C



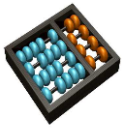
Switches (mais)



Switches (cont)

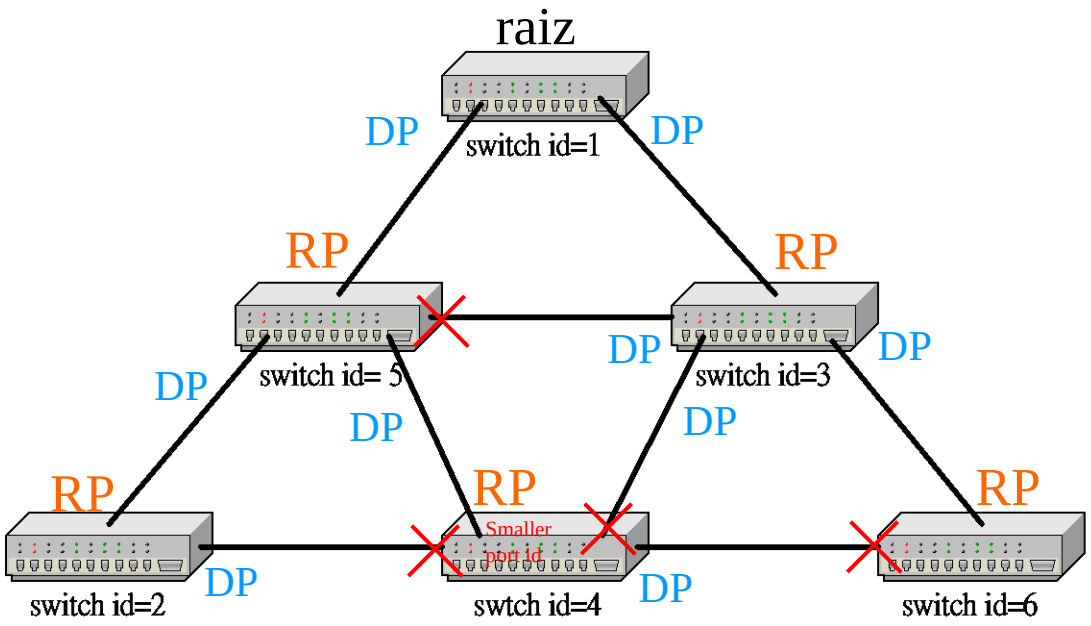
- Alguns switches suportam comutação acelerada (cut-through switching): o quadro é enviado da entrada para a saída sem esperar pela montagem do quadro inteiro
  - ✓ pequena redução da latência

- Switches variam em tamanho, e os mais rápidos incorporam uma rede de interconexão de alta capacidade



Spanning Tree

Objetivo: Resolve loops em bridges conectadas

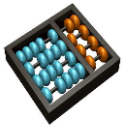


1. A raiz é a switch com menor id
2. Propaga informações de configuração tais como custo dos enlaces em pacotes BPDU para as portas designadas
3. Para cada LAN (switch), a DP (RP) seleciona-se a porta com menor custo para sera a porta designada
4. Se houver empate, seleciona-se a porta com o menor identificador (id)
5. Todas as portas que não são portas designadas são bloqueadas

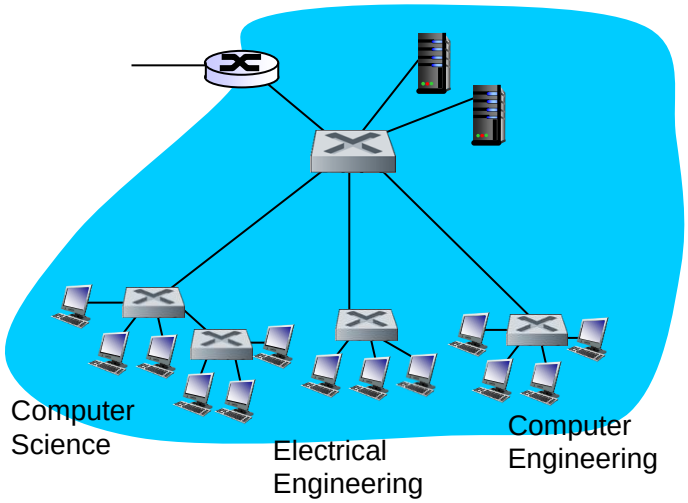
RP: Porta da Raiz (root port)

DP: Porta designada (designated port)

BPDU: Bridge Protocol Data Unit



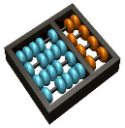
VLANs: motivation



consider:

- CS user moves office to EE, but wants connect to CS switch?
- single broadcast domain:
  - all layer-2 broadcast traffic (ARP, DHCP, unknown location of destination MAC address) must cross entire LAN
  - security/privacy, efficiency issues

UNICAMP

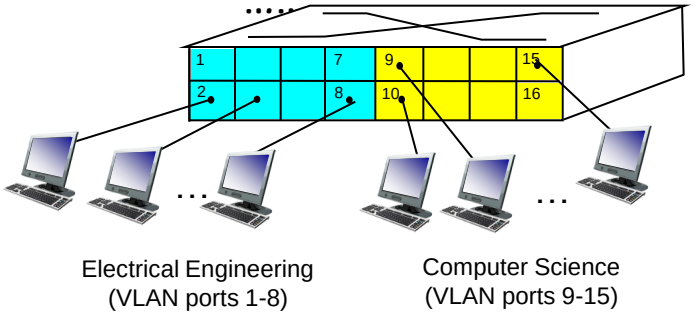


VLANs

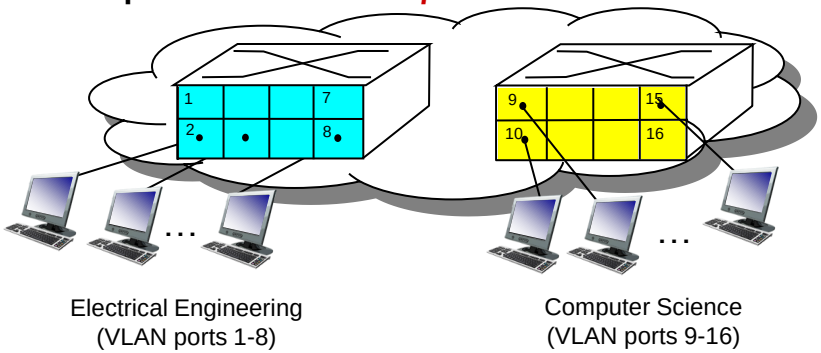
port-based VLAN: switch ports grouped (by switch management software) so that *single* physical switch

Virtual Local Area Network

switch(es) supporting VLAN capabilities can be configured to define multiple ***virtual*** LANS over single physical LAN infrastructure.



