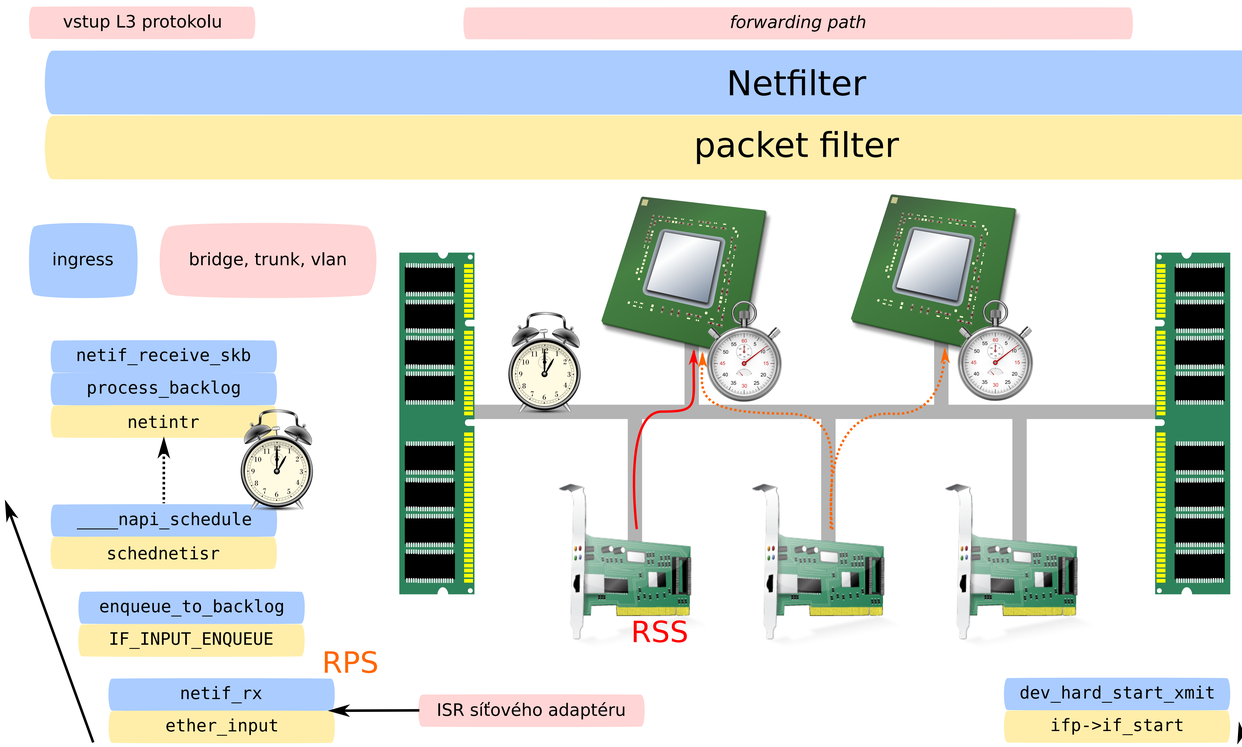
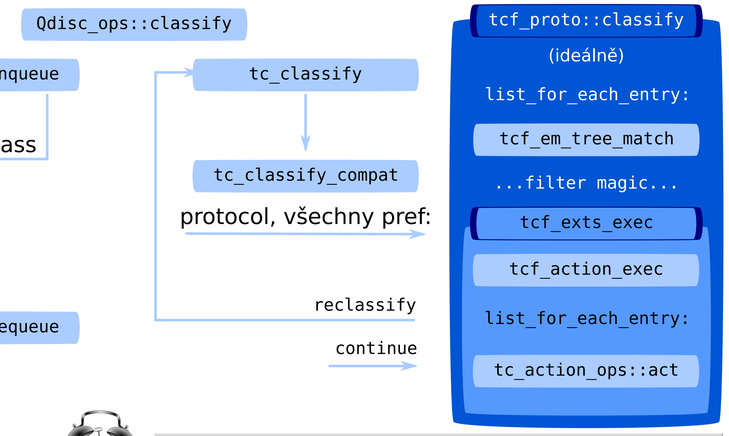
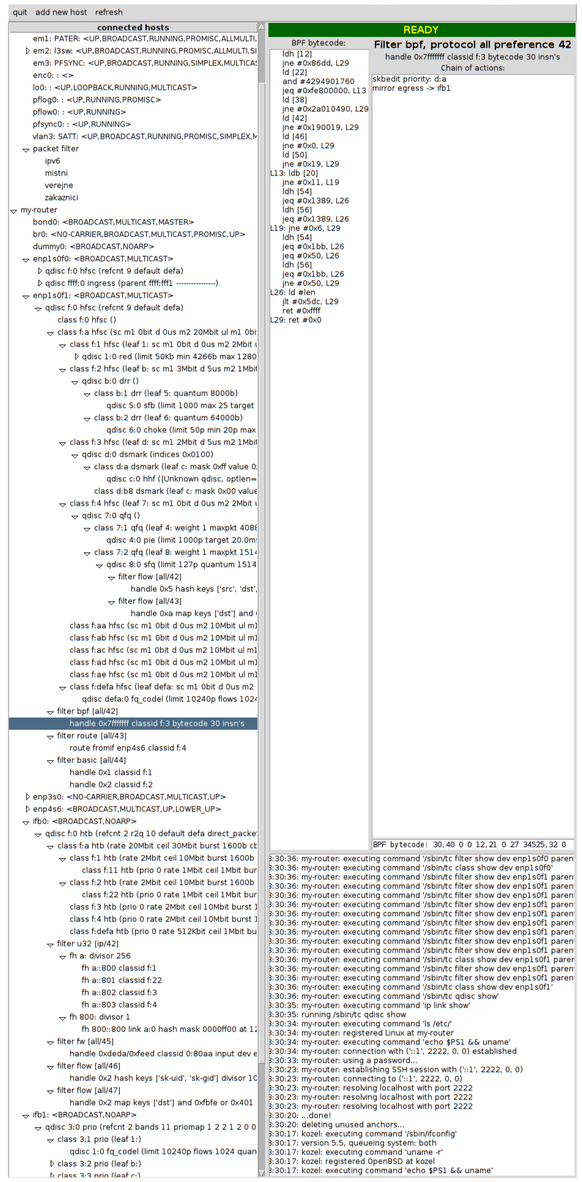
Optimalizace využití kapacity sítí na unixových systémech







