PTC 3450 - Aula 17

3.7 Controle de congestionamento no TCP

4.1 Introdução à camada de rede

4.2 O que há dentro de um roteador?

(Kurose, p. 205 - 241)

(Peterson, p. 103-171)

30/05/2016

Capítulo 3: conteúdo

3.1 serviços da camada de transporte

3.2 multiplexação e desmultiplexação

3.3 transporte sem conexão: UDP

3.4 princípios da transferência de dados confiável

3.5 transporte orientado para conexão: TCP

- estrutura dos segmentos
- transferência de dados confiável
- controle de fluxo
- gerenciamento de conexão

3.6 princípios do controle de congestionamento

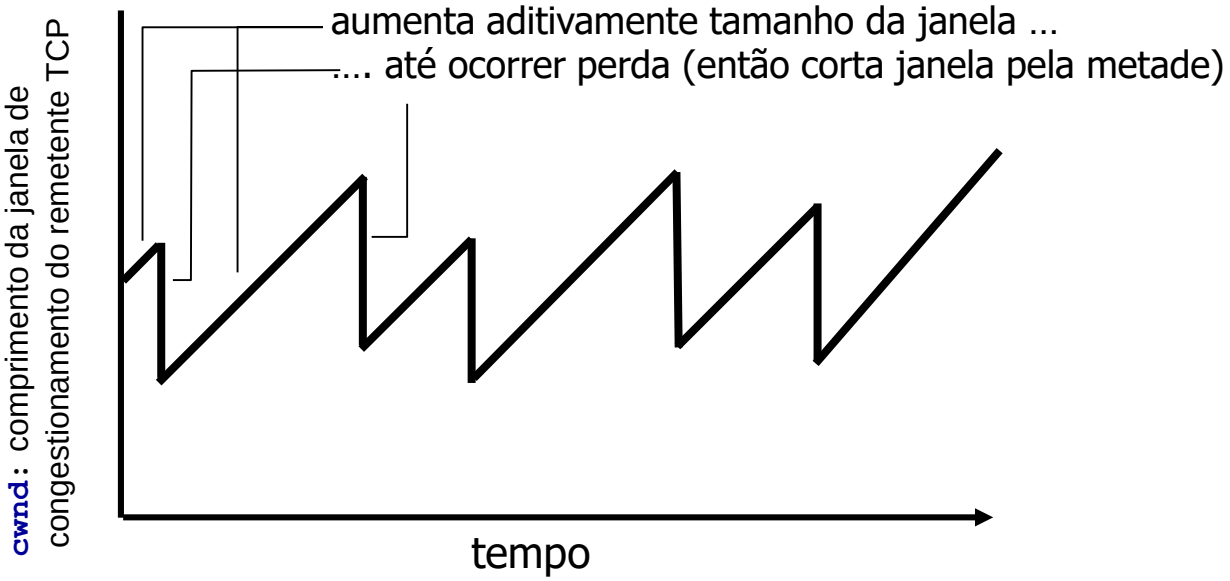
3.7 controle de congestionamento no TCP

TCP: controle de congestionamento

incremento aditivo, decremento multiplicativo (AIMD)

- *Controle fim a fim:* IP não fornece realimentação
- *abordagem:* remetente aumenta taxa de transmissão (compr. janela), sondando por capacidade utilizável, até ocorrer perda [Jacobson, 1988; RFC 5681 - 2009]
  - *aumento aditivo:* aumenta **`cwnd`** em 1 MSS a cada ACK válido (tudo vai bem! ☺) até que perda é detectada (indício de congestionamento ☹)
  - *diminuição multiplicativa:* corta **`cwnd`** pela metade após perda

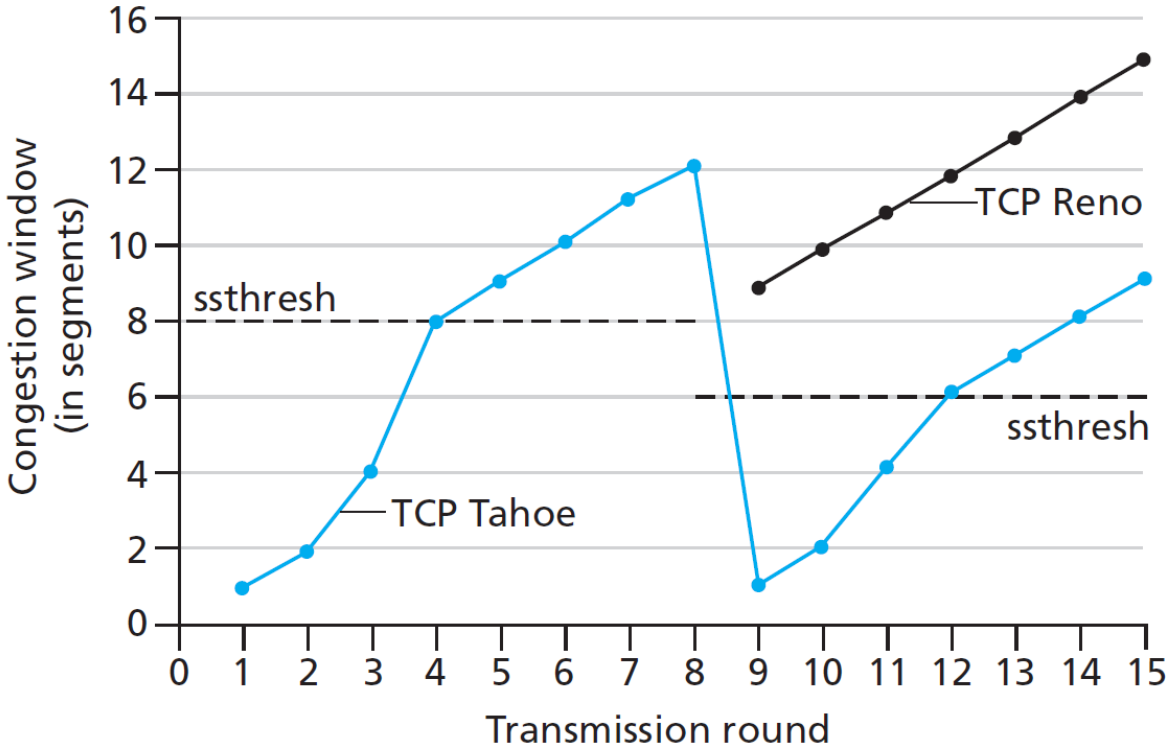
comportamento dente de serra do AIMD: sondando por capacidade



TCP: mudando de partida lenta para CA

Q: quando o aumento exponencial deve mudar para linear?

R: quando **cwnd** chega a 1/2 do seu valor antes do *timeout*.



