MO401

Arquitetura de Computadores I

2006

Prof. Paulo Cesar Centoducatte

ducatte@ic.unicamp.br

www.ic.unicamp.br/~ducatte

MO401

Arquitetura de Computadores I

Network: Definições, Métricas, Protocolos, Roteamento, Wireless, Clusters

"Computer Architecture: A Quantitative Approach" - (Capítulo 8)

Networks

- Terminologia
- ABC de Redes
- Network: Medidas de Desempenho
  - Métricas de Desempenho
- Network Media
  - Comparação entre medias
- Conectando Múltiplos Computadores

Networks

- **Finalidade**: Comunicação entre computadores
- **"Finalidade"**: tratar uma coleção de computadores como um grande computador com compartilhamento de recursos distribuídos.

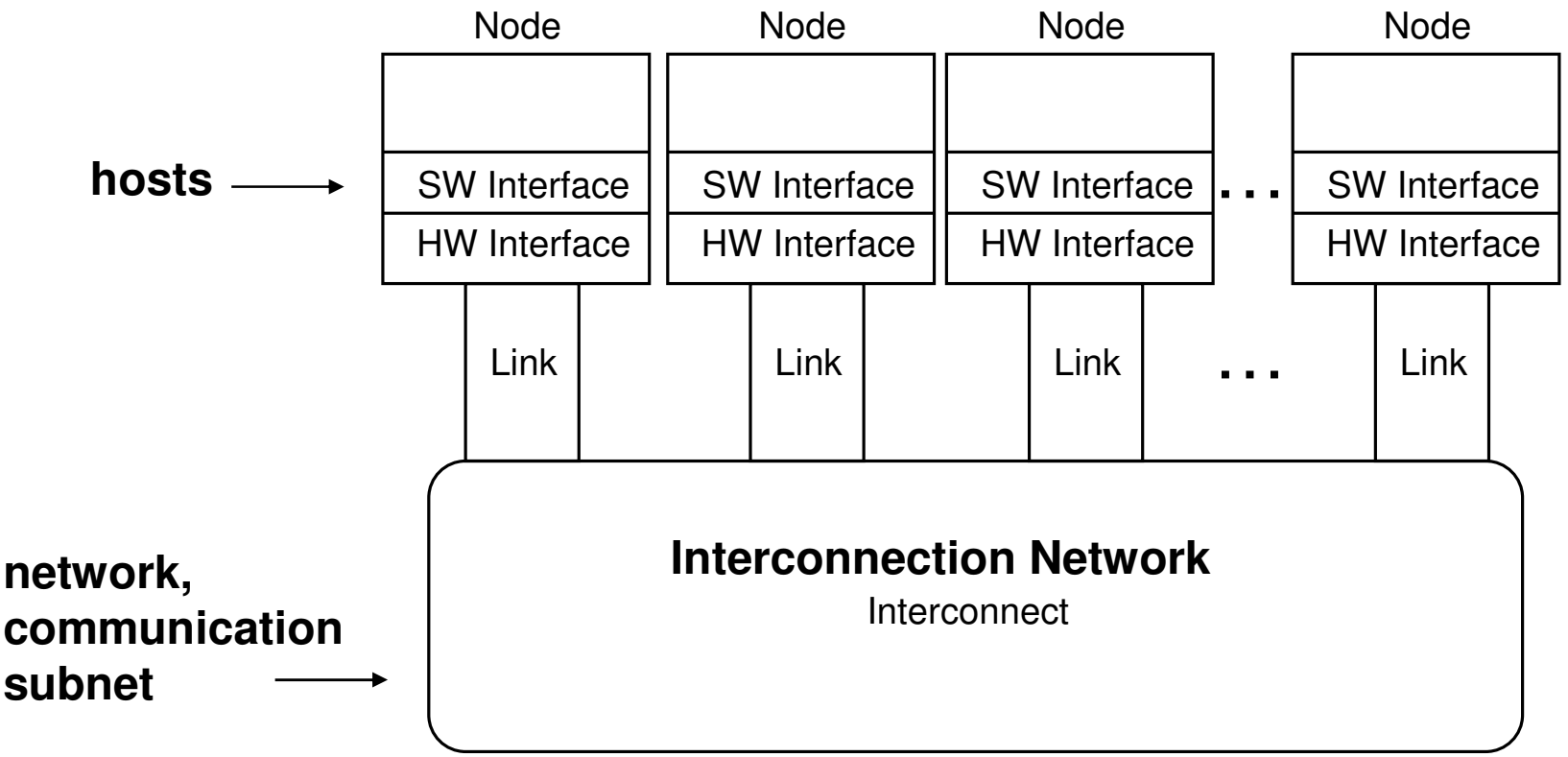
Networks

- O que se fala quando o assunto é redes:
  - direta (point-to-point) vs. indireta (multi-hop)
  - topologia (e.g., bus, ring, DAG)
  - Algoritmos de roteamento
  - Switching, bridge
  - wiring (e.g., par trançado, coaxial, fibra)
- O que é importante:
  - latency
  - bandwidth
  - custo
  - Confiabilidade

Interconexões (Networks)
(Terminologia)

- Wide Area Network - WAN (ATM): 100-1000s nodes; ~ 5,000 km
- Local Area Networks - LAN (Ethernet): 10-1000 nodes; ~ 1-2 km
- System/Storage Area Networks - SAN (FC-AL): 10-100s nodes; ~ 0.025 to 0.1 km por link

Interconexões (Networks)
(Terminologia)



SAN (Storage ou System)
(Terminologia)

- Storage Area Network (SAN): Rede Orientada a block I/O, usada entre servidores de aplicação e storage
  - Ex. Fibre Channel
- Usualmente requer alto Bandwidth, e menos preoucupação com a Latência
  - em 2001: 1 Gbit bandwidth e latência de millisegundos (OK)
- Em geral rede dedicada
  (não é conectada a outras redes)
- Precisa trabalhar bem mesmo quando saturada
- Uma vez que trabalha com grandes blocos, pode ter taxas de bit error (BER) maiores que as LAN

SAN (Storage ou System)
(Terminologia)

- System Area Network (SAN): Uma rede usada para conectar computadores

- Em geral requer alto Bandwidth e baixa Latência.
  - em 2001: > 1 Gbit bandwidth e ~ 10 microsegundos de latência

- Pode fornecer despacho de pacotes em ordem

- Uma vez que trabalha com grandes blocos, pode ter taxas de bit error (BER) maiores que as LAN

Networks
(Terminologias)

- Conexão de 2 ou mais networks: Interconexão

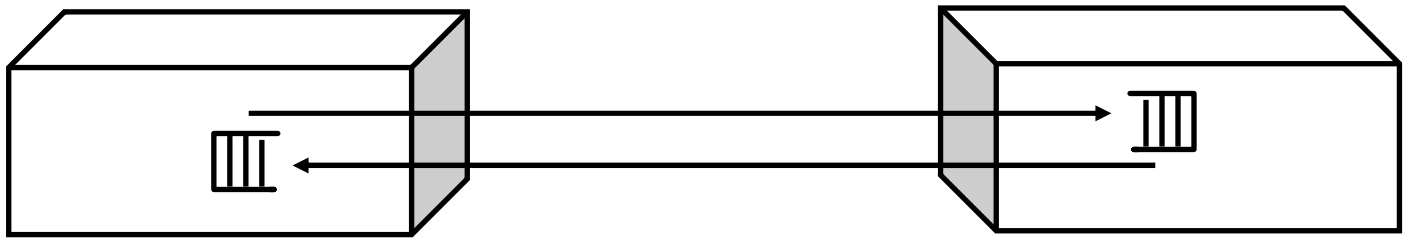
- Para 3 culturas 3 classes de networks

  - WAN: telecomunicações, Internet
  - LAN: PC, workstations, servidores
  - SAN: Clusters, RAID boxes: latency (System A.N.) ou bandwidth (Storage A.N.)