… operates as *multiple* virtual switches



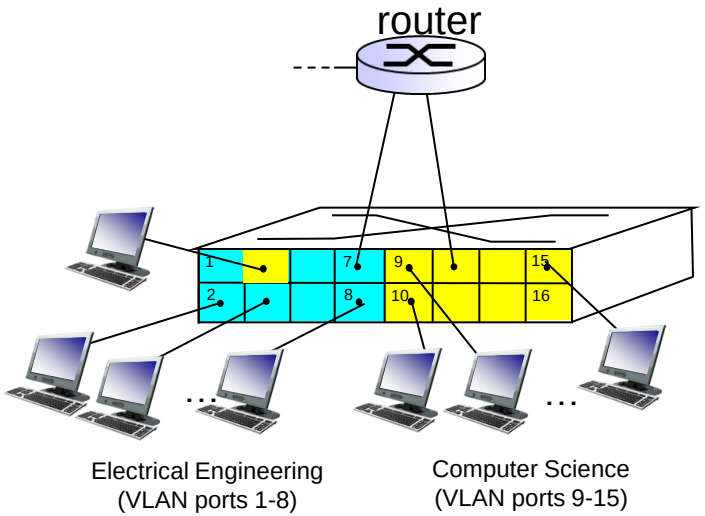
UNICAMP

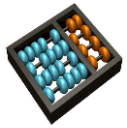
Port-based VLAN

- *traffic isolation:* frames to/from ports 1-8 can *only* reach ports 1-8
  - can also define VLAN based on MAC addresses of endpoints, rather than switch port

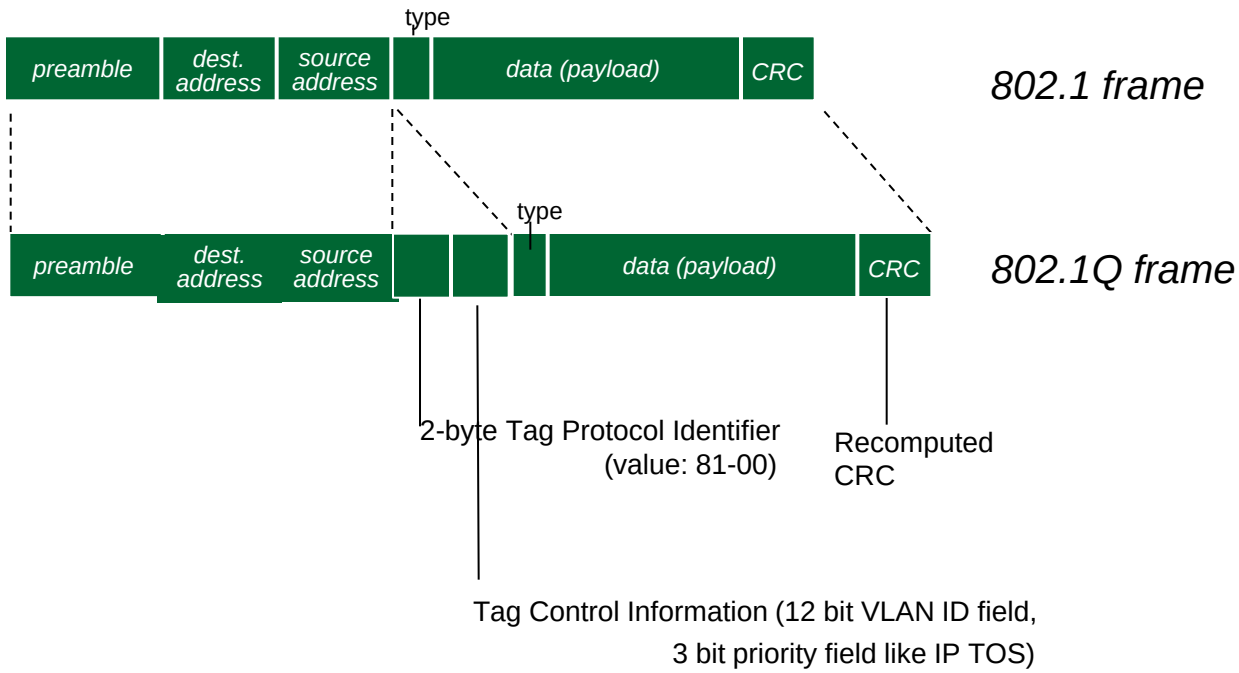
- dynamic membership: ports can be dynamically assigned among VLANs

- forwarding between VLANS: done via routing (just as with separate switches)
  - in practice vendors sell combined switches plus routers





802.1Q VLAN frame format



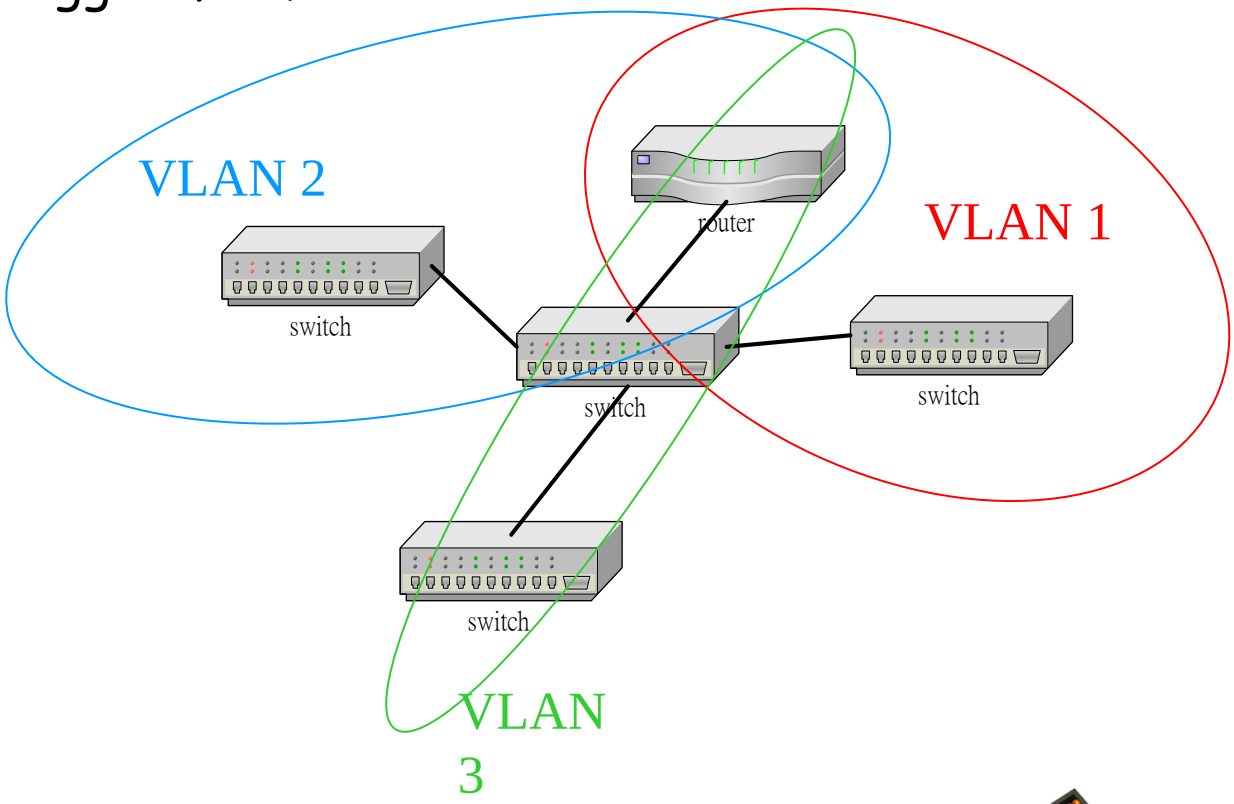
UNICAMP

VLAN

- Especificado no IEEE 802.1Q
- Conectividade lógica
- tagged frame vs. untagged frame

