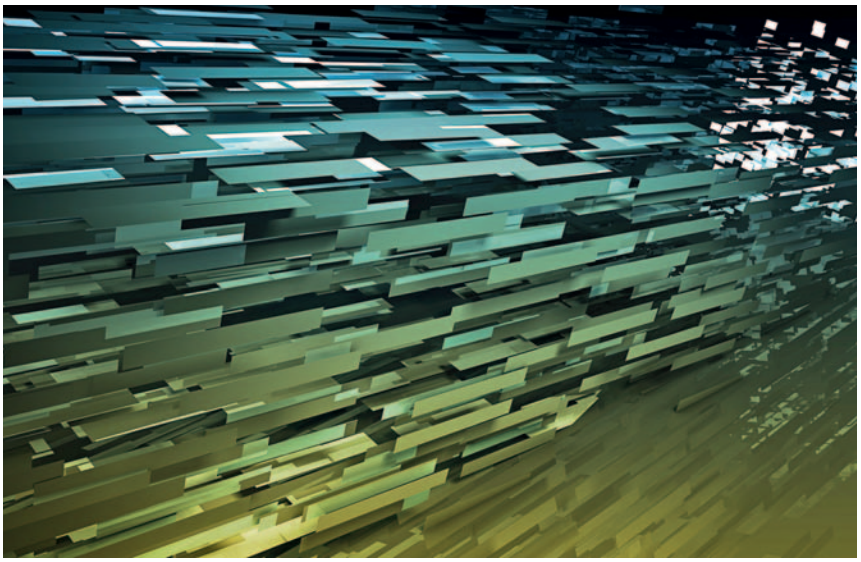
Protokol BGP je dobrý sluha, ale špatný pán

aneb Internet pro pokročilé

Tomáš Hlaváček



Jak vlastně funguje Internet? Co řídí transport dat přes celosvětovou počítačovou síť? A jak se najde cesta pro data mezi miliony sítěmi a miliardami uživatelů v Internetu? Zkusíme si v následujícím článku odpovědět na tyto otázky a dotkneme se i obchodní a organizační stránky Internetu.

Protokol IP

Data se v počítačových sítích přenáší pomocí protokolů. Protokoly jsou standardizované formáty a postupy pro efektivní přenos dat. Hlavním znakem protokolů, které se používají v současném Internetu, je orientace na „packetový přenos“, neboli přenos dat po nepříliš velkých balíčcích s definovanou maximální délkou. Tyto balíčky, které obvykle nepřesahují 1500 bytů, jsou směrovány na základě informací v hlavičce každého packetu. Protokol IP se od počátku zaměřil na požadavek velké robustnosti sítě a ta implikuje, že si každý packet nese kompletní směrovací informaci. To znamená, že pokud existuje cesta k cíli, tak jí lze pro jakýkoliv packet najít nezávisle na předchozí komunikaci. Tím se počítačová síť s protokolem IP liší například od klasické telefonní sítě, kde probíhá fáze vytáčení a navazování spojení, pak je možné přenášet informace a nakonec se spojení ukončí.

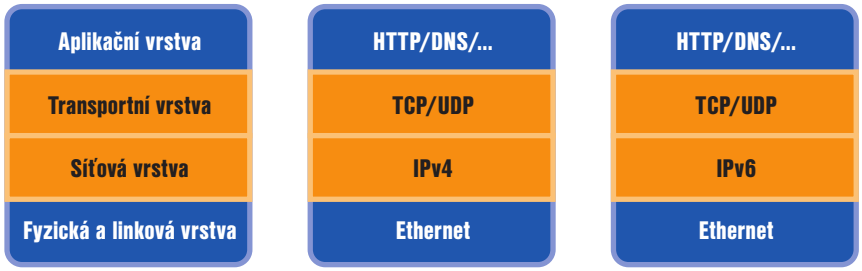
Protokol IP je jen jedním z mnoha stavebních kamenů počítačové sítě. Protokoly se staví nad sebe a každý vykonává jednu konkrétní funkci, přičemž mohou stát na jedné pozici i různé protokoly a zbytek protokolového stacku zůstává stejný, což zajišťuje modularitu. Příkladem modularity je právě rodina protokolů IP, neboť v současnosti jsou v Internetu používány protokoly IPv4 a IPv6 vedle sebe (Obr. 1). IPv4 je starší, trpí řadou problémů a růst Internetu jen s tímto protokolem už není nadále možný kvůli omezením adresního prostoru. Naproti tomu protokol IPv6 byl vyvinut v druhé polovině devadesátých let a vedle dramatického rozšíření adresního prostoru do dimenzí, které takřka není možné nikdy vyčerpat, přináší i další významné technické výhody. Přechod z IPv4 na IPv6 je však pomalý a proto se oba protokoly provozují souběžně, což modulární architektura sítí umožňuje a v současné přechodové fázi to i dává smysl.