- Tentativas de terminologia única

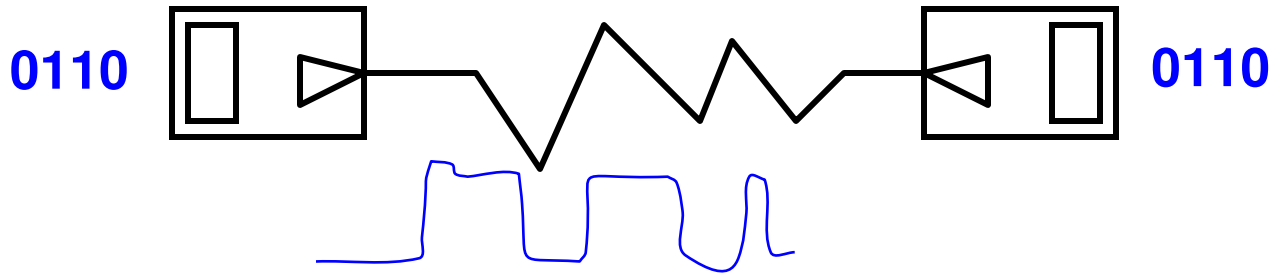
ABC de Redes

- **Ponto de Partida**: Envia bits entre 2 computadores



- Em cada extremidade uma **Fila (FIFO)**
- Informação enviada chamada de "**mensagem**"
- Pode ser enviada nos dois sentidos ("**Full Duplex**")
- Regras para a comunicação? "**protocolo**"

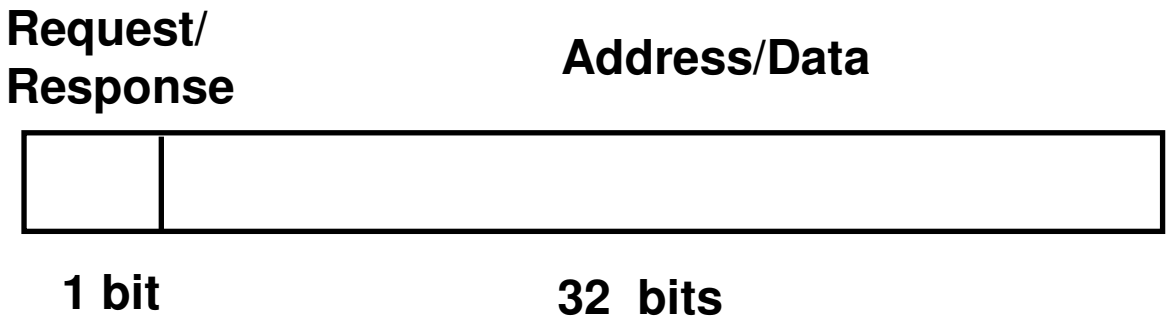
Network



- Link realizado por algum meio físico
  - fio, fibra, ar
- Com um transmitter (tx) em uma ponta
  - Converte símbolos digitais em sinais analógicos e os coloca no link
- E um receiver (rx) na outra ponta
  - Captura os sinais analógicos e os converte para digital
- tx+rx chamados de transceiver

Exemplo (simples)

- Qual o formato das mensagens?
  - Fixo/variável? Número de bits (/máximo)?



0: Request: Envia Dado a partir do Address
1: Response: Pacote contém o Data correspondente à solicitação

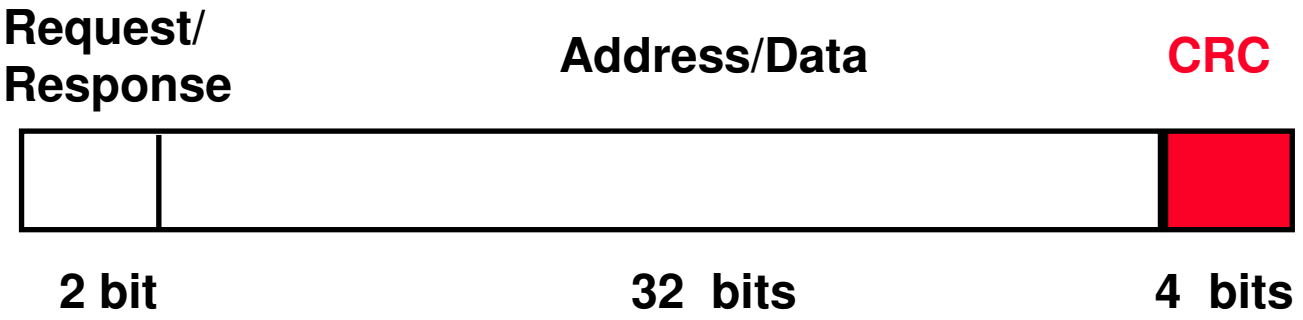
- **Header/Trailer**: informações sobre a mensagem
- **Payload**: Dado na mensagem (acima de uma palavra)

Exemplo (simples)
Questões a serem respondidas

- Se há mais do que 2 computadores querendo comunicar-se?
  - È necessário "address field" (destination) no packet
- Se o packet pode ser modificado na transmissão?
  - Adicionar "error detection field" no packet (e.g., Cyclic Redundancy Chk)
- Se o packet pode ser perdido?
  - Protocolo mais elaborado para detetar perda de pacote (e.g., NAK, ARQ, time outs)
- Se há multiplos processos/máquinas?
  - Queue por processo
- Respostas a questões simples como essas levam a protocolos e formatos para os pacotes mais complexos

Exemplo (simples)
Revisado

- Qual o formato das mensagens?
  - Fixo/variável? Número de bits (/máximo)?



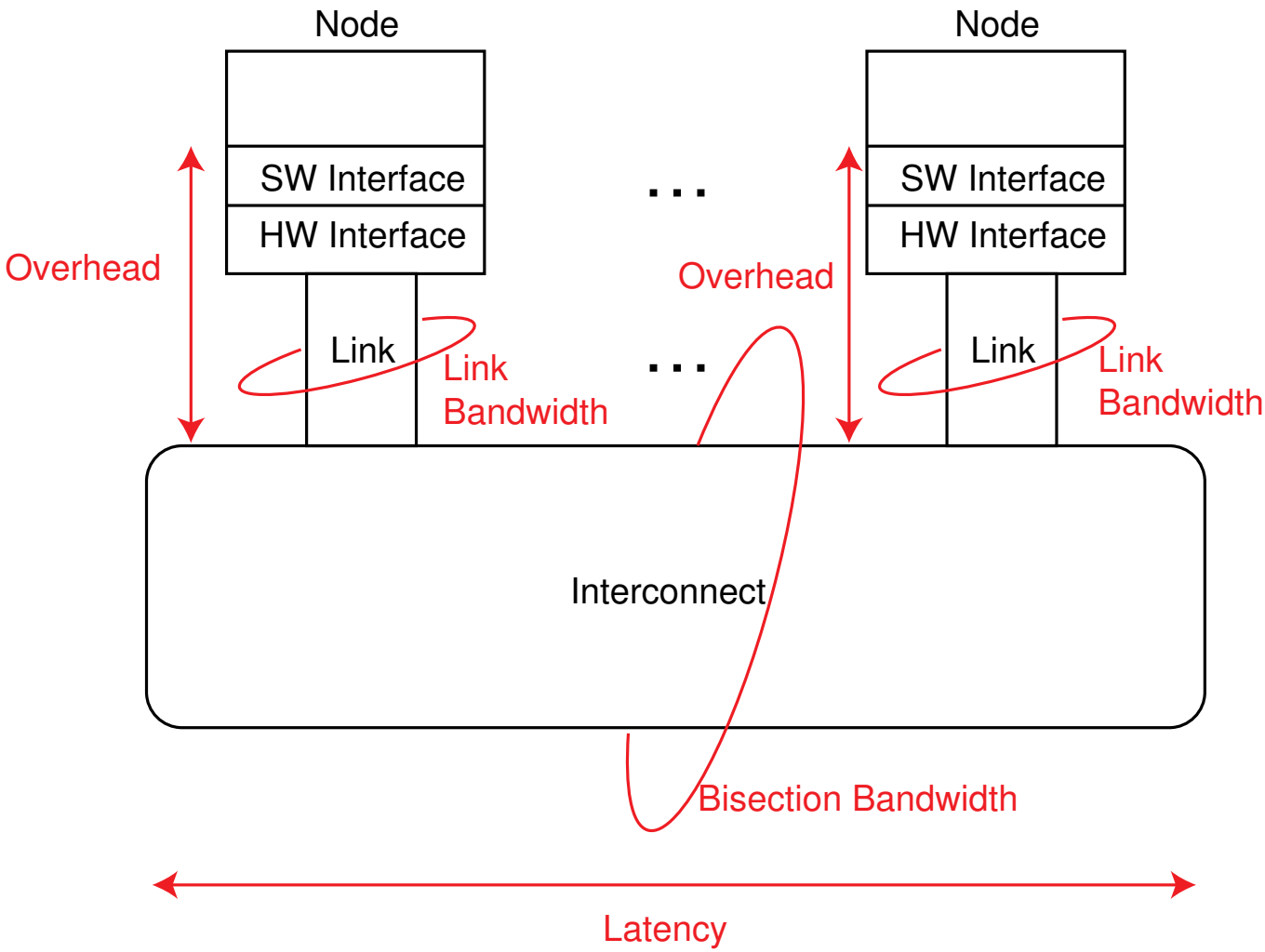
00: Request: Envia Dado a partir do Address
01: Reply— Pacote contém o Data correspondente à solicitação
10: Acknowledge request
11: Acknowledge reply