Plánování provozu

prioritní FIFO: Oba systémy ve standardním nastavení respektují IPv4 bajt *Type of Service* a upřednostňují například provoz označený *lowdelay*. OpenBSD navíc priorizuje BPDU, LACPDU, CARP announcementy a prázdné TCP pakety s flagem ACK, což pomáhá zejména ucpaným asymetrickým DSL připojením udržet rychlost i interaktivitu.

DRR: Každá fronta má *kvantum* bajtů, standardně MTU, o které se zvyšuje její *deficit*. Při odbavení paketu se deficit sníží o jeho délku.

fq_codel: Návrh na výchozí algoritmus pro masové používání k zajištění nízké doby odezvy v interaktivních aplikacích za všech okolností. Minimální variantu SFQ následuje DRR, za nímž je CoDel. Standardně nepodporuje priority, nutno použít *prio* nad ním.

SFQ: Hashovací funkce nastavitelná filtrem `flow` řadí toky do `divisor` front, standardně 1024. Periodická změna této funkce přeuspořádává pakety v rámci spojení, standardní funkce hashuje i čísla portů, pročež plánovač nerozezná více spojení téhož programu a tedy není spravedlivý.

QFQ: Třída k patří do skupiny i podle váhy Φ_k a jejího nejdelšího paketu L_k tak, že $i = \lceil \log_2(L_k/\Phi_k) \rceil$. Tedy pro pakety mezi 64 B a 16 kB je méně než 32 skupin. Číslo skupiny je index do bitových polí, podle kterých se plánuje. Bitové pole *(eligible, ready)* obsahuje pakety, které již měly odejít a nejsou blokovány těmi s vyšším indexem skupiny, které již měly dojít. Pole *(eligible, blocked)* obsahuje bity u skupin, před nimiž je důležitější provoz. *Ineligible* varianty jsou udržovány, aby mohly být včas povýšeny na *eligible*. V rámci skupiny jsou pakety bucket-sortem seřazené podle času odchodu, algoritmus funguje s asymptoticky konstantní složitostí.

HHF: k hashovacích funkcí klasifikuje paket přičtením jeho velikosti do k přihrádek filtru. Toky s velkým objemem dat přetečou `admit_bytes` ve všech svých přihrádkách a jdou na seznam `hh_limit` toků, dokud se po `evict_timeout` sekund neuklidní. Plánovač je vážený *round-robin*, váhy 1 resp. `non_hh_weight`. Zřejmě nemá *false-positiva*.