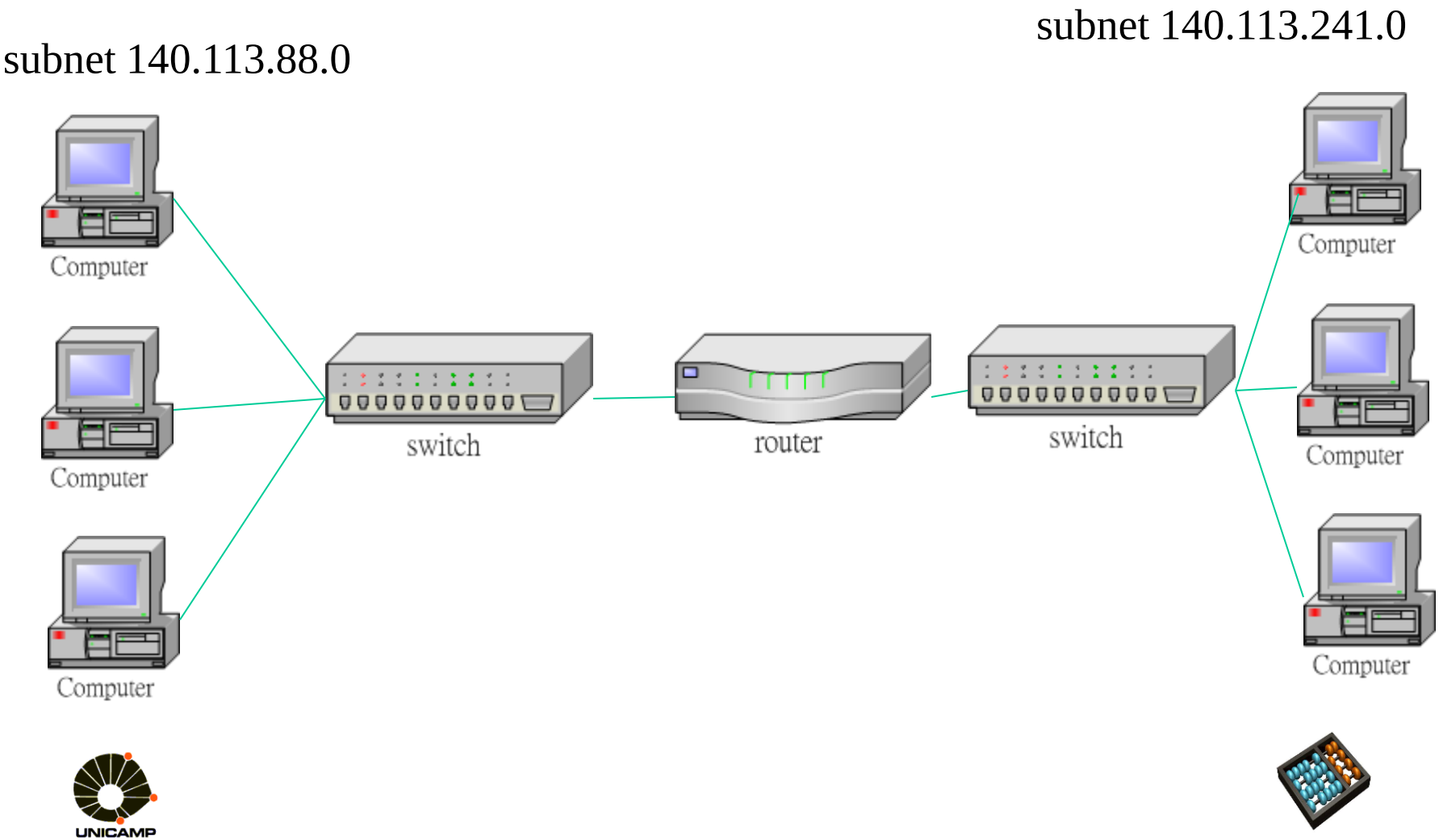
VLAN pode ser associada a

1. Port
2. MAC address
3. Protocolo
4. Subrede IP
5. Application-based

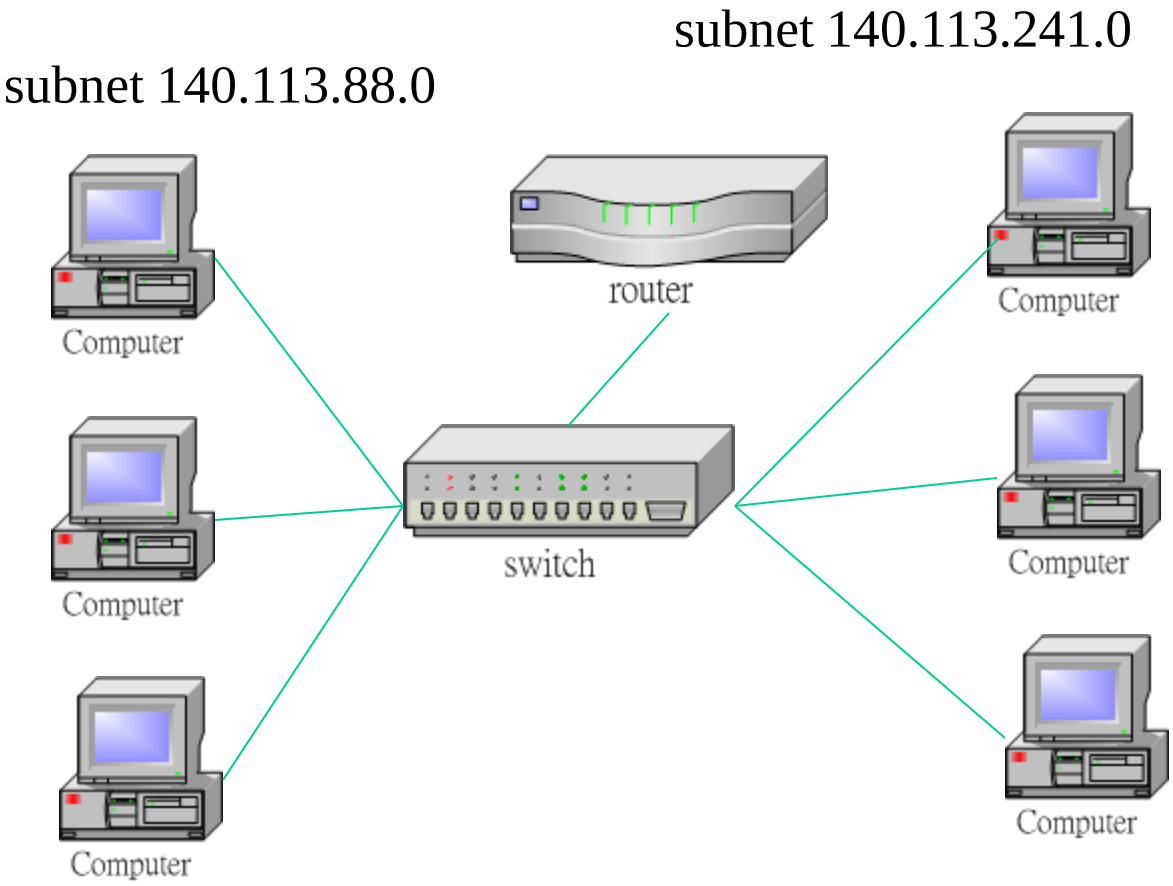


UNICAMP

Exemplo VLAN

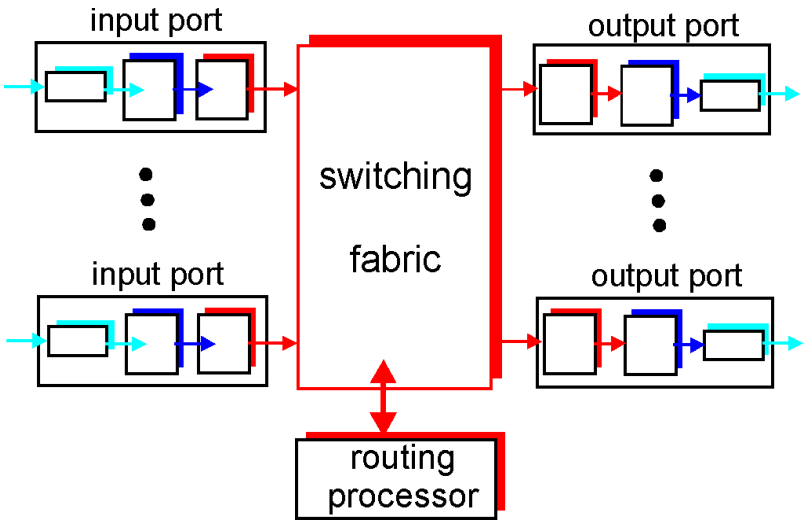


Exemplo VLAN



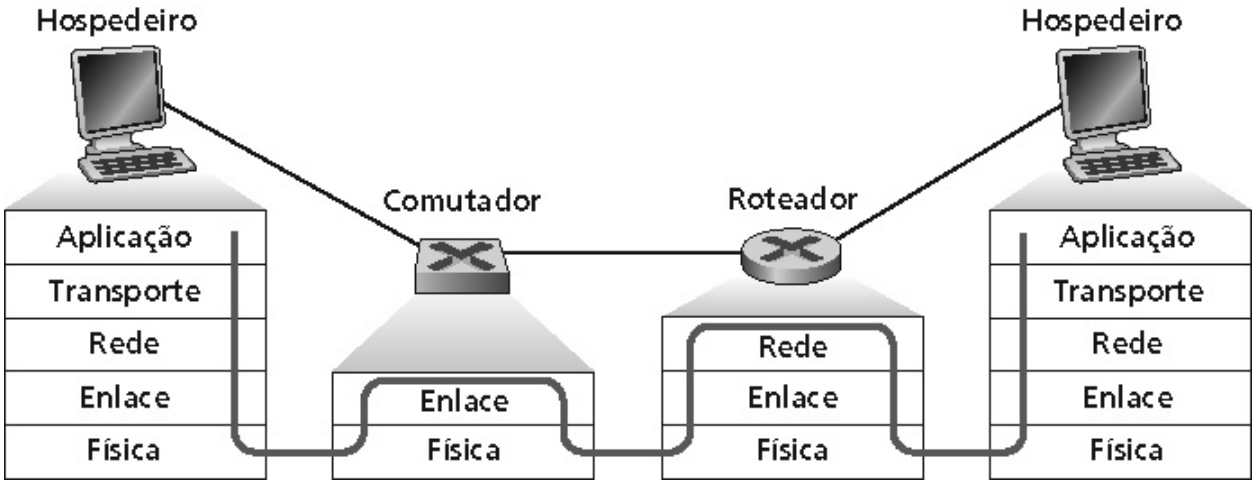
UNICAMP

Switch Layer 2



Switches versus Roteadores

- ambos são dispositivos "armazena e re-encaminha",
  - ✓ roteadores são dispositivos da Camada de Rede (examinam cabeçalhos da camada de rede)
  - ✓ switches são dispositivos da Camada de Enlace
- roteadores mantêm tabelas de rotas e implementam algoritmos de roteamento;
- switches mantêm tabelas, implementam filtragem, são autodidatas e mantêm algoritmos de árvore geradora