Software: Send e Recive

- **SW Send:**
  1: Aplicação copia o dado no buffer do Sistema Operacional (SO)
  2: SO cálcula o checksum, starts timer
  3: SO envia o dado para a network interface e diz start

- **SW Recive:**
  3: SO copia o dado da network interface para o buffer do SO
  2: SO calcula o checksum, se OK envia ACK; se não, *deleta a mensagem* (sender re-envia quando o timer expira)
  1: Se OK, SO copia o dado para o espaço de endereçamento do usuário e sinaliza a aplicação para continuar.

Network: Medidas de Desempenho

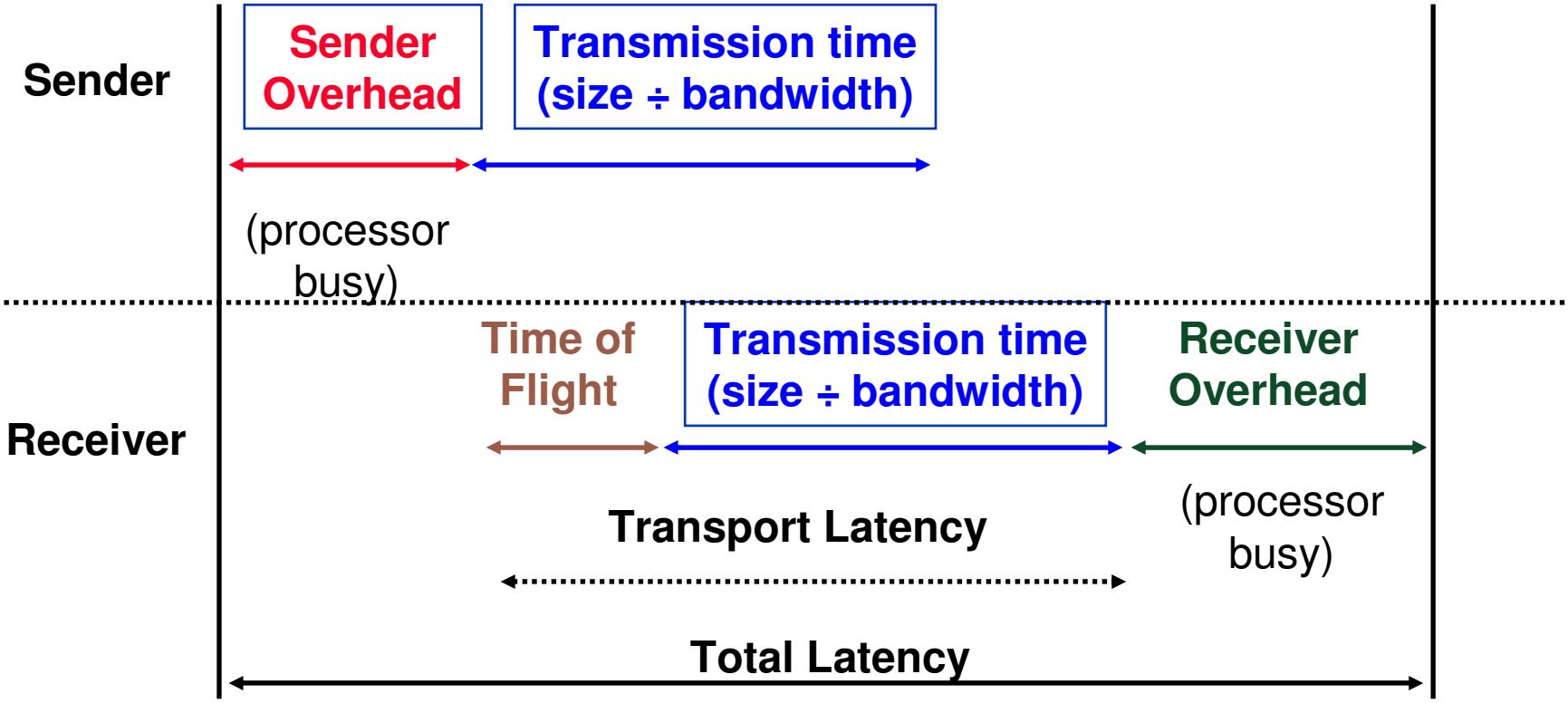


Overhead: latência da interface **Latência**: network

Terminologia

- **Bandwidth** (largura de banda): **bits/s** – taxa máxima de transmisão de "informação" pela rede
- **Time of Flight**: Tempo para um bit percorrer toda a rede de transmissão (inclui os atrasos dos dispositivos da rede – repetidor etc)
- **Transmission Time**: **tam. Mensg/BW** - tempo para toda a mensagem passar pela rede
- **Transport Latency**: Time of Flight + Transmission Time
- **Sender Overhead**: tempo para o processador enviar a mensagem
- **Reciver Overhead**: tempo para o processador receber a mensagem

Métricas de Desempenho (Universal)



Latência Total = Sender Overhead + Time of Flight + Message Size ÷ BW + Receiver Overhead

Latência Total
Exemplo

- BW = 1000 Mbit/seg.
  - sending overhead -> 80 μseg
  - receiving overhead -> 100 μseg.

- Mensagem de 10.000 bytes (incluíndo o header), (permite 10.000 bytes em uma única mensagem)

- 3 situações: distâncias de 1000 km; 0.5 km e 0.01 km
- Velocidade da luz ~ 300,000 km/seg (1/2 em média)

- $Latency_{0.01km}$ =
- $Latency_{0.5km}$ =
- $Latency_{1000km}$ =

Latência Total
Exemplo

- 1000 Mbit/seg. (send overhead -> 80 μseg; rec overhead -> 100 μseg)
- Mensagem de 10.000 byte (incluíndo o header), (permite 10.000 bytes em uma única mensagem)

- $Latência_{0.01km}$ = 80 + 0.01km / (50% x 300,000) + 10000 x 8 / 1000 + 100 = 260 μseg

- $Latency_{0.5km}$ = 80 + 0.5km / (50% x 300,000) + 10000 x 8 / 1000 + 100 = 263 μseg

- $Latency_{1000km}$ = 80 + 1000 km / (50% x 300,000) + 10000 x 8 / 1000 + 100 = 6931μseg

