



VYSOKÉ UČENÍ FAKULTA
TECHNICKÉ INFORMAČNÍCH
V BRNĚ TECHNOLOGIÍ

6

Směrování v Internetu

IPK2021L

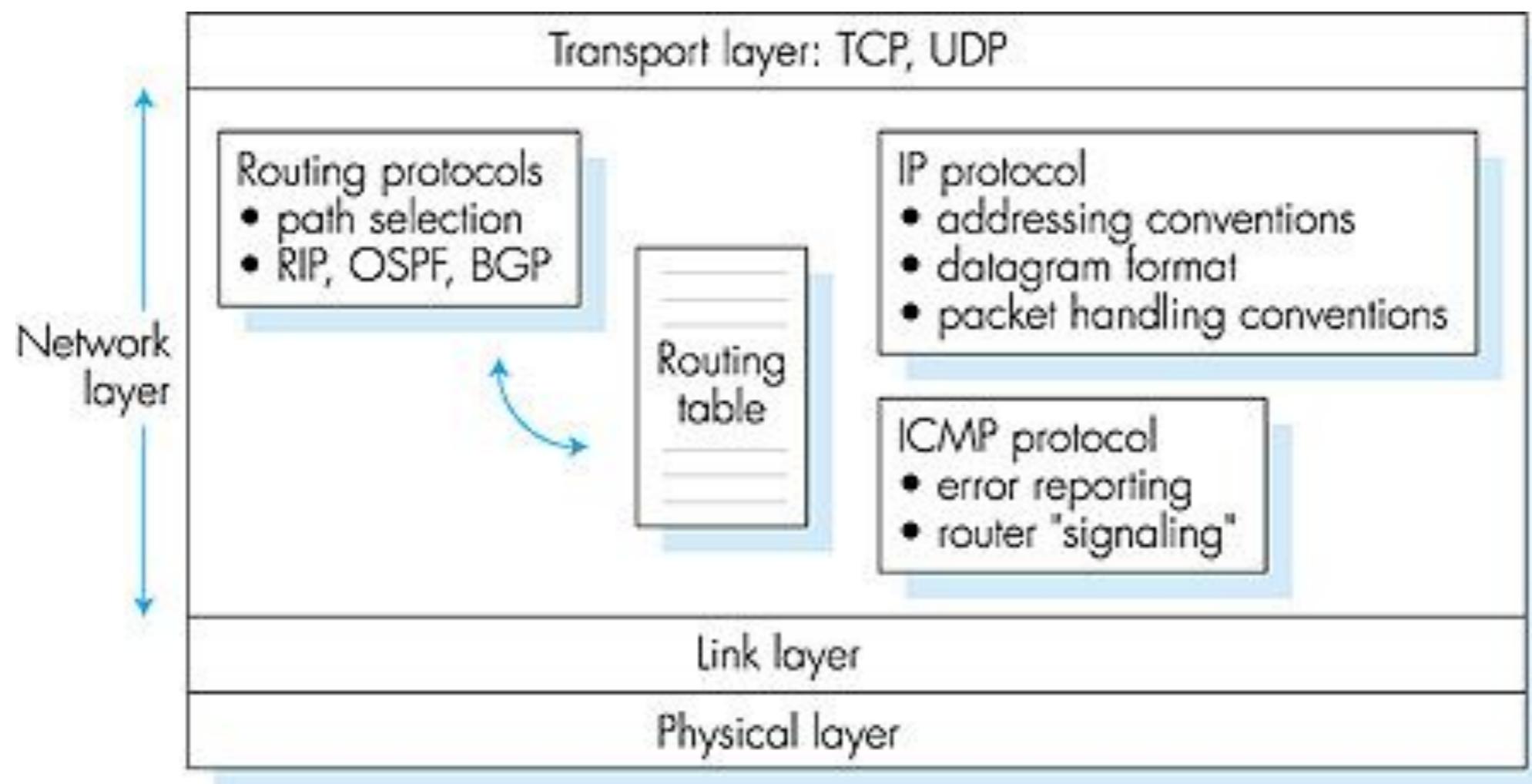
Obsah

- 1) Úvod do směrování
- 2) Směrování paketů
 - Směrovací tabulky
 - Algoritmus výběru
 - Architektura směrovače
- 3) Směrování podle nejkratší cesty
 - Algoritmus Bellman-Ford
 - Algoritmus Dijkstra
 - Path-Vector směrování
- 4) IGP směrovací protokoly
 - RIP
 - OSPF
- 5) EGP směrování
 - BGP

Úvod do směrování

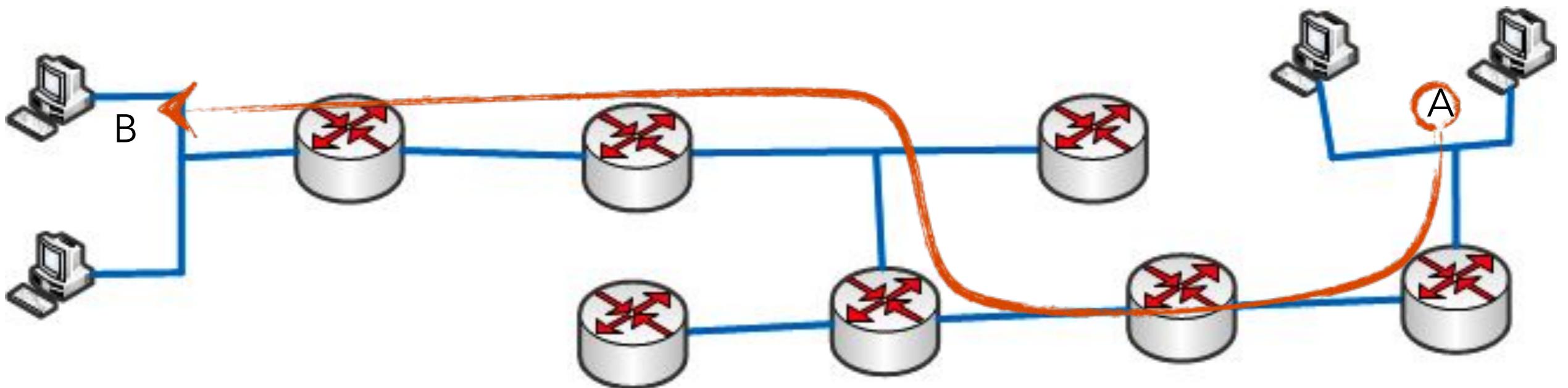
Směrování a síťová vrstva

- Síťová vrstva doručuje pakety koncovým zařízením
- Síťová vrstva zajišťuje šíření směrovacích informací v Internetu pomocí směrovacích protokolů



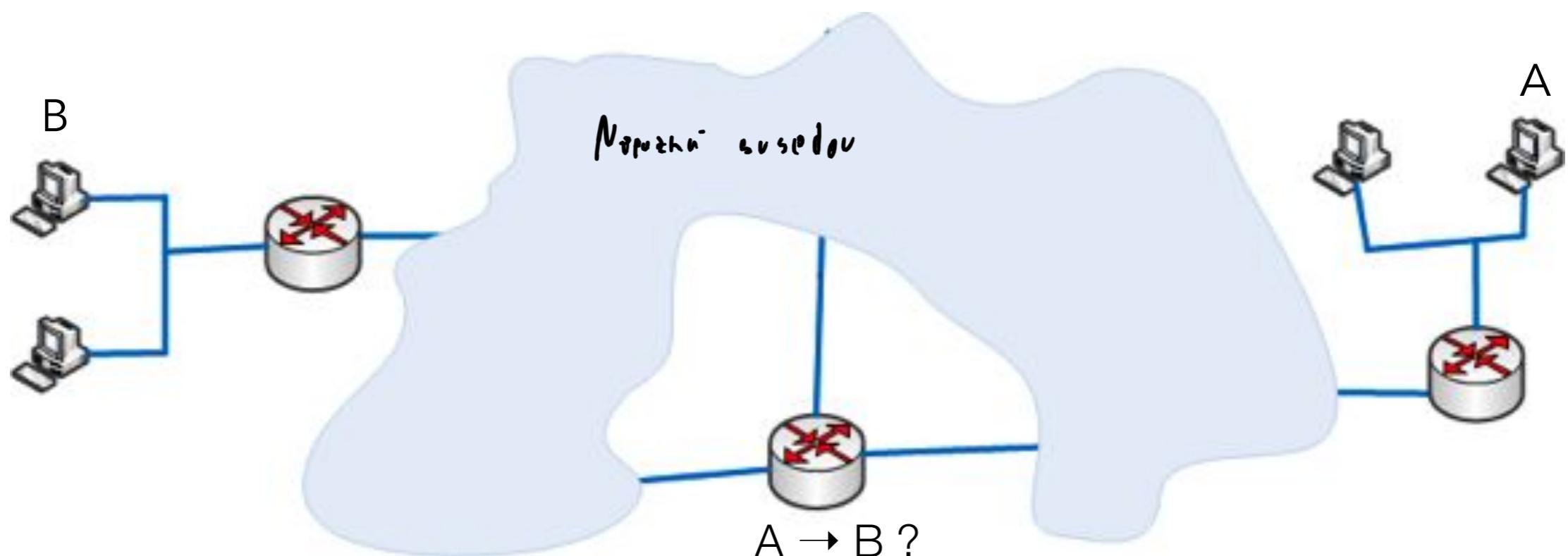
Co je směrování?

Najít vhodnou cestu (tj. posloupnost směrovačů) od zdroje A k cíli B.



Co je směrování?

- Směrovače v síti potřebují vědět:
 - Jaký směrovač X použít pro dosažení cíle B
 - Jaké rozhraní použít pro dosažení směrovače X



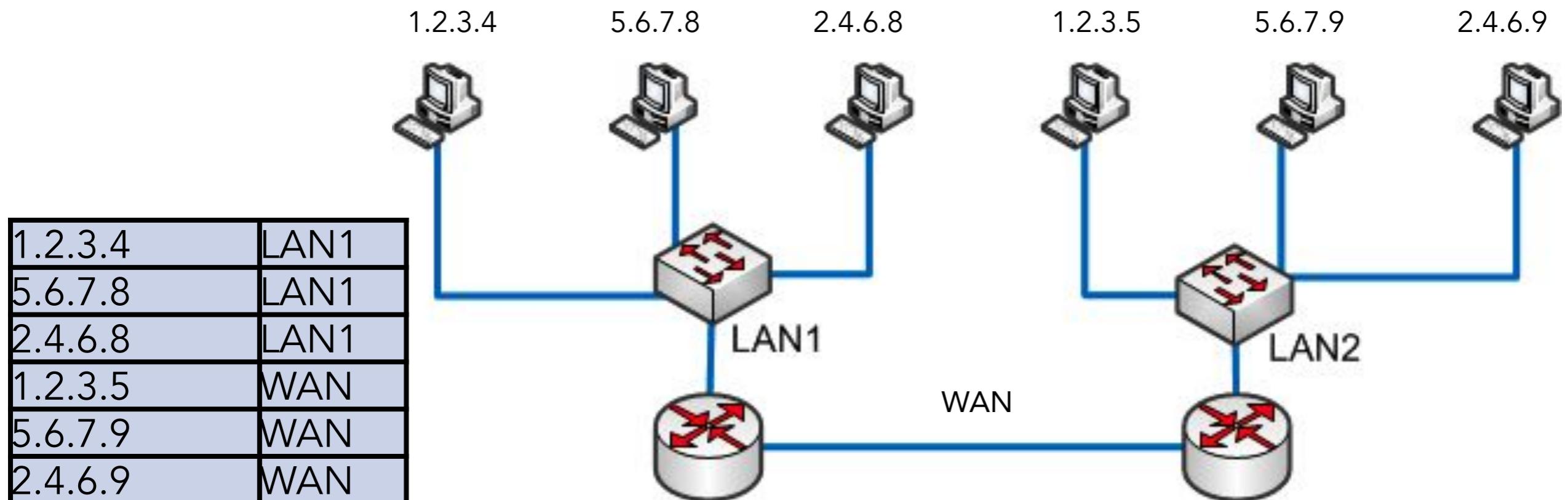
Přepínání a směrování

- Přepínání (forwarding) - vž akce, výsledek směrování
 - směrovač přesune paket, který přichází na vstupní linku na odpovídající výstup
 - lokální akce prováděná směrovačem
- Směrování (routing)
 - síťová vrstva musí nalézt cestu pro paket
 - směrovací algoritmus
 - síťový proces, forma distribuovaného výpočtu, na kterém participují směrovače v síti
 - Směrovač používá směrovací tabulku (forwarding/routing table)

Směrovací tabulky

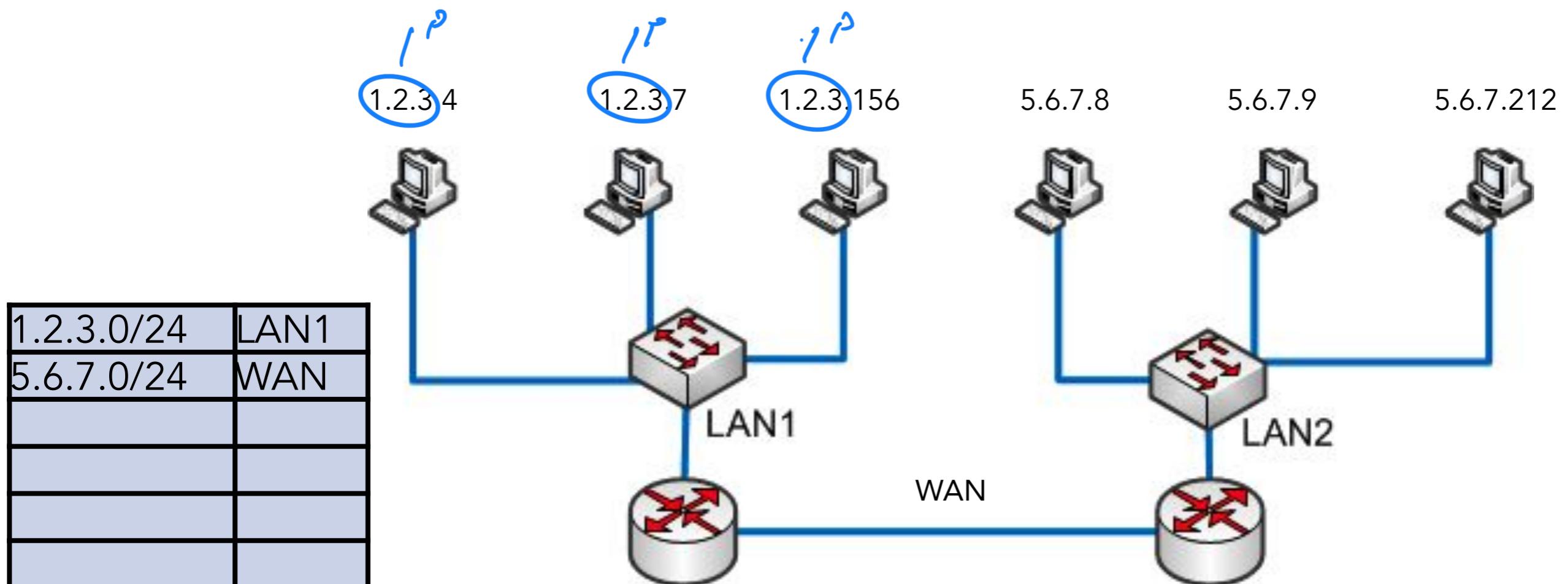
Přepínání paketů

- Směrovač má v tabulce informaci pro každou IP adresu
 - testování shody pro cílovou adresu v paketu
 - jednoznačné určení výstupního rozhraní
- *Velikost směrovacích tabulek!* *32 bitů, $2^{32}-1$ zářežat*



Přepínání paketů

- Směrovač má v tabulce informaci pro 24-bitový prefix
 - testování shody pro prefix cílové adresy v paketu
 - vyžaduje rozdělení IP adresy na síťovou část a uzlu
- Zmenšení velikosti tabulek :) fixní délka prefixu :(



Třídní směrování

- angl. Classful routing
- Položky RT jsou JEN síťové adresy
 - třída adresy definuje prefix (8, 16, 24)
 - třídu lze poznat podle nejvyšších bitů
 - Classful routing protocol (např. RIPv1 či IGRP neposílá masku sítě)

Address Class	Bit Pattern of First Byte	First Byte Decimal Range	Host Assignment Range in Dotted Decimal	
A	0xxxxxxx	32 bits	1 to 127	1.0.0.1 to 126.255.255.254
B	10xxxxxx	11 - 1h	128 to 191	128.0.0.1 to 191.255.255.254
C	110xxxxx	6 - 11	192 to 223	192.0.0.1 to 223.255.255.254
D	1110xxxx	1 - 10	224 to 239	224.0.0.1 to 239.255.255.254
E	11110xxx	2 - 11	240 to 255	240.0.0.1 to 255.255.255.255

Beztrídní směrování (1)

- Classless routing - v 127 je maska sítě třídy určuje až 11 bitů za prefix
- Položky RT jsou síťové adresy + maska sítě
- Význam masky sítě (subnet mask):
 - 1 = bit je součástí NetID
 - 0 = bit je součástí HostID

158	193	138	40
10011110	11000001	10001010	00101000
11111111	11111111	11111111	00000000

Beztrídní směrování (2)

- Hranice mezi NetId a HostId nemusí být na bytech
- Router porovnává dstIP s každým záznam v RT (součástí čehož je i vymaskování dstIP SM)
- $158.193.138.40 \ \& \ 255.255.255.224 = 158.193.138.32$

10011110	11000001	10001010	00101000
AND			
11111111	11111111	11111111	11100000
=			
10011110	11000001	10001010	00100000

Příklad

**Packet
address:**

1.2.3.4

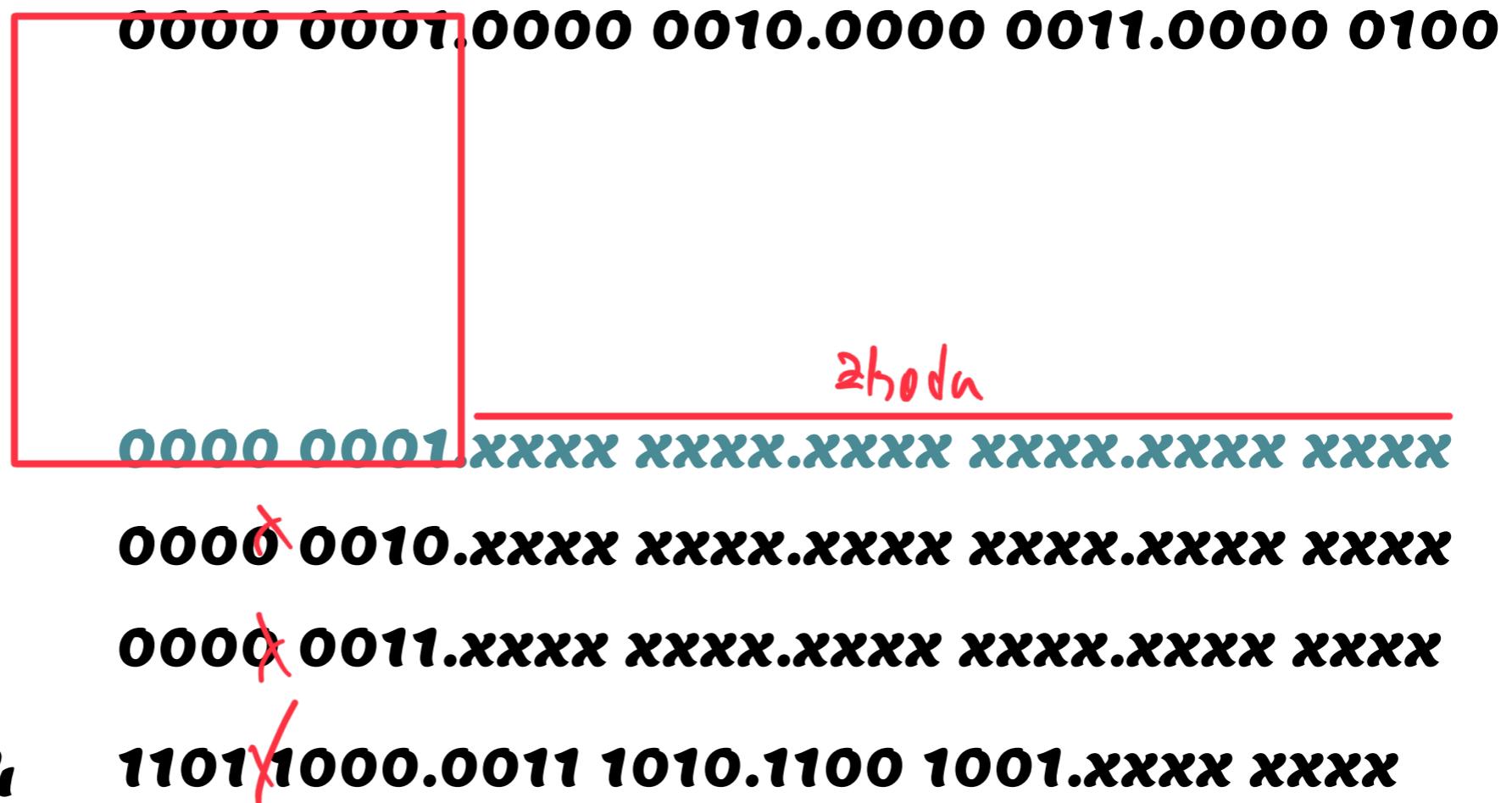
**Routing
Table:**

1.0.0.0/8

2.0.0.0/8

3.0.0.0/8

216.58.201.0/24



CIDR v beztřídním směrování

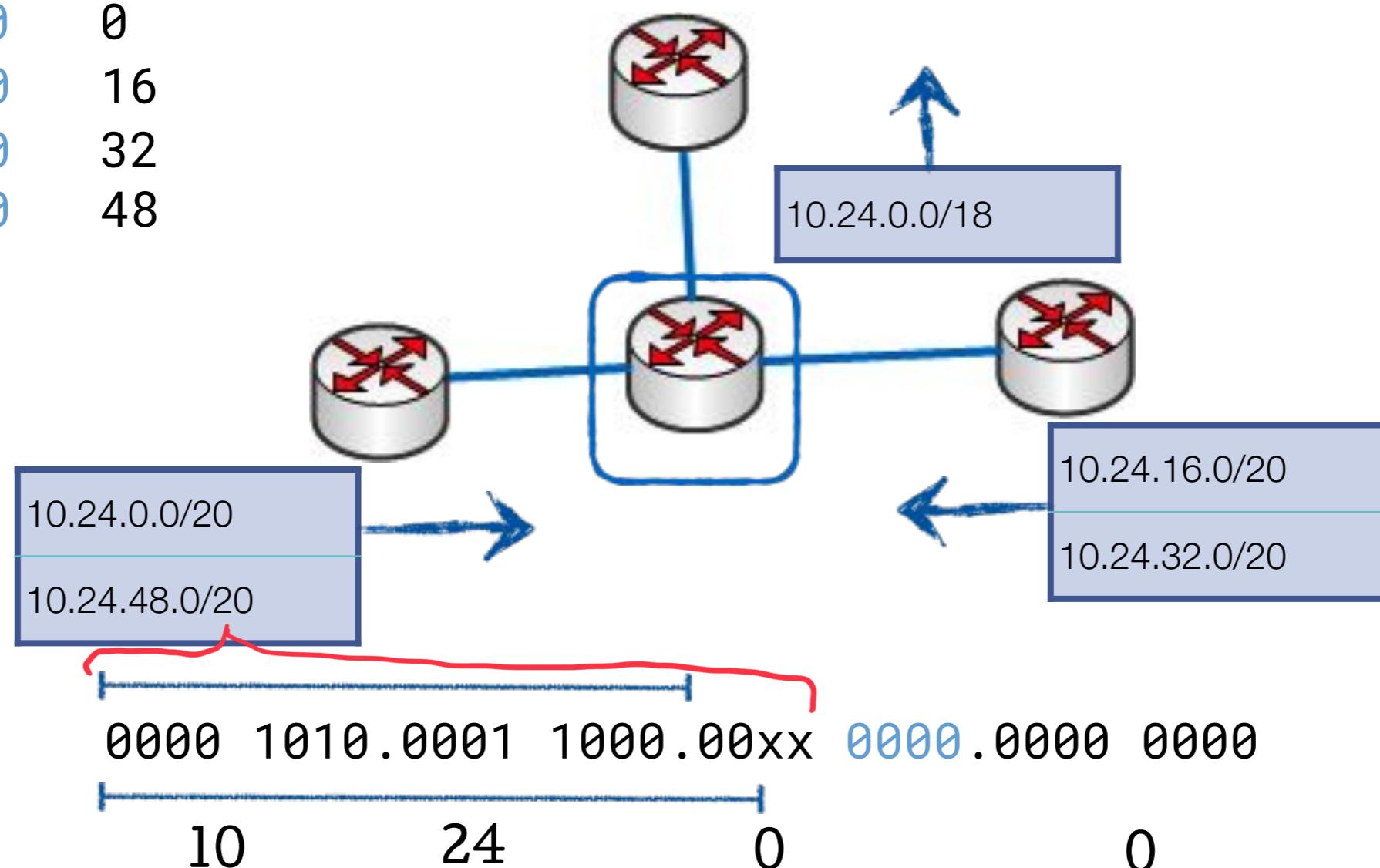
- Počet možných sítí jednotlivých tříd:
 - Třída A: 127
 - Třída B: 16 384
 - Třída C: 2 097 152
- Celkem možných síťových prefixů
 - 2 113 664 (ne všechny prefixy jsou použitelné v Internetu)
 - Kompletní (a velká) směrovací tabulka Internetu!

