Stěhování serverů za provozu a moderní konfigurace sítě (CSNOG 2024)

root.cz/clanky/stehovani-serveru-za-provozu-a-moderni-konfigurace-site-csnog-2024/

V lednu proběhlo ve Zlíně další setkání komunity CSNOG, českých a slovenských správců internetových sítí. Mluvilo se o přesunu serverů mezi datacentry, moderní konfiguraci sítě a sledování dění v BGP.

Tomáš Procházka: Přesun datového centra za plného provozu

Datové centrum Nagano ukončovalo v roce 2023 provoz a Seznam neustále datově a výpočetně roste. “Bylo jasné, že Nagano musíme opustit a někam se s 41 tunami hardware přesunout.” Nakonec vyhrála stavba vlastního datacentra, které umožňuje zlevnit provoz, zbavit se dlouhodobých závazků plynoucích z pronájmu a přizpůsobit si vlastní prostory. “Chtěli jsme také vlastní monitoring, který by nám umožňoval sbírat více dat.”

Seznam.cz provozoval datacentra Kokura a Osaka, nové datacentrum bylo nazváno Nagoja. Zbývala už jen maličkost: najít dostatečně vhodný pozemek, který bude velký a nabídne možnost dobrého napájení. “Přislíbeno máme až 6 MW, zatím máme spotřebu asi 0,6 MW.” PUE se dlouhodobě pohybuje okolo 1,15.

Při stěhování bylo potřeba přesunout 2725 serverů, stěhovat se měly dva racky denně od pondělí do čtvrtka. “Pátek jsme si nechali volný, kdyby nastaly nějaké problémy.” Při stěhování byla část uzlů ve starém datacentru a část už v novém. “Služby byly stále provozovány bez výpadku nebo větší odstávky.”

Každý den ráno byly rozebrány a převezeny dva racky, v cíli bylo třeba je zase poskládat a postupně zapojit a zapnout. “Jako první nám kolegové zapojili switch a my jsme mohli spustit migrační skript.” Ten si zjistil IP adresu, zkontrolovala se změna hostname a vytvořila se minimální konfiguraci přepínače, aby se připojil do managementu. Poté se pustil další skript, který prováděl rekonfiguraci. Pro automatizaci byl využit Ansible a Python.

Nakonec stěhování proběhlo velice klidně, kromě několika drobných problémů se díky intenzivní přípravě všechno podařilo. “Díky automatizaci se nám vše povedlo bez přešlapů.” Propoje mezi datacentry zvládly veškeré synchronizace a komunikace clusteru.

Ondřej Zajíček: BIRD, MPLS a EVPN

MPLS slouží k přenosu paketů jiným způsobem než pomocí IP. “Paket si na vstupu do sítě označíme bezvýznamovým identifikátorem a na výstupu hlavičku zahodíme.” Jde o technologii, která běží mezi linkovou a síťovou vrstvou. Výhodou je, že to umožňuje velmi jemnou práci s provozem, implementace může být velmi rychlá a umožní to explicitní separaci jednotlivých toků.

Nevýhodou MPLS je, že správa takové dynamické sítě může být komplikovanější. “V síti musíte distribuovat informace o tom, kterému toku odpovídá který label.” K distribuci těchto informací slouží protokoly jako LDP, RSVP-TE nebo BGP.

MPLS je v démonu BIRD plně implementován od verze 2.14. “Modulární podoba BIRDu předpokládá, že se pohybujeme ve světě IP. Proto jsme dlouho přemýšleli, jak MPLS uchopit.” Nakonec bylo zvoleno řešení pouze s BGP, v tuto chvíli není k dispozici podpora LDP nebo RSVP-TE. “Je ale možné používat BGP jako interní routovací protokol, pak už nepotřebujete další protokol.” Někdy v budoucnu ale zřejmě dojde na implementaci dalších variant.

K dispozici jsou MPLS tabulky, které vlastně odpovídají IP tabulkám a také umožňují exportovat informace do systémového jádra. “Zavedli jsme route atributy, které dovolují definovat pravidla pro přijetí labelů.”