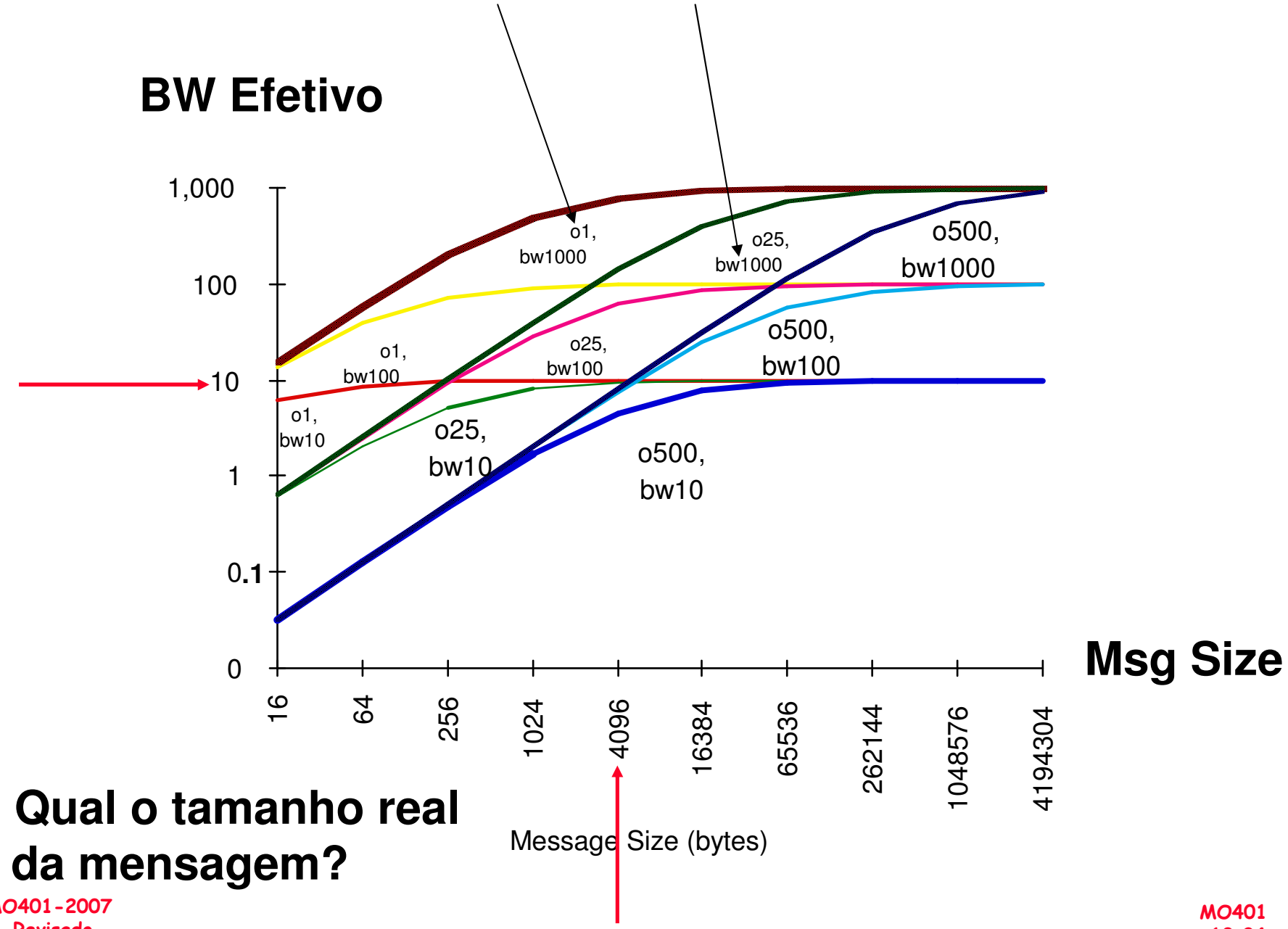
Métricas

- Aplicar recursivamente a todos os níveis do sistema
- Dentro do chip, entre chips em uma placa, entre computadores em um cluster, …
- WAN v. LAN v. SAN

Modelo Simplificado para Latência

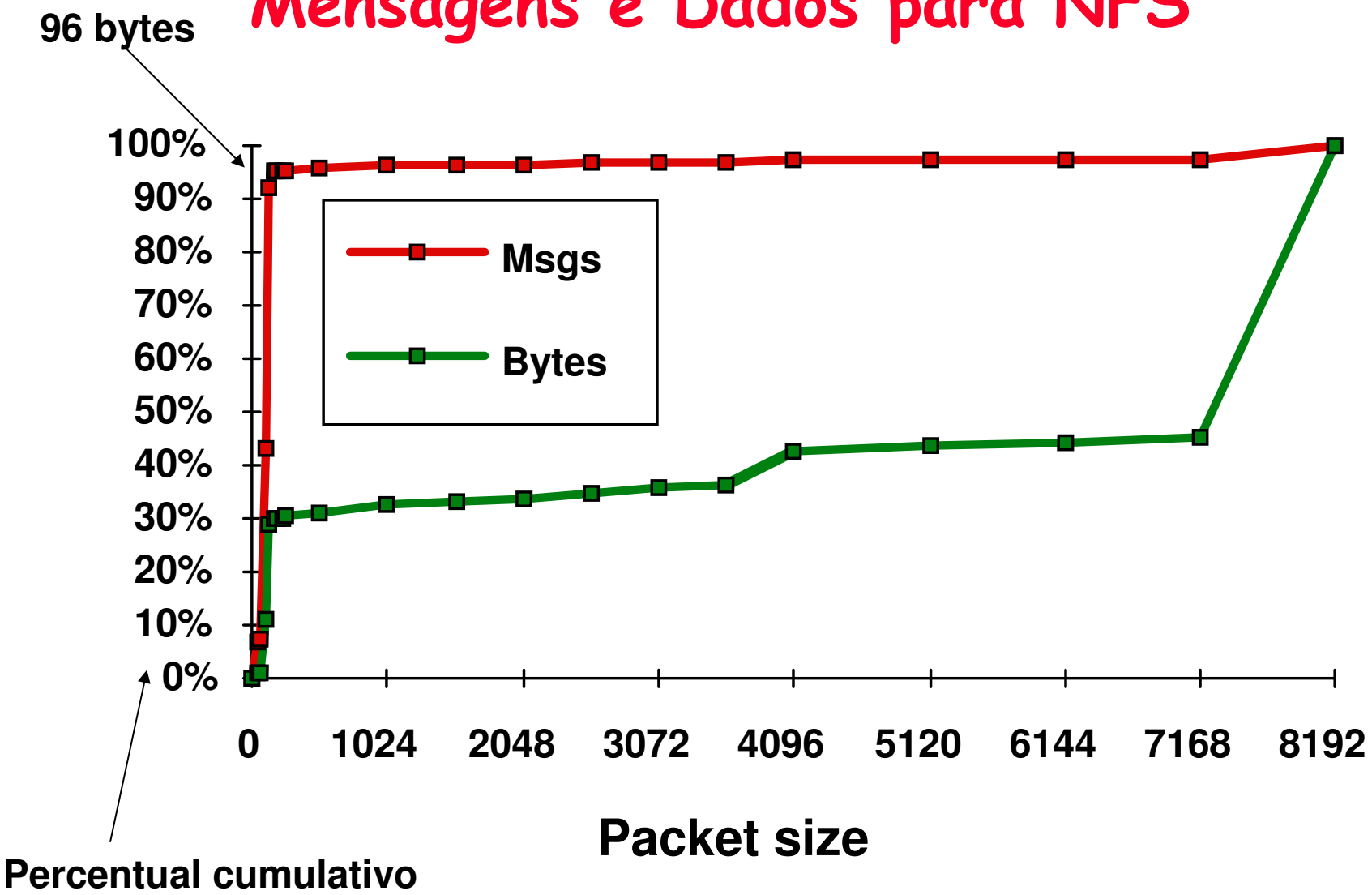
- Latência Total - Overhead + (Message Size / BW)
- Overhead = Sender Overhead + Time of Flight + Receiver Overhead
- BW Efetivo = Message Size / Latência Total
- Exemplo: (valores para vários tipos de redes)
  - Overhead: 1, 25, 500 *µseg*
  - BW: 10, 100, 1000 Mbit/seg (fator de 10)
  - Message Size: 16 Bytes a 4 MB
- Se overhead 500 *µseg*,
  qual o tamanho da mensagem para > 10 Mb/s?