CLASS	IP RANGE (1ST OCTET)	DEFAULT SUBNET MASK	NETWORK/NODE PORTIONS	TOTAL NUMBER OF NETWORKS	TOTAL NUMBER OF USABLE ADDRESSES
A	0–127	255.0.0.0	Net.Node.Node.Node	2^7 or 128	$2^{14} - 2$ or 16,777,214
B	128–191	255.255.0.0	Net.Net.Node.Node	2^{14} or 16,384	$2^{16} - 2$ or 65,534
C	192–223	255.255.255.0	Net.Net.Net.Node	2^{21} or 2,097,151	$2^8 - 2$ or 254
D	224–239	N/A	N/A	N/A	N/A
E	240–255	N/A	N/A	N/A	N/A

CIDR v třídním směrování

- CIDR (classless interdomain routing) ↔ agregace adres pro směrování (někdy supernetting) - ~~neskladovatelné adresy~~
- Směrovače šíří informaci o prefixu směrování

0000	0000	0
0001	0000	16
0010	0000	32
0011	0000	48



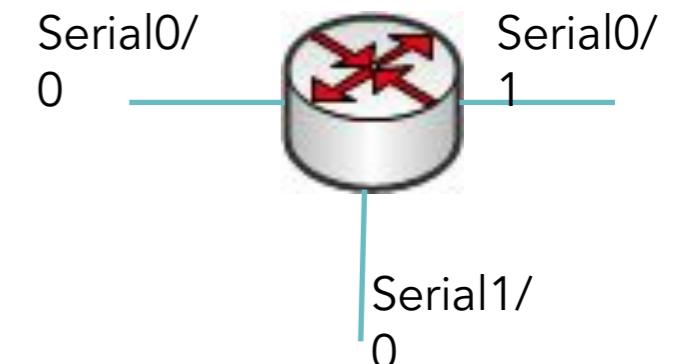
Longest Prefix Match (1)

- Jak se určí, která cesta v RT se vybere?
- Směrovače hledají nejdelší shodu
 - řeší problém možné vícenásobné shody ve směrovací tabulce
 - existují efektivní algoritmy

bit

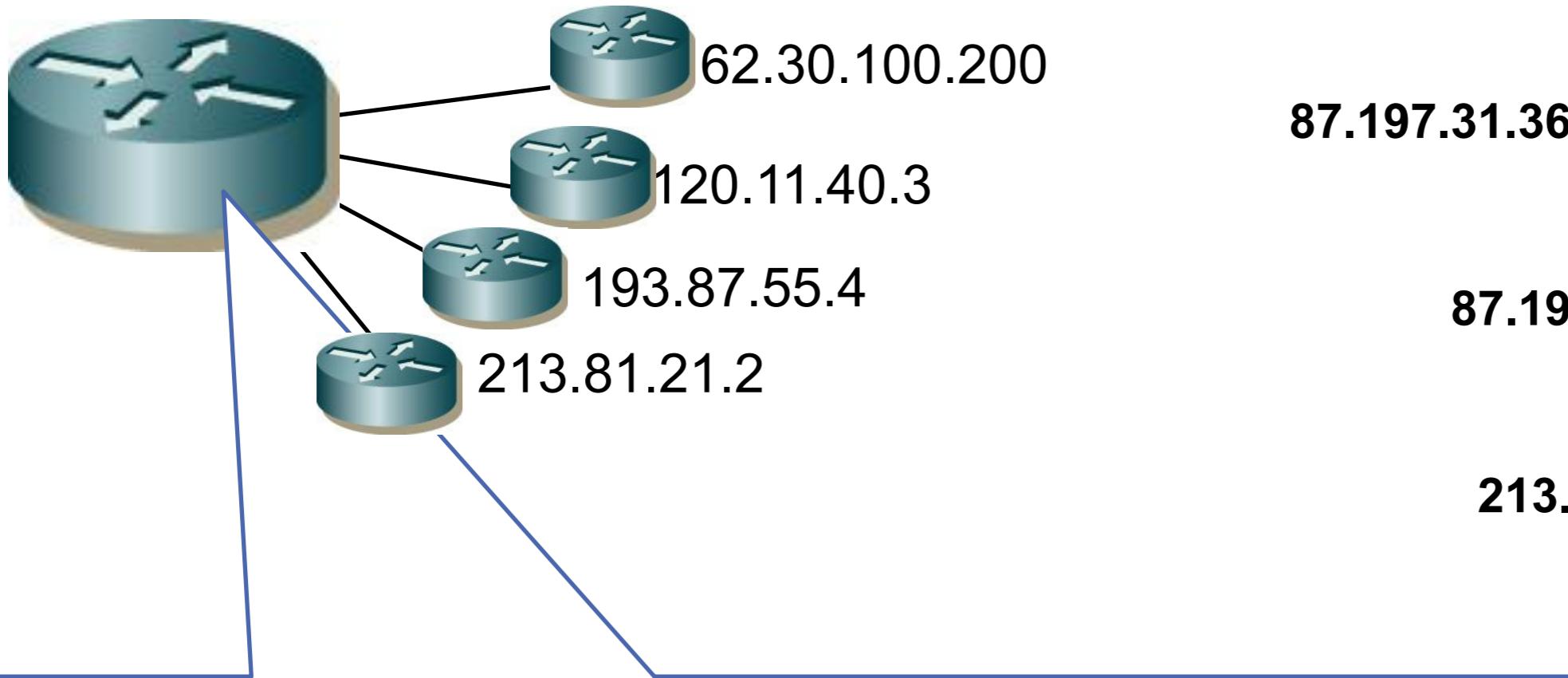
Cílová adresa: 201.10.6.17

1100	1001.0000	1010.0000	0110.0001	0001	
0000	0100.xxxx	xxxx.xxxx	xxxx.xxxx	xxxx	/8
0000	0100.0101	0011.1xxx	xxxx.xxxx	xxxx	/17
1100	1001.0000	1010.0000	0xxx.xxxx	xxxx	/21
1100	1001.0000	1010.0000	011x.xxxx	xxxx	/23
0111	1110.1111	1111.0110	0111.xxxx	xxxx	/24



4.0.0.0/8	Serial0/0
4.83.128.0/17	Serial0/0
201.10.0.0/21	Serial1/0
201.10.6.0/23	Serial0/1
126.255.103.0/24	Serial0/0

Longest Prefix Match (2)



87.197.31.42 & 255.255.255.248 =
87.197.31.40

87.197.31.36 & 255.255.255.240 =
87.197.31.32

87.197.1.1 & 255.255.0.0 =
87.197.0.0

213.81.187.59 & 0.0.0.0 =
0.0.0.0

Mask	NetID	Next hop
255.255.255.248	87.197.31.40	62.30.100.200
255.255.255.240	87.197.31.32	120.11.40.3
255.255.0.0	87.197.0.0	193.87.55.4
0.0.0.0	0.0.0.0	213.81.21.2

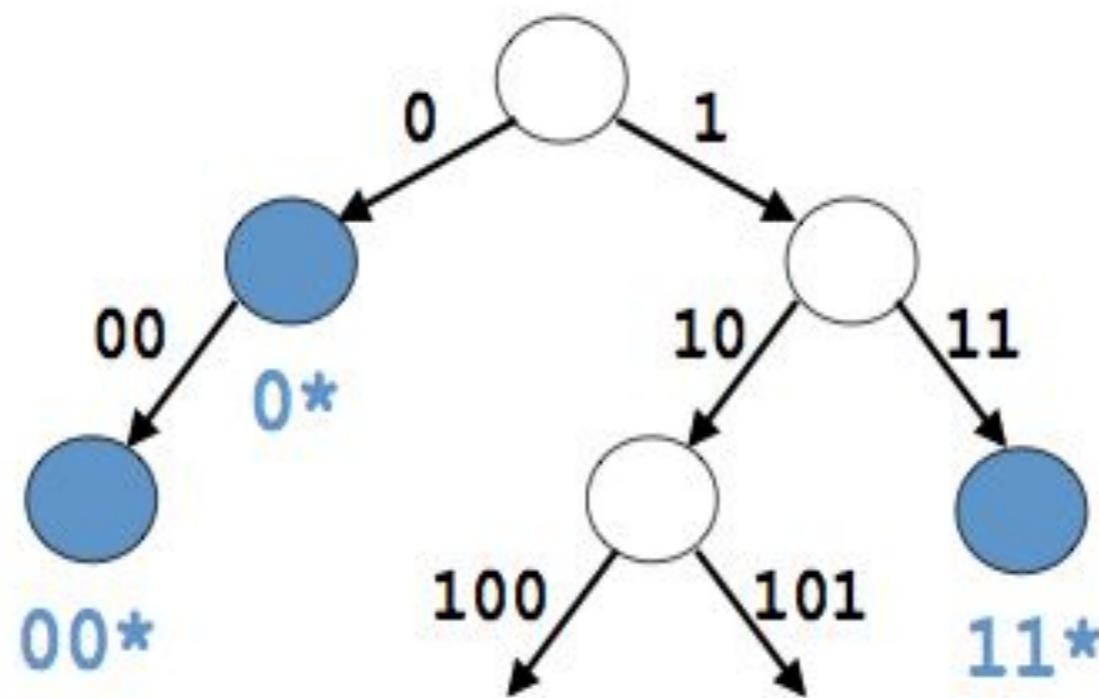
Naivní přístup

- Sekvenční průchod směrovací tabulkou
- Složitost je lineární, nicméně je potřeba lepší algoritmus
- Problém:
 - některé směrovače mohou mít až 350.000 záznamů
 - Čas na zpracování, 10Gbps směrovač, pakety 64B:

$$\frac{10^{10}}{8 \cdot 64} = 19531250 \text{ p/s} \sim 51 \text{ ns/p}$$

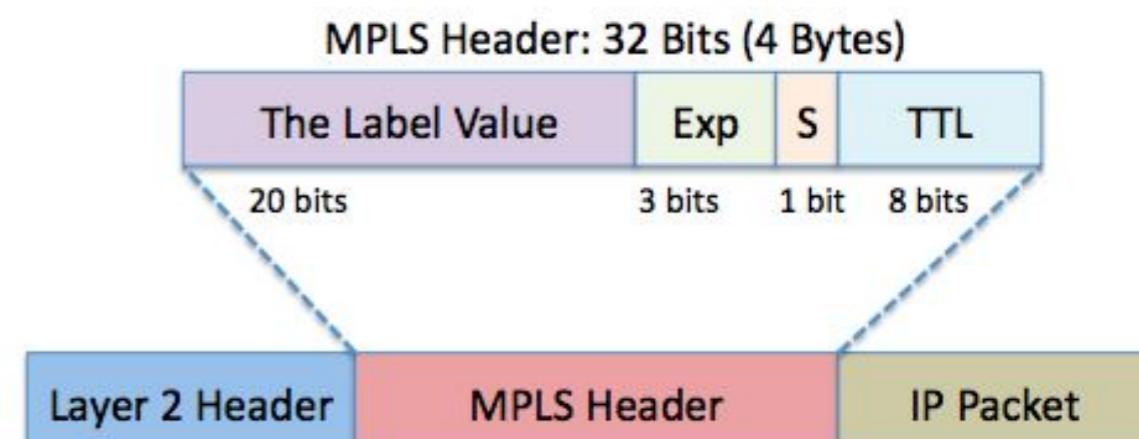
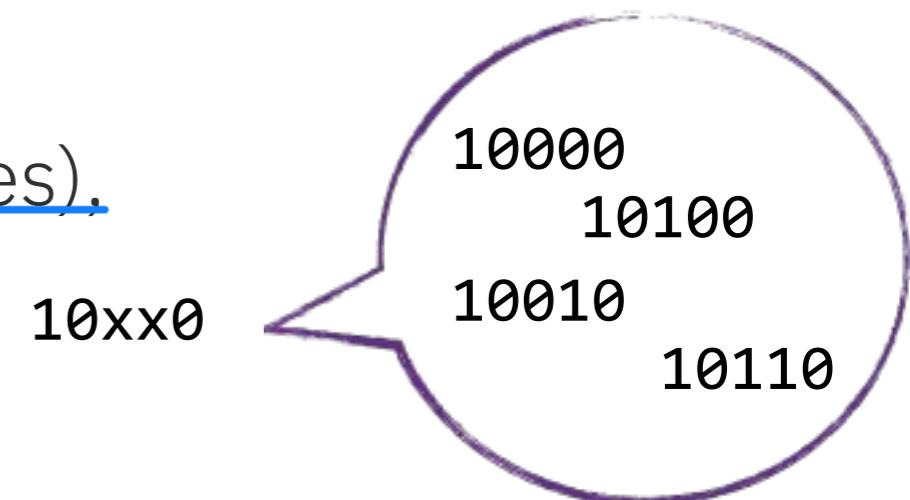
Dobrý přístup

- Prefixové stromy (Patricia Tree, 1968)
 - každá úroveň stromu je jeden bit v adrese
 - některé uzly mají přiřazeny záznamy v tabulce
- Když přijde paket hledá se nejdelší prefix ve stromě
- Listy jsou cesty ve směrovací tabulce



„Nejchytrější“ přístup

- Použiji speciální HW
 - Použití CAM (Content Addressable Memories), které implementuje asociativní pole
 - Dnes nejčastěji jako TCAM (hodnoty 0,1,X)
- *Lze to ještě urychlit?*
 - Použiji jiný koncept směrování než na základě cílové IP
- například MPLS
 - umožňují vytvářet virtuální okruhy na kterých se přepínají data identifikovaná pomocí labelů



Kde se berou záznamy v RT?

- Položky mohou být **staticky** definovány
 - administrátor je musí vložit
 - nejsou aktuální v případě selhání zařízení
 - nejsou dynamické, nereflektují aktuální stav sítě
- Nebo zjištěny pomocí dynamických směrovacích protokolů
 - IGP protokoly pro směrování uvnitř organizace (RIP, OSPF)
v jednotkách ISP
 - EGP protokol pro výměnu informací mezi organizacemi a ISP (BGP)
v rámci ISP

Vhodné pro Impresivní
TCPba Gnatová kohesivní

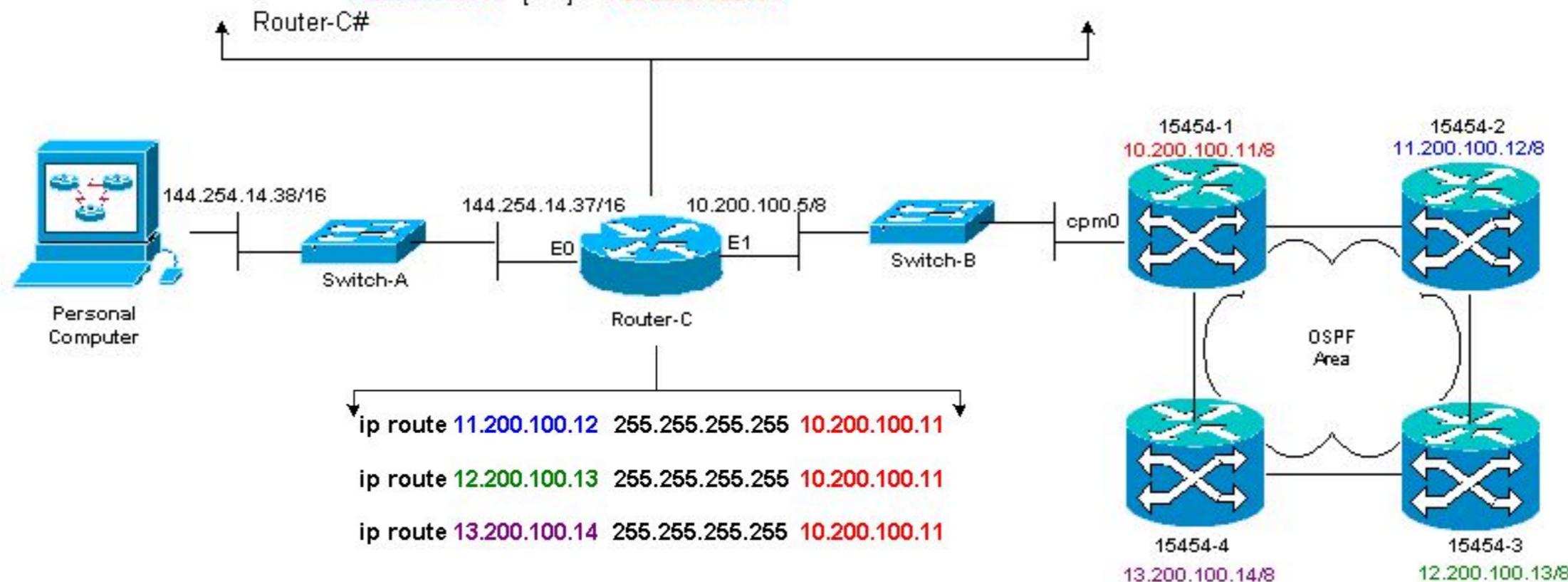
Statické směrování

Router-C# **show ip route**

Codes: C - connected, S - static, I - IGRP, R - RIP, M - mobile, B - BGP
D - EIGRP, EX - EIGRP external, O - OSPF, IA - OSPF inter area
N1 - OSPF NSSA external type 1, N2 - OSPF NSSA external type 2
E1 - OSPF external type 1, E2 - OSPF external type 2, E - EGP
i - IS-IS, L1 - IS-IS level-1, L2 - IS-IS level-2, * - candidate default
U - per-user static route, o - ODR

Gateway of last resort is not set

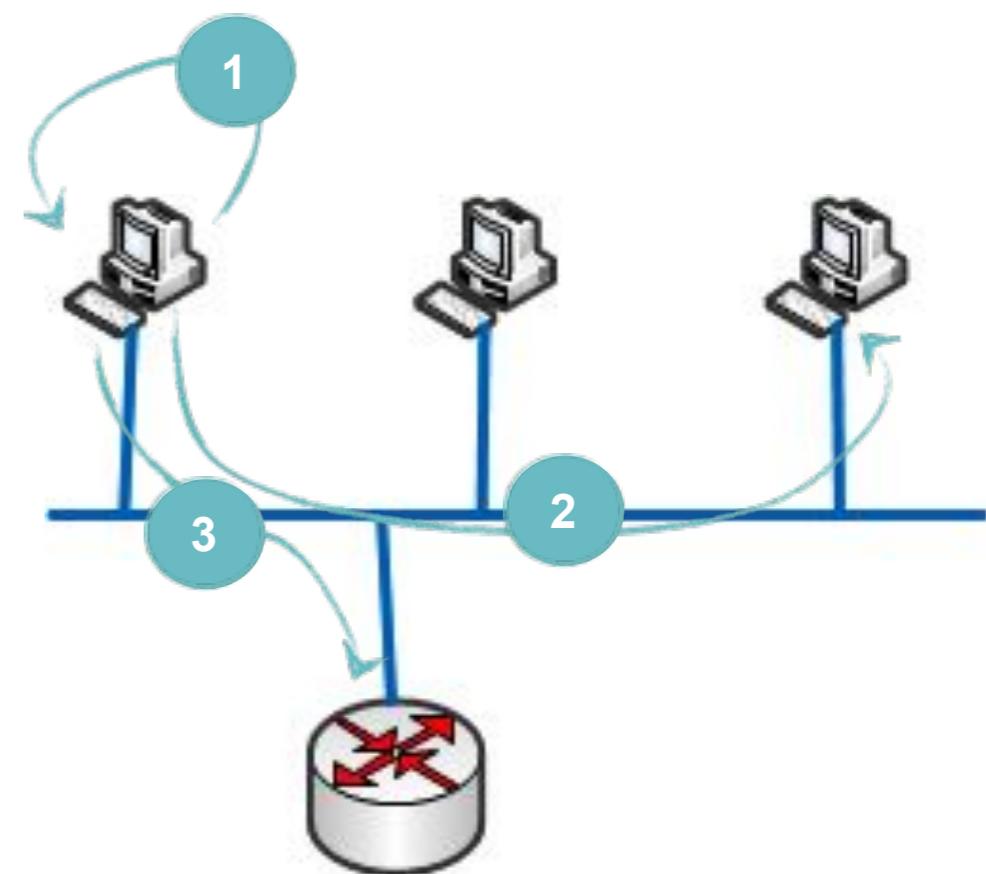
C 10.0.0.0/8 is directly connected, Ethernet0
C 144.254.0.0/16 is directly connected, Ethernet1
S **11.200.100.12** [1/0] via **10.200.100.11**
S **12.200.100.13** [1/0] via **10.200.100.11**
S **13.200.100.14** [1/0] via **10.200.100.11**



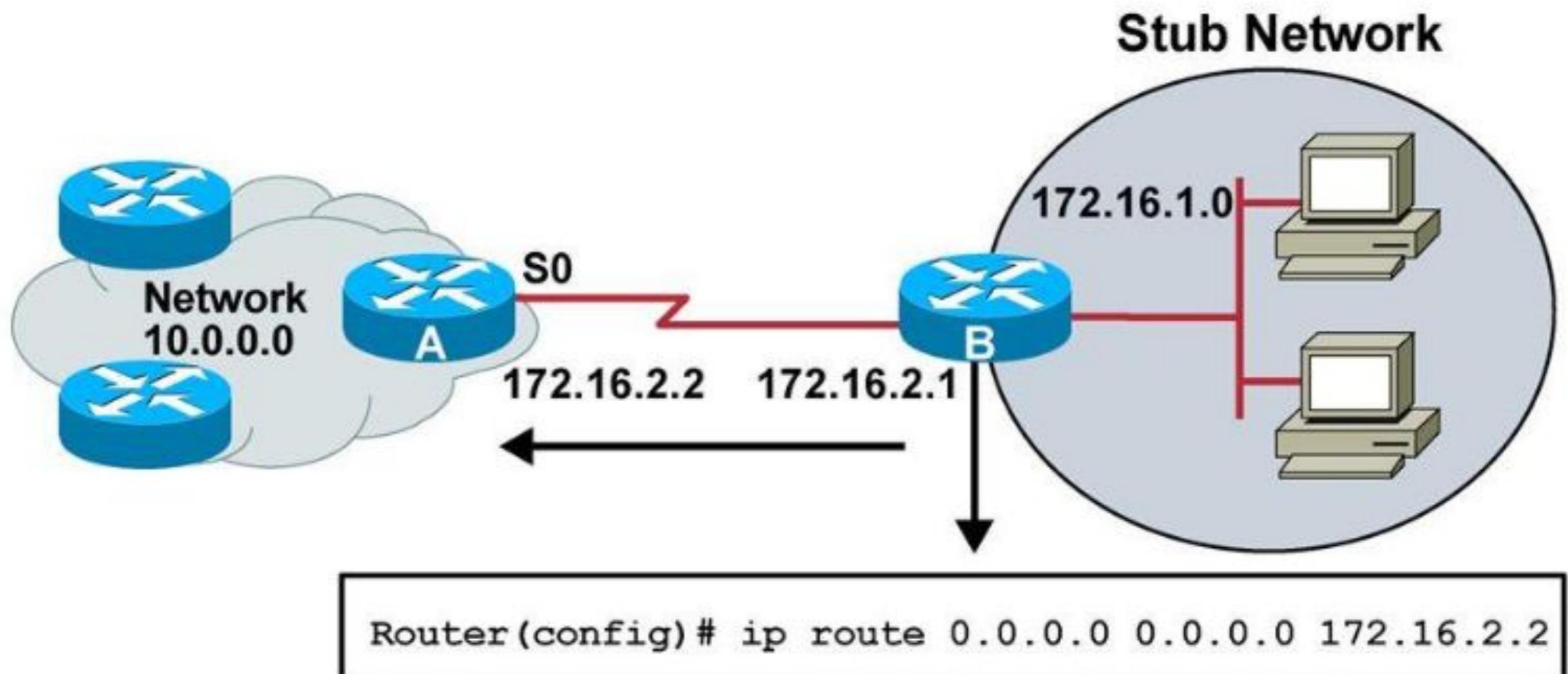
Směrování z pohledu koncové stanice

- V případě, že má koncová stanice pouze jedno rozhraní, nepotřebuje směrování:
 1. Paket na vlastní adresu je doručen lokálně
 2. Paket ostatním uzlům v síti je poslán v Ethernetovém rámci s konkrétní adresou příjemce
 3. Paket vně sítě je poslán na lokální výchozí bránu

```
C:\WINDOWS\system32\cmd.exe
IPv4 Route Table
=====
Active Routes:
Network Destination      Netmask      Gateway      Interface Metric
          0.0.0.0        0.0.0.0    10.20.40.1    10.20.40.47    25
         10.7.0.0   255.255.0.0        On-link     10.7.0.1    291
         10.7.0.0   255.255.0.0  192.168.255.1  192.168.255.4    259
         10.7.0.1   255.255.255.255        On-link     10.7.0.1    291
        10.7.255.255 255.255.255.255        On-link     10.7.0.1    291
         10.20.40.0   255.255.255.0        On-link    10.20.40.47    281
         10.20.40.47   255.255.255.255        On-link    10.20.40.47    281
        10.20.40.255 255.255.255.255        On-link    10.20.40.47    281
         10.100.0.0   255.255.0.0  192.168.255.1  192.168.255.4    259
         10.200.0.0   255.255.0.0  192.168.255.1  192.168.255.4    259
         127.0.0.0    255.0.0.0        On-link    127.0.0.1    331
         127.0.0.1    255.255.255.255        On-link    127.0.0.1    331
        127.255.255.255 255.255.255.255        On-link    127.0.0.1    331
         169.254.0.0   255.255.0.0        On-link  169.254.86.8    291
         169.254.0.0   255.255.0.0        On-link  169.254.125.121    291
         169.254.0.0   255.255.0.0        On-link  169.254.217.213    291
         169.254.0.0   255.255.0.0        On-link  169.254.150.30    291
         169.254.0.0   255.255.0.0        On-link  169.254.83.108    291
         169.254.0.0   255.255.0.0        On-link  169.254.71.147    291
        169.254.71.147 255.255.255.255        On-link  169.254.71.147    291
         169.254.83.108 255.255.255.255        On-link  169.254.83.108    291
         169.254.86.8   255.255.255.255        On-link  169.254.86.8    291
        169.254.125.121 255.255.255.255        On-link  169.254.125.121    291
         169.254.150.30 255.255.255.255        On-link  169.254.150.30    291
        169.254.217.213 255.255.255.255        On-link  169.254.217.213    291
```

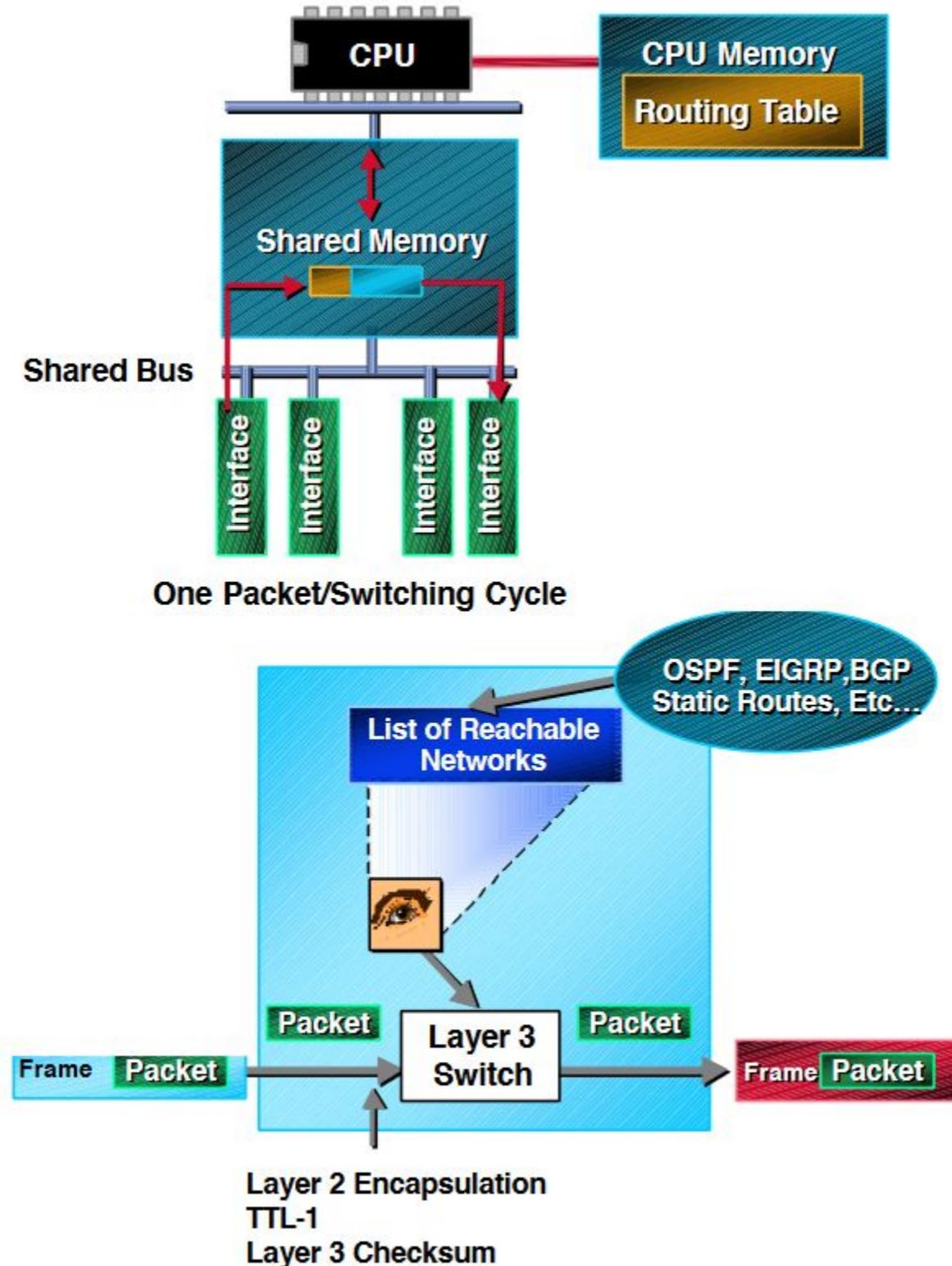


Defaultní cesta



Architektura směrovačů

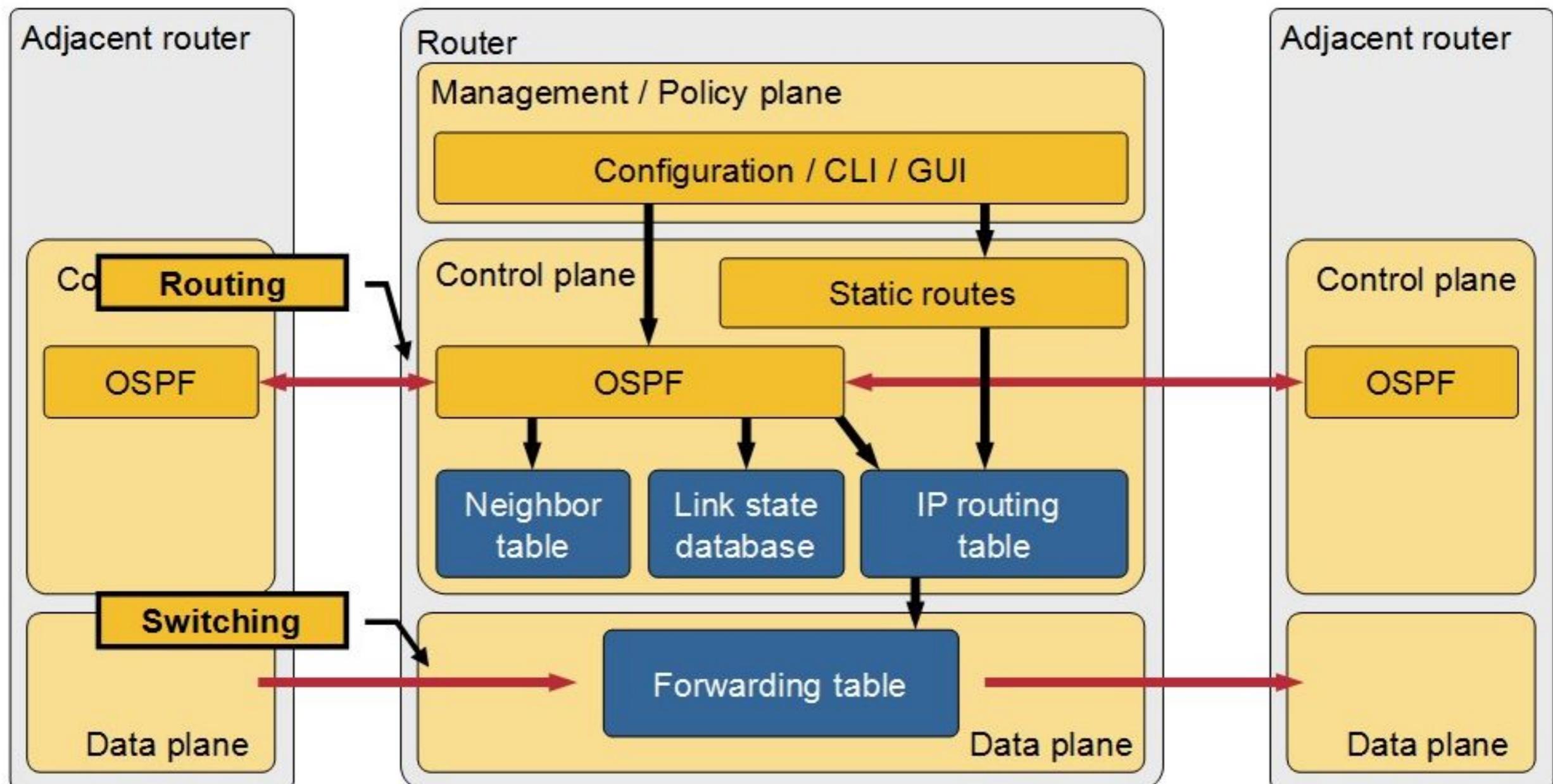
Co je směrovač?