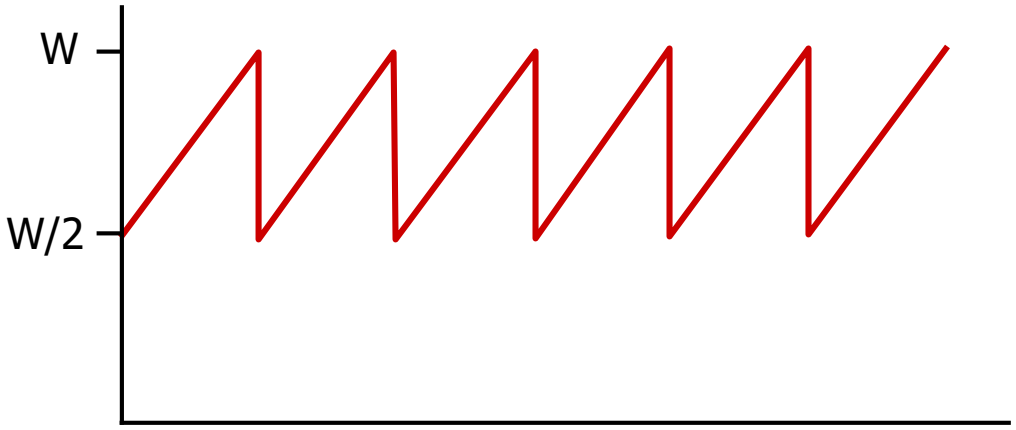
Implementação:

- variável **ssthresh** (slow start threshold)
- em evento de perda, **ssthresh** é ajustada para 1/2 de **cwnd** logo antes do evento de perda

TCP: vazão

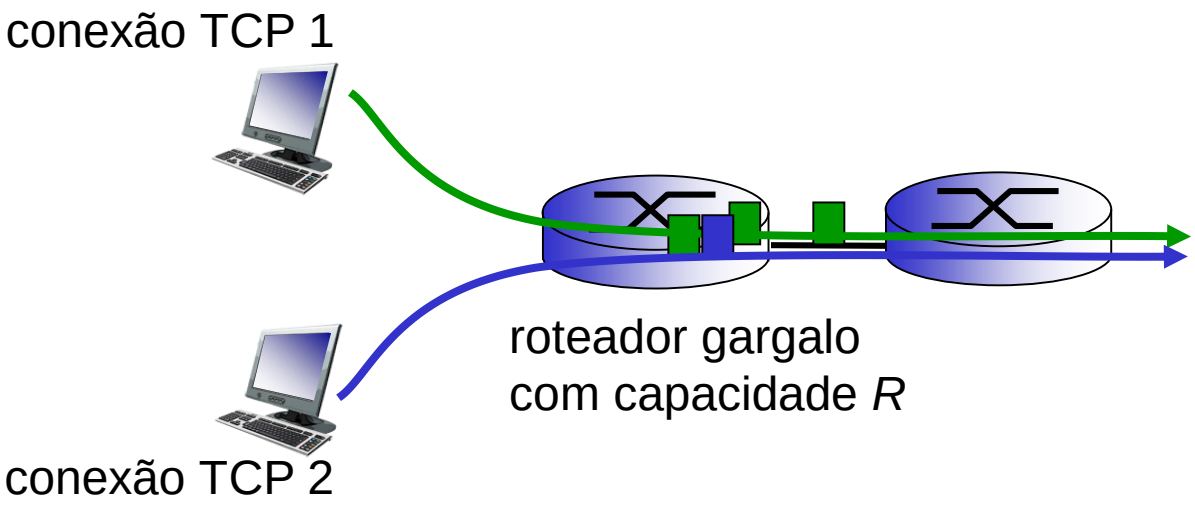
- vazão TCP média em função do comprimento da janela, RTT?
  - hipóteses: ignorar partida lenta (TCP sai dela exponencialmente rápido), assumir que sempre existe dados para enviar, comprimento da janela em que ocorre perda é constante
- W: comprimento da janela (medido em bytes) em que ocorre perda
  - tamanho médio da janela (número de bytes *in-flight*) é ¾ W
  - vazão média é 3/4W por RTT

$$\text{vazão média TCP} = \frac{3}{4}\,\frac{W}{RTT}\ \text{bytes/s}$$



TCP: Equidade (*Fairness*)

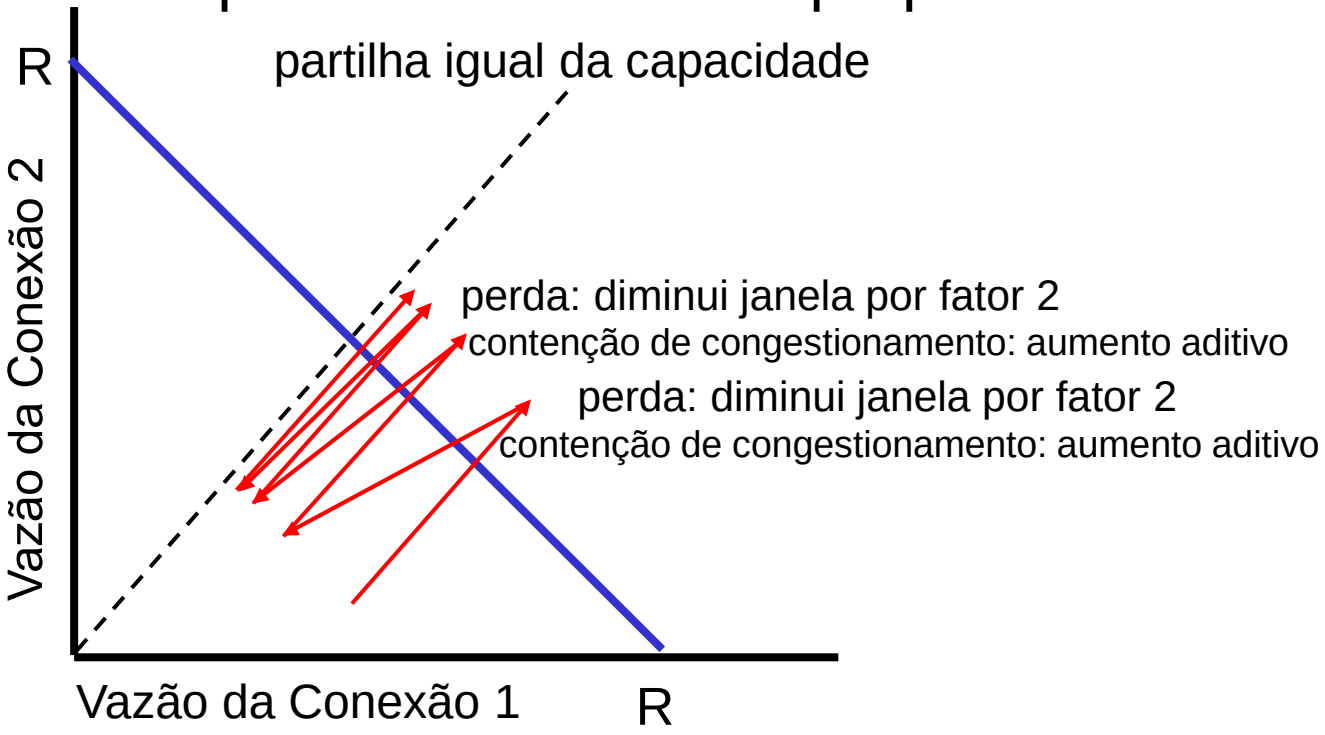
objetivo equidade: se *K* sessões TCP partilham o mesmo enlace gargalo de capacidade *R*, cada um deve ter taxa média de *R*/*K*



Por que o TCP é justo (*fair*)?

2 sessões competindo:

- aumento aditivo resulta em inclinação 1, enquanto vazão aumenta
- diminuição multiplicativa diminui vazão proporcionalmente



Equidade (problemas...)