EVPN je v podstatě distribuovaný bridge, kdy se může síť rozprostřená přes více routerů chovat jako jedno prostředí. Při tom je potřeba signalizovat stav sítě a přenášet propagaci MAC adres. “To uděláte přes BGP, samotná data pak tečou přes nějakou enkapsulaci.” Záznamy mohou obsahovat kromě MAC také informaci o VLAN. “Stále na tom pracujeme, jsou tam pořád ještě nedodělky.”

Ľubor Jurena: Moderní konfigurace sítě pomocí systemd-networkd

Nástroj systemd-networkd je součást celého ekosystému systemd a spouští se jako samostatná služba. “Cílem je snížit závislost na dalších systémových knihovnách a celou konfiguraci připravit na jednom místě.” Umožňuje konfigurovat fyzická rozhraní, ale i virtuální síťová zařízení. “Interaguje s dalšími komponentami systemd jako je například systemd-networkd.”

Konfigurační soubor `/etc/systemd/network` je zapsaný ve specifické syntaxi systemd a je rozdělený na několik sekcí. Základní konfigurace je velmi jednoduchá: nejprve určíme, se kterým rozhraním pracujeme a poté definujeme jeho vlastnosti. Pokud chceme rozhraní zařadit do VRF, nemusíme to dělat žádnými post-skripty, ale můžeme to udělat přímo v konfiguraci. Je možné také použít hvězdičkový záznam, který pokrývá více síťových rozhraní.

Podporován je režim běhu pouze s IPv6, je možné použít delegaci prefixu a jednoduše zapínat například předávání paketů mezi rozhraními nebo použít maškarádu. K ovládání je možné použít řádkový nástroj `networkctl` , který umožňuje načítat konfiguraci, měnit volby a zobrazovat si například informace o DHCP. “Prostým spuštěním zjistíme, která rozhraní máme k dispozici a které z nich networkd spravuje.”

Pavel Šimerda: Linux: Podpora hardwarových switchů

V roce 2008 přišla do linuxového jádra funkce DSA, která umožňuje označovat pakety odcházející z CPU do switche, který je například součástí desky. “Máme pak k dispozici metadata, která nám umožňují používat výstupní porty jako samostatná síťová rozhraní.” Později vznikl takzvaný bridge offloading, který umožňuje maximum konfigurace přenést na hardware. “I když to hardware nepodporuje, stále je k dispozici softwarový bridge.” Ten samozřejmě nemá stejný výkon jako hardware.

V roce 2017 přišlo velké nadšení ohledně linuxových distribucí určených pro nasazení na switche. “Nadšení postupně ochladlo a tolik se toho nakonec v následujících letech nedělo.” V roce 2021 přišla podpora DSA i do OpenWRT.

Režim DSA umožňuje rozhodnout, jak který provoz zpracujete. Zda jde o rámce, které chcete zpracovat v procesoru nebo je chcete nechat zpracovat switch čipem. Zvláštní typ provozu představují protokoly jako STP, RSTP, MSTP, LLDP nebo třeba LACP. Ty je potřeba zpracovat v CPU a ne standardním zpracováním s použitím VLAN.

V roce 2022 přibyla možnost nastavovat stav MSTI na jednotlivých portech. To řeší problém odpojení některých linek v rámci běžného STP. S použitím MSTP je možné rozdělit provoz podle VLAN a poslat různý provoz různými linkami, které tak dokážete rovnoměrně zatížit. “Původní implementace byla velmi nevhodná z hlediska standardu i hardware.”

Václav Nesvadba: Alternativy ladění sítě

Ještě stále musíme používat IPv4 a těch není k dispozici mnoho. “Snažíme se tedy je používat co nejefektivněji.” Pokud například propojujeme v síti dva prvky, obvykle se použije prefix /30, kde ale přijdeme o polovinu vyčleněných adres. Proto je lepší použít /31, ale ne všechny prvky toto nastavení podporují. “MikroTik na to má nedokumentované řešení, v Linuxu to funguje automaticky.”