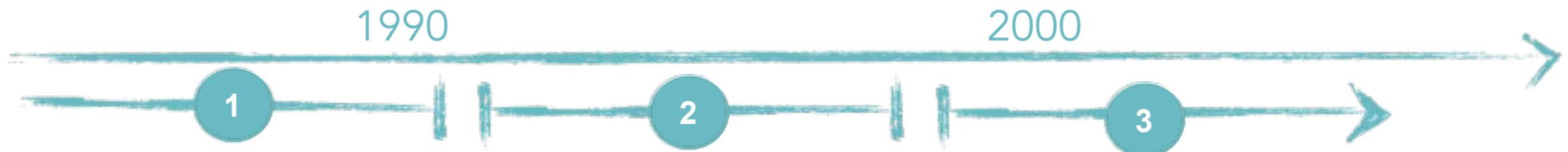
Funkce směrovače

- Zpracování směrovacích informací
 - výpočet nejlepší cesty
 - správa směrovací tabulky
 - provoz směrovacích protokolů
- Přepínání paketů
 - zjištění cílové adresy paketu
 - nalezení výstupního portu
 - kontrola stáří paketu (TTL)
 - výpočet kontrolního součtu
- Speciální funkce
 - transformace paketů **hlavička**

Struktura směrovače



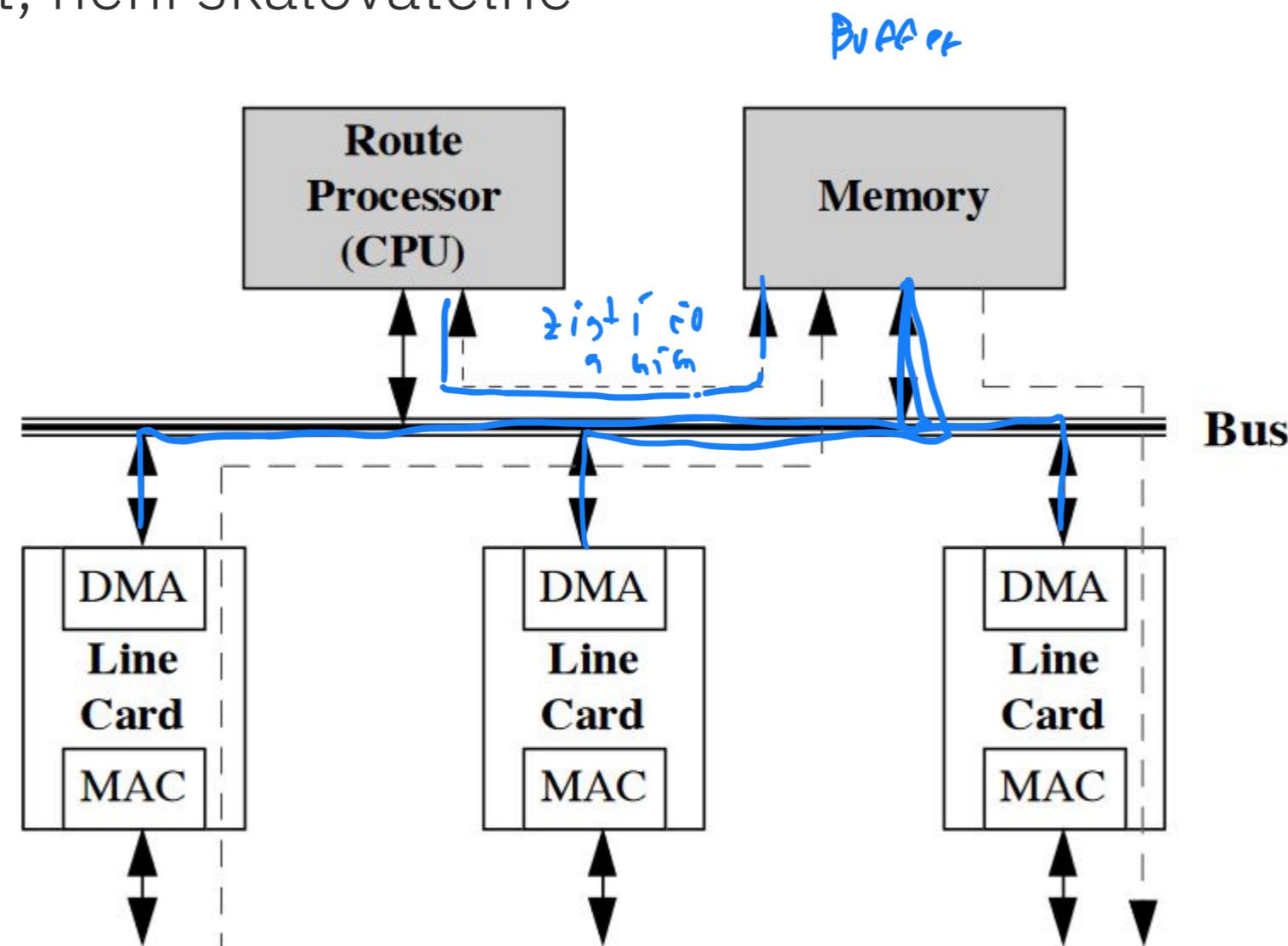
Architektury směrovačů: vývoj



- 1.generace
 - softwarové směrovače ← *přímá*
 - standardní PC
- 2.generace
 - sběrnice pro vnitřní komunikaci
 - paralelní zpracování na rozhraní
- 3.generace
 - přepínač pro vnitřní komunikaci
 - distribuovaná architektura

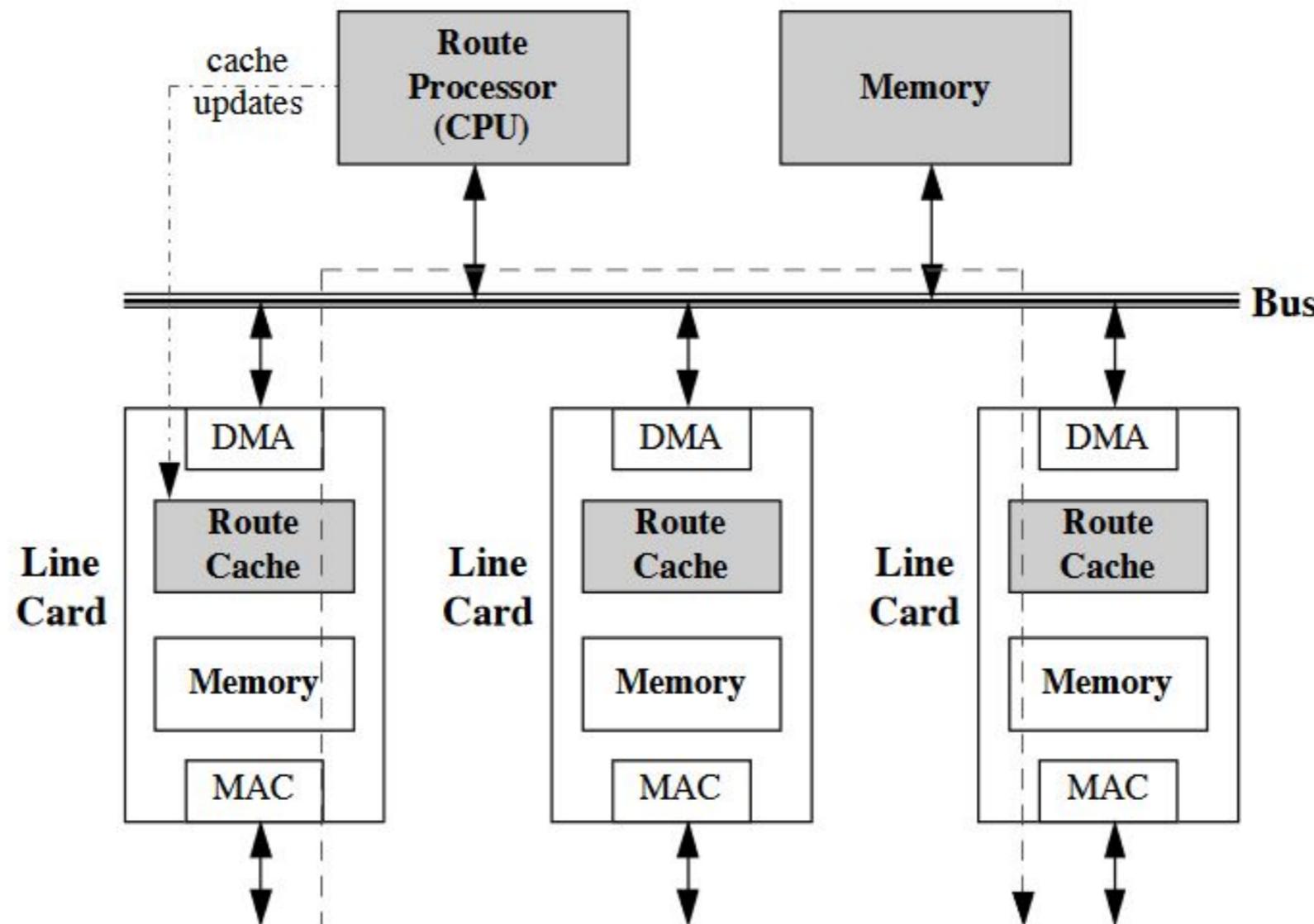
Sběrnice s centrálním CPU

- všechny pakety jsou posílány do CPU, které je analyzuje
- CPU musí dělat i další věci
- malá výkonnost, není škálovatelné



Sběrnice s lokální pamětí cache

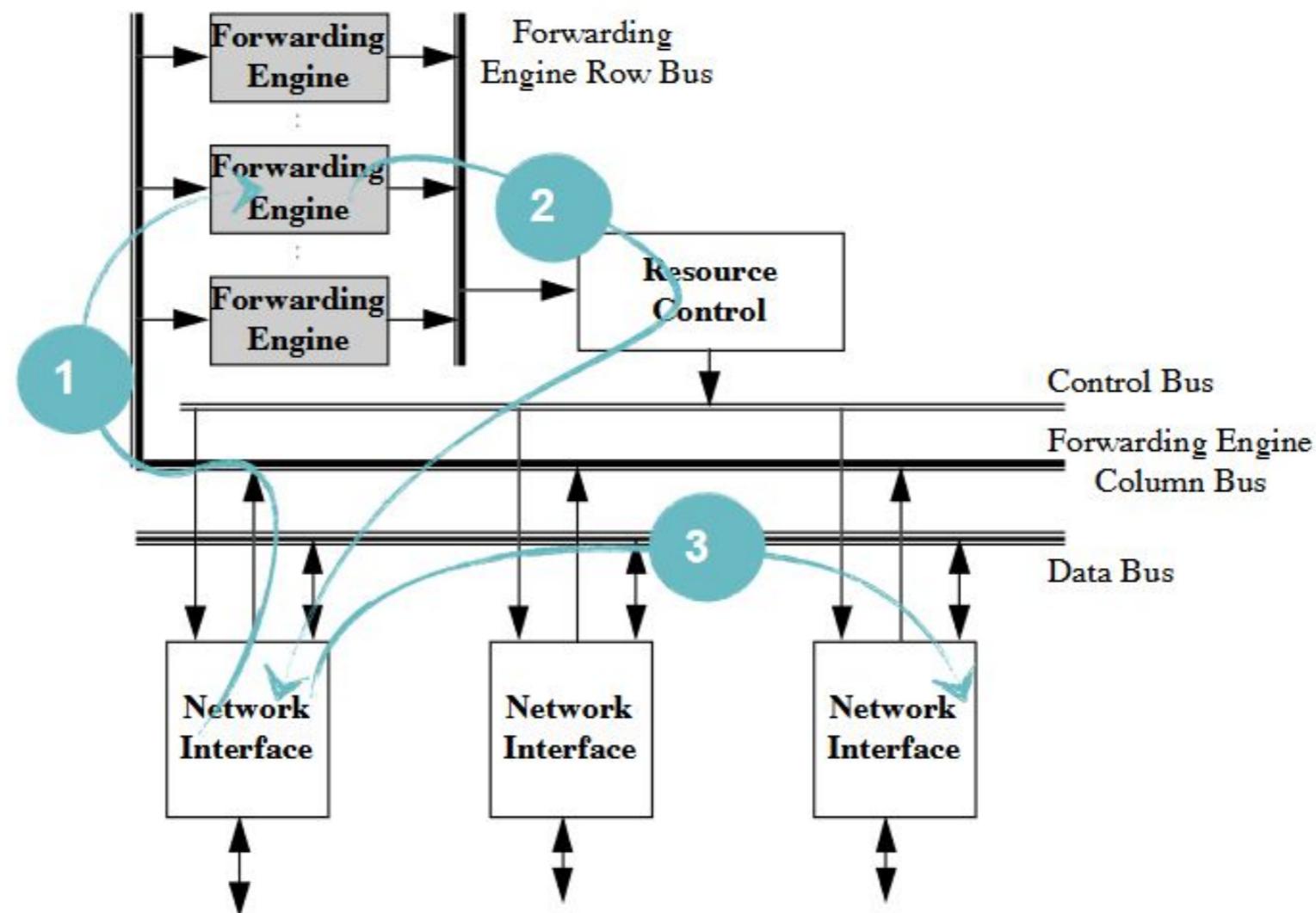
- lokální CPU na kartách obsahují částečné kopie centrální směrovací tabulky
- sběrnice přenáší data mezi kartami
- pakety, které není možné zpracovat na kartě se posílají do centrálního CPU



Sběrnice s paralelním zpracováním

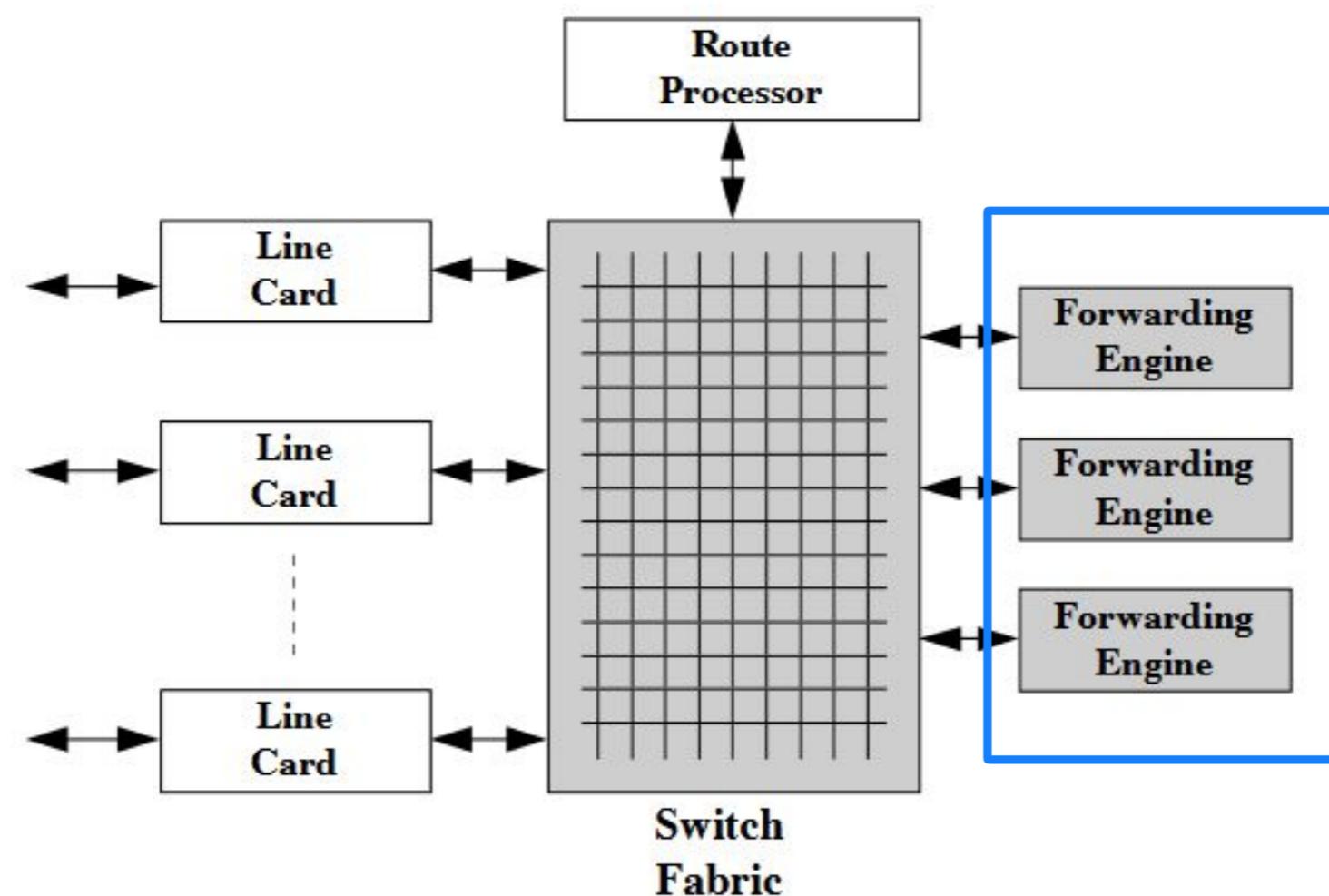
1. síťová karta oddělí hlavičku a pošle ji do FE
2. FE určí z hlavičky výstupní rozhraní pro paket
3. paket uložený v buffer vstupního rozhraní je pak přesunut na výstupní rozhraní

Předpokládá se, že ne všechny porty jsou vždy maximálně zatíženy

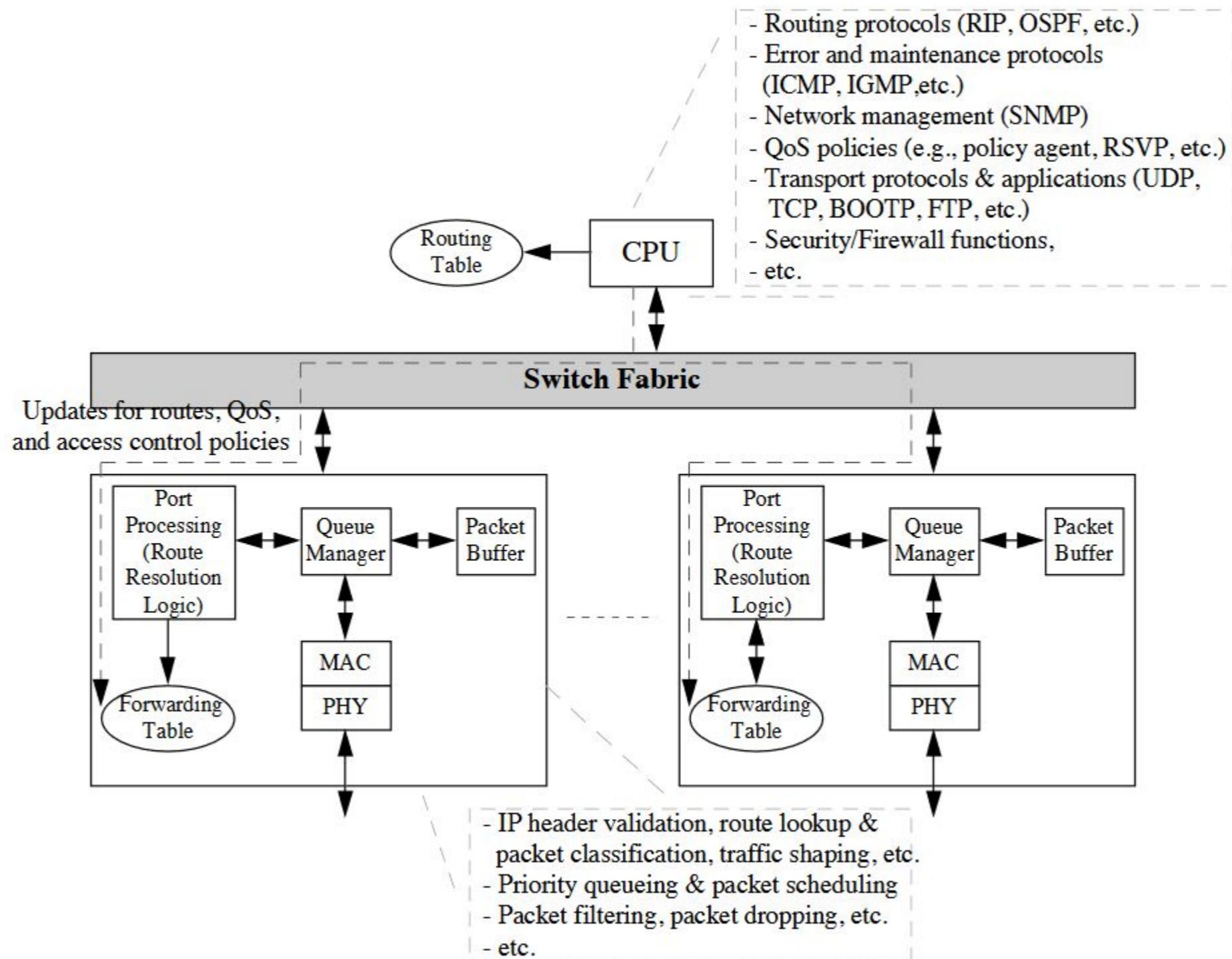


Přepínač s více procesory

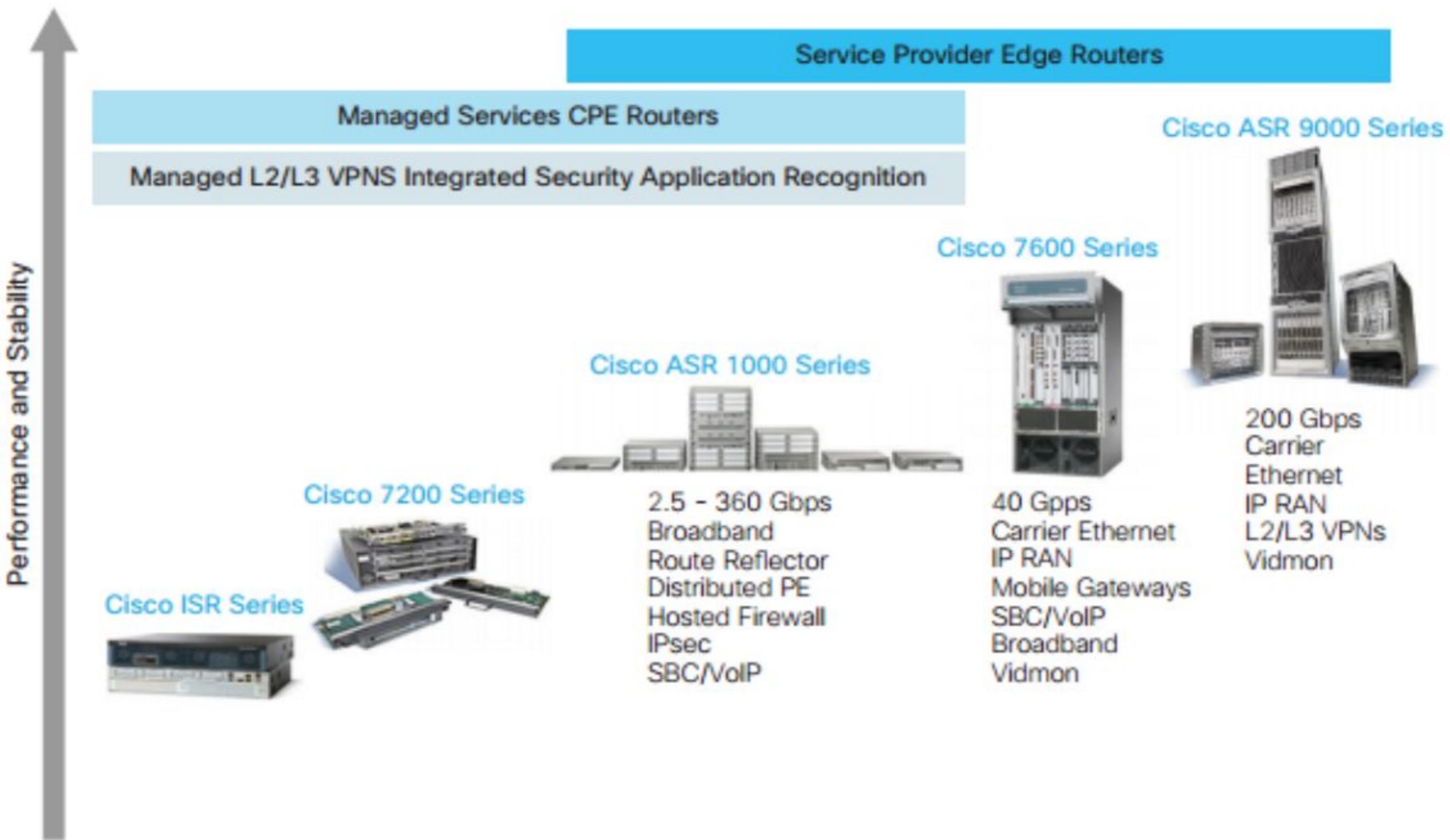
- podobně jako u sběrnice s paralelním zpracováním, nicméně větší rychlosť přenosu mezi vnitřními komponentami
- fast-path: informace je nalezena v cache FE
- slow-path: při výpadku v cache je nutné najít informaci v hlavní směrovací tabulce
- velikost cache je limitujícím faktorem, enterprise směrovače mohou “obsluhovat” statisíce aktivních toků



Distribuovaná architektura



Jaký router potřebuji?