Equidade e UDP

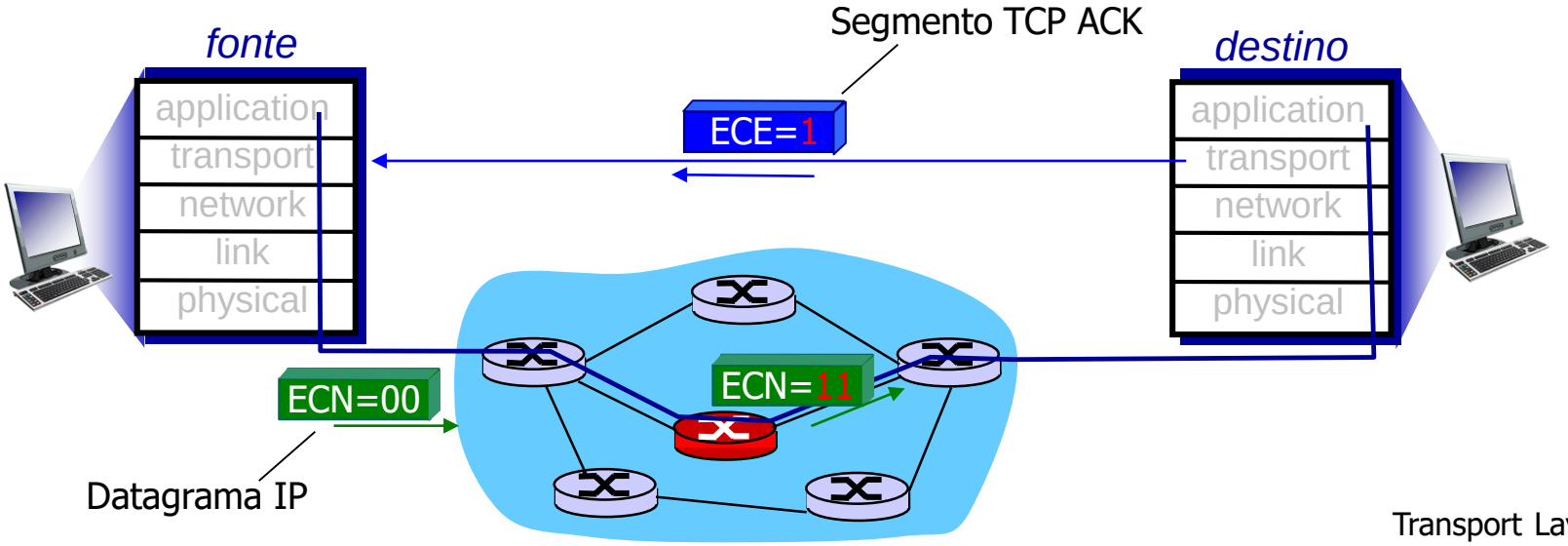
- *apps* multimídia muitas vezes não usam TCP
  - não querem taxa reduzida devido a controle de congestionamento
- ao invés, usam UDP:
  - enviar áudio/vídeo a uma taxa constante, tolerante a perda de pacotes
- Como tráfego TCP reduz sua taxa em congestionamento e UDP não, UDP pode expulsar tráfego TCP!
- Resolver esse problema é tema de pesquisa!

Equidade, conexões paralelas TCP

- aplicativos podem abrir múltiplas conexões paralelas entre 2 *hosts*
- navegadores *web* fazem isso
- e.g., enlace de taxa *R* com 9 conexões existentes:
  - novo *app* pede 1 TCP, ganha taxa R/10
  - novo *app* pede 11 TCPs, leva mais do que R/2

Explicit Congestion Notification (ECN)

- Desde a década de 1980 TCP implementou controle de congestionament fim a fim
- Mais recentemente: *Controle de congestionamento assistido pela rede [RFC 3168]:*
- 2 bits no cabeçalho IP (campo ToS) marcado *por roteador da rede* para indicar congestionamento
- Indicador de congestionamento chega ao *host* receptor
- Receptor (vendo o indicador de congestionamento no datagrama IP) liga o bit ECE (*Explicit Congestion Notification Echo*) no segment ACK do receptor-para-transmissor para notificar o transmissor de congestionamento



Capítulo 3: resumo

- princípios por trás dos serviços da camada de transporte:
  - multiplexação, desmultiplexação
  - transferência de dados confiável
  - controle de fluxo
  - controle de congestionamento
- exemplificação, implementação na Internet
  - UDP
  - TCP
- Outros protocolos estão em estudo: DCCP (*Datagram Congestion Control Protocol* – *RFC 4340*), SCTP (Stream Control Transmission Protocol – RFC 4960), ...

a seguir:

- deixamos a "borda da rede" (camadas de aplicação e transporte)
- vamos para o "núcleo" da rede

Capítulo 4
Camada de Rede: o plano de dados

Capítulo 4: conteúdo

4.1 Introdução à camada de rede
- Plano de dados
- Plano de controle

4.2 o que tem dentro de um roteador?

4.3 IP: *Internet Protocol*
- formato do datagrama
- fragmentação
- endereçamento IPv4
- NAT
- IPv6

4.4 Repasse generalizado e SDN
- Casamento
- Ação
- Exemplos OpenFlow de casamento-mais-ação em andamento

Capítulo 4: camada de rede

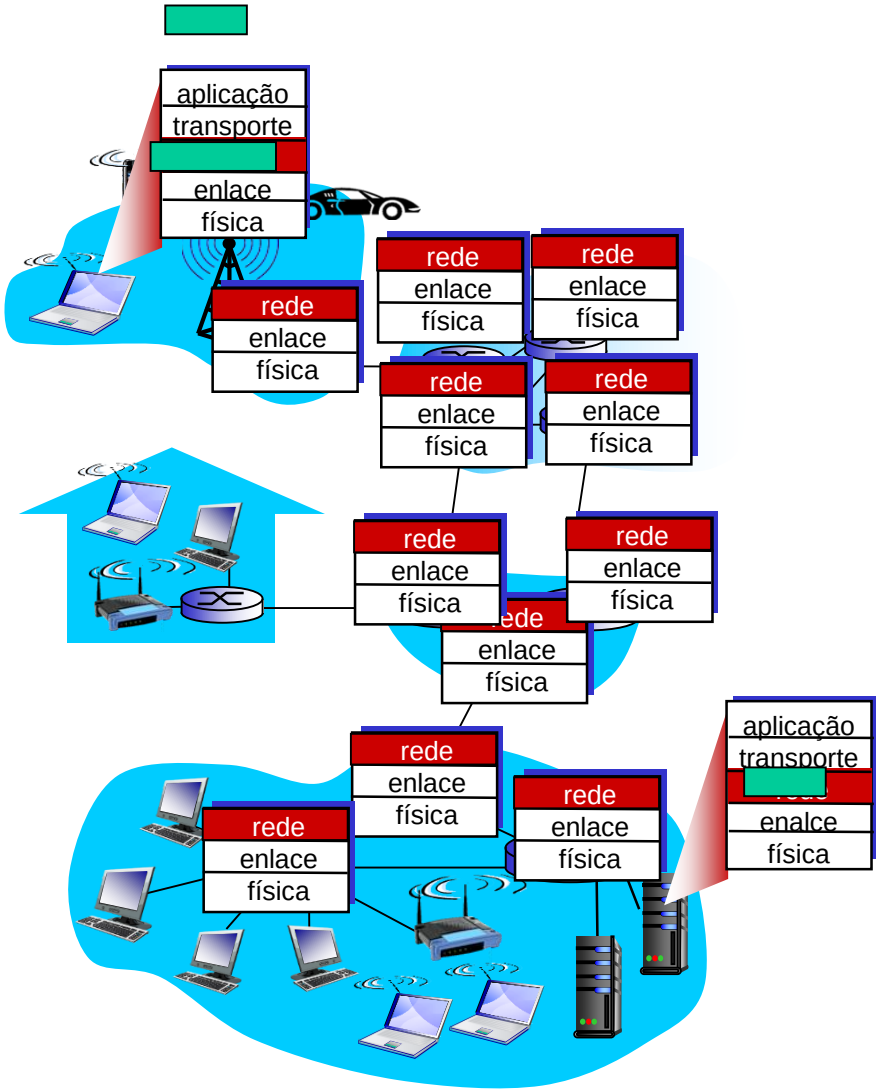
objetivos do capítulo:

- entender princípios subjacentes aos serviços da camada de rede, focando no plano de dados
  - modelos de serviço da camada de rede
  - repasse *(forwarding)* versus roteamento
  - como um roteador funciona
  - Repasse generalizado

- exemplo prático, implementação na Internet

Camada de rede

- transporta segmentos do *host* fonte para o *host* destino
- no lado fonte encapsula segmentos em datagramas
- no lado recebendo, entrega segmentos para camada de transporte
- protocolos da camada de rede em *todo host*, roteador
- roteador examina campos de cabeçalho em todos os datagramas IP que passam por ele



Duas funções chaves da camada de rede

- *repasse (forwarding):* mover pacotes da entrada do roteador para saída apropriada (local)
  - *Hardware – escala de nanosegundos*
- *roteamento:* determinar rota tomada por pacotes da fonte ao destino (global)
  - *Software – escala de segundos*
  - *algoritmos de roteamento*

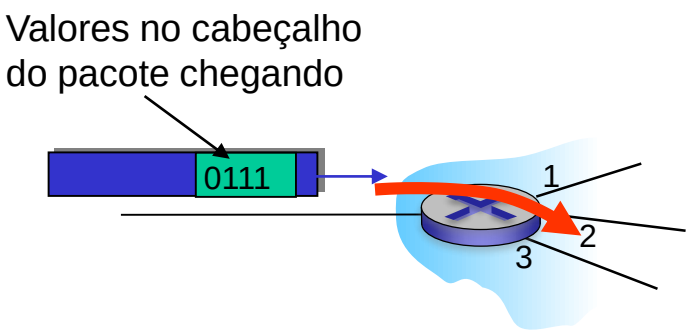
analogia:

- *roteamento:* processo de planejar viagem da origem ao destino
- *repasse:* processo de passar por uma única encruzilhada

Camada de rede: plano de dados e plano de controle

Plano de Dados

- função local, por cada roteador
- Determina como um datagrama chegando por uma porta de entrada do roteador é repassado para uma porta de saída
- função de repasse, basicamente

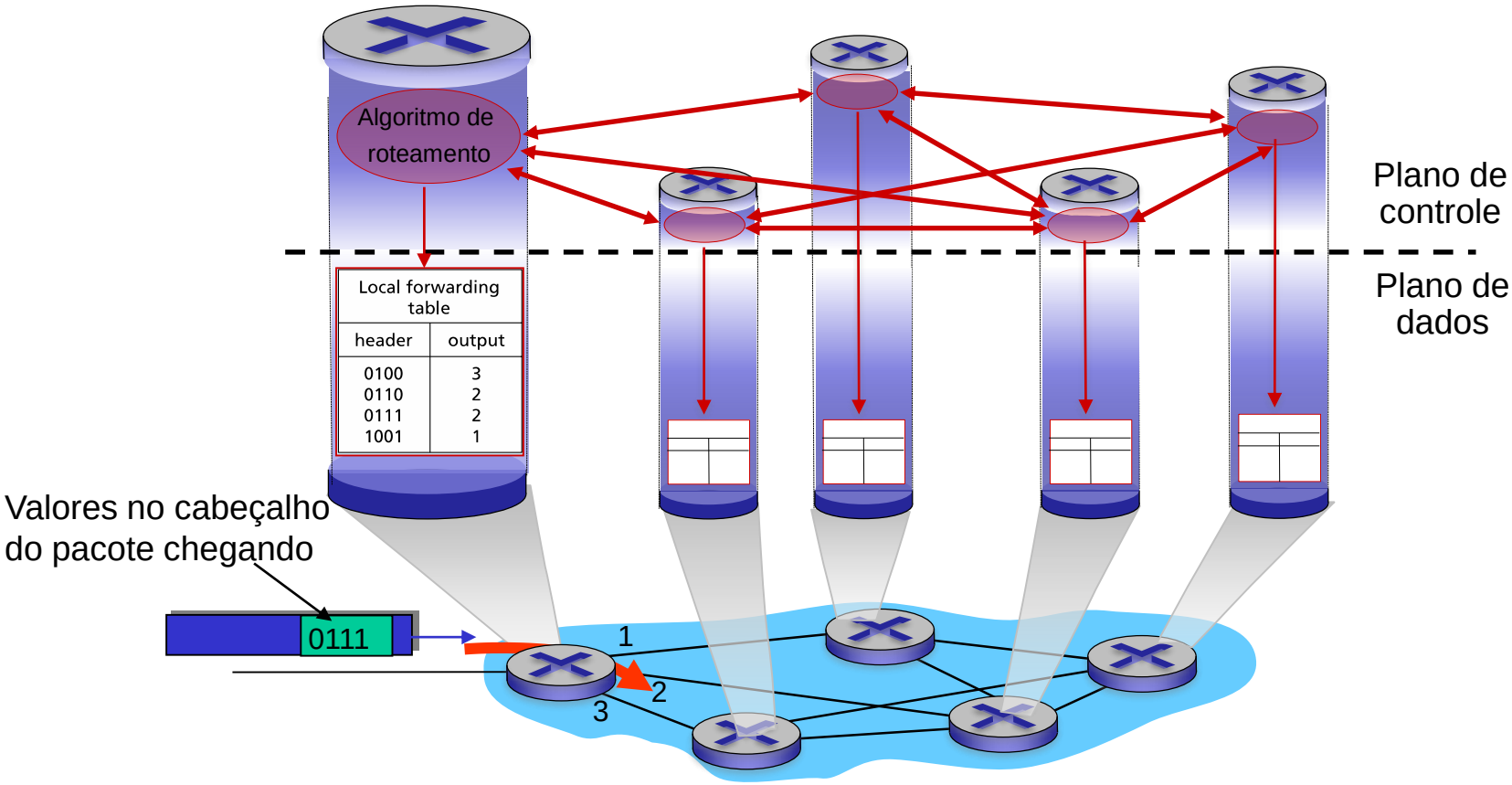


Plano de Controle

- Lógica para rede inteira
- Determina como datagrama são roetados entre roteadores ao longo do caminho fim-a-fim do *host* fonte até o *host* destino
- 2 abordagens para plano de controle:
  - *Algoritmos de roteamento tradicionais:* implementados nos roteadores
  - *Software-Defined Networking (SDN)*: implementados em servidores (remotos)