Overhead, BW, Size



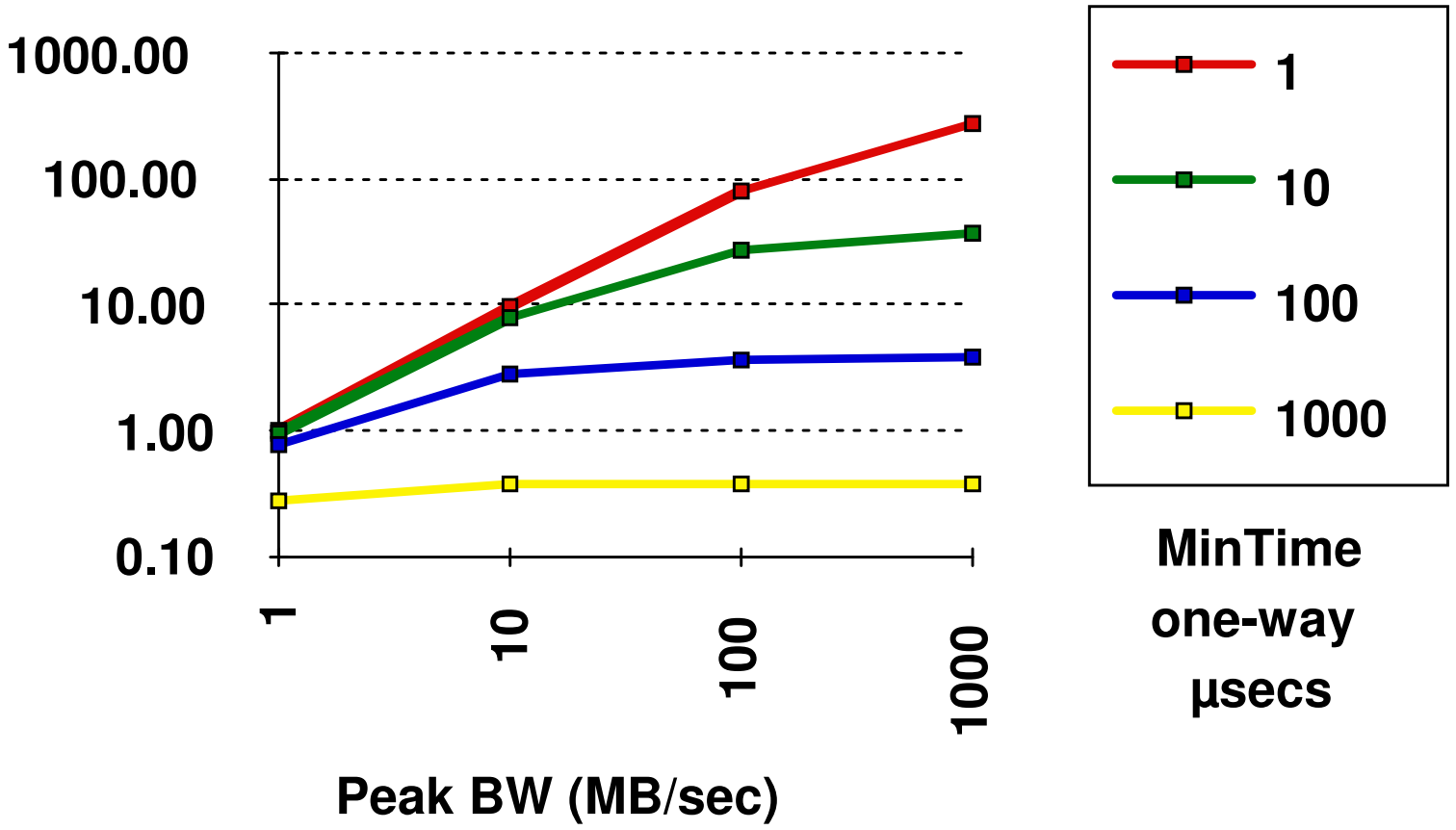
- **Qual o tamanho real da mensagem?**

Mensagens e Dados para NFS



- 95% Msgs, 30% bytes por packets ~ 200 bytes
- > 50% data transfered em packets = 8KB

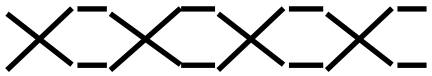
Impacto do Overhead no BW Efetivo



- Modelo BW: Time = overhead + msg size/peak BW

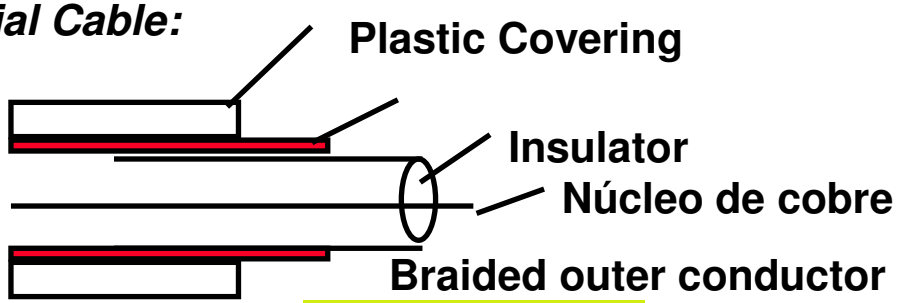
Network Media

Twisted Pair:



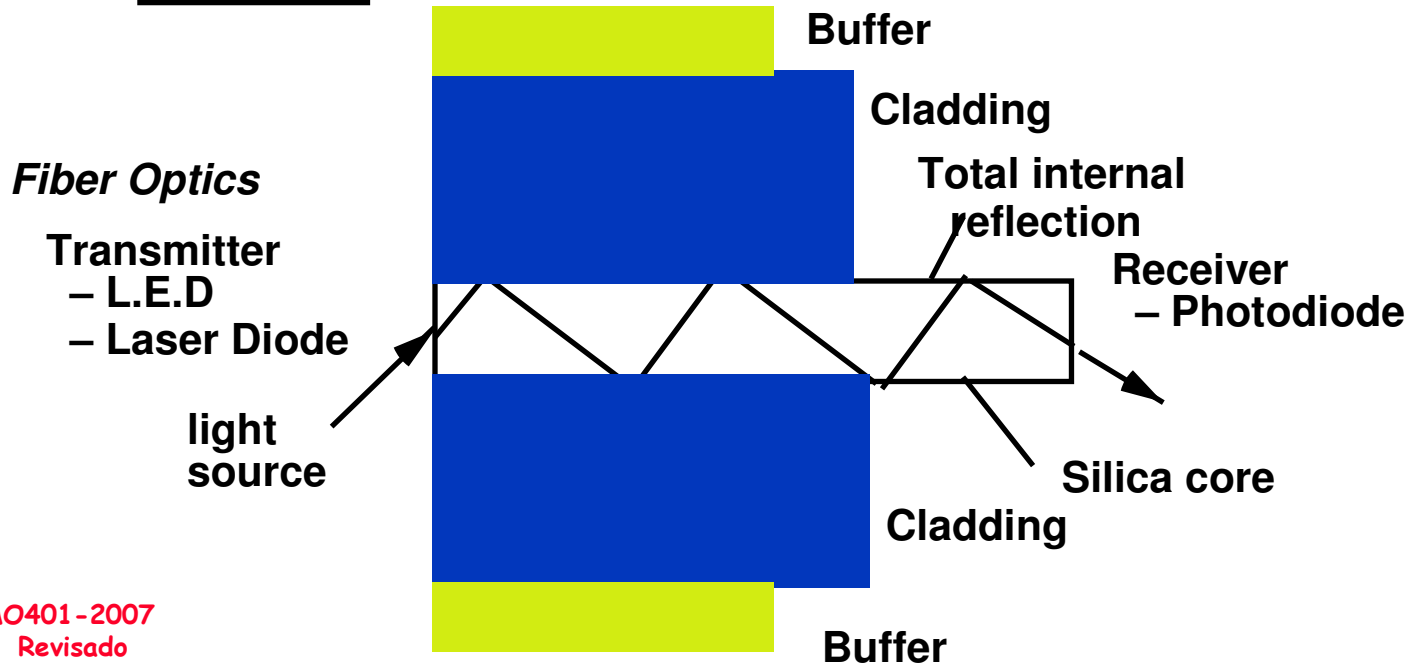
Cobre, 1mm, trançado para evitar efeito antena.
"Cat 5" -> 100 M bits/seg (100m)