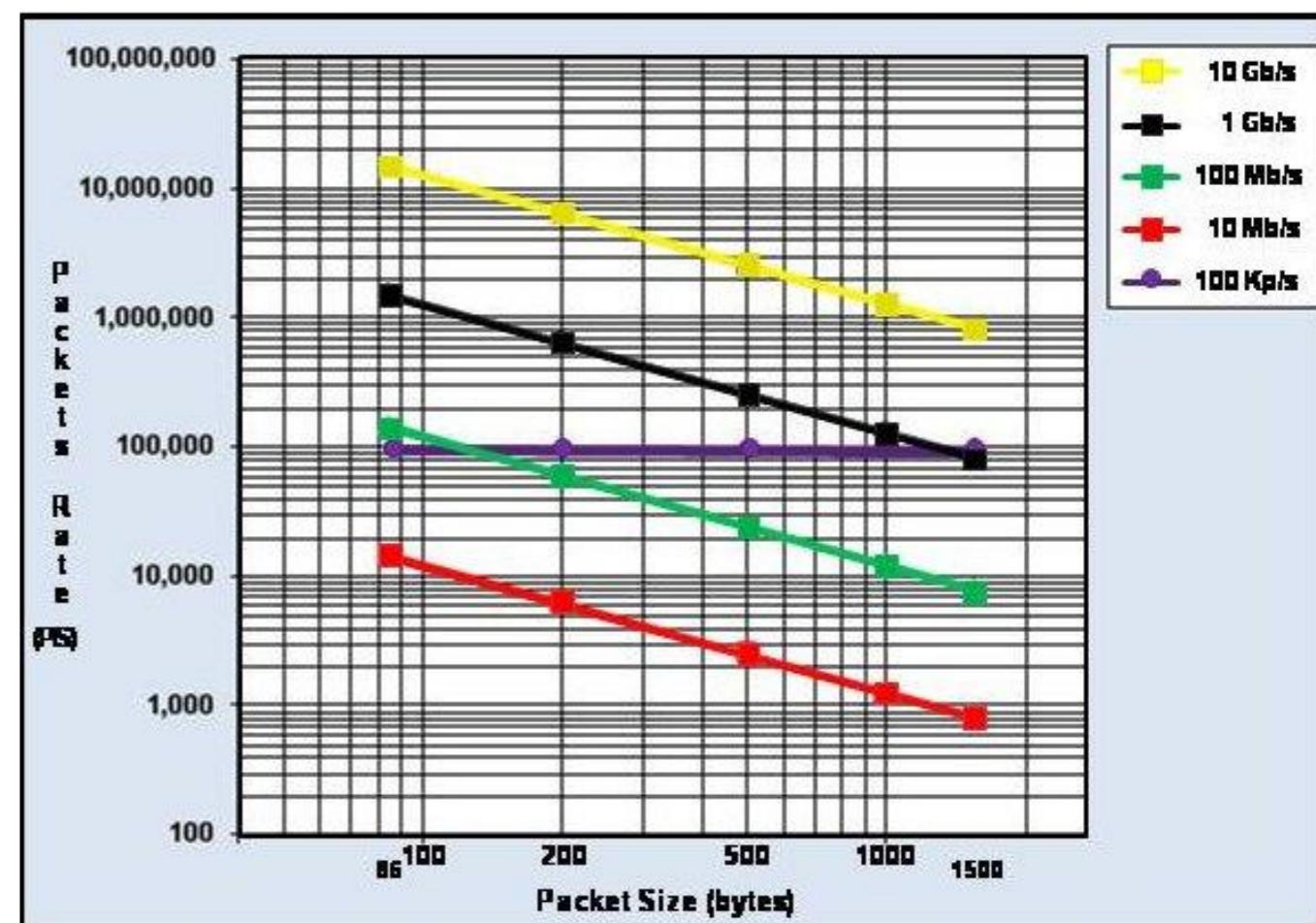


Performance Metrics

- Někdo nám tvrdí:
 - „Cisco ASR 1000 Series Router, is capable of forwarding packets at up to 16 Mp/s with services enabled, it can support the processing of the equivalent of 10 Gb/s of traffic at line rate?“
 - Je to hodně nebo málo?
- Maximum Frame Rate
(Minimum Frame Size)
- Maximum Throughput
(Maximum Frame Size)

$[1,000,000,000 \text{ b/s} / (84 \text{ B} * 8 \text{ b/B})] == 1,488,096 \text{ f/s (maximum rate)}$

$[1,000,000,000 \text{ b/s} / (1,538 \text{ B} * 8 \text{ b/B})] == 81,274 \text{ f/s (minimum rate)}$



A kolik mě to bude stát?



Cisco Linksys E4200 v2 Maximum Performance Dual-Band N900 router

Part Number: E4200V2

\$188.49

MSRP: \$199.99

(-6%)



Cisco ISR 4431 - router - rack-mountable

Part Number: ISR4431/K9

\$5,164.11 to \$5,374.99



Cisco 7604 - router - desktop, rack-mountable - with Cisco 7600 Series Route Switch Processor 720 with 10 Gigabit Ethernet (RSP720-3CXL-10GE)

Part Number: 7604-RSP7XL-10G-P

1 Related Model

MSRP: \$54,000.00

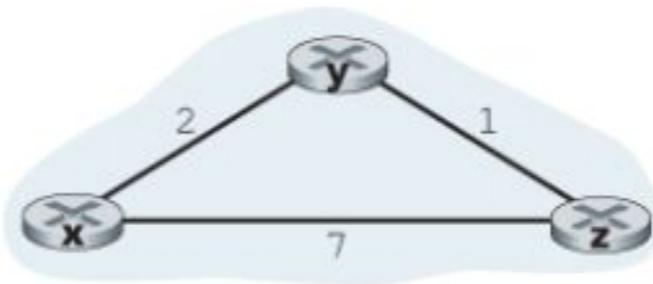
Směrovací algoritmy

Bellman-Fordův algoritmus

- Používá se u distance-vector protokolů
- Každý uzel komunikuje se svými sousedy (*routing by rumor*)
- Každý uzel počítá na základě dostupných informací vlastní nejkratší cestu k cíli *Klínovitá káčka měří toutatně výplň a RT*
- Vlastnosti
 - Iterativní - Běží pokud jsou nové informace vyměňovány mezi sousedy
 - Asynchronní - Není potřeba synchronizace mezi uzly pro zajištění správnosti výpočtu
 - Reaktivní - změna ceny lokální linky a zpráva od sousedního uzlu s aktualizací *když se něco změní posílí svědectví*
 - Distribuovaný
 - Každý uzel oznamuje změnu (hodnotu Distance Vector)
 - Změny mohou způsobit změnu na sousedním uzlu, který ji potom šíří dále

Algorithmus

```
1 Initialization:
2     for all destinations y in N:
3          $D_x(y) = c(x,y)$  /* if y is not a neighbor then  $c(x,y) = \infty$  */
4     for each neighbor w
5          $D_w(y) = ?$  for all destinations y in N
6     for each neighbor w
7         send distance vector  $\mathbf{D}_x = [D_x(y): y \text{ in } N]$  to w
8
9 loop
10    wait (until I see a link cost change to some neighbor w or
11        until I receive a distance vector from some neighbor w)
12
13    for each y in N:
14         $D_x(y) = \min_v\{c(x,v) + D_v(y)\}$ 
15
16    if  $D_x(y)$  changed for any destination y
17        send distance vector  $\mathbf{D}_x = [D_x(y): y \text{ in } N]$  to all neighbors
18
19 forever
```



Node x table

		cost to		
		x	y	z
from	x	0	2	7
	y	∞	∞	∞
		cost to		
		x	y	z
from	x	0	2	3
	y	2	0	1
		cost to		
		x	y	z
from	x	0	2	3
	y	2	0	1
from	x	0	2	3
	y	2	0	1
from	x	0	2	3
	y	3	1	0
from	x	0	2	3
	y	3	1	0

Node y table

		cost to		
		x	y	z
from	x	∞	∞	∞
	y	2	0	1
		cost to		
		x	y	z
from	x	0	2	7
	y	2	0	1
		cost to		
		x	y	z
from	x	0	2	3
	y	2	0	1
		cost to		
		x	y	z
from	x	0	2	3
	y	2	0	1
		cost to		
		x	y	z
from	x	0	2	3
	y	3	1	0
		cost to		
		x	y	z
from	x	0	2	3
	y	3	1	0

Node z table

		cost to		
		x	y	z
from	x	∞	∞	∞
	y	∞	∞	∞
		cost to		
		x	y	z
from	x	0	2	7
	y	2	0	1
		cost to		
		x	y	z
from	x	0	2	3
	y	2	0	1
		cost to		
		x	y	z
from	x	0	2	3
	y	3	1	0
		cost to		
		x	y	z
from	x	0	2	3
	y	3	1	0

Time

Init. Shift.

x	x	5	2
x	<u>0</u>	<u>2</u>	<u>7</u>
5	-	-	-
2	-	-	-

x	x	5	2
x	<u>0</u>	<u>2</u>	<u>7</u>
5	<u>2</u>	$0+2$	$7+2$
2	<u>7</u>	<u>7</u>	<u>0</u>

x	x	5	2
x	<u>0</u>	<u>2</u>	<u>7</u>
5	<u>2</u>	<u>0</u>	<u>7</u>
2	<u>7</u>	<u>0</u>	<u>7</u>

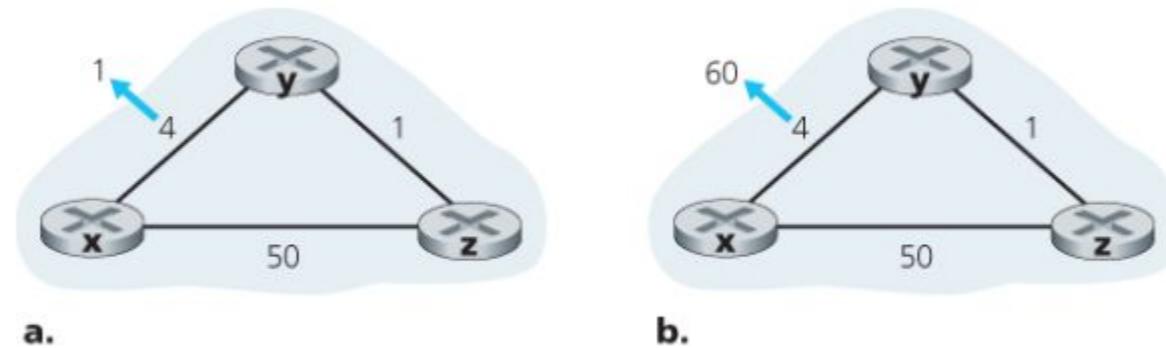
x	x	5	2
x	-	-	-
5	<u>2</u>	<u>0</u>	<u>7</u>
2	-	-	-

x	x	5	2
x	-	-	-
5	-	-	-
2	<u>7</u>	<u>1</u>	<u>0</u>

x	x	5	2
x	<u>0</u>	<u>2</u>	<u>7</u>
5	<u>2</u>	<u>0</u>	<u>7</u>
2	<u>7</u>	<u>1</u>	<u>0</u>

lempisio 3c7

Změna linky



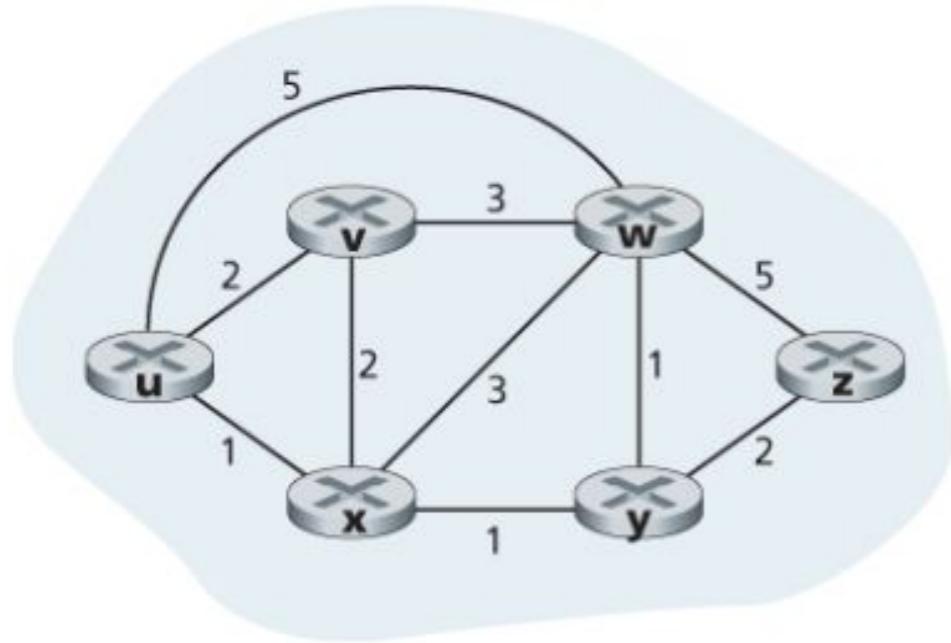
Domácí úkol – vyzkoušejte si co se stane...

Dijkstrův algoritmus

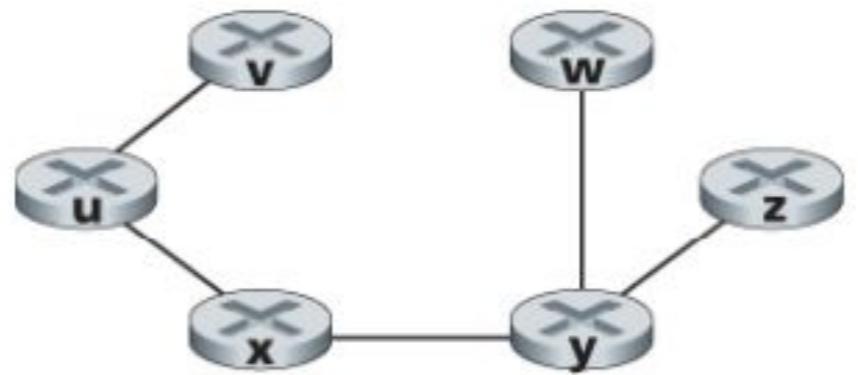
- Používá se u link-state protokolů
- Předpoklady
 - Topologie sítě včetně ceny všech linek je předem známa každému uzlu
 - Všechny uzly mají stejnou informaci o stavu sítě
- Každý směrovač vypočítá cestu s nejnižší cenou pro každou cílovou síť
 - Začátek cesty je vždy aktuální směrovač
 - Iterativní výpočet

Algorithmus

```
1 Initialization:  
2   N' = {u}  
3   for all nodes v  
4     if v is a neighbor of u  
5       then D(v) = c(u,v)  
6     else D(v) = ∞  
7  
8 Loop  
9   find w not in N' such that D(w) is a minimum  
10  add w to N'  
11  update D(v) for each neighbor v of w and not in N':  
12    D(v) = min( D(v), D(w) + c(w,v) )  
13  /* new cost to v is either old cost to v or known  
14    least path cost to w plus cost from w to v */  
15 until N' = N
```



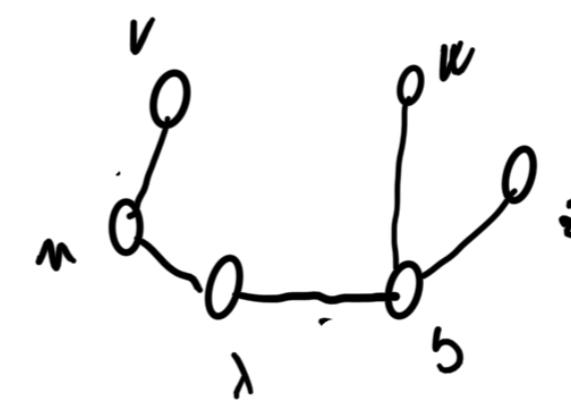
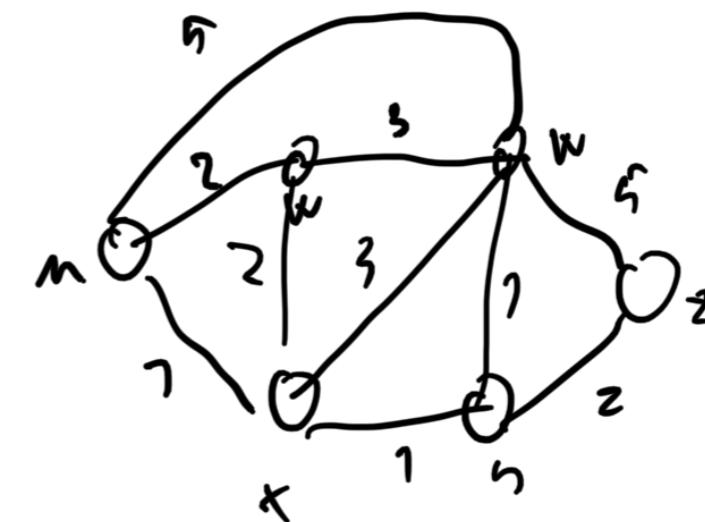
step	N'	$D(v), p(v)$	$D(w), p(w)$	$D(x), p(x)$	$D(y), p(y)$	$D(z), p(z)$
0	u	2, u	5, u	1, u	∞	∞
1	ux	2, u	4, x		2, x	∞
2	uxy	2, u	3, y			4, y
3	uxyv		3, y			4, y
4	uxyvw					4, y
5	uxyvwz					



Destination	Link
v	(u, v)
w	(u, x)
x	(u, x)
y	(u, x)
z	(u, x)

(n)

	N	V	W	X	g	z
M	2, m	5, m	7, u	-	-	
m, x	2, m	9, x	7, v	2, x	-	
m, x, g	2, m	3, g	7, m	2, x	9, g	
m, x, g, v	2, m	3, g	7, m	2, x	9, g	
m, x, g, v, w	2, m	3, g	7, u	2, x	9, g	



V	m, v
W	v, x
X	v, x
g	v, x
z	v, x

Porovnání LS a DV

Složitost

- LS musí šířit zprávy záplavou - počet zpráv je $N \cdot E$
- DV šíří zprávy pouze mezi sousedy, několik iterací

Rychlosť konvergencie

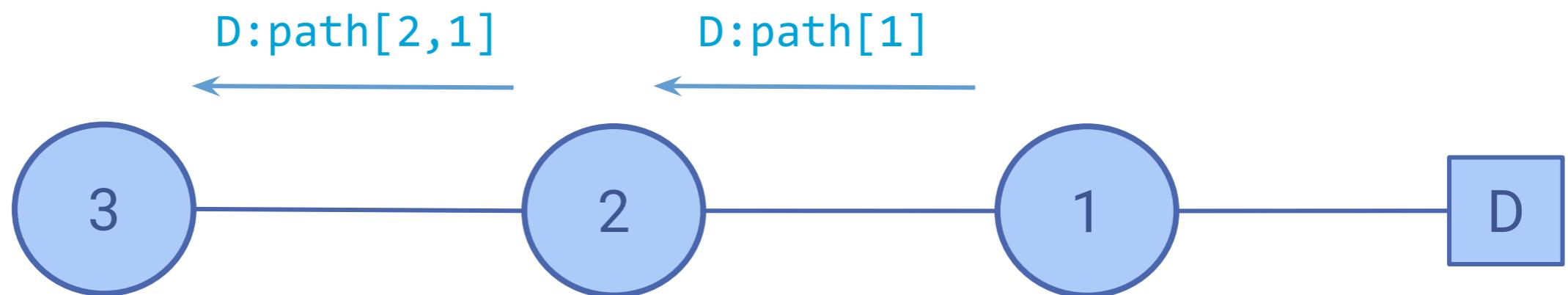
- LS je kvadratický algoritmus
- DV může konvergovat pomalu a trpí na počítání do nekonečna (částečně řešeno pomocí mechanismu poisson reverse)

Robustnosť

- každý uzel provádí svůj výpočet, nicméně je zde pořád možnost šíření nesprávné informace
- uzly spoléhají na informace od sousedů

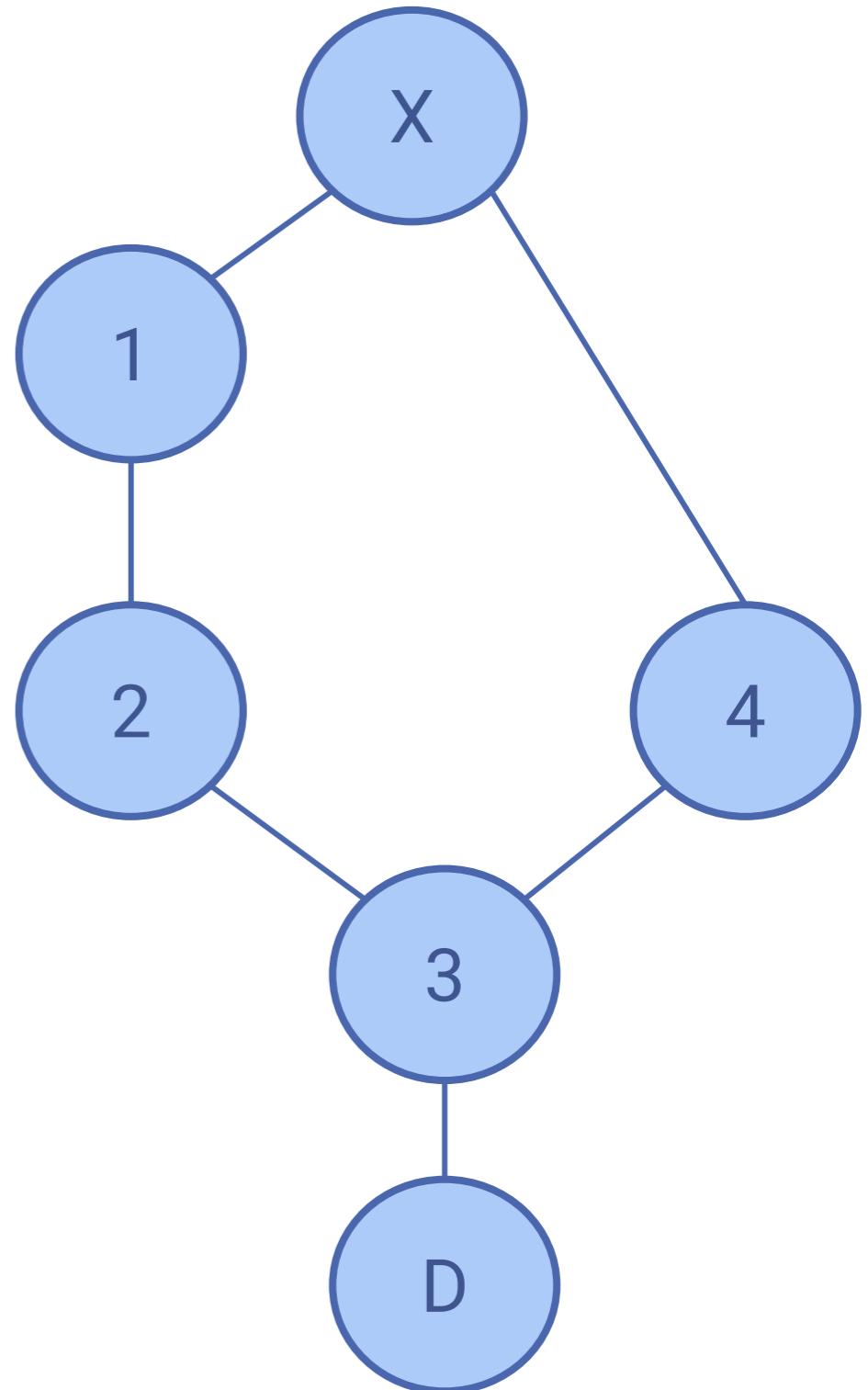
Path Vector směrování

- Základní myšlenka \leftrightarrow oznamovat celou cestu ke koncové síti (místo vzdálenosti)
 - DV: má pro každý cíl D jeho vzdálenost
 - PV: má pro každý cíl D jeho celou cestu
 - je jednoduché detekovat a eliminovat smyčky

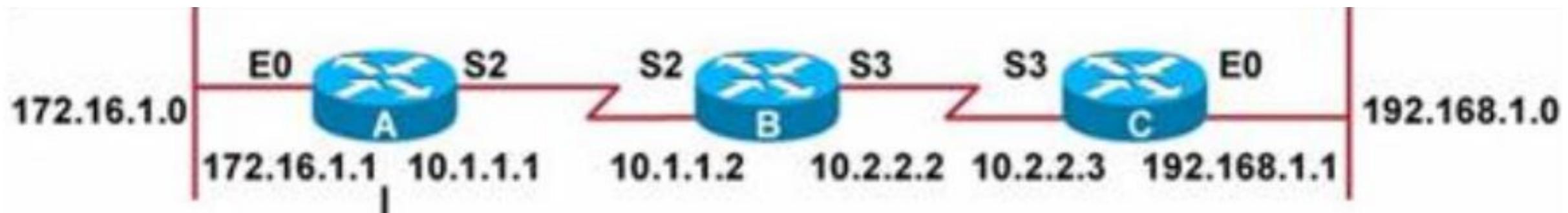


Flexibilní politiky

- Každý uzel může aplikovat lokální politiku
 - Výběr cesty
 - Oznamování cesty
- Například:
 - AS X preferuje [1,2,3] místo [4,3]
 - AS 3 nechce říct AS 2 o cestě [4,X]
- Použito v BGP
 - Path odpovídá AS
 - Politiky odpovídají smlouvám mezi ISP



Co je tedy v RT?



```
RouterA#show ip route
Codes: C - connected, S - static, I - IGRP, R - RIP, M - mobile, B - BGP
      D - EIGRP, EX - EIGRP external, O - OSPF, IA - OSPF inter area
      N1 - OSPF NSSA external type 1, N2 - OSPF NSSA external type 2
      E1 - OSPF external type 1, E2 - OSPF external type 2, E - EGP
      i - IS-IS, L1 - IS-IS level-1, L2 - IS-IS level-2, * - candidate
      default
      U - per-user static route, o - ODR
      T - traffic engineered route

Gateway of last resort is not set

      172.16.0.0/24 is subnetted, 1 subnets
C        172.16.1.0 is directly connected, Ethernet0
      10.0.0.0/24 is subnetted, 2 subnets
R        10.2.2.0 [120/1] via 10.1.1.2, 00:00:07, Serial2
C        10.1.1.0 is directly connected, Serial2
R        192.168.1.0/24 [120/2] via 10.1.1.2, 00:00:07, Serial2
```