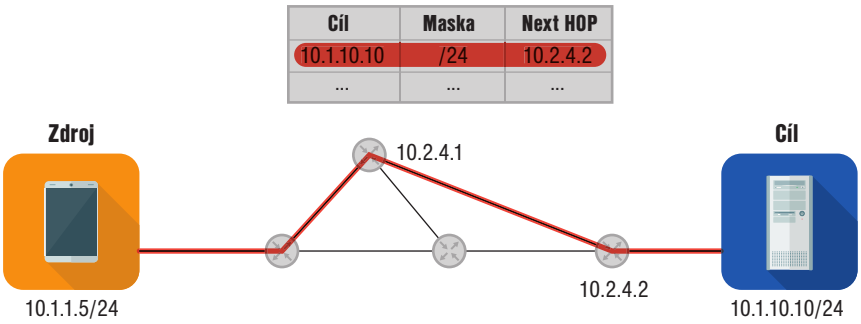
Bez ohledu na to, zda se jedná o protokol IPv6 se 128 bitovou adresou a nebo protokol IPv4 s 32 bitovou adresou vždy je přenos informací procesem, kde se packet ze zdrojového bodu předává mezi směrovači po cestě až k destinaci, která je určená cílovou adresou a tu si packet nese v hlavičce. Směrovače mají vyplněné směrovací tabulky, které obsahují řádky se síťovou adresou, k nim příslušnou síťovou maskou a adresou dalšího směrovače v cestě. Packety, jejichž adresa bitově vynásobená (AND operace) síťovou maskou odpovídá síťové adrese v daném řádku, se posílají ve směru dalšího směrovače v cestě (Obr. 2).

Směrovací protokoly

Vyplňovat směrovací tabulky lze ručně, ale to je představitelné jen v jednoduchých situacích, kdy máme jednotky sítí a jednotky směrovačů. Například na domácích routerech či v malých firmách je to obvyklý postup. Větší síť však takto udržovat nelze, protože sebemenší změna v topologii a nebo v připojených subnetech na okraji sítě se musí promítnout do směrovacích tabulek všech směrovačů, které by potenciálně mohly přenášet data pro danou destinaci, kde došlo ke změně. Proto se používají směrovací protokoly, které informace o dostupných sítích a nejlepších cestách k nim naplní do směrovacích tabulek téměř automaticky. Směrovacích

Obr. 1: IP stack





Obr. 2: Směrování

protokolů existuje více a mají dokonce vlastní taxonomii podle typu informací, které přenášejí. Výsadní postavení má však protokol BGP, který slouží pro přenos směrovacích informací v Internetu a to pro protokoly IPv4 i IPv6. Specifikem protokolu BGP je právě schopnost úsporně a efektivně přenášet obrovská množství informací o statisících sítí. Cenou za to je relativně pomalá konvergence a přirozeně i nutnost správně dimenzovat hardware pro provoz protokolu BGP, zejména pokud jde o paměť a pak schopnost naplnit FIB - tabulku, podle které se skutečně směrují packety, všemi záznamy, které protokol BGP vygeneruje.