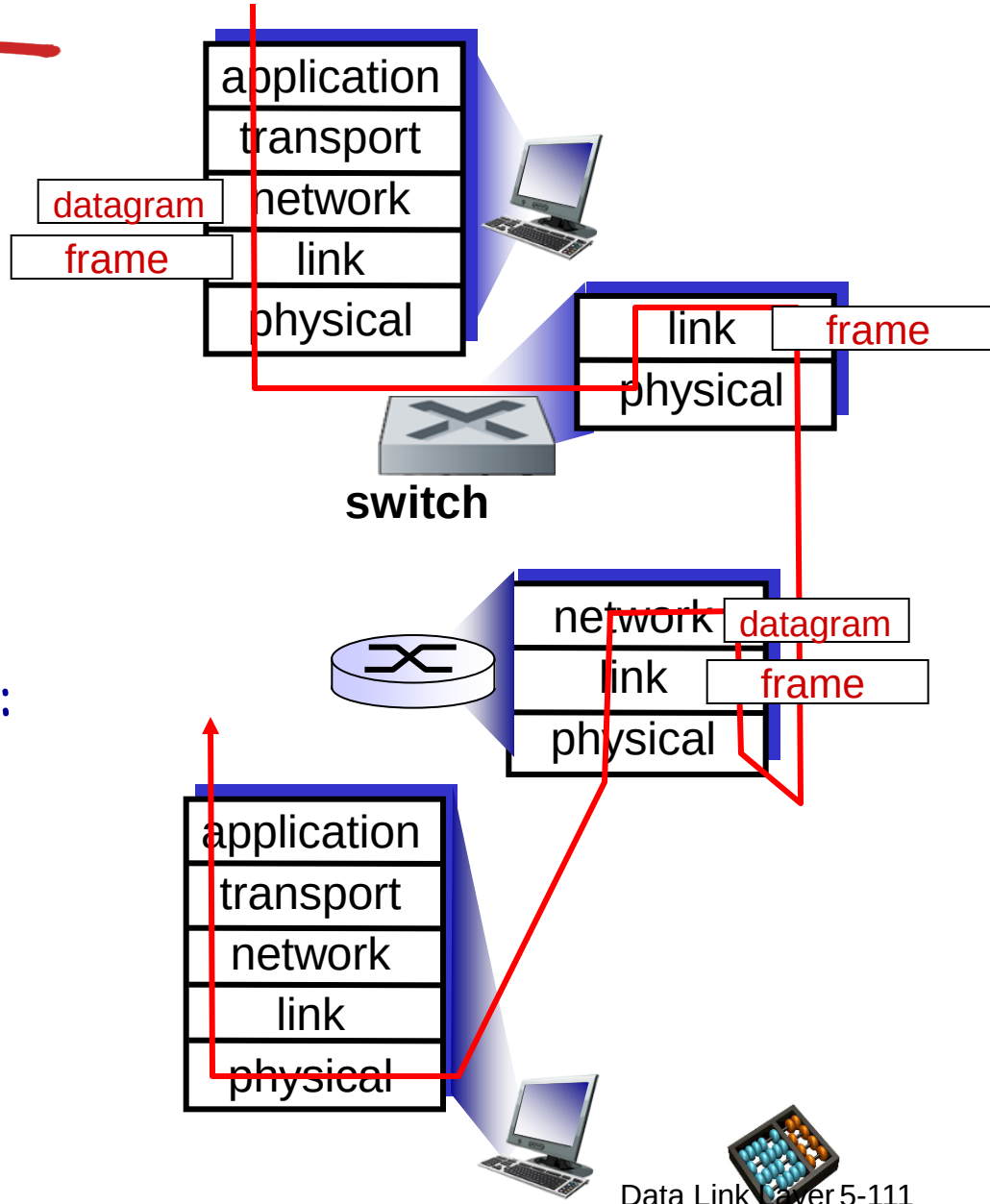
Switches vs. routers

both are store-and-forward:

- *routers:* network-layer devices (examine network-layer headers)
- *switches:* link-layer devices (examine link-layer headers)

both have forwarding tables:

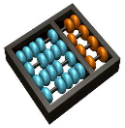
- *routers:* compute tables using routing algorithms, IP addresses
- *switches:* learn forwarding table using flooding, learning, MAC addresses



Switches versus Roteadores

Switches + e -

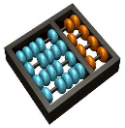
+ Operação de um switch é mais simples requerendo menor capacidade de processamento
+ Tabelas de swicthes são autodidatas;
- Topologias são restritas com switches: uma árvore geradora deve ser construída para evitar ciclos
- Switches não oferecem proteção contra tempestades de difusão ("broadcast storms"): difusão contínua feita por um nó será espalhada por um switch



Switches versus Roteadores

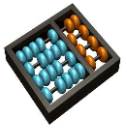
Roteadores + e -

+ São suportadas topologias arbitrárias, ciclos são limitados por contadores TTL (e bons protocolos de roteamento)
+ Provêem proteção "parede corta-fogo" contra tempestades de difusão
- Requerem configuração de endereços IP (não são "plug and play")
- Requerem maior capacidade de processamento



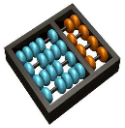
Comparação

	hubs	roteadores	comutadores
isolamento de tráfego	não	sim	sim
plug & play	sim	não	sim
roteamento ótimo	não	sim	não
comutação acelerada	sim	não	sim



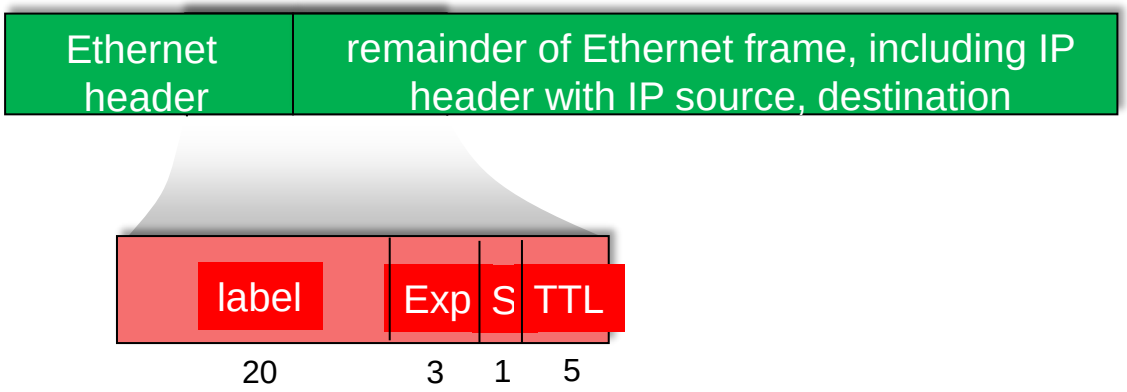
Camada de Enlace de Dados

- 5.1 Introdução e Serviços
- 5.2 Correção e detecção de
- 5.3protocolos Múltilpo Acesso
- 5.4 Endereçamento
- 5.5 Ethernet
- 5.6 Switches
- 5.7 PPP
- 5.8 Virtualização: MPLS
- 5.9 Data Center Networks



Multiprotocol label switching (MPLS)

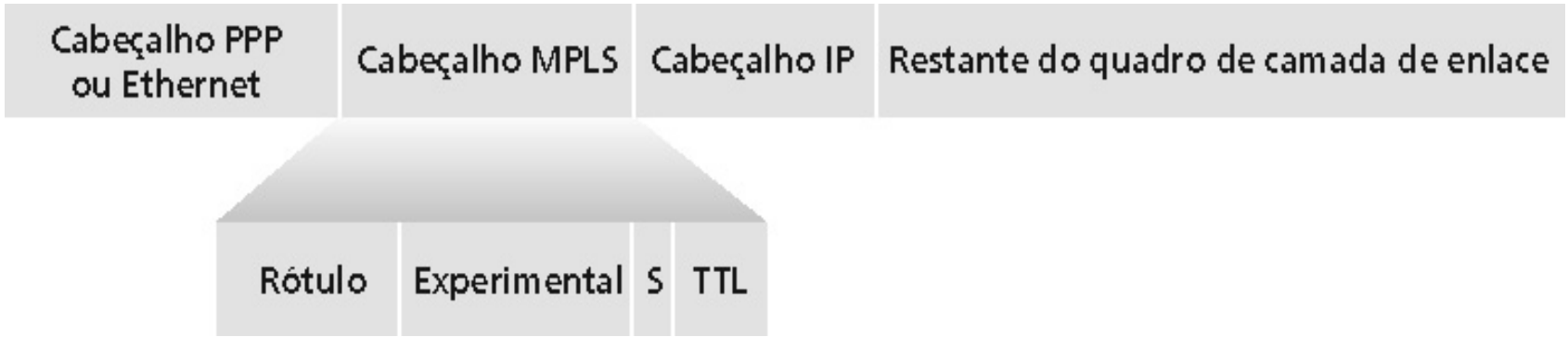
- goal: high-speed IP forwarding among network of MPLS-capable routers, using fixed length label (instead of shortest prefix matching)
  - faster lookup using fixed length identifier
  - borrowing ideas from Virtual Circuit (VC) approach
  - but IP datagram still keeps IP address!