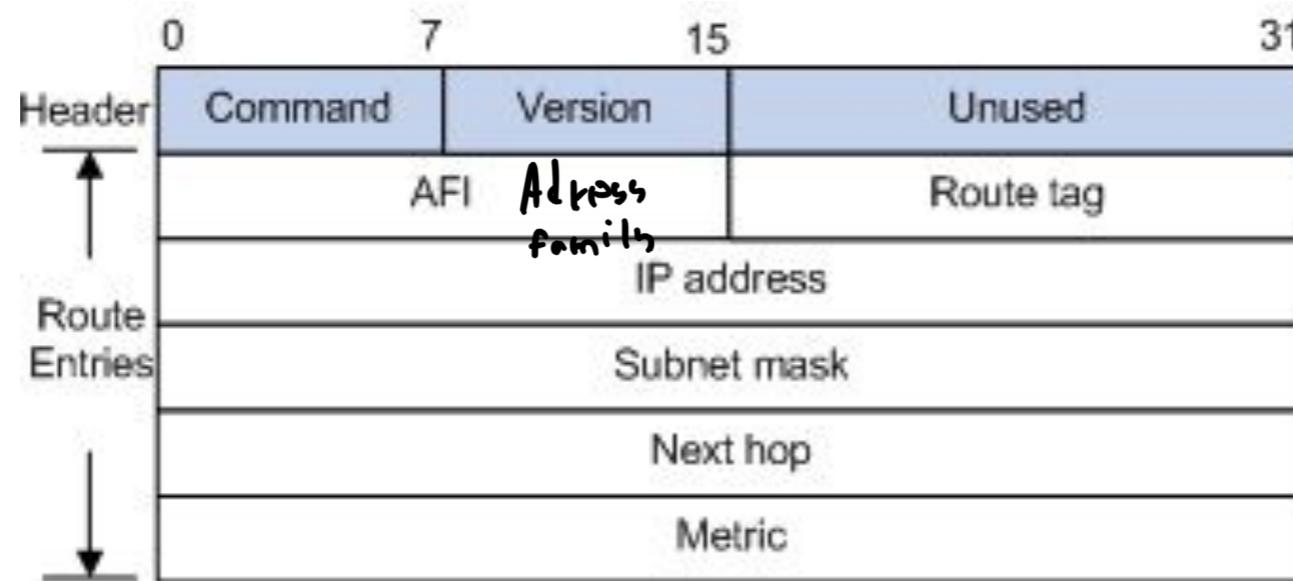
Směrovací protokoly

Protokol RIP

- Principy definovány v roce 1969 pro ARPANET and CYCLADES
- V 70-tých letech použit v Xerox PUP sítích a posléze v Xerox Network System jako XNS RIP
- XNS RIP se poté uplatnil v IPX RIP, AppleTalk RTMP and IP RIP
- 1982 byl RIP implementován v BSD UNIXu jako routed daemon
- 1988 byl vydán standard RFC 1058 (Charles Hedrick)

Charakteristika

- Classful (verze 1), classless (verze 2)
- Metrikou je počet skoků (hop-count)
- Hop-count > 15 označuje nedosažitelnou cestu
- Periodické aktualizace každých 30s
- RIP zprávy jsou neseny v UDP datagramu a posílány broadcastem na port 520



RIP zprávy

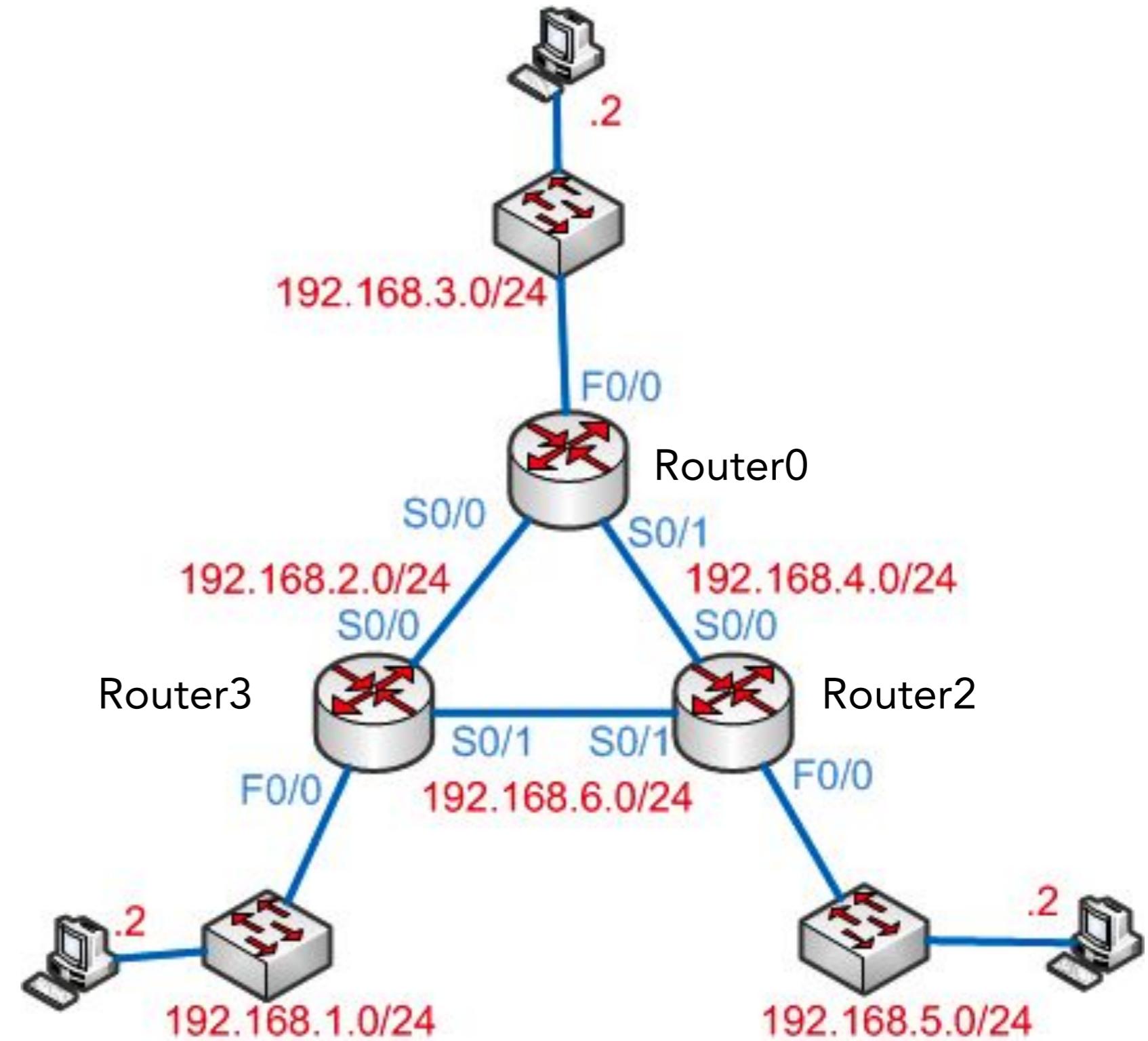
- Response message
 - Až 25 směrovacích záznamů
 - Periodické aktualizace
 - Spuštěné aktualizace
 - Odpověď na Request message
- Request message
 - Zasíláno při spuštění směrovače
 - Požadavek na zaslání kompletní či částečné směrovací tabulky

Aktualizace stavu linky

- Jestliže nepřijde po dobu 180s aktualizace od souseda je označen za neplatný
 - Cesty vedoucí přes tohoto souseda jsou zneplatněny
 - Jsou informováni sousedé
 - Sousedé aktualizují svá data a pokud došlo ke změně ve směrovací tabulce posílají oznámení
- Selhání linky/směrovače je takto šířeno celou sítí
- Poison Reverse je použit pro zabránění vytváření směrovacích smyček

RIP Demo

```
debug ip rip  
router rip  
network  
show ip route  
show ip rip database
```



RIP Packets

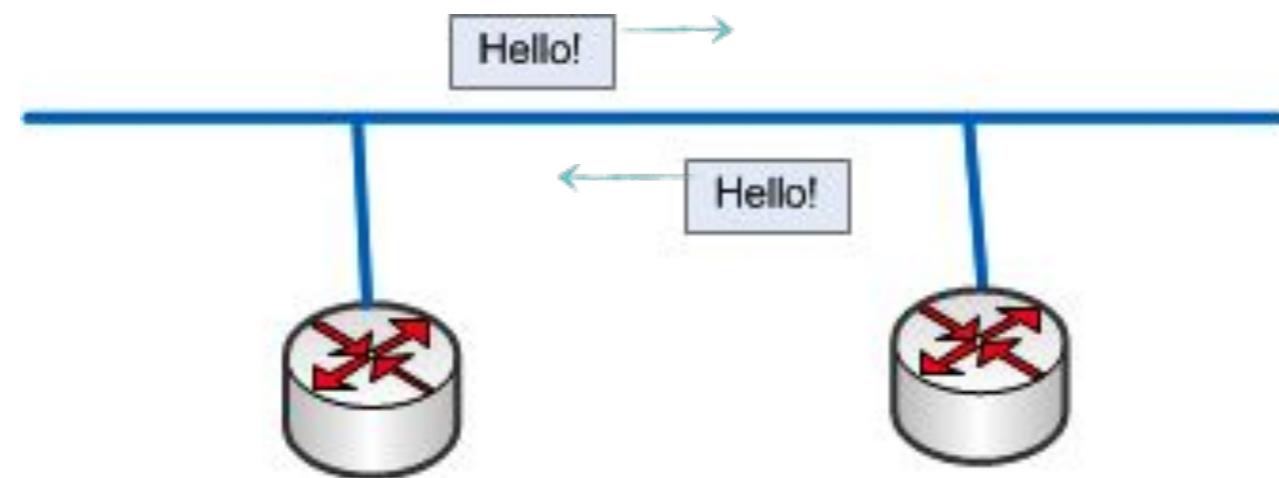
- RIPv1 Updates
<https://www.cloudshark.org/captures/00d58e1f4dd5>
- RIPv2 Updates
<https://www.cloudshark.org/captures/00bdca4b449a>
- RIPv2 Unreachable Update
<https://www.cloudshark.org/captures/016c88c0e465>

OSPF

- Open Shortest Path First
- Link state protocol
- Šíří informaci o změně v celé oblasti (LSA flooding)
- Dijkstrův algoritmus pro výpočet nejlepší cesty
- Ceny linek definuje administrátor (hop-count, bandwidth)
- Informace jsou posílány okamžitě v případě změny, či alespoň jednou za 30 minut.
- OSPF protokol je nesen IP protokolem (má číslo 89)
- V režii OSPF protokolu je zajištění spolehlivého přenosu a broadcastu
- OSPF testuje stav linky pomocí HELLO paketů

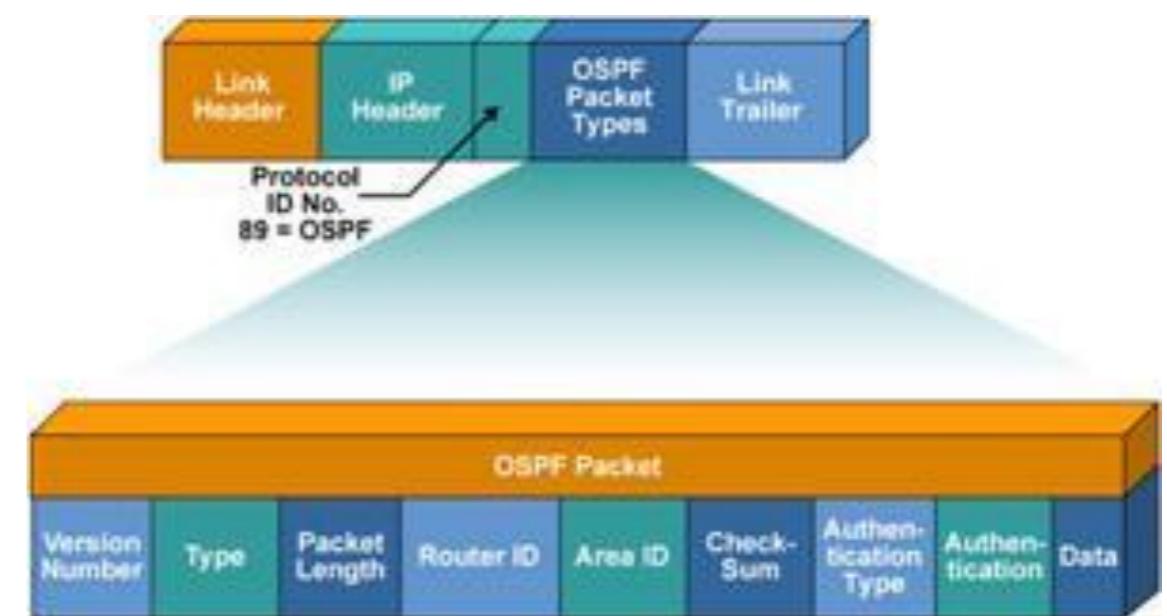
Detekce změny topologie

- Beaconing
 - pravidelné posílání krátkých zpráv oběma směry
 - detekce selhání při ztrátě několika těchto zpráv
- Není ideální
 - rychlosť detekce odpovídá intervalu zpráv
 - režie posílání zpráv bez datového obsahu
 - možnost mylné detekce



Šíření informací

- Používá záplavové šíření (flooding)
 - směrovač posílá LS informace všem sousedům
 - sousedé šíří tyto informace dál
- Vyžaduje spolehlivé šíření informací
 - každý směrovač musí informaci dostat
 - všichni musí mít stejnou informaci
- Používá přímo IP paket pro přenos dat, musí řešit:
 - ztrátu paketů
 - přijetí v jiném pořadí
- Řešení
 - ACK a znovaodesílání
 - sekvenční čísla

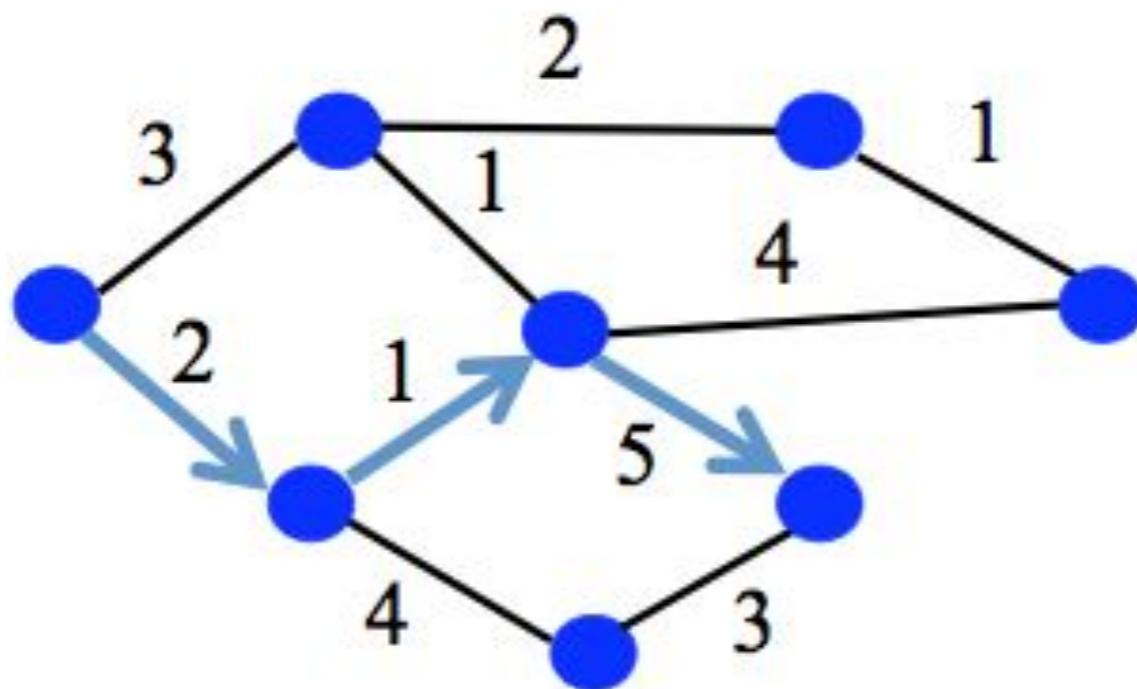


Kdy se posílají aktualizace

- Změna topologie
 - selhání linky/sousedů
 - nalezení linky/sousedů
- Změna konfigurace
 - změna nastavené ceny linky
- Periodicky
 - Většinou každých 30 minut
 - Korekce stavu
 - Pro jistotu, že všichni mají stejné informace

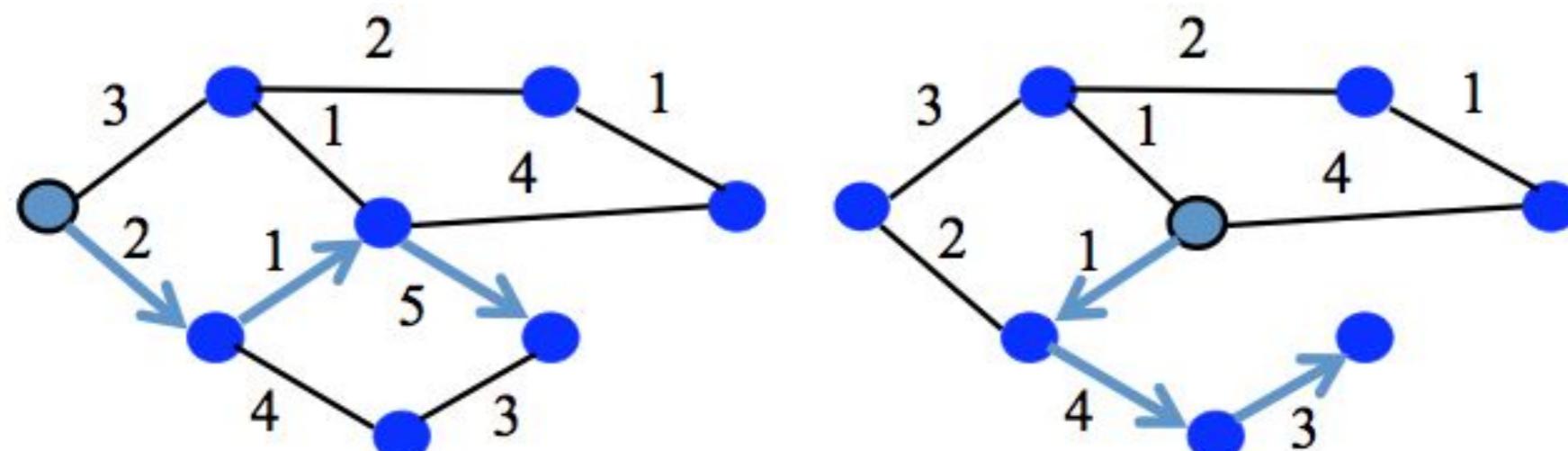
Konvergence

- Konzistentní informace ve všech uzlech
 - všechny uzly mají stejnou LS databázi
- Směrování je konzistentní v konvergovaném stavu
 - všechny uzly mají stejnou LS databázi - stejná topologie
 - pakety jdou nejkratší cestou



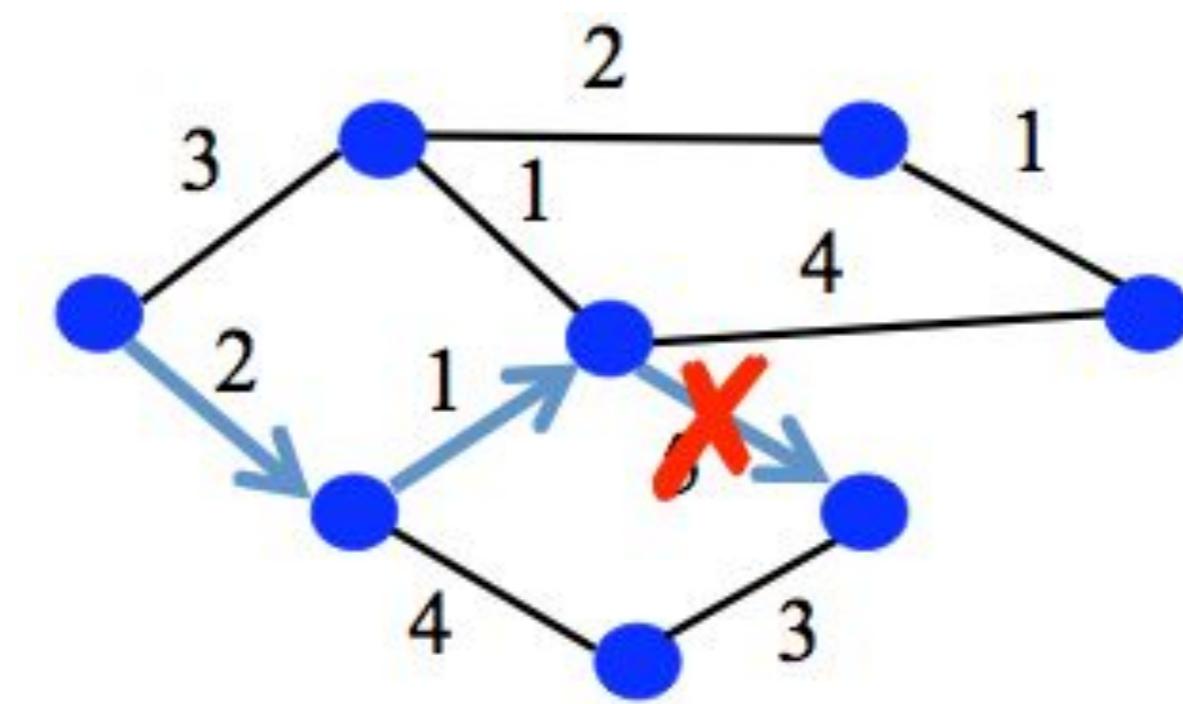
Problémy

- Nekonzistentní databáze
 - některé směrovače mají jinou informaci
 - mohou vědět o selhání zatímco jiné ještě ne
 - může způsobit
 - zvýšenou ztrátovost
 - nemožnost komunikovat
 - dočasnou směrovací smyčku



Problémy

- Opožděná detekce selhání
 - data jsou posílány na nefunkční linku/zařízení
 - cílová síť je nedostupná



Doba konvergence

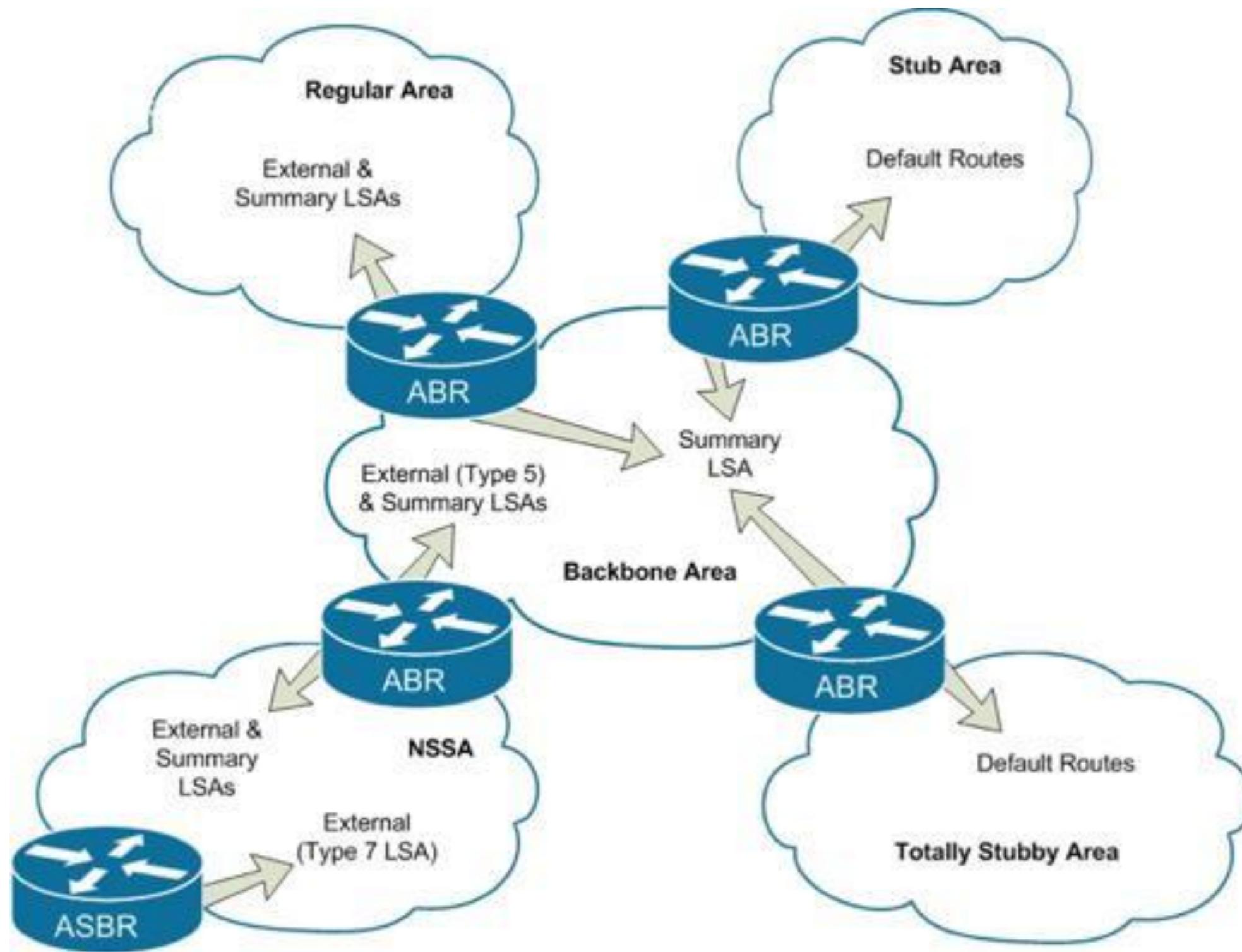
- Zdroje zpoždění
 - detekce
 - záplava LS informací
 - výpočet SP
 - naplnění FT
- Důsledky
 - ztracené pakety
 - nedoručitelné pakety zabírají zdroje
 - pakety mimo pořadí
 - citlivé aplikace (VoIp, video)

Doba konvergence

Opatření

- Rychlejší detekce
 - kratší hello interval
 - detekce na L2
- Rychlejší šíření LS
 - okamžité informování
 - priorita pro LS pakety
- Rychlejší výpočet
 - rychlejší HW
 - inkrementální DA
 - inkrementální změna FT