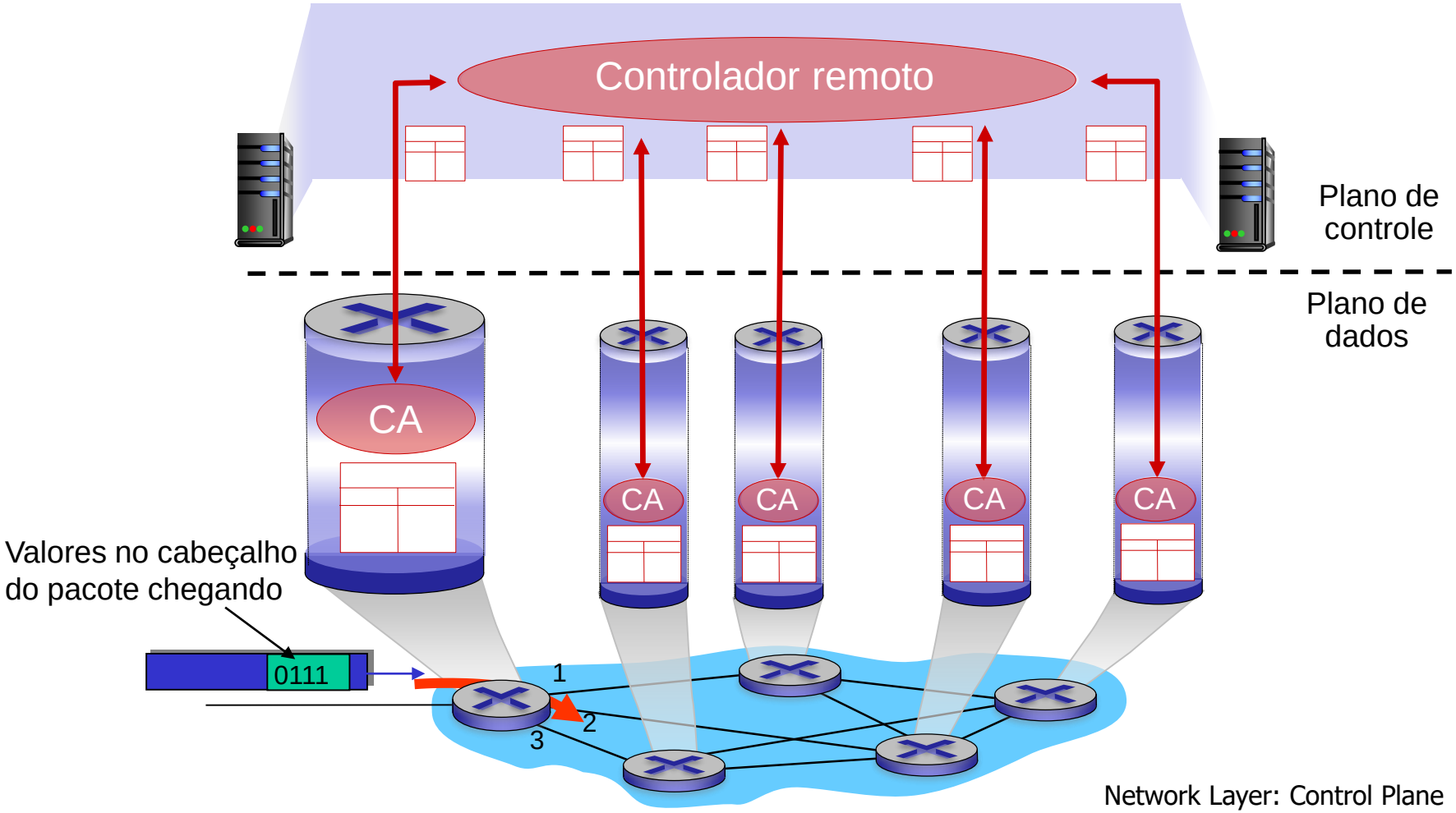
Plano de controle por roteador

Componentes do algoritmo de roteamento individual em *cada roteador* interagem no plano de controle



Plano de Controle centralizado logicamente

Um controlador distinto (tipicamente remote) interage com os agentes controle locais (*Control Agents – CA*)



Modelos de serviços de rede

Q: Qual *modelo de serviço* para "canal" transportando datagramas do remetente para destinatário?

exemplos de serviços para datagramas individuais:

- entrega garantida
- entrega garantida com atraso *(timing)* menor do que 40 ms

exemplos de serviços para um fluxo de datagramas:

- entrega dos datagramas em ordem
- taxa mínima garantida para o fluxo
- restrições nas mudanças de espaçamento entre pacotes (*jitter*)
- segurança

Modelos de serviços de rede

Arquitetura de rede	Modelo de serviço	Garantias?				Indicação de congestionamento
		Taxa	Perdas	Ordem	*Timing*	
Internet	Melhor esforço	nehuma	não	não	não	não (inferido via perdas)
ATM	CBR	taxa constante	sim	sim	sim	não há congestionamento
ATM	VBR	taxa garantida	sim	sim	sim	não há congestionamento
ATM	ABR	mínima garantida	não	sim	não	sim
ATM	UBR	nenhuma	não	sim	não	não

Capítulo 4: conteúdo

4.1 Introdução à camada de rede
- Plano de dados
- Plano de controle

4.2 o que tem dentro de um roteador?

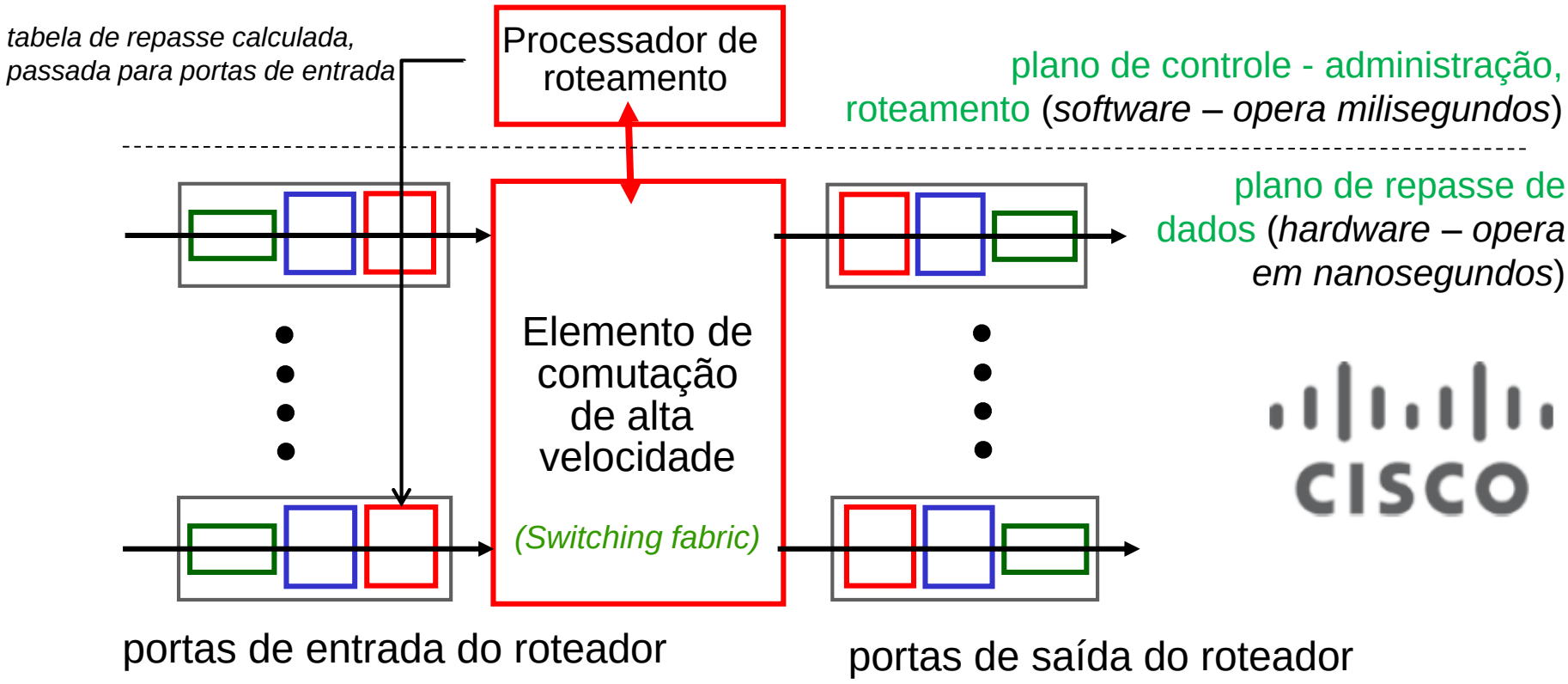
4.3 IP: *Internet Protocol*
- formato do datagrama
- fragmentação
- endereçamento IPv4
- NAT
- IPv6