UNICAMP

Multiprotocol label switching (MPLS)

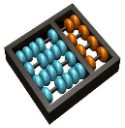
- Objetivo inicial: aumentar a velocidade de encaminhamento IP usando labels de tamanho fixo (em vez de endereço IP)
  - ✓ Mesma idéia do método de circuito virtual (VC)
  - ✓ Mas o datagrama IP ainda mantém o endereço IP!



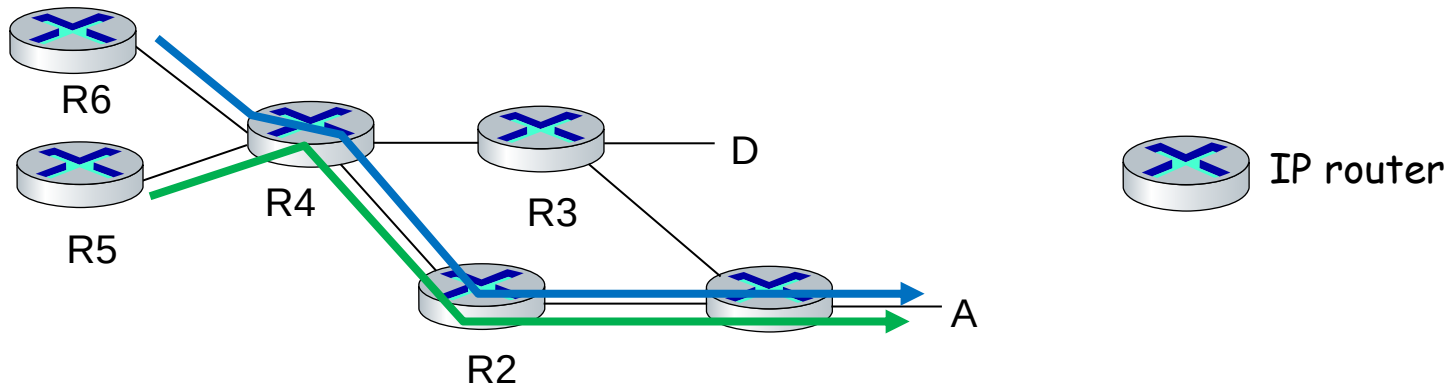
MPLS capable routers

- a.k.a. label-switched router
- forward packets to outgoing interface based only on label value (*don't inspect IP address*)
  - MPLS forwarding table distinct from IP forwarding tables
- *flexibility:* MPLS forwarding decisions can *differ* from those of IP
  - use destination *and* source addresses to route flows to same destination differently (traffic engineering)
  - re-route flows quickly if link fails: pre-computed backup paths

UNICAMP

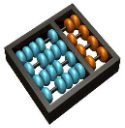


MPLS versus IP paths

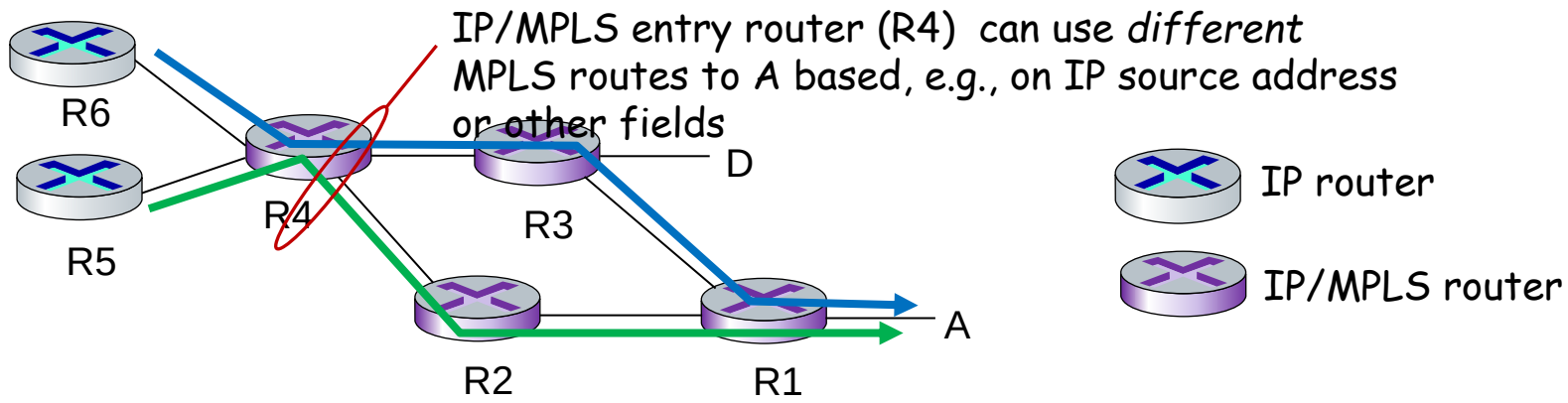


- IP routing: path to destination determined by destination address alone

UNICAMP

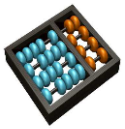


MPLS versus IP paths



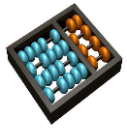
- IP routing: path to destination determined by destination address alone
- MPLS routing: path to destination can be based on source *and* destination address
  - flavor of generalized forwarding (MPLS 10 years earlier)
  - *fast reroute:* precompute backup routes in case of link failure

UNICAMP



Roteadores MPLS

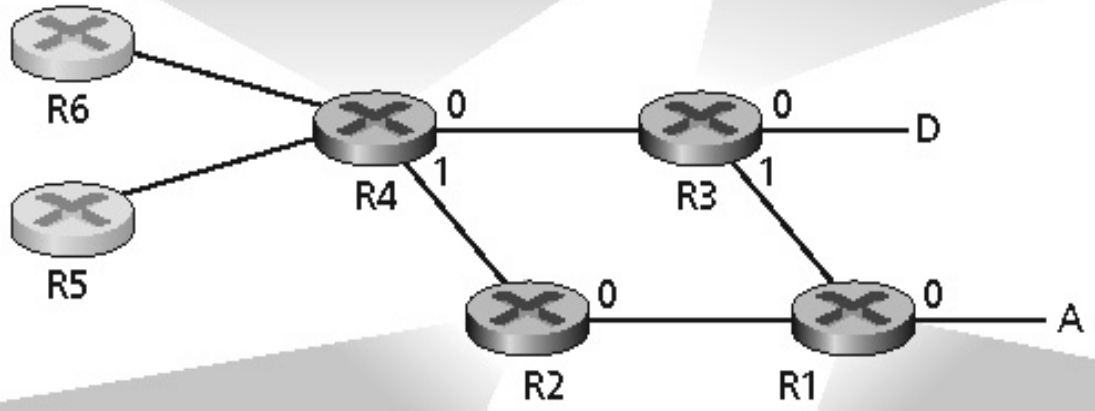
- Roteador faz a função de comutador de rótulo
- Pacotes encaminhados para interface de saída com base apenas no valor do rótulo (não inspeciona o endereço IP)
  - ✓ Tabela de encaminhamento MPLS distinta das tabelas de encaminhamento IP
- Protocolo de sinalização necessário para estabelecer o encaminhamento
  - ✓ RSVP-TE
  - ✓ Encaminhamento é possível por caminhos que o IP sozinho não pode usar (ex.: roteamento de especificado pela origem)!!
  - ✓ Use MPLS para engenharia de tráfego
- Deve coexistir com roteadores unicamente IP