Škálovatelnost OSPF: Oblasti

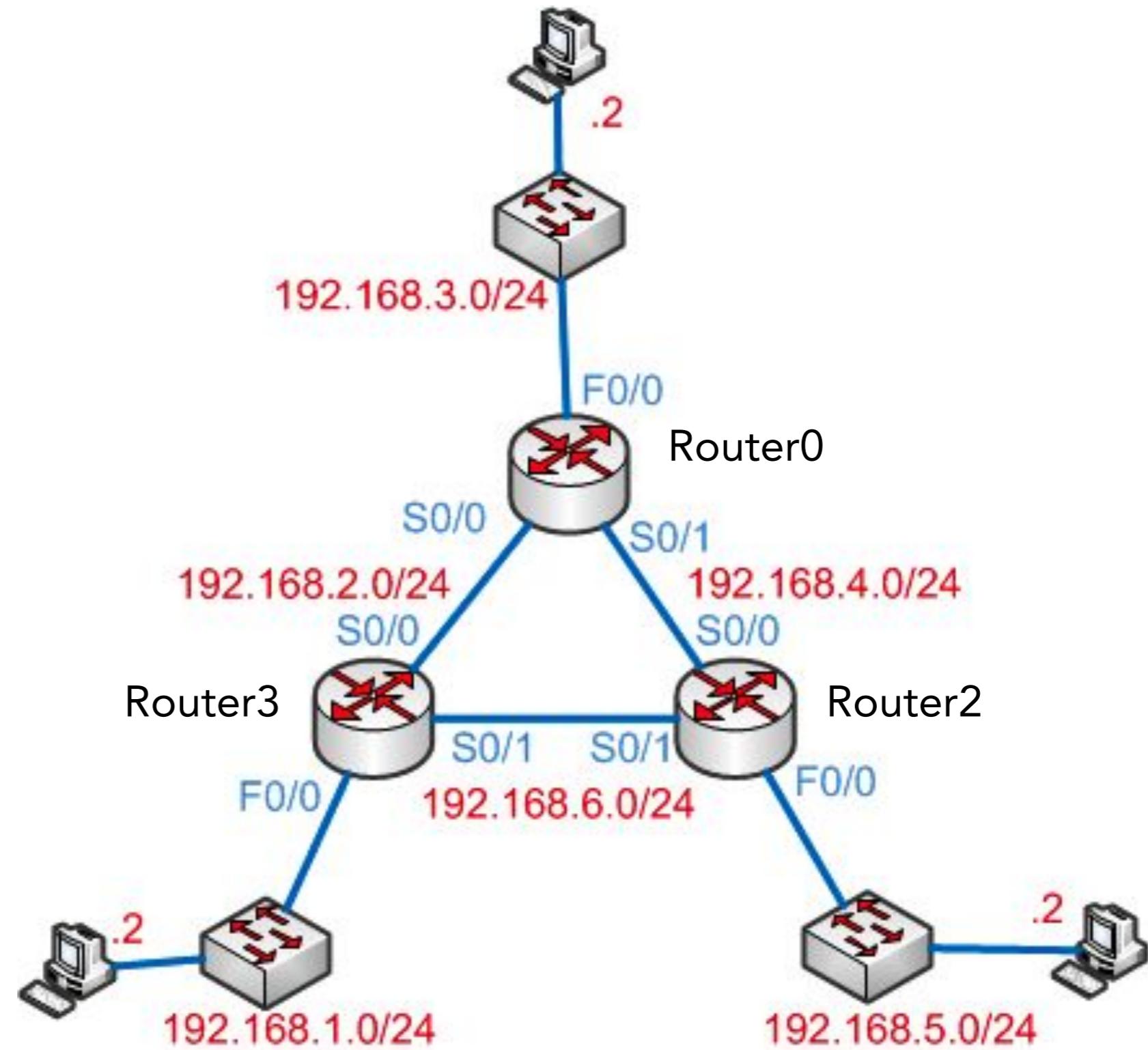


OSPF Vlastnosti

- Bezpečnost
 - Aktualizace mohou být autentizovány
 - Zabránění vkládání nesprávných informací
- Load-balancing
 - Více cest se stejnou cenou může být současně použito pro přenos dat
- Podpora pro multicastové směrování
 - Multicast OSPF
 - Přidává multicast LSA, jinak používá mechanismy OSPF
- Podpora pro intra-autonomous system hierarchické směrování (oblasti)

OSPF Demo

```
debug ip ospf  
router ospf  
network  
show ip route  
show ip ospf database  
show ip ospf neighbors
```



OSPF Messages

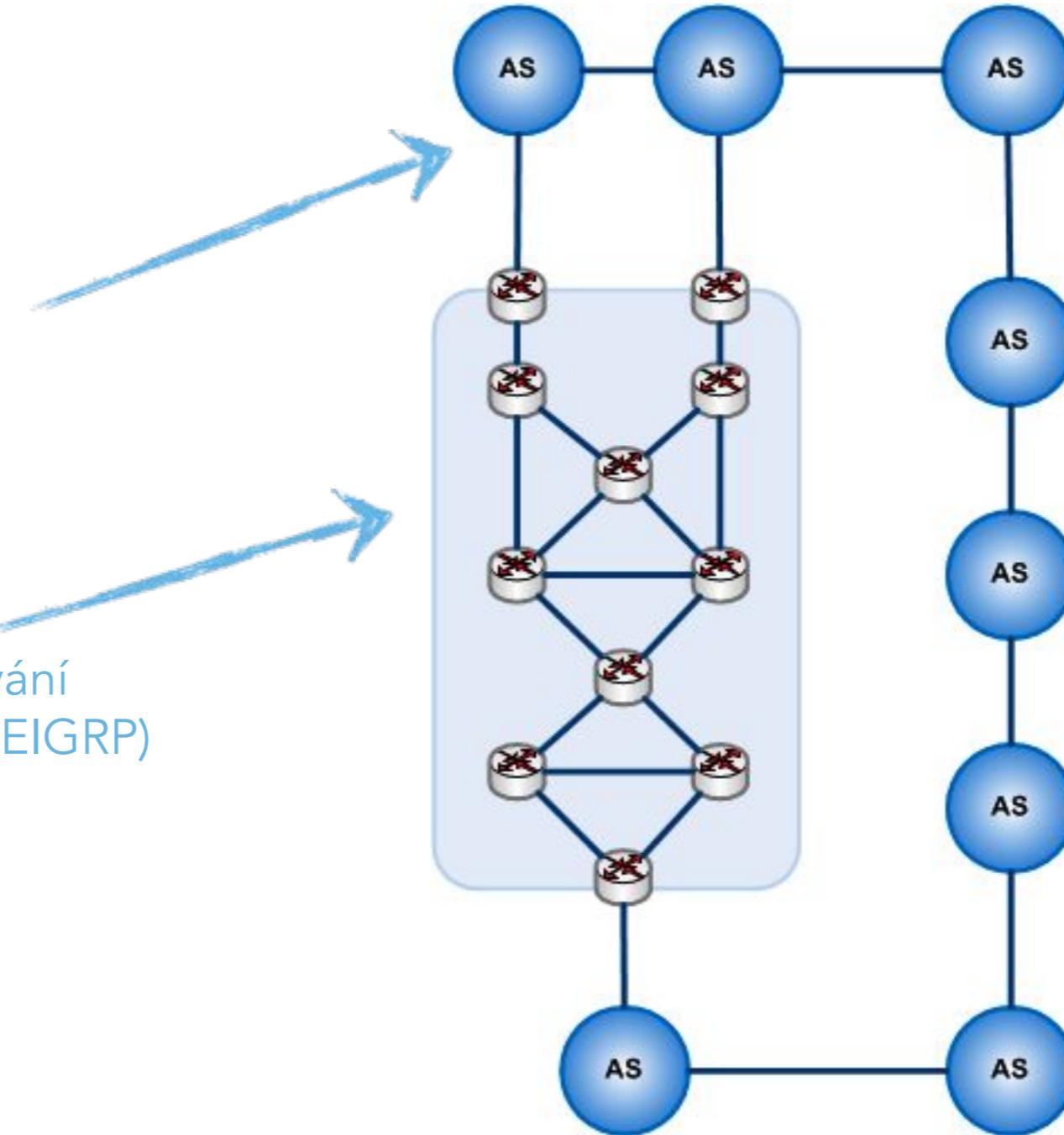
- OSPF LSA Types
<https://www.cloudshark.org/captures/0062204357ab>
- OSPF With Authentication
<https://www.cloudshark.org/captures/007bba156585>

Směrování v Internetu

Směrování v Internetu: 2 vrstvy

Interdomain směrování
mezi domény
(BGPv4)

Doména
IGP směrování
(RIP, OSPF, EIGRP)



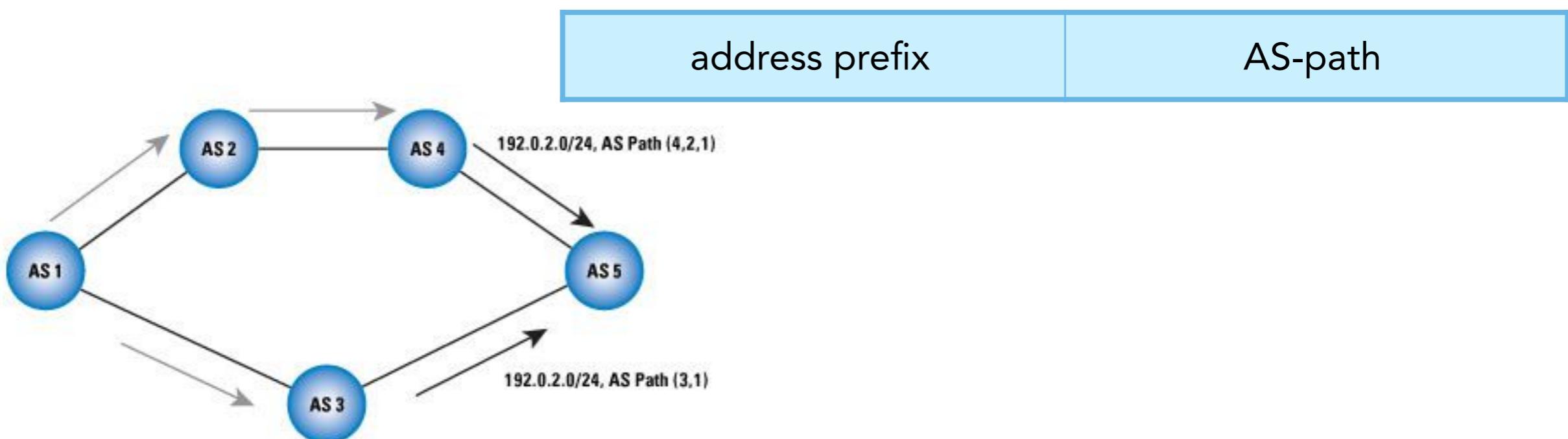
Internet je soubor domén

- Je rozdělen do domén zvaných autonomní systémy
 - oblasti jsou samostatně spravovány
 - sítě spravované jednou institucí
 - poskytovatelé služeb, firmy, univerzity, ...

An AS is a group of IP networks operated by one or more network operator(s) that has a single and clearly defined external routing policy. Exterior routing protocols are used to exchange routing information between ASes.
[RFC1930]

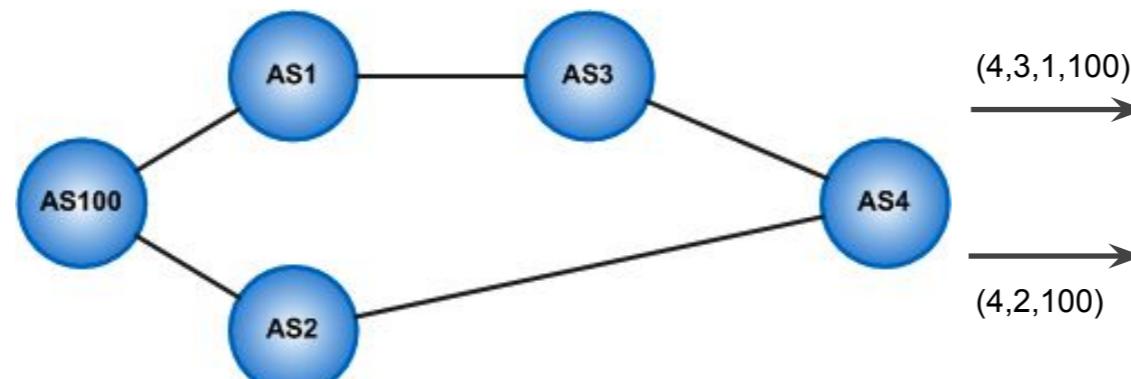
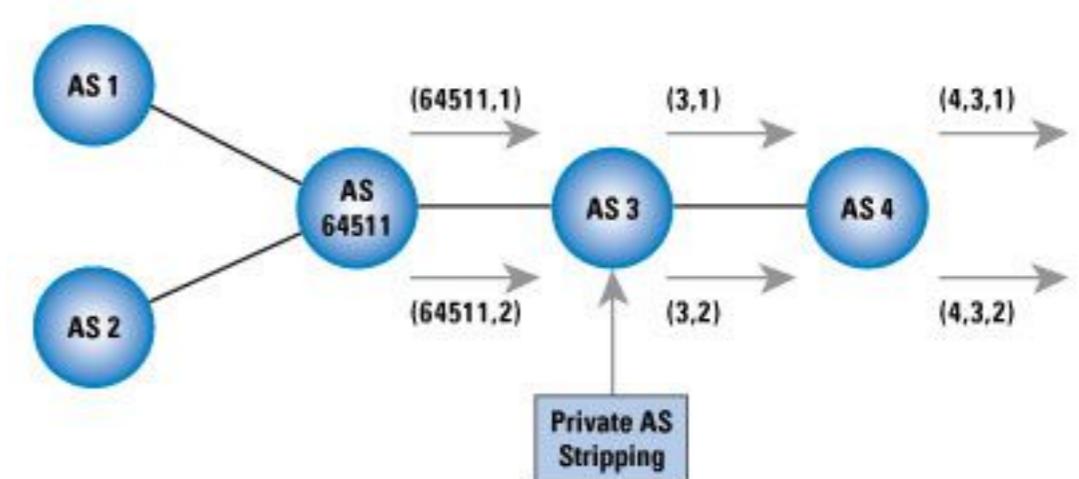
ASN

- AS mají různou velikost, identifikovány pomocí ASN
- ASN je 16 bitové číslo:
 - 1-64511: veřejné AS
 - 64512-65534: privátní AS
- Použití v interdomain směrování:



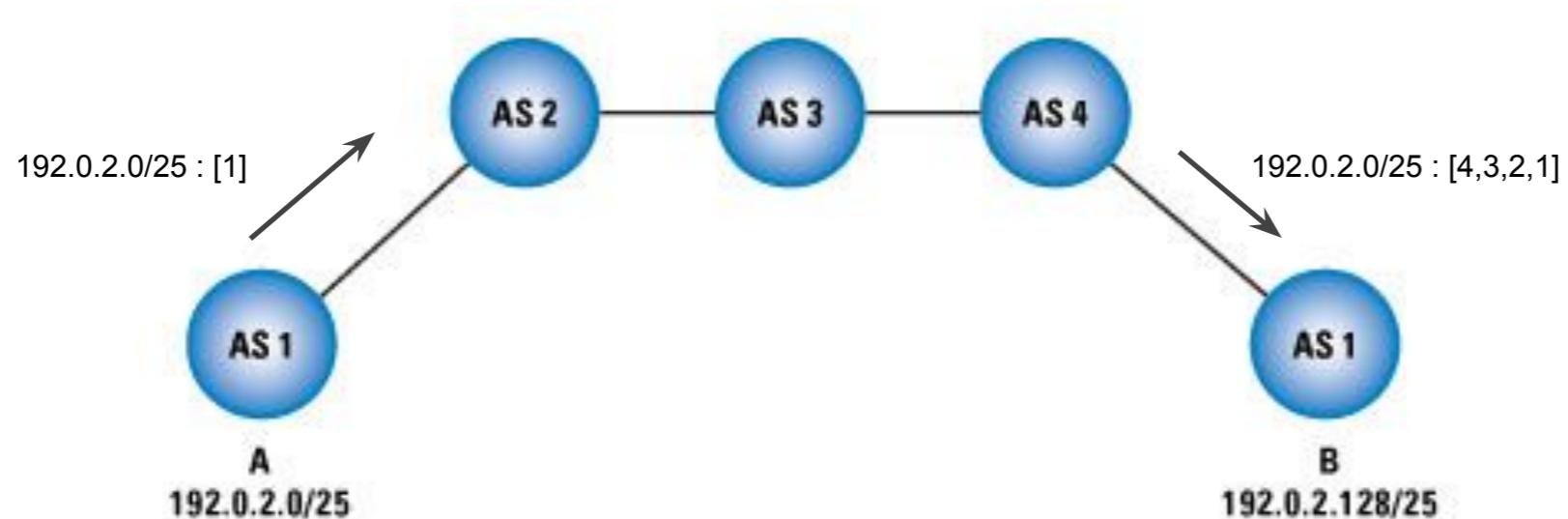
Kdo potřebuje ASN?

- ASN identifikují sítě s rozdílnými směrovacími politikami
- běžní zákazníci mají ASN providerů
- v případě použití BGP lze přiřadit privátní ASN
 - Private ASN Stripping
- veřejné ASN zákazníka
 - multihoming
 - nutnost rozlišit ve směrování více možných



ASN a administrativní domény

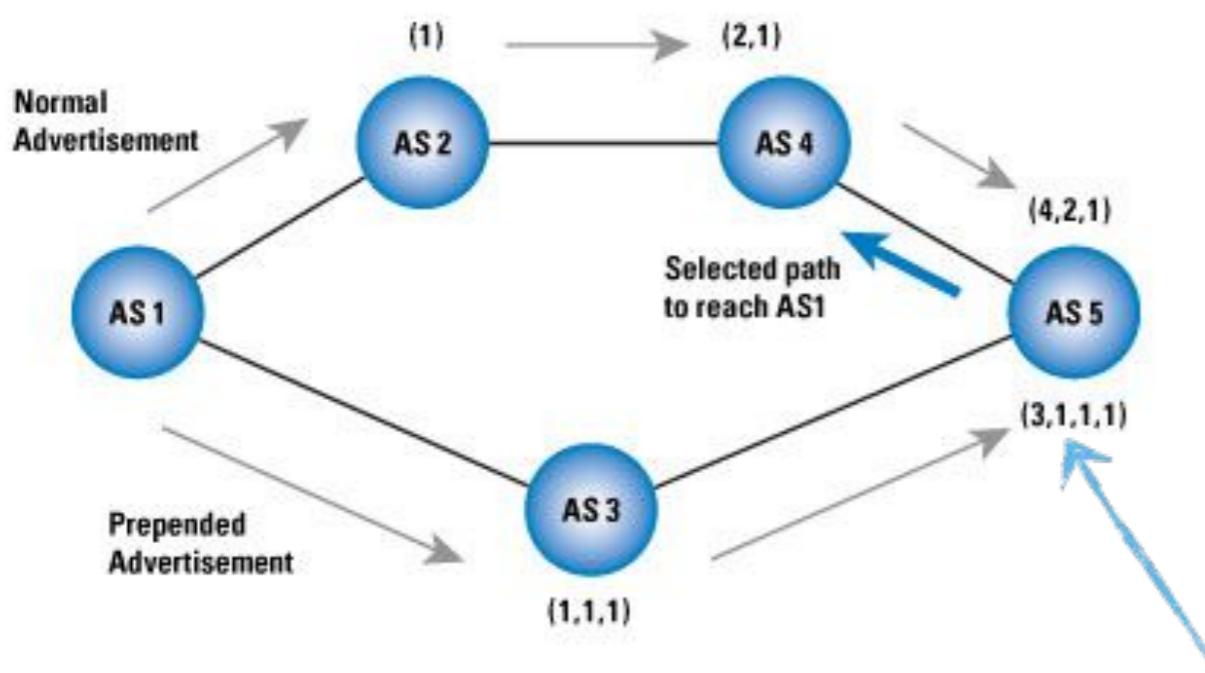
- sítě jedné domény mohou být rozděleny do mnoha lokalit
- problém pro směrování:
 - BGP směrovač v AS1(B) odmítne informaci o 192.0.2.0/25
 - vyžaduje použití staticky nakonfigurovaných cest



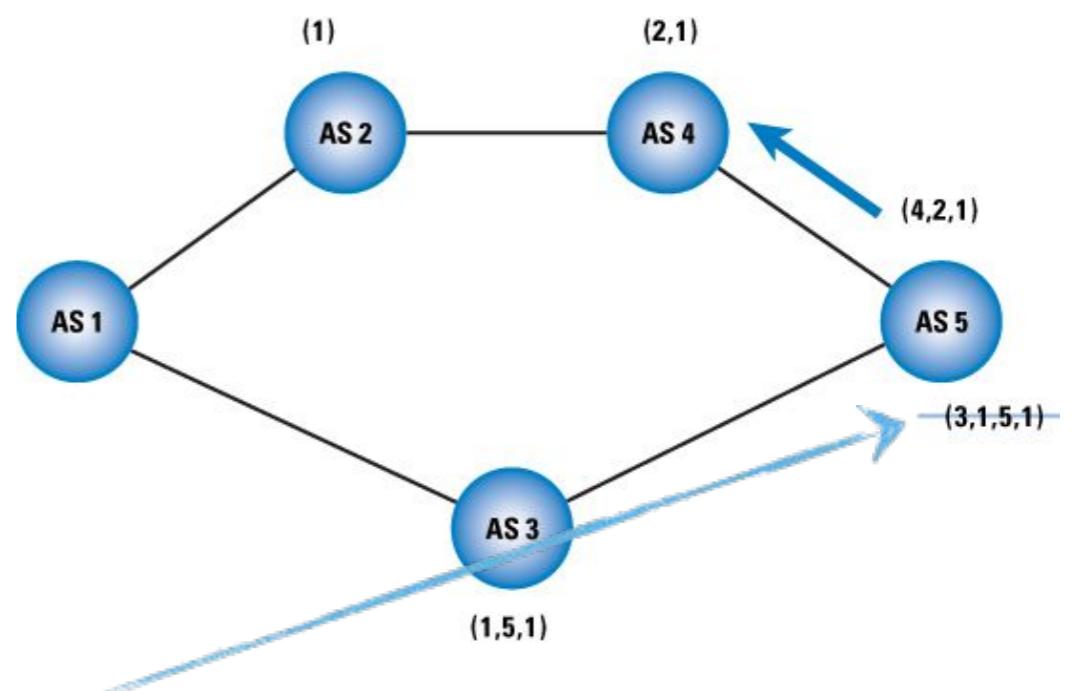
Výběr nejlepší cesty podle ASN

- BGP implicitně vybírá cestu podle nejkratší AS-Path
- Modifikace tohoto chování je možná:
 - jednoduché, ale může způsobit komplikace
 - lepší řešení BGP community

AS path prepending



AS path poisoning



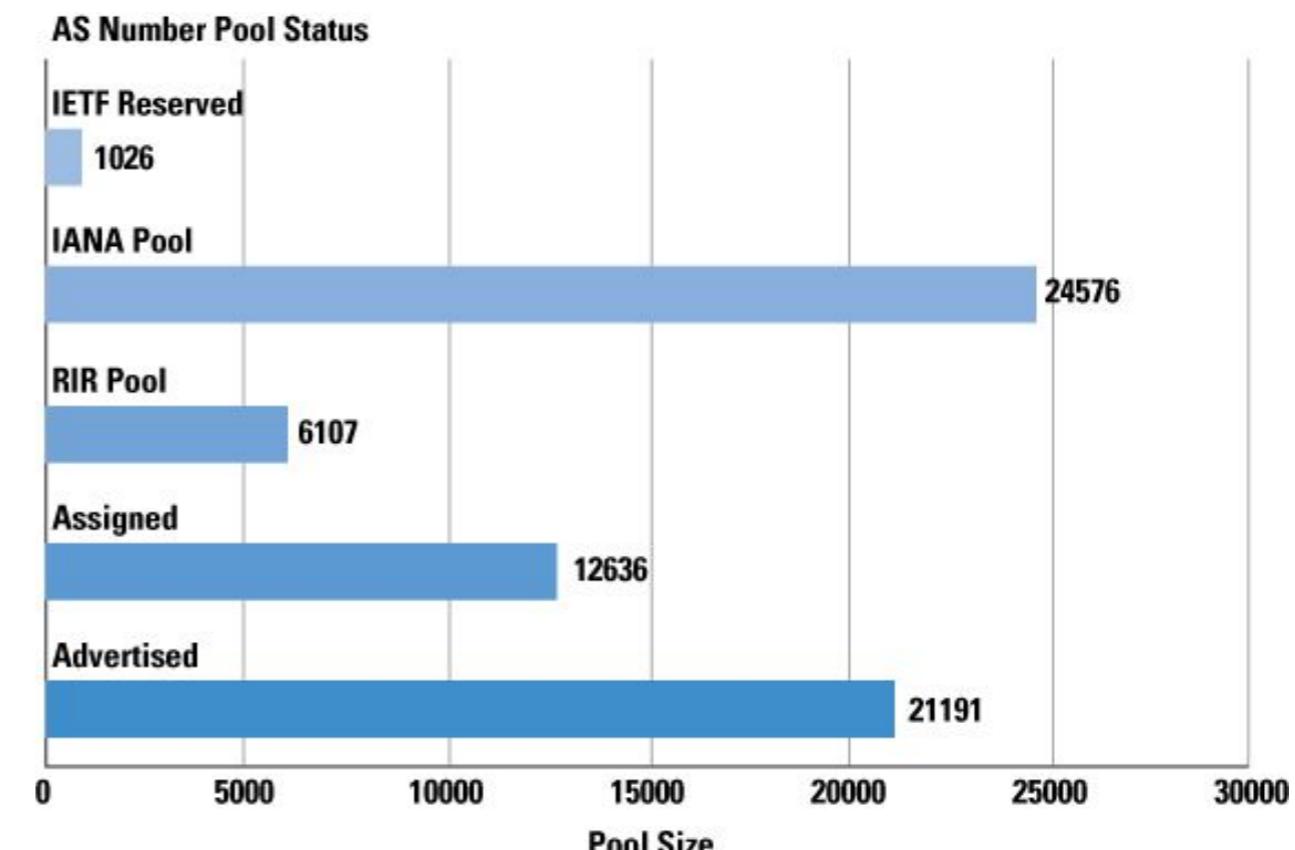
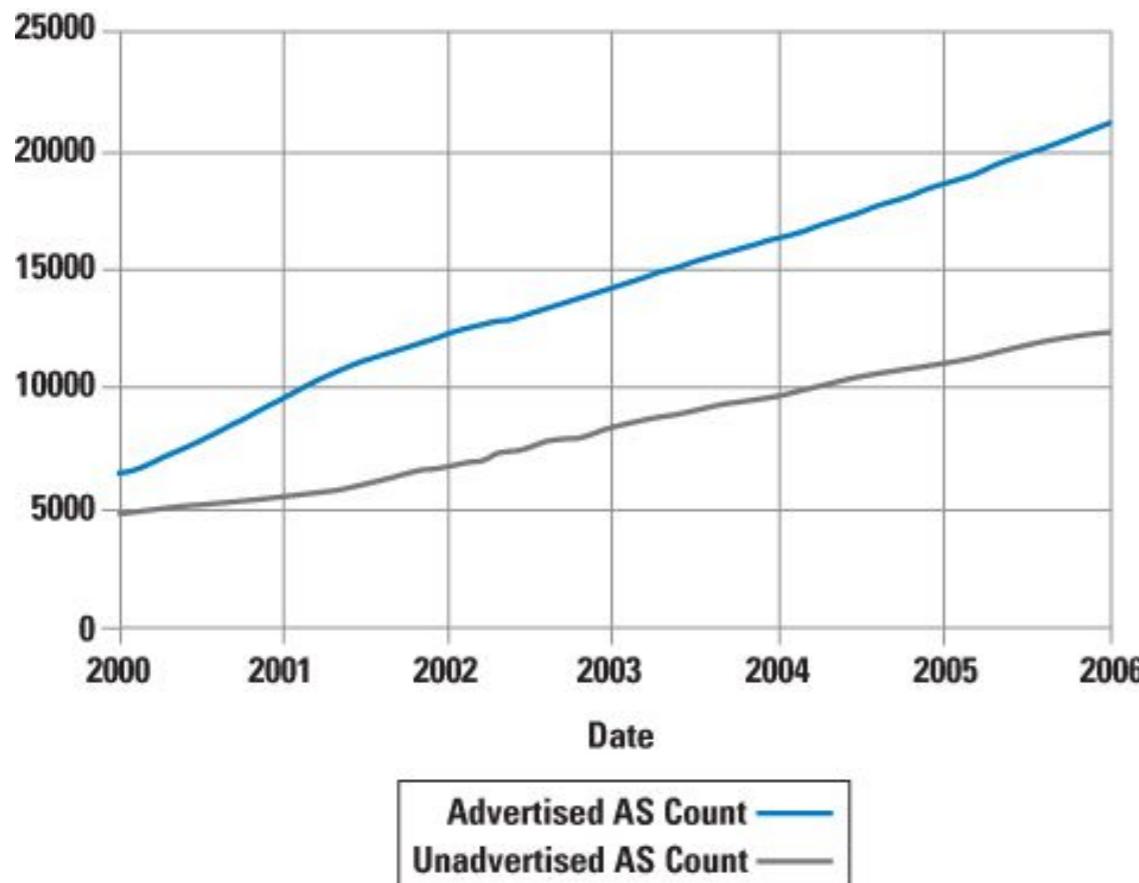
jaký je zde rozdíl?