Coaxial Cable:



Alto BW, boa imunidade a ruídos

Fiber Optics



Unidirecional; (2 para full duplex)

Fibra

- **Fibra Multimodo**: ~ 62.5 micron de diâmetro; comprimento de onda de 1.3 micron (luz infra vermelha). Devido a problemas de disperção, limitada a 1000 Mbits/s para 0.1 km, e 1-3 km a 100 Mbits/s. Usa LED como fonte de luz

- **Fibra Monomodo**: fibra "single-wavelength" (8-9 microns) usa **laser diodes**, 1-5 Gbits/s para 100´s kms
  - Mais cara, mais restrições quanto a curvaturas
  - Custo, bandwidth e distância afetados pela potência da fonte de luz.
  - Tipicamente fibra de vidro, já que tem melhores características do que fibras de plástico (mais baratas)

Fibra
Multiplexação de Ondas

- Envia N independentes streams em uma única fibra
- Usa-se diferentes Comprimentos de ondas no envio e demultiplexa no recebimento

- Em 2000: 40 Gbit/s usando 8 comprimentos de ondas

- Evolução: 10X/4 anos, ou 1.8X por ano

Comparação entre Medias

- Assuma 40 discos de 2.5"; 25 GB; mover 1 km
- Compare: Cat 5 (100 Mbit/s), fibra Multimodo (1000 Mbit/s), monomodo (2500 Mbit/s), e carro (50 kph)

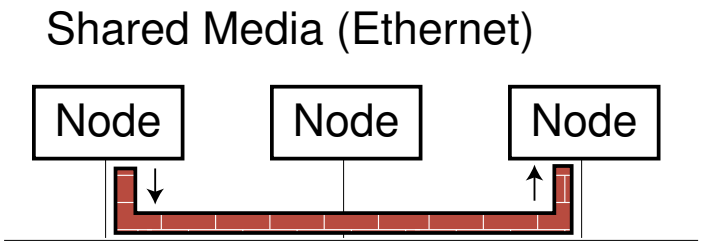
- 40 discos => 40 x 25 = 1000 GB

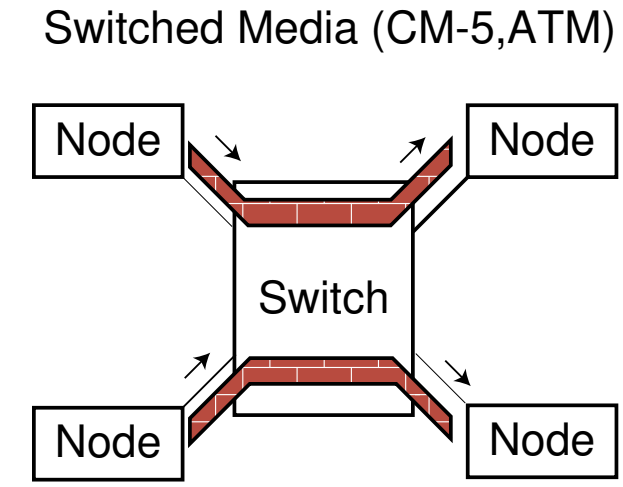
- Cat 5: 1000 x 1024 x 8 Mb / 100 Mb/s = 23 hrs
- Mult: 1000 x 1024 x 8 Mb / 1000 Mb/s = 2.3 hrs
- Mono: 1000 x 1024 x 8 Mb / 2500 Mb/s = 0.9 hrs
- Carro: 5 min + 1 km / 50 kph + 10 min = 0.27 hrs

- Carro-Disk= media de melhor BW

Conectando Múltiplos Computadores

- Media Compartilhada vs. Switched: o par comunica-se ao mesmo tempo: conexão "ponto-a-ponto"
- A BW agregada na switched network é melhor que na compartilhada
  - ponto-a-ponto rápida (não há arbitragem, interface simples)
- Arbitragem em Shared network?
  - Arbitro Central para LAN?
  - Listen to check if being used ("Carrier Sensing")
  - Listen to check if collision ("Collision Detection")
  - Reenvio Randômico para evitar colisões repetidas; arbitragem não amigável;
  - OK se a utilização é baixa





Connection-Based vs. Connectionless

Connection-Based ("circuit switching")

- Ex.: Telefone: operador realiza a conexão entre quem chama (caller) e quem recebe (receiver)
  - Uma vez estabelecida a conexão a conversação pode continuar por horas.
- Compartilham linhas de transmisão longas usando switches para multiplexar várias conversações na mesma linha
  - "Multiplexação por divisão de tempo" divide BW da linha de transmisão em um número fixo de slots, com cada slot associado a uma conversação
  - "Multiplexação por divisão de freqüências" divide o BW da linha em um número fixo de freqüências e cada comunicação é realizada em uma freqüência
- Vantagem: reserva de bandwidth
- Desvantagem: Não é adequada à comunicação de dados. A ocupação da linha é baseada no número de conversções e não pela quantidade de informações enviadas.

Connection-Based vs. Connectionless

Connectionless ("packet switching")

- Cada pacote é roteado da origem ao destino
  - Todo pacote de informação deve ter um endereço

  => packets

  - Cada pacote é roteado até seu destino com base em seu endereço
  - Analogia, sistema postal (envio de uma carta)
  - Também chamado de "Statistical multiplexing"

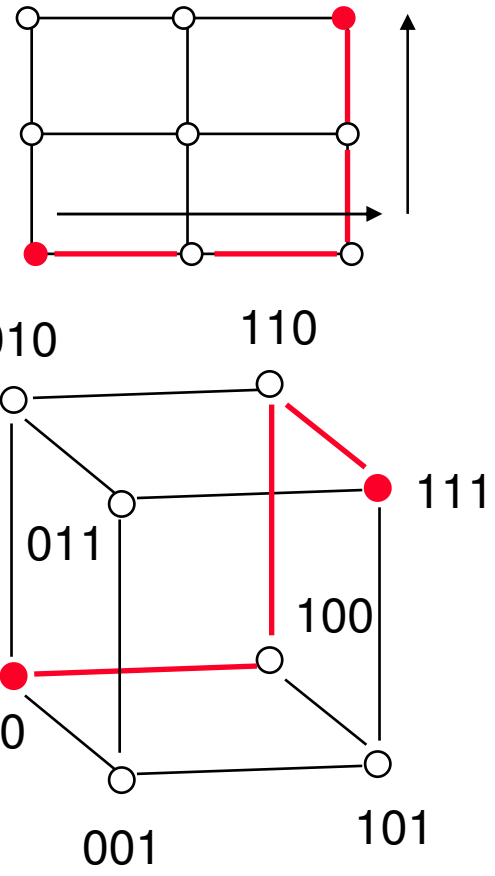
Roteamento de Mensagens

- Media Compartilhada
  - Broadcast para todos

- Switched Media – necessita de roteamento. Opções:
  - Source-based routing: a mensagem especifica o caminho até o destino
  - Virtual Circuit: um circuito é estabelecido da fonte até o destino, a mensagem segue esse caminho
  - Destination-based routing: a mensagem especifica o destino, switches informam o caminho
    - » deterministic: sempre seguem o mesmo caminho
    - » adaptive: usam diferentes caminhos evitando congestionamentos, falhas, …
    - » Randomized routing: usam vários caminhos entre os melhores objetivando o balanceamento dacarga na rede

Roteamento Deterministico: exemplo

- mesh: dimension-order routing
  - (x1, y1) -> (x2, y2)
  - first Δx = x2 - x1,
  - then Δy = y2 - y1,
- hypercube: edge-cube routing
  - X = xox1x2 . . .xn -> Y = yoy1y2 . . .yn
  - R = X xor Y
  - Traverse dimensions of differing address in order
- tree: ancestral comum
- Deadlock free?