Tabelas de encaminhamento MPLS

rótulo de entrada	rótulo de saída	destino	interface de saída
	10	A	0
	12	D	0
	8	A	1

rótulo de entrada	rótulo de saída	destino	interface de saída
10	6	A	1
12	9	D	0



rótulo de entrada	rótulo de saída	destino	interface de saída
8	6	A	0

rótulo de entrada	rótulo de saída	destino	interface de saída
6	–	A	0

Datacenter networks

10's to 100's of thousands of hosts, often closely coupled, in close proximity:

- e-business (e.g. Amazon)
- content-servers (e.g., YouTube, Akamai, Apple, Microsoft)
- search engines, data mining (e.g., Google)

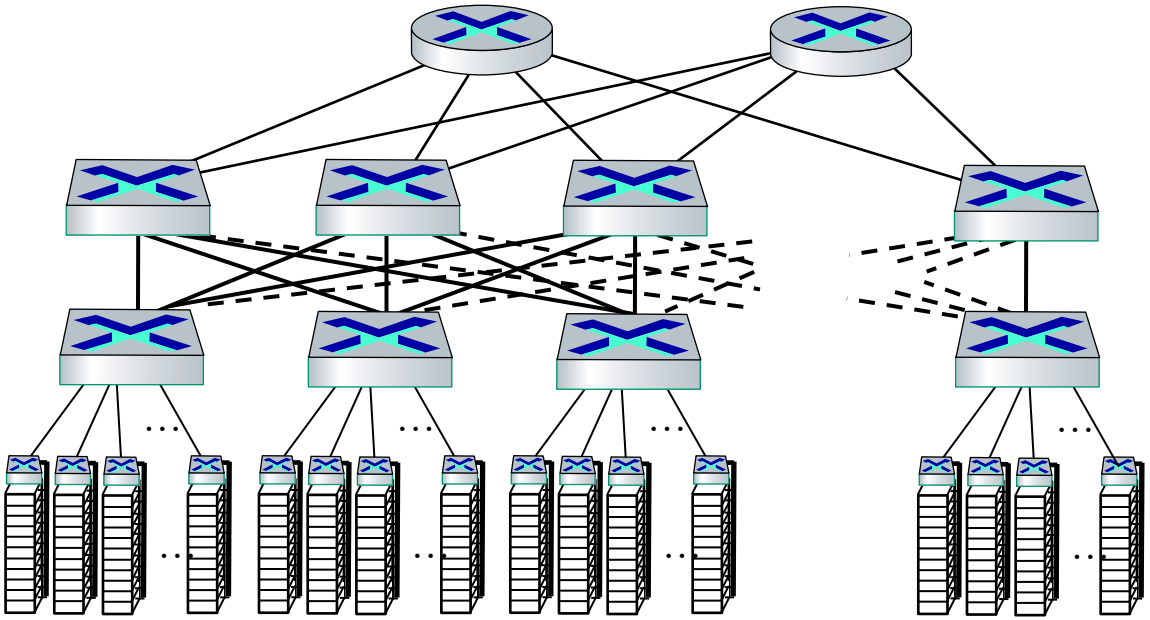
challenges:

- multiple applications, each serving massive numbers of clients
- reliability
- managing/balancing load, avoiding processing, networking, data bottlenecks

Inside a 40-ft Microsoft container, Chicago data center

UNICAMP

Datacenter networks: network elements



Border routers
- connections outside datacenter

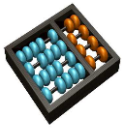
Tier-1 switches
- connecting to ~16 T-2s below

Tier-2 switches
- connecting to ~16 TORs below

Top of Rack (TOR) switch
- one per rack
- 100G-400G Ethernet to blades

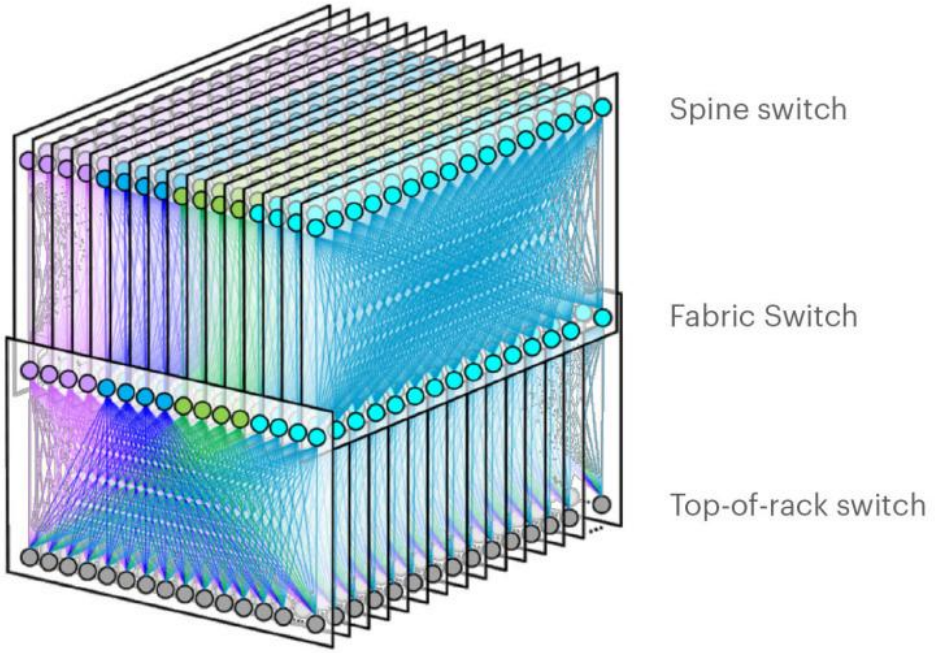
Server racks
- 20- 40 server blades: hosts

UNICAMP



Datacenter networks: network elements

Facebook F16 data center network topology:

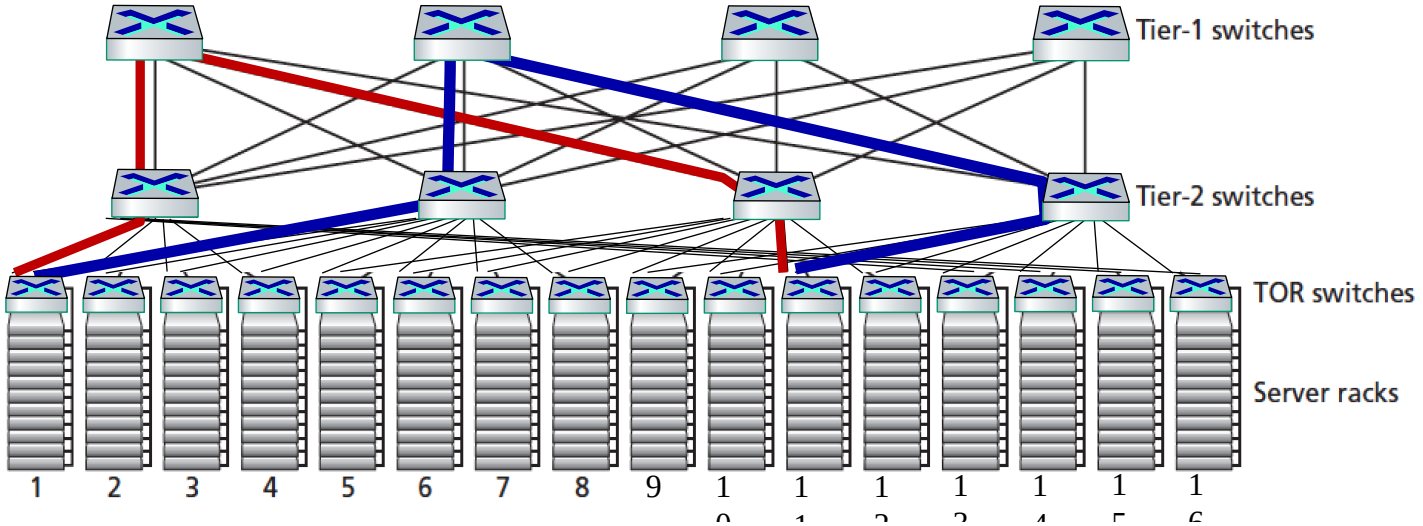


https://engineering.fb.com/data-center-engineering/f16-minipack/ (posted 3/2019)

UNICAMP

Datacenter networks: multipath

- rich interconnection among switches, racks:
  - increased throughput between racks (multiple routing paths possible)
  - increased reliability via redundancy

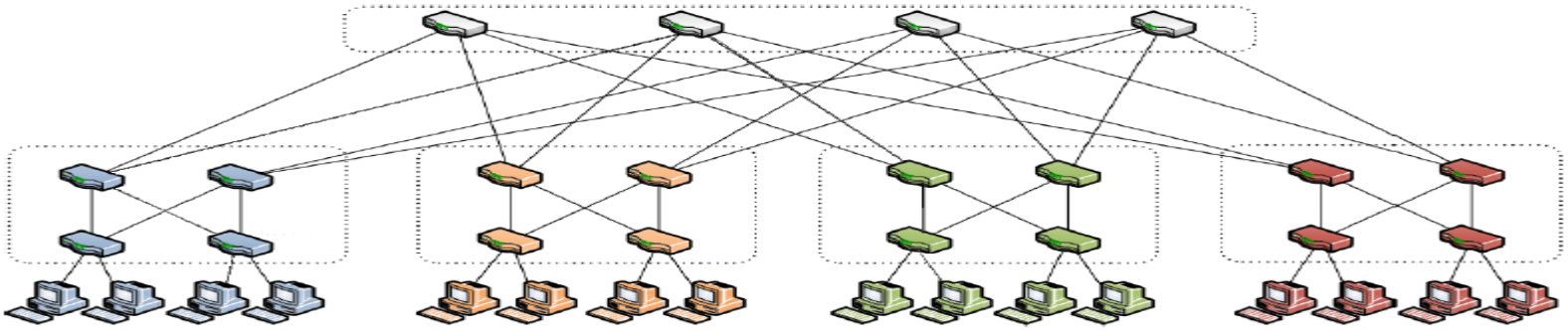


two disjoint paths highlighted between racks 1 and 11

UNICAMP

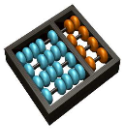
Fat Tree

- Every level is fully connected to lower and upper levels
- Provides higher fault-tolerance and richer connectivity
- Theoretical achievable 1:1 oversubscription with multi-path routingcabling complexity



Fat-Tree topology (adapted from [Al-Fares et al., 2008])

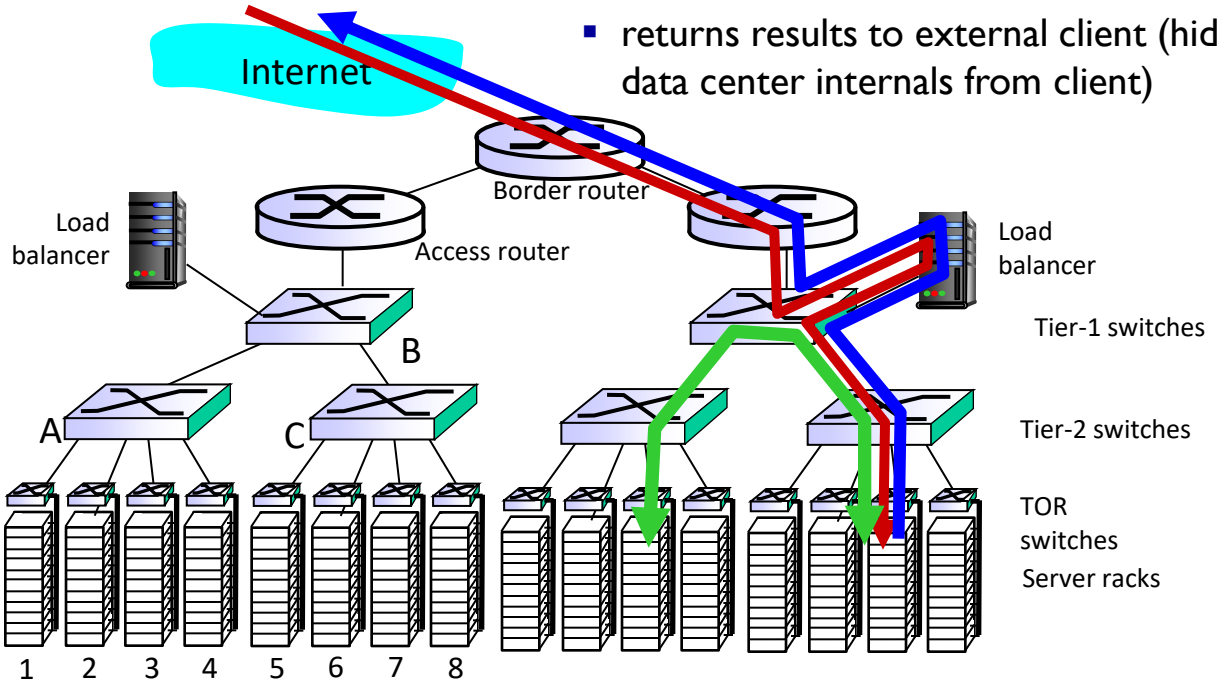
UNICAMP



Data center networks

load balancer: application-layer routing

- receives external client requests
- directs workload within data center
- returns results to external client (hiding data center internals from client)



UNICAMP