4.4 Repasse generalizado e SDN
- Casamento
- Ação
- Exemplos OpenFlow de casamento-mais-ação em andamento

Arquitetura de roteadores (CISCO/Huawei/Alcatel-Lucent/Juniper)

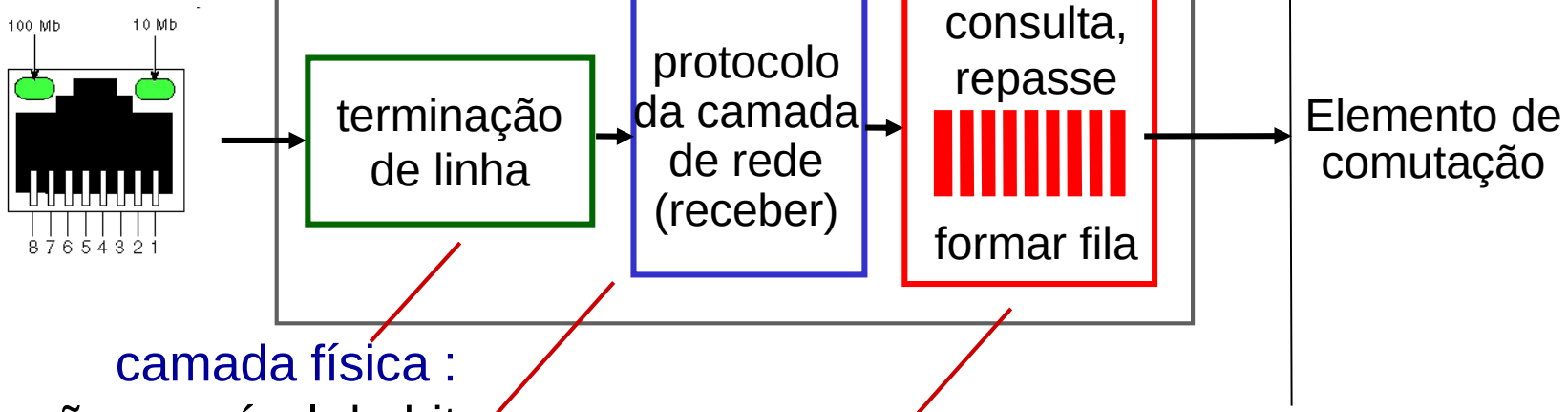
2 funções chaves do roteador:

- rodar protocolos e algoritmos de roteamento (RIP, OSPF, BGP)
- *repassar ou comutar* datagramas de enlace de entrada para enlace de saída



Exemplo: Arquitetura da Família Cisco MDS 9000

Funções da porta de entrada



camada física :
recepção em nível de bit

camada de enlace de dados:
e.g., Ethernet
veja cap. 6

comutação descentralizada:

- dados dest. do datagrama, consulta porta de saída usando tabela de repasse na memória da porta de entrada *("match plus action" – casamento mais ação)*
- objetivo: completar processamento da porta de entrada em "velocidade de linha" (ns)
- fila: se datagramas chegam mais rápido do que taxa de repasse para elemento de comutação

Exemplo: Arquitetura da Família Cisco MDS 9000

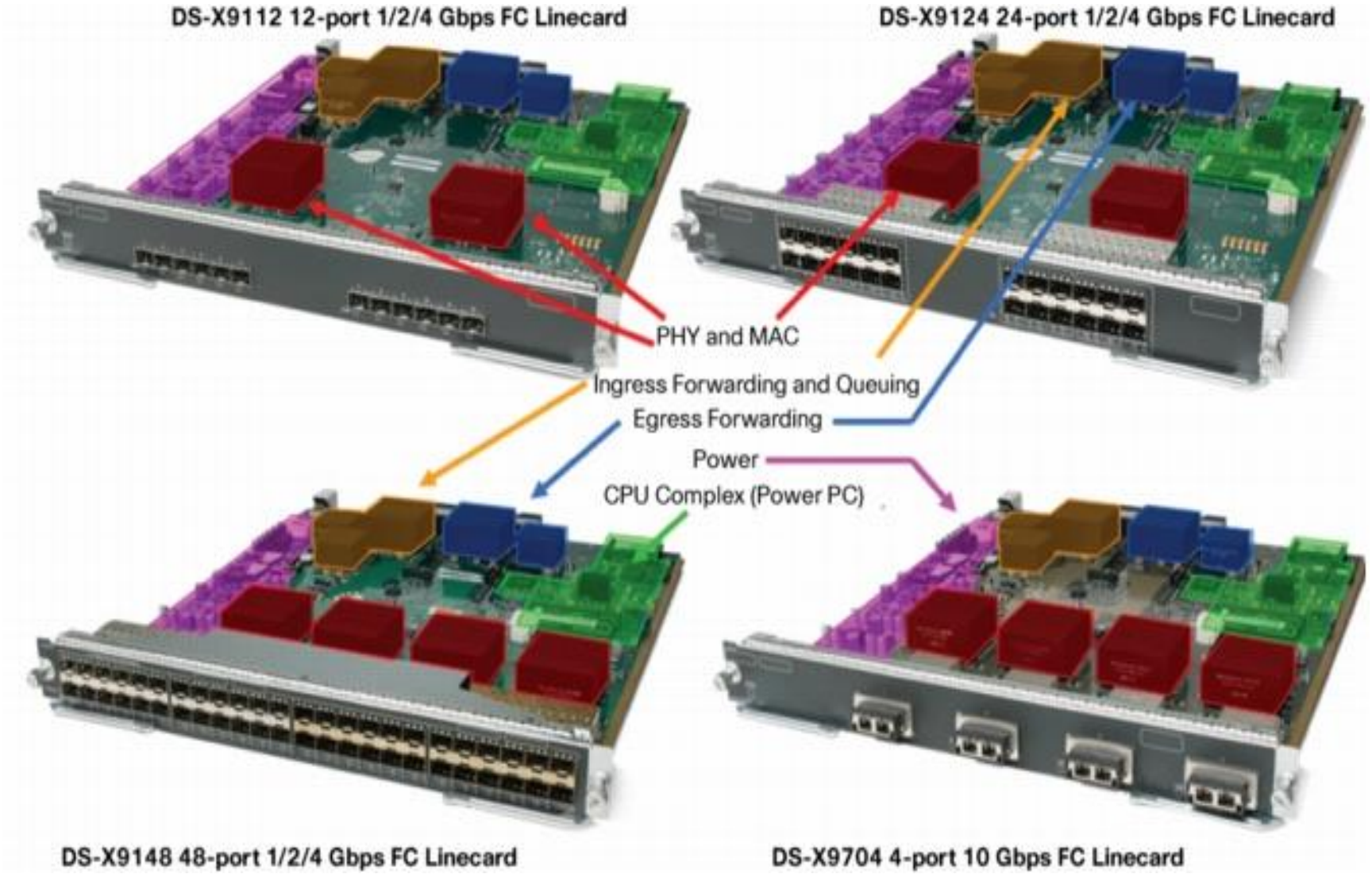


Tabela de repasse de datagramas

Intervalo do endereço de destino	Interface do Enlace
11001000 00010111 00010000 00000000 até 11001000 00010111 00010111 11111111	0
11001000 00010111 00011000 00000000 até 11001000 00010111 00011000 11111111	1
11001000 00010111 00011001 00000000 até 11001000 00010111 00011111 11111111	2
caso contrário	3

Q: mas o que acontece se intervalos não dividem-se assim tão agradavelmente?