Autonomní systémy

BGP protokol přenáší informace o dostupných sítích mezi autonomními systémy. Autonomní systém je síť či skupina sítí s jednotnou technickou a administrativní správou a jednotnou směrovací politikou. Autonomní systémy mezi sebou navazují dvoubodové BGP relace na základě obchodních vztahů, které mají však i technologický a „politický“ aspekt (Obr. 3). Přenesené informace o dostupných sítích se stávají vstupem do algoritmu na výběr nejlepší cesty, která se nakonec vloží do FIB a může být redistribuována dalším sítím. Podkladem pro navázání BGP sessiony mezi dvěma autonomními systémy je v první řadě fyzická konektivita, která může být realizovaná například optickým vláknem, propojem v rámci datacentra a nebo připojením obou AS do propojovacího centra - IXP (Internet Exchange Point, v případě České republiky je takovým IXP sdružení NIX.CZ). Druhým předpokladem pro navázání BGP sessiony je uzavření dohody o propojení, která definuje zejména jaké prefixy si budou sítě oznamovat protokolem BGP. Prerekvizitou uzavření dohody je předpoklad oboustranné užitečnosti takového propojení. V případě, kdy obě sítě mají symetrické postavení a propojení bude pro oba partnery přibližně stejně užitečné, se obvykle hovoří o „peeringu“ a peeringové dohody bývají zejména v Evropě neformální záležitostí, která obvykle nepředpokládá placení poplatků za peering. Přesto poplatky za peering nejsou vyloučené, zejména v situaci, kdy je postavení sítí asymetrické a tedy jedna síť bude mít z propojení významně větší prospěch. Naproti tomu stojí zpravidla placená Internetová tranzitní konektivita, kdy tranzitní síť zprostředkovává konektivitu do všech ostatních sítí v Internetu.

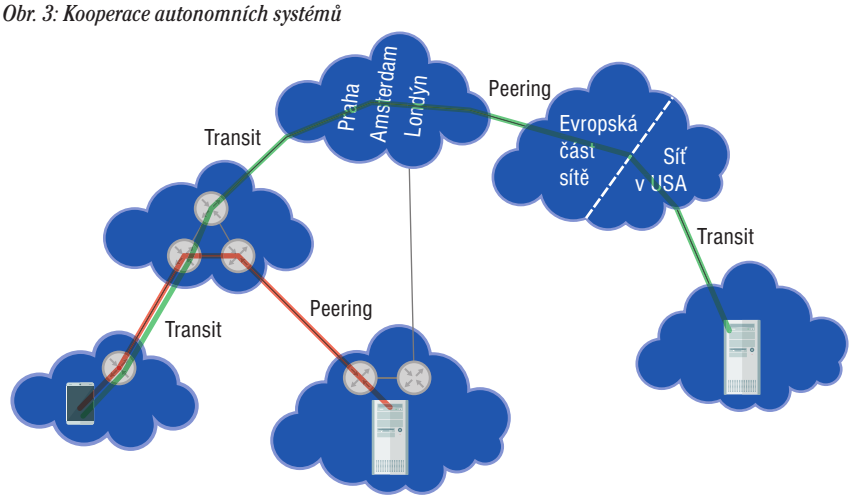
Připojení k Internetu ve vlastní režii

Firmy, které považují své působení na Internetu za stěžejní záležitost, zpravidla časem překročí rámec obvyklé nabídky ISP a nebo narazí na limity při vyjednávání o podmínkách Internetových služeb s jedním jediným partnerem. Pro takové firmy je alternativou převzít technickou, administrativní i obchodní stránku věci pod svou kontrolu. K tomu je zapotřebí získání prostředků v podobě přidělených bloků IP adres a čísla autonomního systému. V Evropě tyto prostředky přiděluje organizace RIPE NCC. V současné situaci je získání IPv4 adres poněkud omezené a vedle možnosti „pronajmout“ si je za nepříliš výhodných podmínek je jedinou možností stát se LIR (Local Internet Registry), tedy jakýmsi členem organizace RIPE NCC. S tím souvisí i nutnost přijmout všechny povinnosti člena. Nejedná se sice o velký finanční náklad, ale přináší to nutnost vzdělání zaměstnanců o politikách a regulacích, kterými se řídí RIPE NCC, aby byli schopni smysluplně jednat s hostmastery RIPE NCC a aby nedocházelo k porušování pravidel, které komunita RIPE zavedla a RIPE NCC je aplikuje. Nicméně i LIR má v současnosti k IPv4 adresám omezený přístup, neboť RIPE NCC má už jen poslední /8 blok IPv4 adres a pro něj platí speciální politika. Ta je zaměřena na nově příchozí firmy a má jim umožnit nouzové připojení k Internetu přes IPv4 na mezidobí, než dojde k plnému přechodu na protokol IPv6.