Store and Forward vs. Cut-Through

- Política "Store-and-forward": cada switch espera que todo o pacote chegue na switch antes de envia-lo para a próxima switch (bom para WAN)

- "Cut-through routing" ou "worm hole routing": a switch examina o header do pacote, decide para onde enviar, e inicia seu envio imediatamente
  - worm hole routing, quando o head da mensagem é bloqueado, a mensagem fica bloqueada na rede, potencialmente bloqueia outras mensagens (requer um buffer pequeno).
  - Cut through routing deixa o final continuar quando o head é bloqueado (Requer um buffer grande, suficiente para guardar o maior pacote).

Cut-Through vs. Store and Forward

- Vantagem
  - A Latência reduz em função do:

    número de switches intermediárias X tamanho dos pacotes

    para

    tempo para a 1ª parte do pacote negociar as switches + (tamanho do pacote ÷ BW da conexão)

Controle de Congestionamento

- **Packet switched networks:** não reserva bandwidth; isto leva à ***contenção***

- Solução: previnir a entrada de pacotes até que a contenção seja reduzida

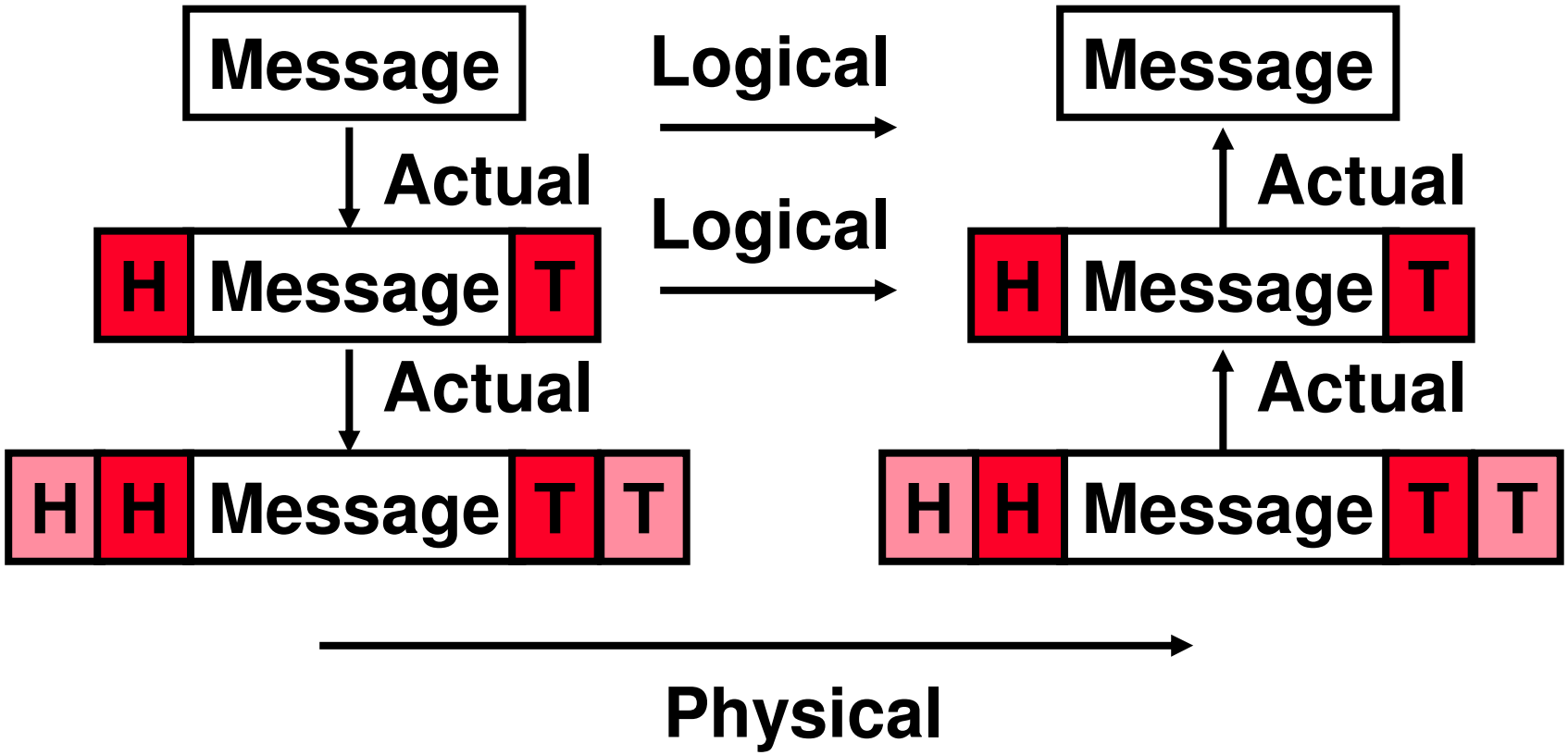
Controle de Congestionamento

- **Opções:**
  - Packet discarding: Se o pacote chega à switch e não há espaço no buffer, o pacote é descartado (ex. UDP)
  - Flow control: entre pares de receivers e senders; usa feedback para avisar quando sender pode enviar o próximo pacote
    - » Back-pressure: fios separados de controle para stop
    - » Window: o sender tem direito de enviar N pacotes antes de pegar permissão para enviar mais; sobrepõe latências de interconexão com overhead para enviar e receber pacotes (ex., TCP), window ajustável.
  - Choke packets: Cada pacote recebido por uma switch no estado de warning é reenviado ao source via choke packet. O Source reduz o tráfego ao destino de um % fixo (ex., ATM)

Protocolos: Interface HW/SW

- Internetworking: permite que computadores em redes independentes e incompátiveis comuniquem-se de forma confiável e eficientemente;
  - Tecnologia disponível: SW padrões que permitem comunicação confiável usando-se redes não confiáveis
  - Camadas hierarquicas de SW: cada camada tem responsbilidade por uma porção das tarefas de toda a comunicação, denominado fámilia de protocolos ou protocol suites
- Transmission Control Protocol/Internet Protocol (TCP/IP) – padrão Internetworking mais popular
  - Está fámilia de protocolos é a base da Internet
  - IP faz o esforço para fazer a entrega; TCP garante a entrega
  - TCP/IP é usado mesmo em comunicações locais : NFS usa IP nas comunicações em LAN homogeneas

Fámilia de Protocolos: Conceitos

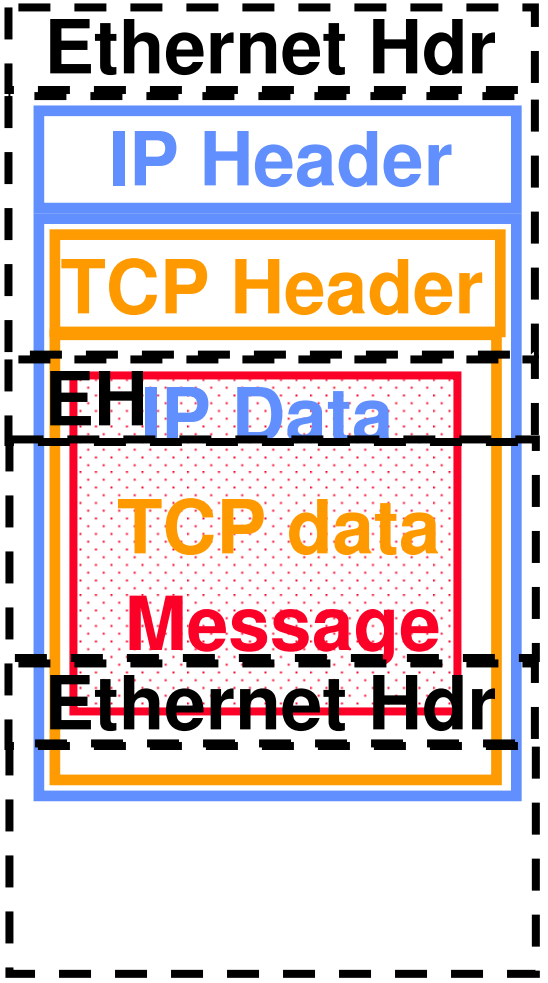


Fámilia de Protocolos: Conceitos

- Em uma fámilia de protocolos a comunicação ocorre de forma lógica no mesmo nível do protocolo, denominado peer-to-peer,
- E é implementado via serviços do nível inferior
- Encapsulamento: às mensagens são adicionadas/ retiradas informações relativas ao nível ("envelope")
- Fragmentação: pacotes são quebrados em múltiplos pacotes menores e remontados