Casamento com prefixo mais longo

casamento com prefixo mais longo

quando busca-se entrada de tabela de repasse para dado endereço de destino (DA), usar prefixo de endereço *mais longo* que casa com endereço de destino.

Intervalo de Endereços de Destino	Interface do enlace
11001000 00010111 00010*** *********	0
11001000 00010111 00011000 *********	1
11001000 00010111 00011*** *********	2
caso contrário	3

exemplos:

DA: 11001000 00010111 00010110 10100001 qual interface?

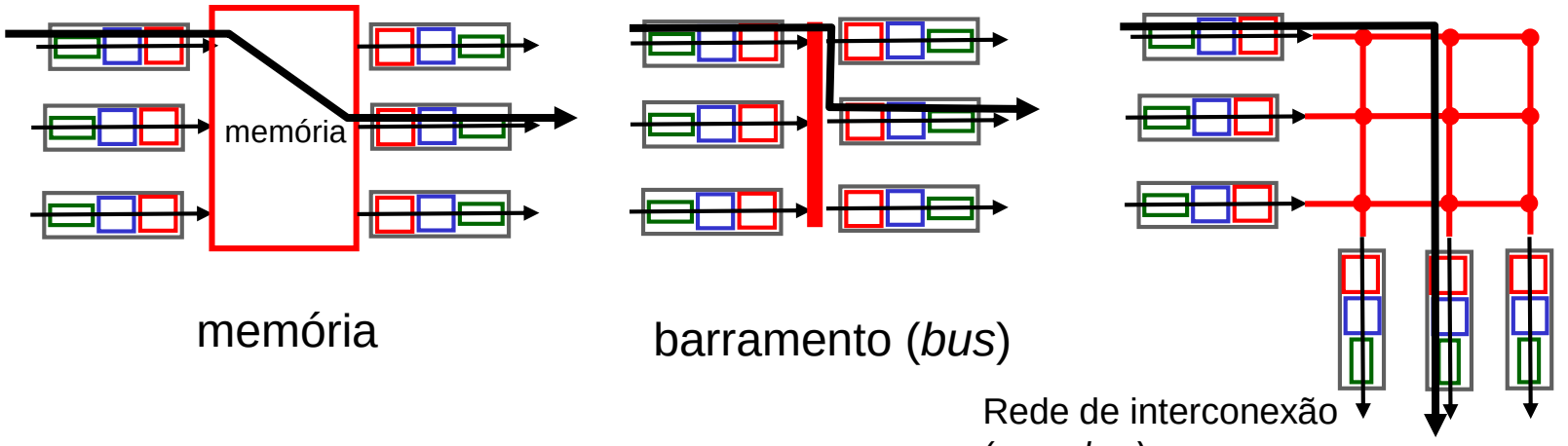
DA: 11001000 00010111 00011000 10101010 qual interface?

Casamento com prefixo mais longo

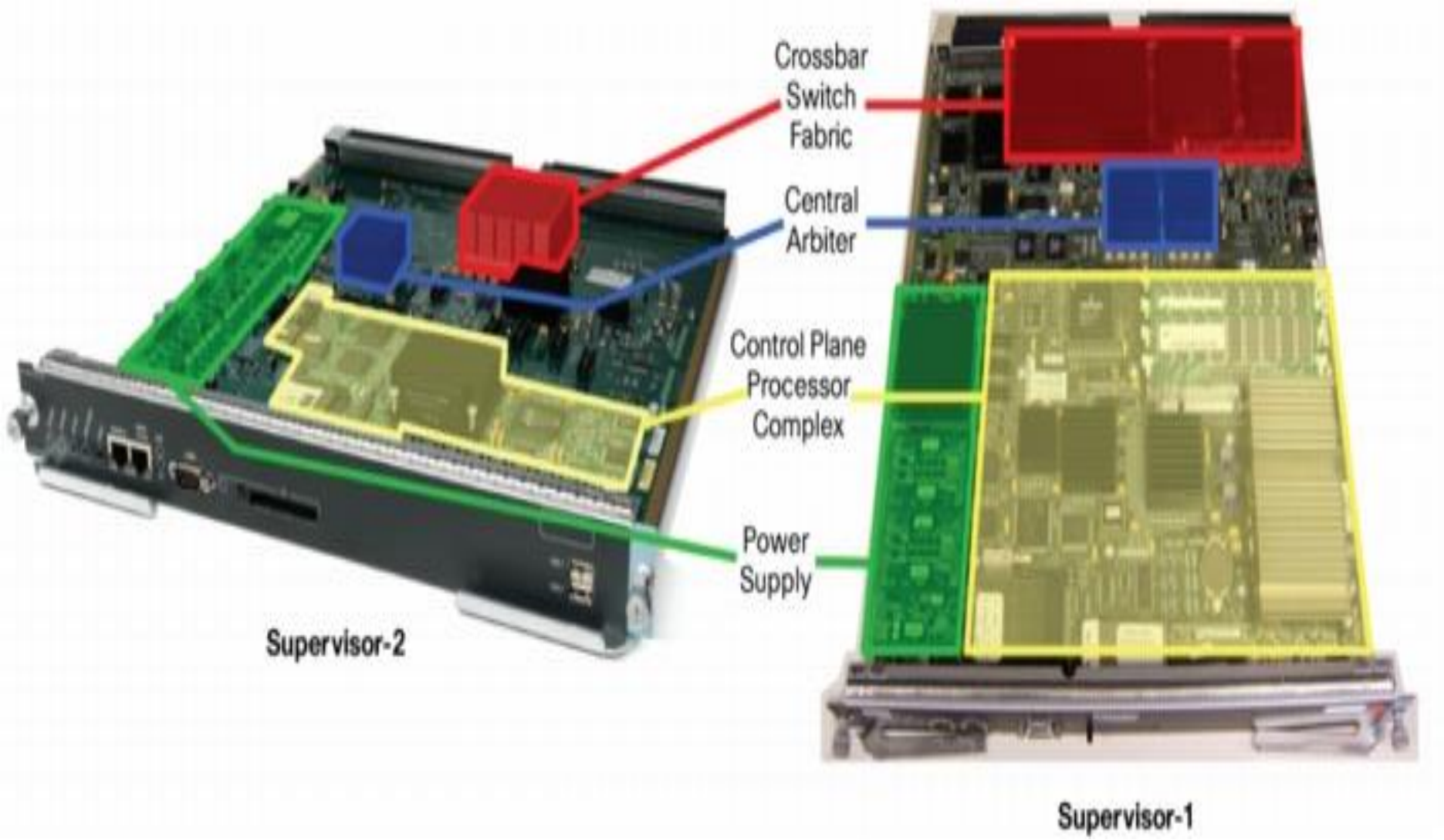
- Veremos em breve porquê o csaamento com prefix mais longo é usado em breve, quando estudarmos endereçamento
- Casamento com prefix mais longo – baixos tempo de acesso a memória fundamental – usual *Ternary Content Addressable Memories* (TCAMs) – exemplo de pesquisa
  - *Endereçamento de conteúdo (content addressable):* apresentado um endereço para a TCAM, obtém-se o resultado em um ciclo de relógio, independente do comprimento da tabela
  - Cisco Catalyst 6500 e 7600: suportam tabela de roteamento com 1 milhão de entradas em TCAM