Využití AS

- počet AS odpovídá počtu aktivních ISP
- neočekává se masivní nárůst počtu aktivních ISP
 - platí ekonomické pravidlo, že větší vyhrává
- nárůst ASN cca 3500 každý rok
 - především způsoben multihomingem
 - použitím v MPLS VPN
 - ISP mají více ASN pro vyjádření různých politik směrování
 - pro privátní klientelu
 - pro firemní klientelu
 - pro různé regiony
 - nefunguje recyklace ASN

ASN Pool

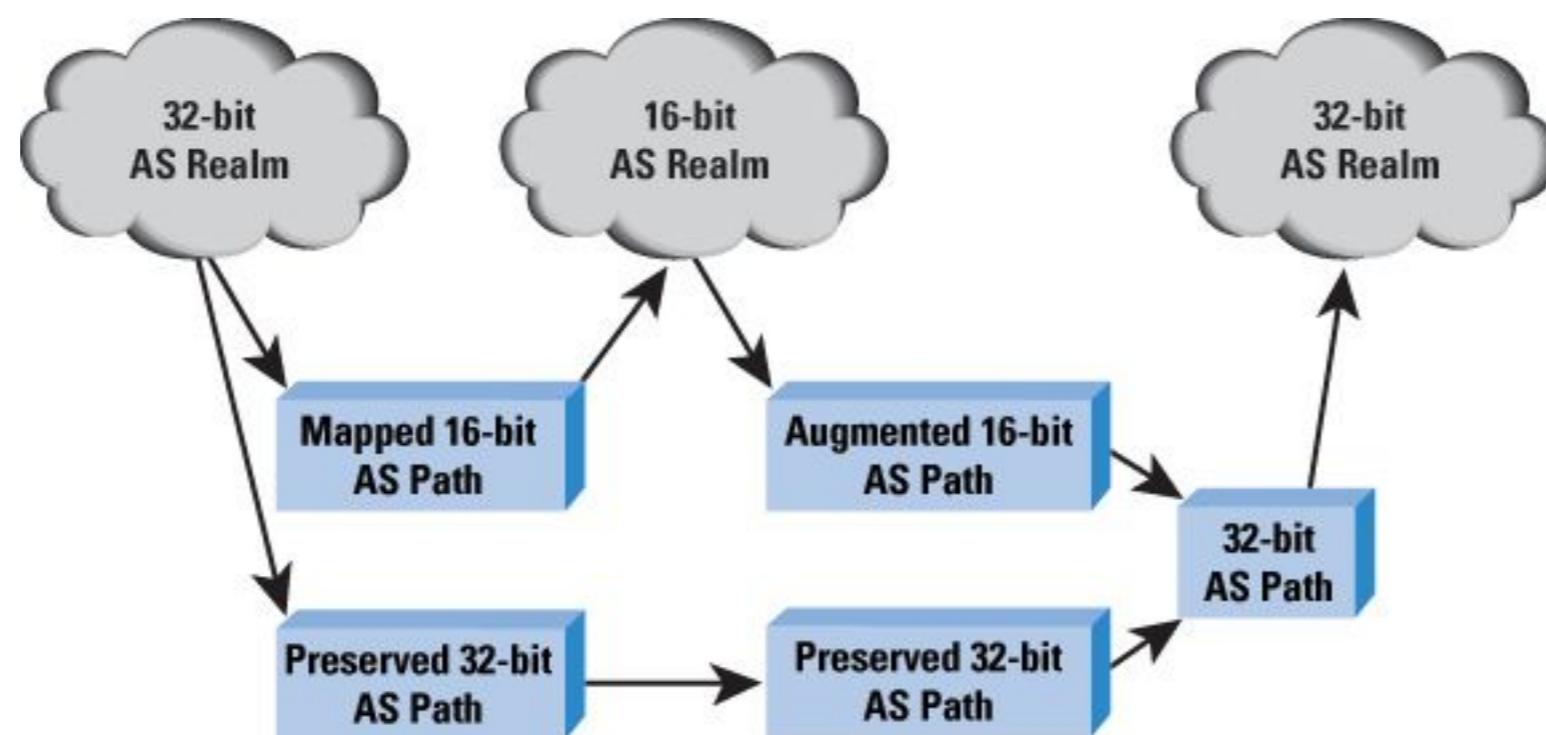
- ASN spravuje IANA, která alokuje bloky po 1024 jednotlivým RIR
- RIR alokuji ASN pro ISP a koncové sítě
- pouze cca 2/3 jsou aktivně využívány pro BGP směrování



Leden, 2006

Potřeba většího počtu ASN

- podle předpokladů budou ASN vyčerpány
- změna AS z 16-ti bitů na 32 bitů: X.Y [RFC4893, 2007]
- vyžaduje změnu v BGP protokolu
- v Internetu musí spolu komunikovat BGP pro 16-bitů AS a BGP pro 32-bitů AS

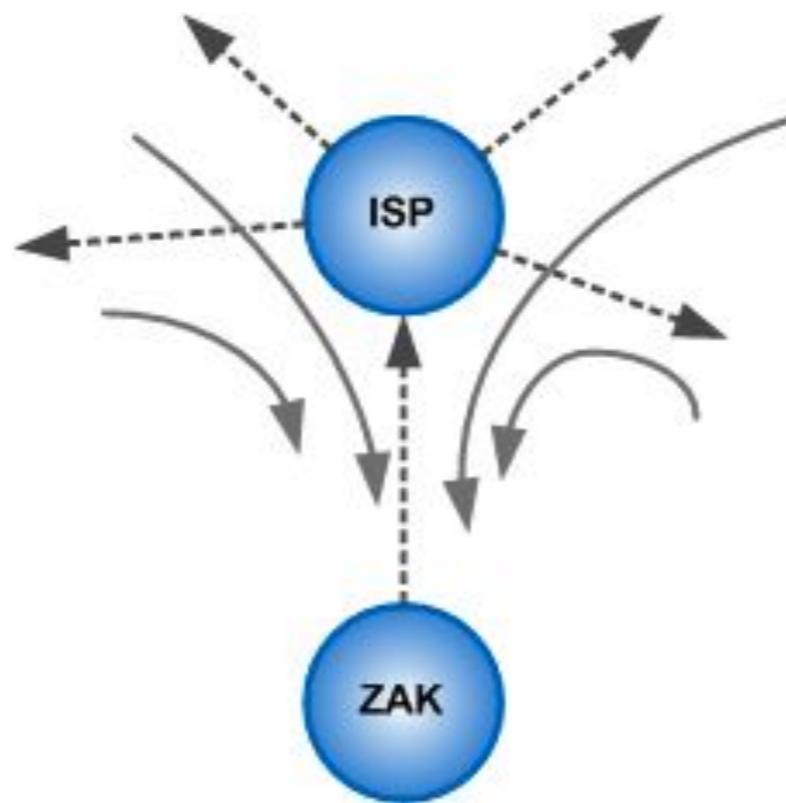


Vztah mezi AS

- sousední AS mají mezi sebou smluvní vztah
 - kolik dat budou přenášet
 - do jakých cílových cílů budou doručovat data
 - kolik se za to bude platit
- typické vztahy:
 - zákazník-poskytovatel
 - VUT je zákazník CESNETU
 - ČD-Telematika je národní ISP pro lokální ISP
 - peer-peer
 - CESNET je peer Telefonica O2 CZ (NIX.CZ)
 - AT&T je peer Sprintu

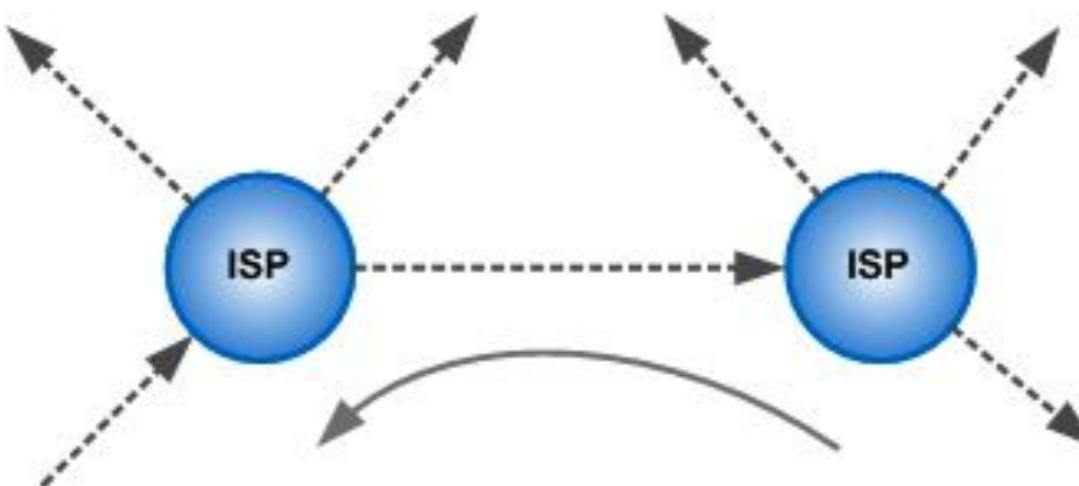
Zákazník-poskytovatel

- Zákazník vyžaduje dostupnost
 - poskytovatel oznamuje sousedům informace o zákazníkovi
- Zákazník nechce dostávat data, která mu nepatří
 - transit traffic



Peer-peer vztah

- vyměňují si data svých zákazníků
 - pouze sítě zákazníků jsou šířeny ostatním
 - sítě ostatních peerů jsou oznamovány zákazníkům
 - často bez finančního vyrovnání

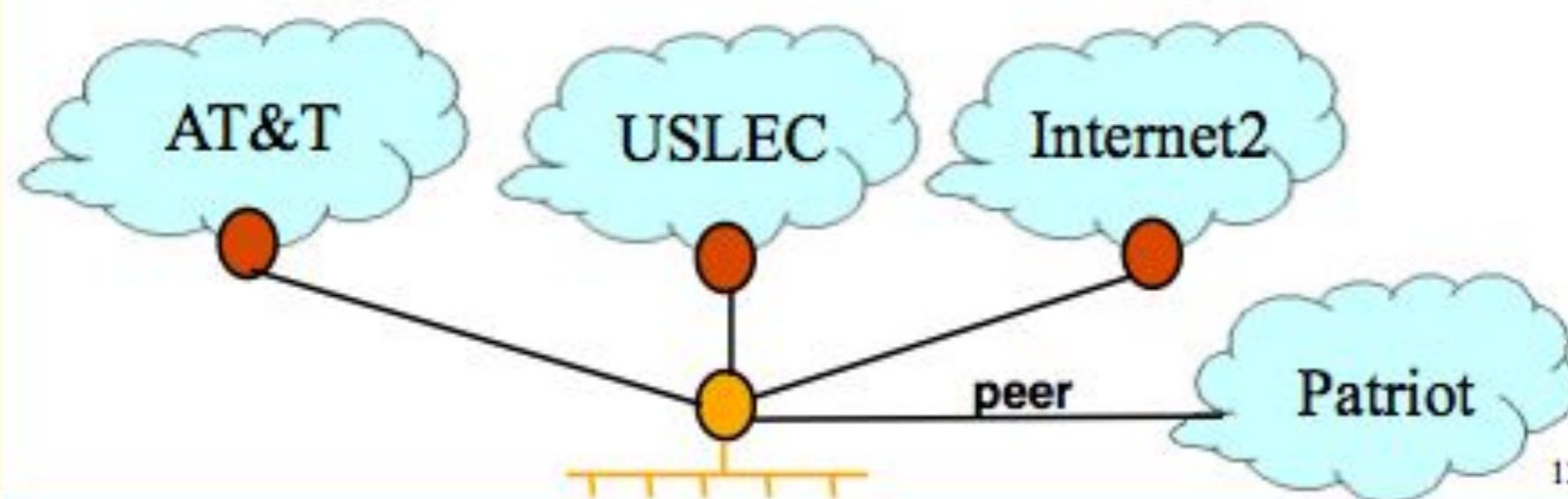


Vícenásobné vztahy

Princeton Example



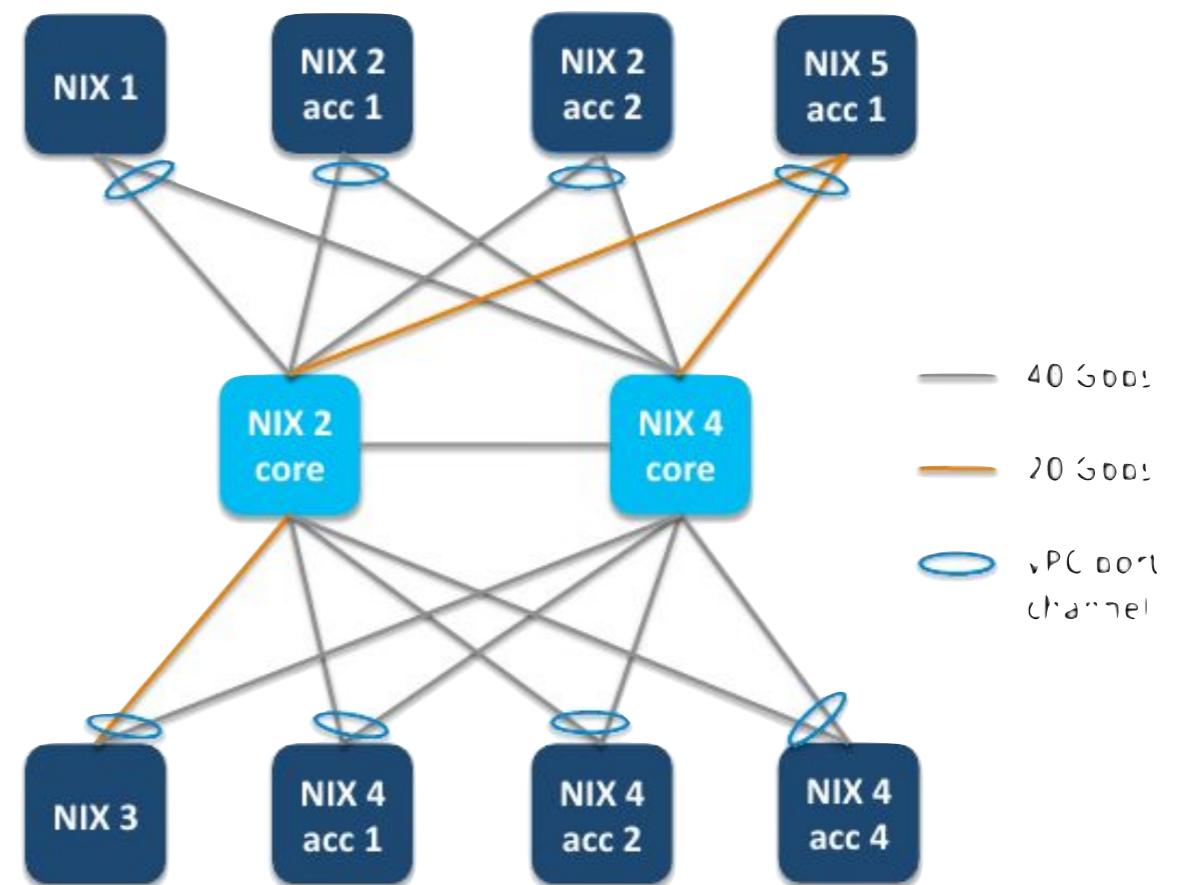
- Internet: customer of AT&T and USLEC
- Research universities/labs: customer of Internet2
- Local residences: peer with Patriot Media
- Local non-profits: provider for several non-profits



12

Peeringová centra

- Propojují různé ISP
- NIX.CZ (www.nix.cz)
- Podmínky připojení:
 - vlastní ASN
 - konektivita do NIX
 - POP
- Cena
 - dle typu připojení
 - náklady za konektivitu do centra

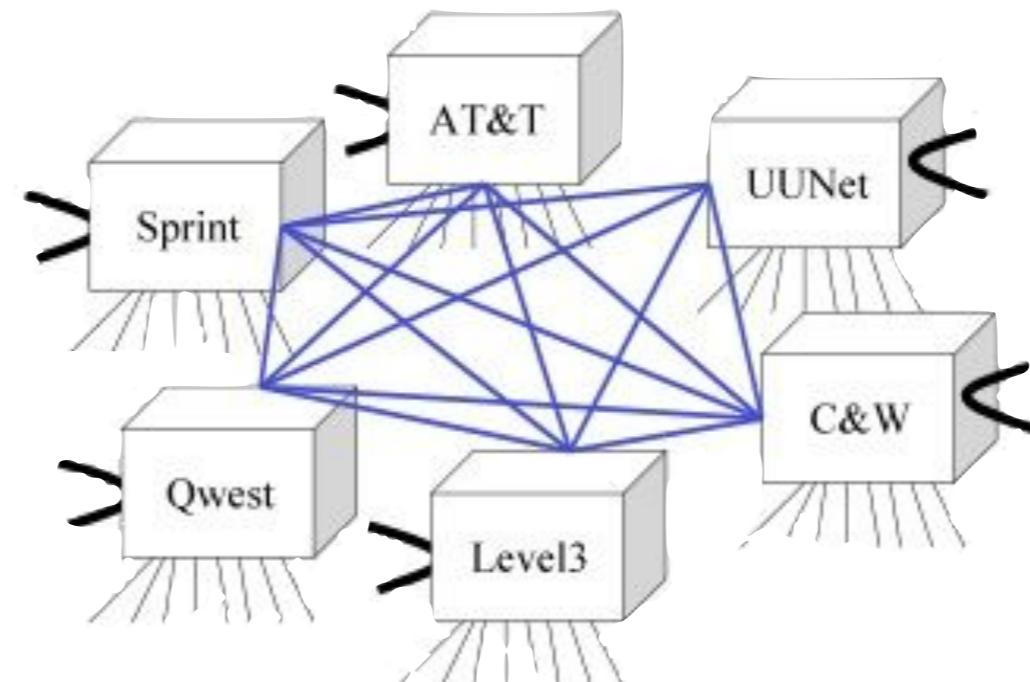


Příklad CESNET

- Podmínky pro peering
 - Partner musí mít vlastní NOC (Network Operating Centre) v nepřetržitém provozu.
 - Partner musí provozovat vlastní internetovou síť s homogenní směrovací politikou.
 - Partner musí být LIR (Local Internet Authority), musí mít vlastní adresní prostor a autonomní systém.
 - Směrovací tabulky musí být maximálně agregovány (mechanismem CIDR), CESNET2 nepřijímá síťové prefixy delší než /24.
 - Partner nesmí šířit defaultní cestu (route of last resort) přes propojení se sítí CESNET2.
 - Všechnen provoz mezi sítěmi CESNET2 a partnera musí být směrován přímým propojením mezi sítěmi.
 - Sdružení umožňuje propojení infrastrukturou NIX.CZ. Připojení k infrastruktuře NIX.CZ realizuje každá strana samostatně.

Struktura Internetu: Tier-1

- Není ve vztahu zákazník s žádným jiným ISP
- Má vlastní páteřní síť
- Plný peering mezi Tier-1 ISP



- Internet Health Map projekt mapuje stav páteře Internetu

Border Gateway Protocol

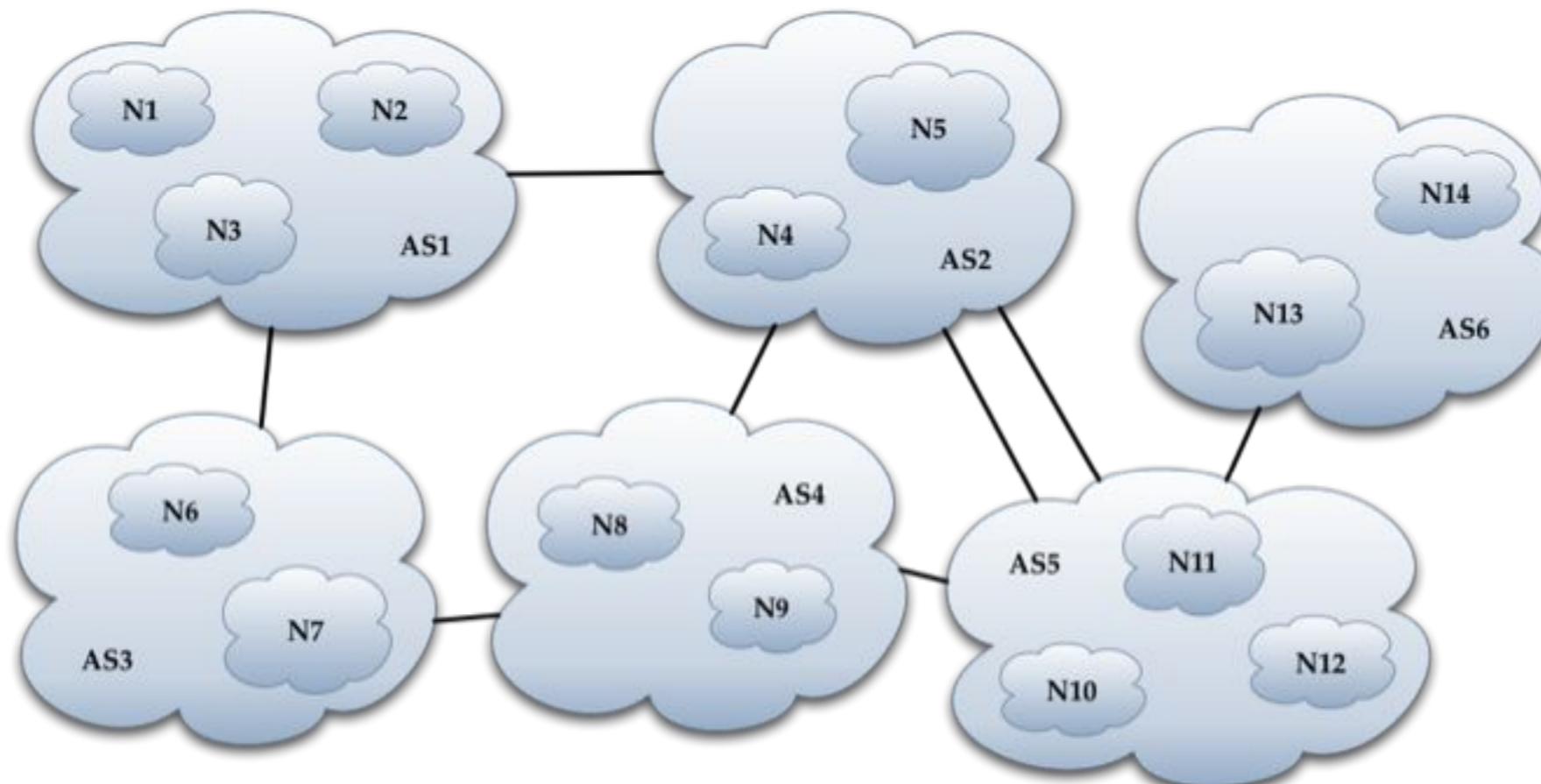
Border Gateway Protocol

- vytvořen koncem 1980/začátkem 1990
- nyní ve verzi BGP-4
- definováno v RFC 4271
- informace jsou neseny protokolem TCP, defaultní port 179
- Příklad BGP tabulky:
 - <http://bgp.potaroo.net/as2.0/bgptable.txt>

BGP

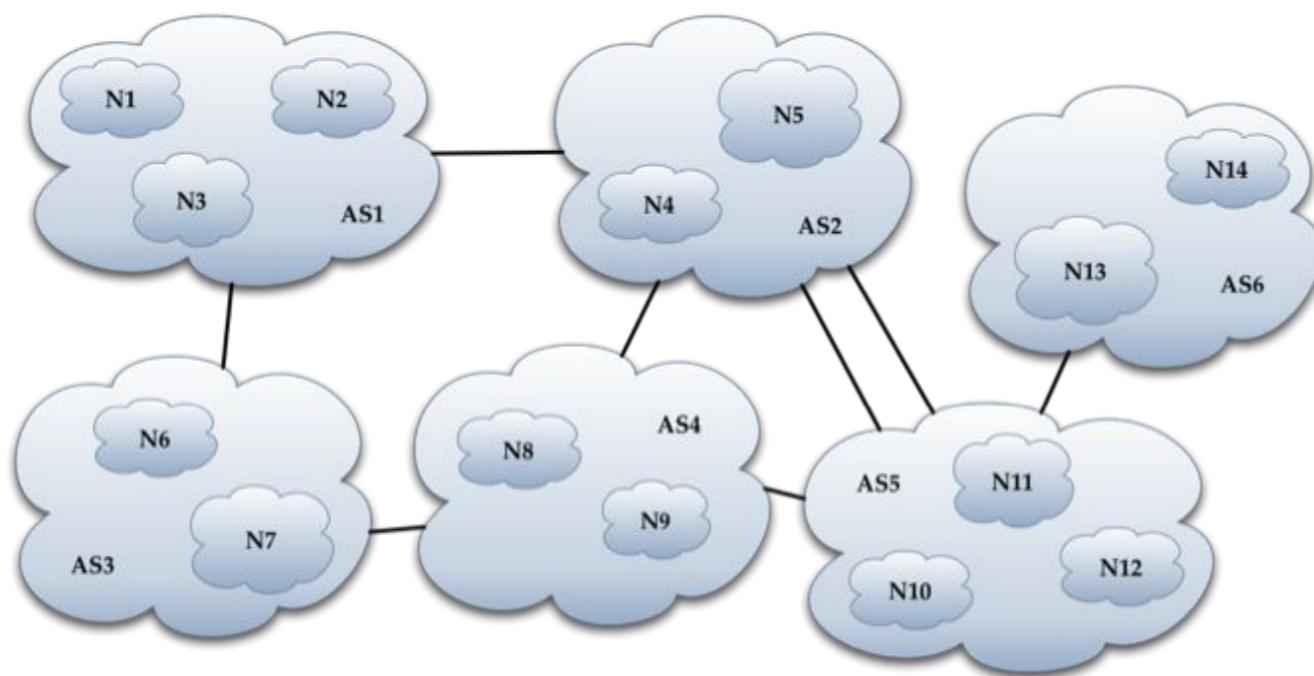
- Co je BGP?
 - vyměňuje si informace o sítích spravovaných v rámci AS
 - síť je zde myšlen blok IP prefixů
- BGP pracuje v rámci propojených AS
 - AS představují superuzel pro BGP
 - BGP agent pro tento superuzel komunikuje s agentem pro sousední superuzel
 - agent = BGP speaker
 - spojení mezi agenty = BGP session
 - cena spojení = AS-hop cost

BGP struktura



BGP komunikace

- AS6 posílá AS5, že vlastní dva adresové bloky
 $(AS6) \rightarrow \{N13, N14\}$
- Každý AS přidá svoje ID a informuje sousedy