S IPv6 adresami a čísly AS není žádný problém a lze je získat buď jako LIR přímo od RIPE NCC a nebo z pozice strany, která sice žádá o vlastní prostředky, ale nepodniká na Internetu tak intenzivně, aby se stala LIR. Dělící čára je počet prostředků každého typu - jedno AS a jeden IPv6 prefix, případně jeden IPv4 prefix z minulosti, nevytváří nutnost stát se LIR. Potřeba více prostředků jednoho typu však tuto nutnost už přináší.

Jakmile má síť k dispozici vlastní adresy a číslo autonomního systému, je povinná začít je v určité lhůtě používat, což znamená oznamovat tyto prostředky protokolem BGP do DFZ (Default Free Zone). DFZ je synonymem pro Internet. K tomu je zapotřebí jednak technologické vybavení, které si lze představit v mnoha podobách od virtuálního serveru s Linuxem a OSS implementací BGP až po skříň o objemu několika metrů krychlových, která ukrývá výkonný směrovač pro velký počet 100 Gb/s linek. Dále je třeba zajistit nejméně jednu tranzitní konektivitu, přes kterou se nově vzniklá síť dostane do Internetu. Zpravidla se však navazuje spojení více - nejméně dvě tranzitní, kvůli robustnosti

Obr. 3: Kooperace autonomních systémů



a vedle toho se obvykle navazují peeringy se sítěmi, které jsou nejčastějšími komunikačními partnery, pokud to geografická lokalizace umožňuje a pokud to dává ekonomicky smysl. V neposlední řadě je zapotřebí správců, kteří se postarají nejen o konfiguraci a bezproblémový provoz směrovačů, ale udržují i technickou dokumentaci, reagují na technické, organizační i administrativní změny v RIPE a u peeringových partnerů.

Nástrahy protokolu BGP

Přestože vlastní Internetová konektivita a s ní spojená volnost vyjednávat si ceny spojení s kýmkoliv, kombinovat různé nabídky a dosáhnout tak významných úspor zní lákavě, přináší to i řadu problémů. Jedním z nejmarkantnějším problémů je komplexita routovacího systému a nutnost proškolení správců. Paradoxně z hlediska profesionálních správců sítí je BGP jen jeden z mnoha směrovacích protokolů, které konfigurují a udržují v běhu. Naproti tomu pro nováčky je BGP zpravidla velký problém k pochopení a udržení v provozu. Chyby v konfiguraci BGP mohou vést nejen k výpadkům vlastní sítě a z toho plynoucím ztrátám, ale za jistých podmínek lze napáchat škody třetím stranám v bezprostředním okolí i na druhé straně planety. Takové chyby často způsobí ostudu v technické komunitě a u větších případů i ztrátu dobrého jména u široké veřejnosti.

Nejtypičtější nástrahy BGP spočívají v tom, že vložení prefixů do BGP, jejich úspěšnost a další rozšíření závisí na nastavení filtrů. Rozmanité druhy filtrů lze aplikovat na každou BGP session v příchozím i odchozím směru a očekává se, že správci budou filtry používat podobně, jako programátoři používají kontroly při defenzivním programování: Ověřovat všechny vstupy a maximálně ručit za vlastní výstupy. Přesto se občas stává, že uniknou do DFZ prefixy, které tam nemají co dělat. To může vést jednak k přilákání packetů cizích sítí do míst, kam by jinak nezavítaly a v důsledku pak k zahlcení chybně nakonfigurované sítě a nebo k přerušení komunikace sítí třetích stran. Aby se předcházelo podobným incidentům, směrovací politiky jednotlivých AS se zveřejňují ve směrovacích databázích, v IXP dochází obvykle k automatické kontrole a správci tranzitních sítí zpravidla vyžadují podrobnou provozně-technickou evidenci a zveřejněnou směrovací politiku, podle které pak konfigurují filtry na vstupu od svých zákazníků.

V každém případě platí, že BGP je dobrý sluha, ale špatný pán. Pokud provoz BGP dává pro organizace smysl a už se jednou spustí, je třeba mu věnovat stálou pozornost a zejména udržovat vyškolený tým, který je schopen postihnout všechny technické i organizační aspekty a vyhnout se případným problémům. ■

Tomáš Hlaváček

Autor článku pracuje jako programátor pro výzkum a vývoj ve sdružení CZ.NIC. V Akademii CZ.NIC vede kurz „Směrovací protokol BGP“.