Pacote TCP/IP, pacote Ethernet, Protocolos

- Aplicação envia mensagens
- TCP quebra em segmentos de 64KB, adiciona 20B de header
- IP adiciona 20B de header, envia para a rede
- Se Ethernet, quebra em pacotes de 1500B com headers, trailers (24B)
- Todos os Headers, trailers têm campo de tamanho, destino, ...



Network: Exemplo

- Ethernet: media compartilhada; 10 Mbit/s proposta em 1978, carrier sensing com deteção de colisão (exponencial)
- 15 anos sem melhorias; BW alta?
- Múltiplas Ethernets com dispositivos permitindo que as Ehternets operem em paralelo
- Sucessores da 10 Mbit Ethernet?
  - FDDI: media compartilhada
  - ATM
  - Switched Ethernet
  - 100 Mbit Ethernet (Fast Ethernet)
  - Gigabit Ethernet
  - 10 Gigabit Ethernet em 2002

Conectando Networks

- **Bridges**: conecta LANs, passa tráfego de um lado para o outro dependendo do endereço do pacote.
  - opera no nível do protocolo Ethernet
  - usualmente mais simples e barato que roteadores
- **Routers ou Gateways**: conectam LANs a WANs ou WANs a WANs e resolve incompatibilidade de endereçamento.
  - Geralmente mais lentos que as bridges, operam no nível de protocolo de internetworking (IP)
  - Routers dividem a interconexão em subnets pequenas e separadas
  - Cisco é o maior fornecedor;
    básicamente são computadores de proposito especial

Comparendo Networks

	SAN		LAN			WAN
	FC-AL	Infini-band	10 Mb Ethernet	100 Mb Ethernet	1000 Mb Ethernet	ATM
Length (meters)	30/1000	17/100	500/2500	200	100	
Data lines	2	1, 4, 12	1	1	4/1	1
Clock (MHz)	1000	2500	10	100	1000	155/ 622
Switch?	Opt.	Yes	Optional	Opt.	Yes	Yes
Nodes	<=127	~1000	<=254	<=254	<=254	~10000
Material	Copper / fiber	Copper /fiber	Copper	Copper	Copper /fiber	Copper /fiber

Comparando Networks

	SAN		LAN			WAN
	FC-AL	Infini-band	10 Mb Ethernet	100 Mb Ethernet	1000 Mb Ethernet	ATM
Switch?	Opt.	Yes	Optional	Opt.	Yes	Yes
Bisection BW (Mbits /sec)	800 shared or 800 x switch ports	(2000 - 24000) x switch ports	10 shared or 10 x switch ports	100 shared or 100 x switch ports	1000 x switch ports	155 x switch ports
Peak link BW(Mbits /sec)	800	2000, 8000, 24000	10	100	1000	155/ 622
Topology	Ring or Star	Star	Line or Star	Line or Star	Star	Star

Comparando Networks

	SAN		LAN			WAN
	FC-AL	Infiniband	10 Mb Ethernet	100 Mb Ethernet	1000 Mb Ethernet	ATM
Connec-tionless?	Yes	Yes	Yes	Yes	Yes	No
Store & forward?	No	No	No	No	No	Yes
Conges-tion control	Credit-based	Back-pressure	Carrier sense	Carrier sense	Carrier sense	Credit based
Standard	ANSI Task Group X3T11	Infiniband Trade Associa-tion	IEEE 802.3	IEEE 802.3	IEEE 802.3 ab-1999	ATM Forum

Wireless

- A Media pode ser o ar tão bem quanto cobre e vidro
- Ondas de Rádio são ondas eletromagnéticas propagadas por uma antena
- Ondas de Radio são moduladas
- Ondas de Radio têm um comprimento de onda ou freqüência: medidas por seu tamanho ou por MHz
  - Ondas longas => baixas freqüências,
  - Ondas curtas => altas freqüências

- As freqüências podem ser ajustadas a diferentes valores
  - Estações FM transmitem na banda de 88 MHz a 108 MHz usando modulação em freqüência (FM)

Wireless

- **Wireless** => móvel => a network deve se rearranjar dinamicamente
- **Power**
  - Os dispositivos tendem a serem alimentados por baterias
  - antenas irradiam potência para a comunicação e só uma pequena parcela chega ao receptor
- **bit error rates** (BER) depende da potência do sinal recebido, nível de ruído (diversas fontes) => muito mais freqüente que em transmissão no cobre

Wireless: 2 Arquiteturas

- *Estação Base*
  - Conectadas por linhas de longa distâncias e as unidades móveis conectam-se à estação base.
  - Exemplo: telefone celular

- *Peer-to-peer*

Telefonia Celular

- Usa a mesma freqüência em locais diferentes (sem interferência)
- Divide as regiões em células hexagonais sem sobreposições, que usam diferentes freqüências se vizinhas, reusando em celuas distantes (sem interferência)
- Interseção de três células hexagonais formam uma estação base com transmisores e antenas

Telefonia Celular

- **Original**
  - Sinal análogico com freqüências diferentes paras as direções
  - 869.04 a 893.97 MHz, (*forward path*)
  - 824.04 MHz a 848.97 MHz, (*reverse path)*

- ***Sucessores***
  - *Code division multiple access* (*CDMA*)
  - *time division multiple access* (*TDMA*)
  - *global system for mobile communication* (*GSM*)
  - *International Mobile Telephony* 2000 (IMT-2000)

Aspectos Práticos em Redes

- Conectividade: número máximo de máquinas que afetam a complexidade da rede e os protocolos

- Conexão da Interface de Rede com o computador
  - Em qual posição na hierarquia de barramentos?
    - » Memory bus?
    - » Fast I/O bus?
    - » Slow I/O bus?
    - » (Ethernet -> Fast I/O bus, Inifiband -> Memory bus já que ele é o Fast I/O bus)
  - Interface de SW: é necessário "flush caches" para consistência para sends ou receives?
  - Programmed I/O vs. DMA?

Aspectos Práticos em Redes

- Vantagens de padronizações :
  - Baixo custo (muitos fornecedores, alta escala de produção)
  - estabilidade (muitos fornecedores, competição)
- Desvantagens da padronização:
  - Tempo para os comitês chegarem a um acordo
  - Quando padronizar?
    - » Antes de se produzir? => O comitê já tem uma decisão?
    - » Se muito cedo, pode inibir inovações
- Confiabilidade (disponibilidade) da interconexão

Aspectos Práticos em Redes

Interconexão	SAN	LAN	WAN
Exemplo	Inifiband	Ethernet	ATM
Standard	sim	sim	sim
Tolerância a Falhas?	sim	sim	sim
Hot Insert?	sim	sim	sim

- Tolerância a Falhas (Fault Tolerance): nós falham e mensagens continuam circulando entre os outros nós?

- Hot Insert: Se a interconexão sobrevive a uma falha, ela pode continuar operando enquando o nó que falhou é substituído?