Elementos de comutação (*Switching fabrics*)

- transfere pacote do *buffer* de entrada para o *buffer* de saída apropriado
- taxa de comutação: taxa em que pacotes podem ser transferidos de entradas para saídas
  - muitas vezes medidos em múltiplos da taxa de linha de entrada/saida
  - *N* entradas: desejável taxa de comutação *N* vezes taxa de linha
- 3 tipos de elementos de comutação



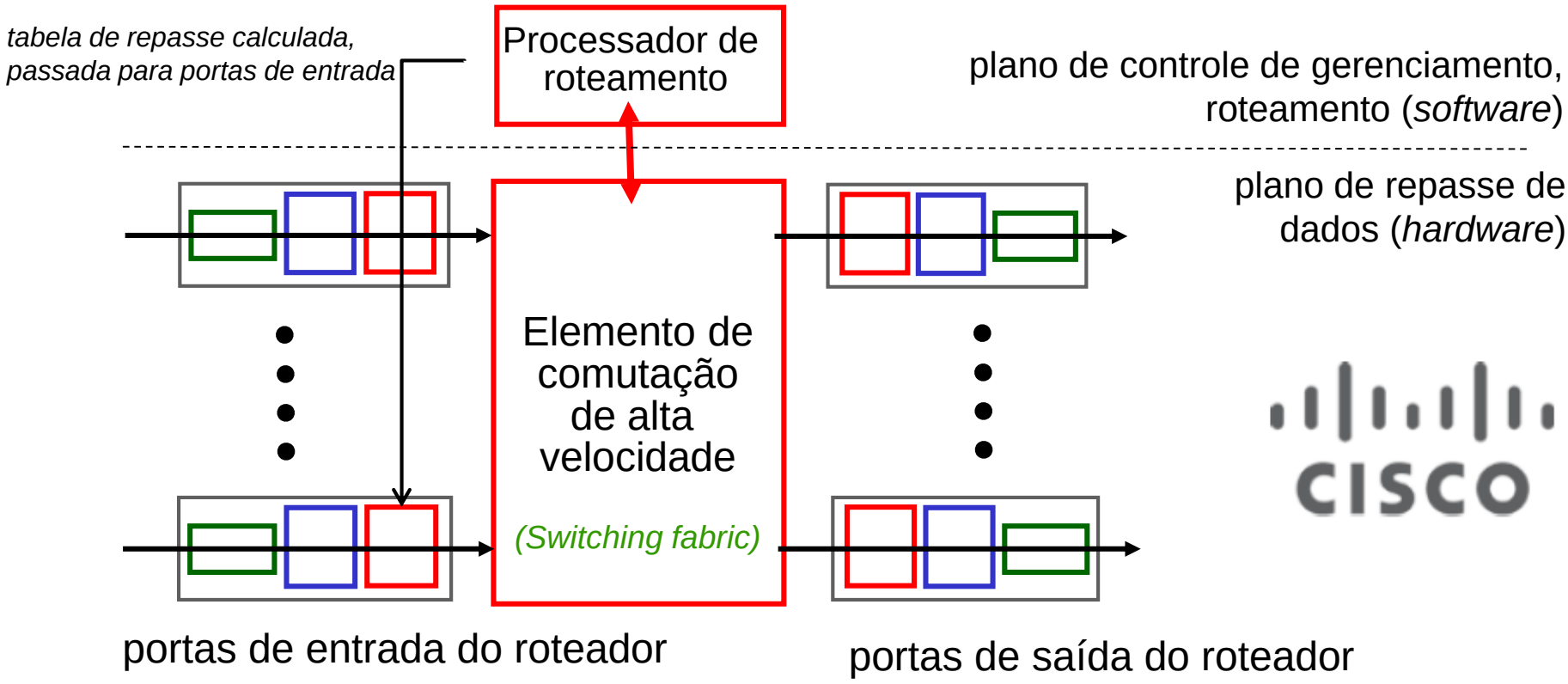
Exemplo: Arquitetura da Família Cisco MDS 9000



Arquitetura de roteadores (CISCO/Huawei/Alcatel-Lucent/Juniper)

2 funções chaves do roteador:

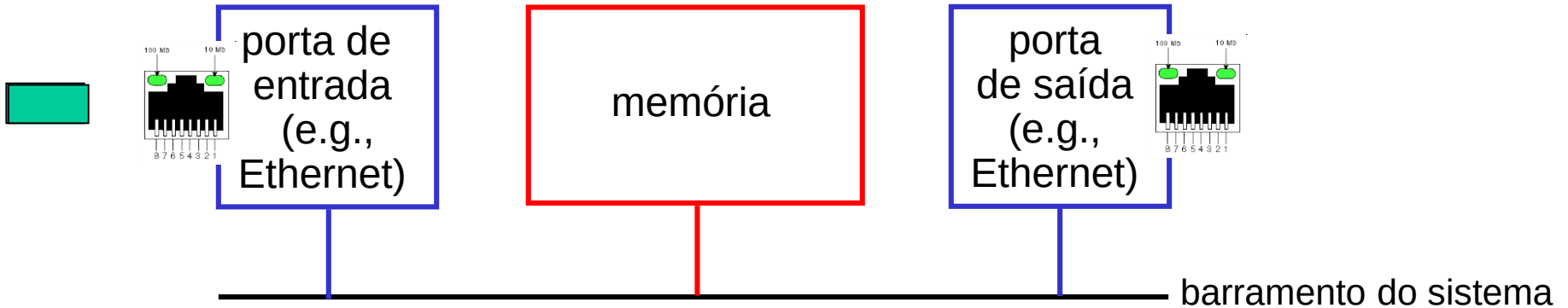
- rodar protocolos e algoritmos de roteamento (RIP, OSPF, BGP)
- *repassar ou comutar* datagramas de enlace de entrada para enlace de saída



Comutação via memória

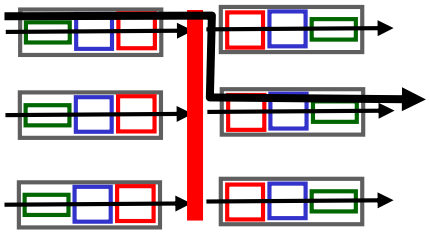
1ª geração de roteadores:

- computadores tradicionais com comutação sob controle direto da CPU
- pacote copiado para memória do sistema (1 de cada vez)
- velocidade limitada pela velocidade da memória (B/2 pacotes/s; B = velocidade da memória)
- atuais também podem usar; mas consulta realizada por processadores nas placas de linha de entrada – 8500 Catalyst Cisco



Comutação via barramento

- datagrama da memória da porta de entrada para porta de saída via barramento
- cabeçalho interno define porta de saída
- apenas 1 pacote pode cruzar o barramento de cada vez
- *contenção pelo barramento:* velocidade de comutação limitada pela velocidade do barramento
- Cisco 5600 barramento de 32 Gbps: velocidade suficiente para roteadores de acesso e de empresas



barramento

Comutação via *crossbar*

- *2N* barramentos ligando *N* entradas a *N* saídas – cruzamento pode ser aberto ou fechado pelo controlador do *switch fabric*

- supera limitações de velocidade do barramento - capaz de repassar vários pacotes em paralelo

- *banyan networks*, *crossbar*, e outras redes de interconexão inicialmente desenvolvidas para conectar processadores em arquiteturas multiprocessadores

- projeto avançado: fragmentando datagramas em células de comprimento fixo, comuta células através da rede

- Cisco 12000: comuta 60 Gbps através da rede de interconexão

- Obs: Cuidado com erros na versão em português do livro!!