Má smysl šetřit IPv6 adresy? “Jsou jich miliardy, ale stejně to dává smysl kvůli rychlejšímu přechodu a možnosti vyhnout se zbytečnému plýtvání.” Důležité je to také pro útoky typu *neighbor cache*. Útočník se při něm snaží vytvořit provoz na všechny adresy z daného rozsahu. “Může dojít k vypadnutí legitimních adres, které jsou pak nedostupné. Zkoušeli jsme to a funguje to.”

Někdy chce zákazník jen jednu IPv6 adresu, ale obvykle se mu dává obrovský rozsah. “Mně to přijde škoda, že použije jen jednu.” Je tedy možné vytvořit rozsah /120 a z ní přidávat zákazníkům třeba /124. “Pokud by zákazníků bylo víc, jen roztáhneme masku.” Při hledání problémů může také pomoci používat na konci IPv4 a IPv6 adres stejné hodnoty. “Je to přehlednější, než používat na rozhraní úplně jiné adresy.”

Ondřej Caletka: Pohled do zákulisí sítě pro RIPE meeting

RIPE meeting je setkání, které probíhá pravidelně dvakrát do roka pro více než 600 účastníků z celého světa. “Vozíme si celou svou Wi-Fi síť s vlastními IP adresami.” K dispozici je vlastní autonomní systém AS2121.

Uživatelé si často stěžují na problémy s geolokací. “Geolokační služby počítají s tím, že se AP na Wi-Fi nepohybují.” Tady se ale síť stěhuje a soukromé společnosti si udržují spoustu seznamů IP adres a jejich polohu.

RIPE se dohodl s Googlem a začal na svém webu zveřejňovat soubor `google.csv`, kde jsou zveřejněny rozsahy adres a jejich poloha. “Bylo to tak populární, že na to vzniklo RFC 8805.” Stejně je ale nutné obeslat největší poskytovatele geolokačních dat a oslovit je při každé změně.

Základ konferenční sítě tvoří dva malé servery SuperMicro, na kterých běží 25 virtuálních serverů zajišťujících směrování, firewally, DHCP servery, DNS resolvery a Wi-Fi kontrolery. “Zbytek už pak tvoří jen L2 switche od firem Juniper, Zyxel a MikroTik.” Páteř je tvořena 10GE porty, obvykle se ale provoz pohybuje ve špičkách okolo 800 Mbits.

Celá síť běží na open source: router dělá BIRD, firewall je provozován pomocí nftables a dále uvnitř běží Knot Resolver, Kea, Jool a další nástroje. “Vše je to orchestrováno pomocí Ansible.”

Veřejná síť je provozována v režimu IPv6-mostly, kdy moderní zařízení nemusejí čerpat IPv4 adresy. V praxi se tím ušetří většina IPv4 adres, místo šesti stovek jich pak stačí přibližně jedna stovka. Dále je k dispozici síť nabízející pouze IPv6 a klasická dual-stacková síť. Pak je tu několik sítí pro správu, samostatná síť pro prvky a malá oddělená síť pro video streaming. “Je to docela dost různých sítí.”

Lubomír Prda: OpenBMP – co se to sakra dělo v mé síti

BMP je zkratka pro BGP Monitoring Protocol a umožňuje dostat informace o BGP ze směrovače do analyzátoru. “Umožňuje to zjistit, že máte v síti nějaký problém.” Jeden z kolektorů pro data se jmenuje OpenBPM a původně se jmenoval SNAS. “Ještě je možné se s tím setkat u některých komerčních zařízení, u kterých je uvedeno, že je možné je napojit na SNAS.”

Jedním z výstupů je *Looking Glass*, který vypíše podrobnosti o zadané IP adrese. “Stahuje si to informace i z různých databází jako je geolokace a další.” Tyto informace se stahují na pozadí, takže jsou dostupné, i když není původní zdroj dostupný.

Je možné analyzovat i podrobnosti týkající se BGP, kdy je možné sledovat například počet aktualizací od jednotlivých peerů. K dispozici je kompletní historie, takže lze zpětně zkoumat, jak docházelo k různým změnám a proč se chování sítě mění.

(Autorem fotografií je Petr Krčmář.)