$(AS2, AS5, AS6) \rightarrow \{N13, N14\}$

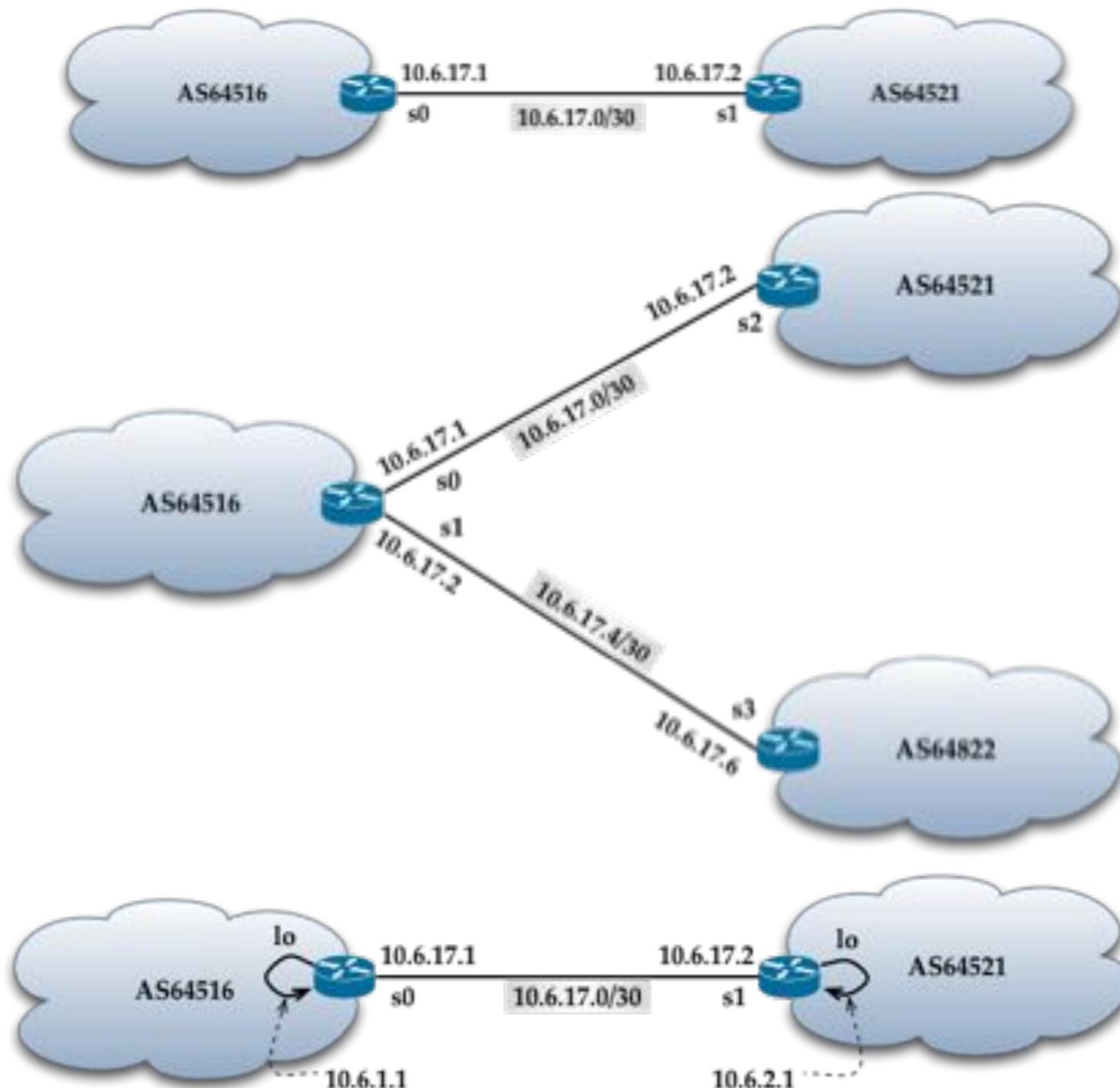
$(AS3, AS4, AS5, AS6) \rightarrow \{N13, N14\}$

lepší dle AS-path

BGP operace

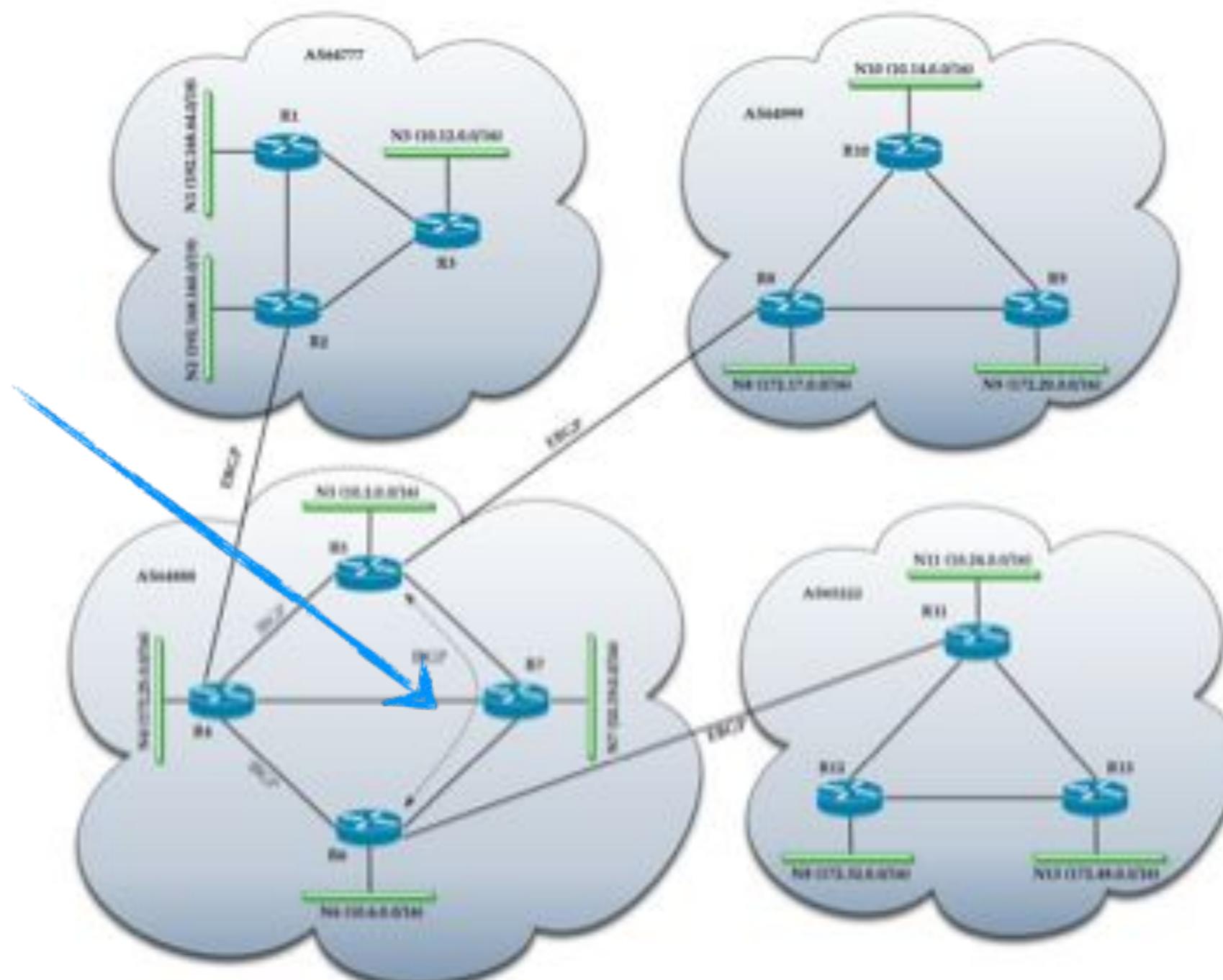
- Po ustanovení TCP spojení mezi dvěma BGP uzly:
 - OPEN: zahájení komunikace, definuje hold time
 - UPDATE: výměna informací o IP prefixech
 - KEEPALIVE: periodické udržování spojení
 - NOTIFICATION: korektní ukončení relace
 - ROUTE-REFRESH: aktivní dotazování informace (novinka BGP-4)

BGP konfigurace



I-BGP

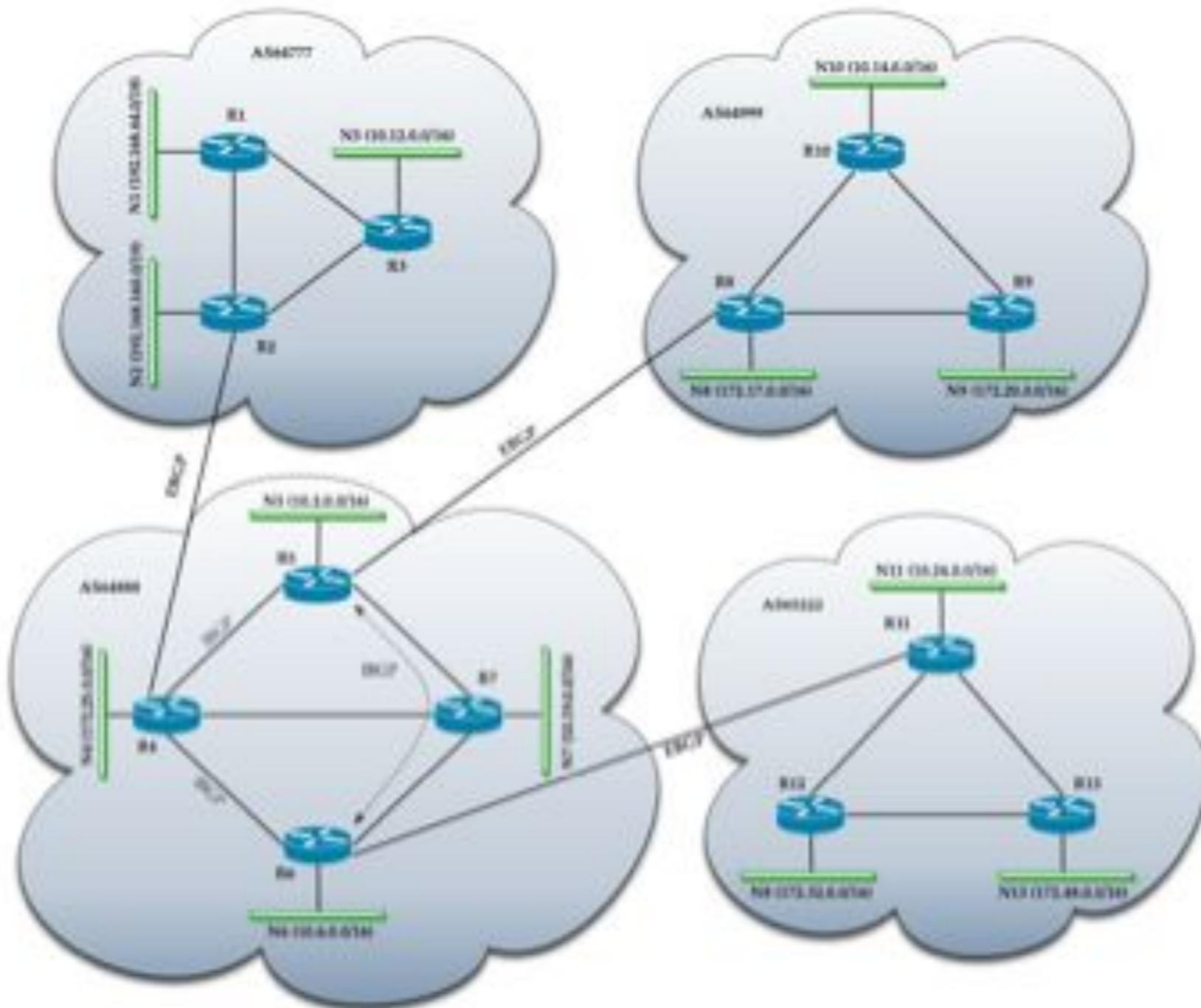
- Směrovače potřebují znát cestu k sousedním AS ze svého AS



I-BGP a E-BGP

- intra-AS ~ I-BGP
- inter-AS ~ E-BGP
- I-BGP
 - uvnitř AS je potřeba vyměňovat si informace mezi různými BGP uzly
- Pravidla
 - BGP uzel může oznamovat prefix, který se naučil od E-BGP všem I-BGP sousedům
 - BGP může oznamovat prefix, který se naučil od I-BGP všem E-BGP sousedům
 - I-BGP nemůže oznamovat prefix naučený od I-BGP jiným I-BGP sousedům (mohla by vzniknout smyčka)

I-BGP a E-BGP

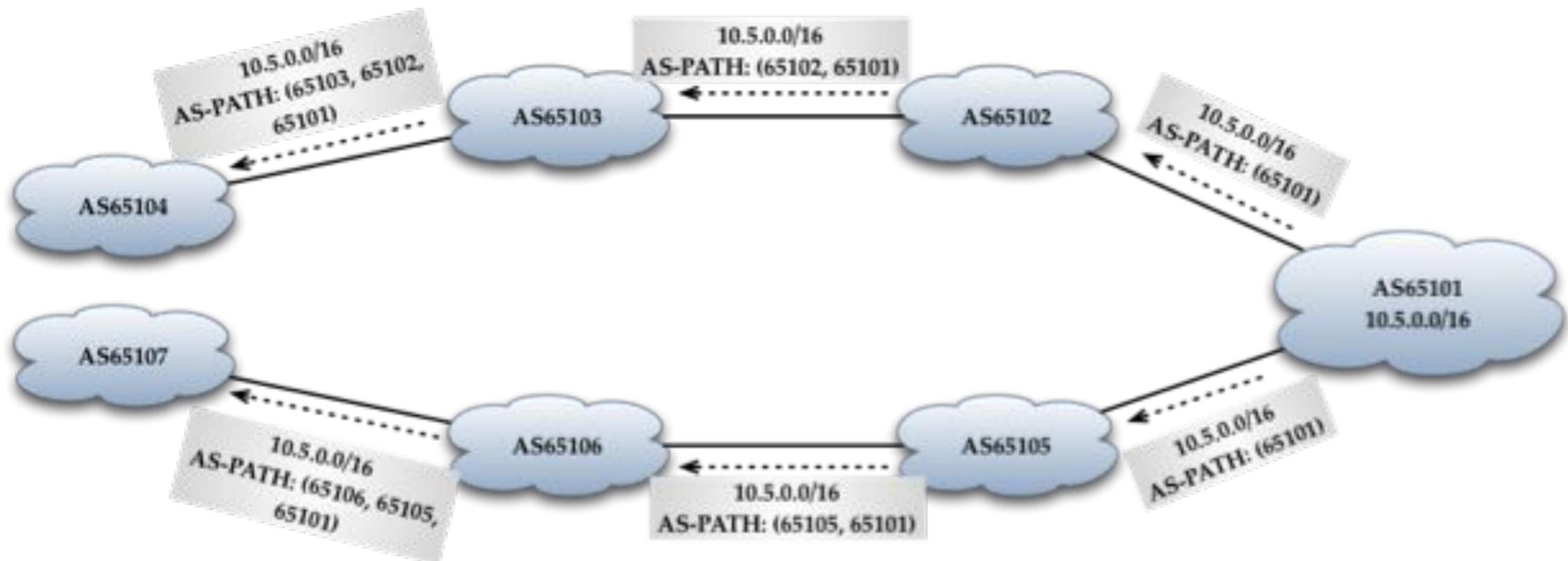


R4 se dozví o N11, N12 a N13 od AS65222 od R6
R4 nemůže informovat R5 o N11, N12 a N13

BGP atributy

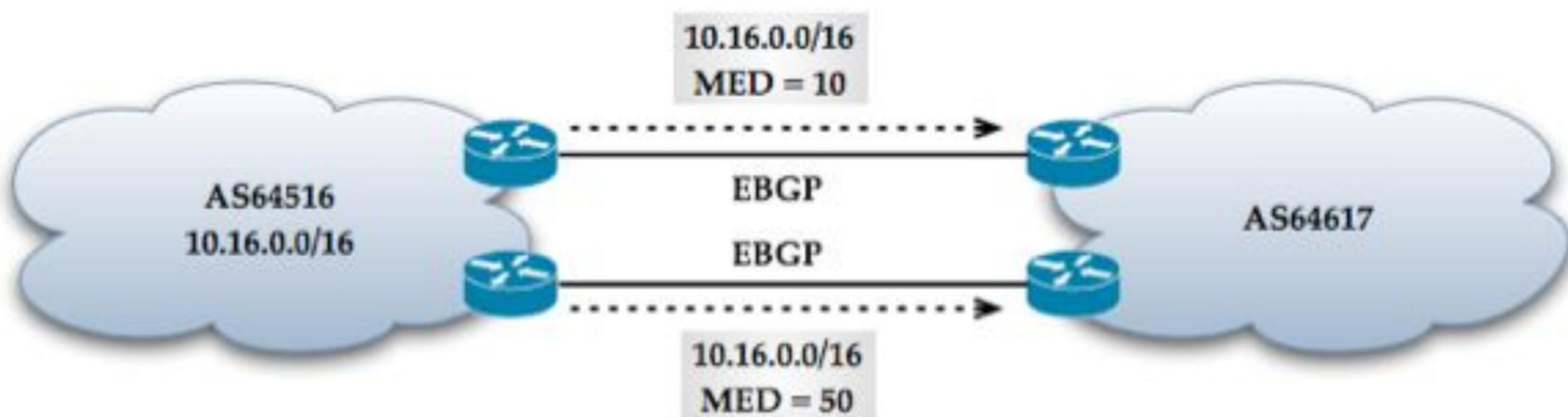
- Kritéria pro rozhodnutí o nejlepší cestě:
 - AS-PATH
 - MED
 - Local PREF
 - Route Aggregation

AS-PATH

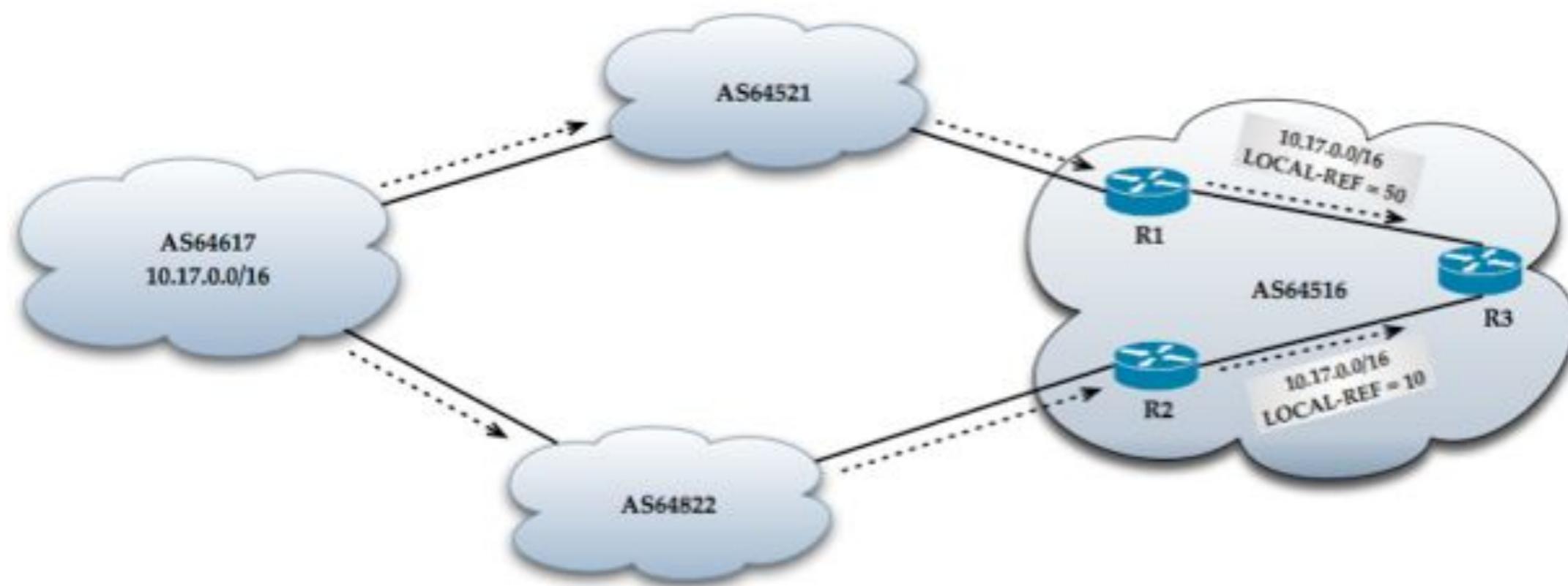


MED

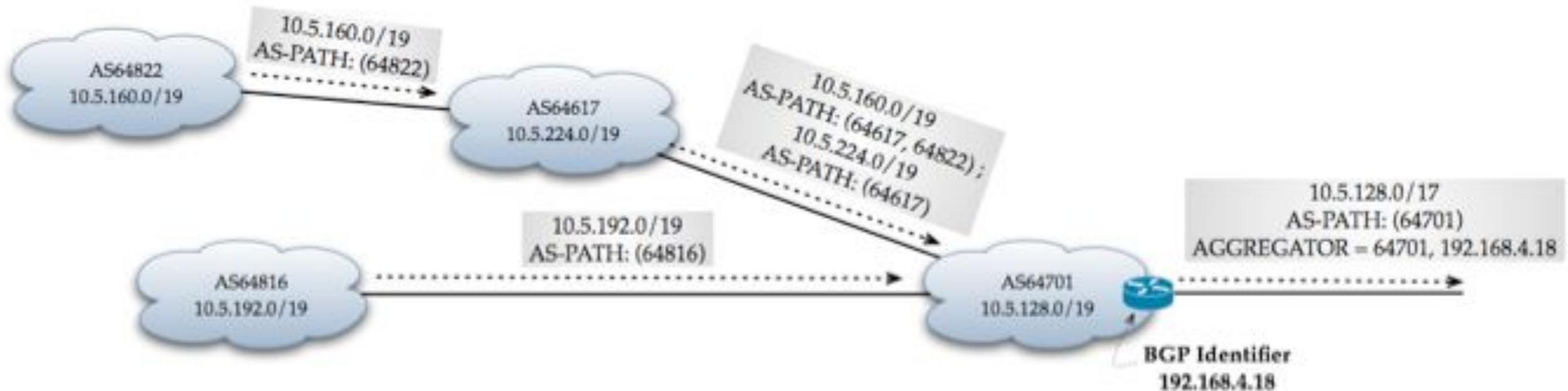
- Multi Exit Discriminator
- The exit point with the lowest MED is preferred



Local PREF



Route Aggregation



Výběr cesty

Existuje-li více cest pro daný prefix:

1. Nejvyšší LOCAL-PREF
2. Nejlepší AS-PATH
3. Nejmenší ORIGIN
4. Nejmenší MED
5. Nejmenší cena do NEXT-HOP

Policy

- Pravidla, která se použijí při zpracování BGP informace na směrovači

TABLE 8.1 Examples of import and export policies at a BGP speaker.

Import Policy	Export Policy
<ul style="list-style-type: none">– Do not accept default 0.0.0.0 from AS64617.– Assign 192.168.1.0/24 coming from AS64617 preference to receiving it from AS64816.– Accept all other IP prefixes.	<ul style="list-style-type: none">– Do not propagate default route 0.0.0.0 except to internal peers.– Do not advertise 192.168.1.0/24 to AS64999.– Assign 172.22.8.0/24 a MED metric of 10 when sent to AS64999.

BGP Message

- Cloud Shark Links

BGP Update

<https://www.cloudshark.org/captures/0224f4ab8f63>

Forming BGP Adjacency

<https://www.cloudshark.org/captures/00249be4441f>

Shrnutí

- Směrování a přepínání paketů
 - směrovací tabulky
 - algoritmus výběru
 - architektura směrovače
- Směrování podle nejkratší cesty
 - Algoritmus Dijkstrův
 - Algoritmus Bellman-Fulkerson-Fordův
 - Path-Vector směrování
- IGP směrovací protokoly
 - RIP
 - OSPF

Shrnutí

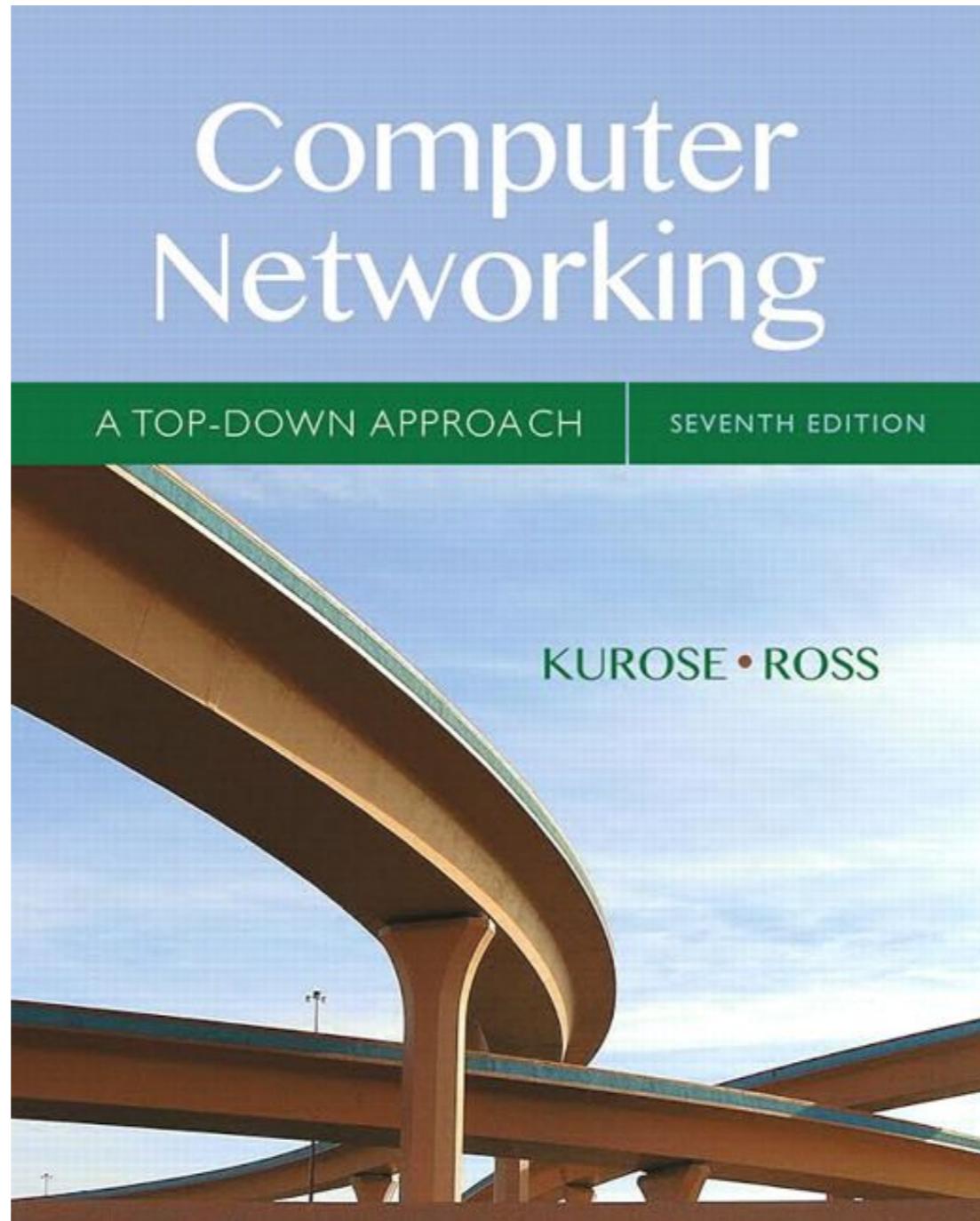
- Internet je složen z autonomních systémů
 - každý systém má své unikátní ASN
- Směrování je na dvou úrovních:
 - EGP - mezi autonomními systémy
 - IGP - uvnitř autonomních systémů
- BGP je protokol pro EGP
 - Path-vector protokol
 - jiné požadavky a tudíž atributy než IGP
 - implementuje směrovací politiky

Literatura

- Kurose J.F., Ross K.W.: Computer Networking, A Top-Down Approach Featuring the Internet. Addison-Wesley, 2003.
- Rita Pužmanová: Routing and Switching: Time of convergence? Addison-Wesley, 2002.
- A.Zinin: Cisco IP Routing: Packet Forwarding and Intra-domain Routing Protocols. 2001.
- Příslušná RFC...

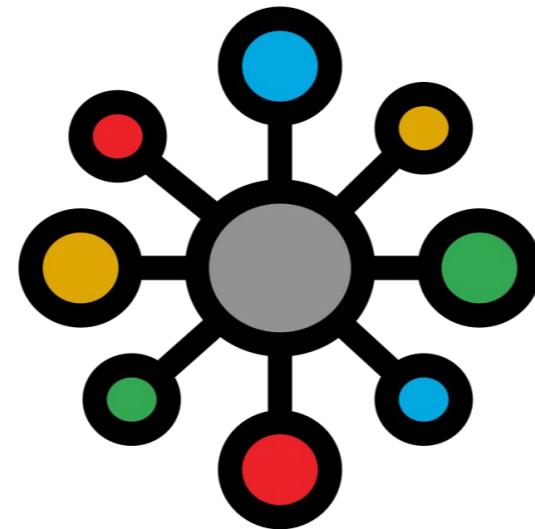
Domácí úkol

- Kapitola o směrování, sekce problémy a otázky.



Ondřej Ryšavý rysavy@fit.vutbr.cz

Vladimír Veselý veselyv@fit.vutbr.cz



<https://nesfit.github.io>