Rozdělování kapacity linky

CBQ: Strom tříd s omezením rychlosti, prioritou a možností půjčovat kapacitu ostatním nebo od ostatních. Plánovač měří, jak dlouho byla linka v klidu, a pamatuje si EWMA limitovaný `maxburst` pakety o průměrné velikosti `avpkt`. Odesílá `minburst` paketů najednou, kvůli kompenzaci rozlišení časovače. Třídy jsou plánovány váženým *round-robin* a provoz je pouze v listech. Vnitřní uzly slouží k půjčování kapacity mezi podstromy.

HFSC: Strom tříd s po dvou částech lineárními křivkami omezujícími dobu, za kterou bude nejdříve možné data odeslat. Provoz je pouze v listech.
Real-time plánovač obsluhuje třídu s nejbližším možným časem odeslání, zbytek je rozdělen podle nejmenšího virtuálního času na dané úrovni.
Úroveň uzlu při plánování nehraje roli, pouze virtuální čas spočtený z křivek.

HTB: Strom tříd se zaručenou a maximální rychlostí, prioritou na dané úrovni a kvantem k půjčování třídám níže. Uzel je zelený, má-li připravená data k odeslání a tehdy je na *globálním seznamu* v poli indexovaném prioritou a úrovní. Žlutý uzel má půjčenou kapacitu od svého rodiče a je na jeho *interním seznamu*. Červený uzel přečerpal dostupnou kapacitu a čeká ve frontě na dané úrovni na čas nejbližší změny barvy.
Úroveň uzlu tedy při plánování hraje roli; uzly výše mají přednost.

Active Queue Management

IEEE 802.3x Flow Control: Zahlcená strana pošle *pause frame* s dobou, po kterou nechce být rušena. Fronta u úzkého hrdla zůstane malá, ale na tranzitním portu se zdrží i interaktivní provoz. Globální synchronizace.

RED: Je-li EWMA délky fronty větší než $\min_{th}$ paketů, s pravděpodobností p_a zahazuje algoritmus provoz; p_a se časem zvyšuje. Váha průměru se nastavuje pomocí odhadu burstu paketů zadané *průměrné velikosti*. Frontě s příliš velkým průměrem se zahazuje vše. Nic neříkající problémová magie.

SFB: Délka fronty nehraje roli, stavy jsou *prázdná* a *plná*. Pravděpodobnost zahození paketů se zvyšuje při zahození a snižuje v klidu. Varianta SFB umí stochasticky zohlednit datové toky Bloomovým filtrem.

CoDel, vysl. [kodl]: Maximální tolerovaná doba zpoždění vzniklá na daném stroji, `target`, je standardně 5 ms. Pokud má po `interval` (100 ms) všechen provoz zpoždění větší než `target`, algoritmus stále více zahazuje. Pomalému připojení algoritmus `target` posune podle MTU. Je třeba měřit dostatečně přesně čas všem zpracovaným paketům.

PIE: Odhaduje zpoždění po blocích a uvažuje i jeho tendenci růst. Krátkodobé bursty nemají zásadní vliv, neboť pravděpodobnost zahazování se mění po malých krocích; přesto je povolen `burst_time` (100 ms) upravovaný jako deficit u DRR. Měření času je tedy třeba méně často.

Klasifikace

OpenBSD používá ke klasifikaci dokumentovaný firewall pf. Linux kromě cílů systému Netfilter (např. CONNMARK, MARK, CLASSIFY) obsahuje navíc nedokumentovaný subsystém *filtrů*. Vyhoví-li, mohou kromě přiřazení čísla třídy navíc vykonávat *akce*.

basic: Klasifikuje pouze pomocí *extended matches*, což jsou jednoduché klasifikátory s logickými spojkami schopné kromě porovnávání obsahu několika způsoby též používat metadata poskytovaná jádrem. Původně měly sloužit k nahrazení jednodušších klasifikátorů a jako doplněk složitých.

bpf: Berkeley Packet Filter jádro na některých architekturách umí přeložit JIT do nativního kódu a výpočetní omezenost jazyka zaručuje bezpečnost takovýchto programů. Čtení ze záporných adres obsahuje užitečná metadata, ale filtr zatím bohužel není plnou náhradou za všechny ostatní.

flow: Vrací třídu jako výsledek `baseclass` + x % `divisor`, kde x je buď hash zadaných klíčů nebo mapování aritmetického výrazu nad některým.

fw: Mapuje značku systému Netfilter na konkrétní třídu/qdisc.

route: Klasifikace na základě rozhraní či realmu zdrojové/cílové cesty.

tcindex: Původně navržen jako doplněk disciplíny **dsmark** umí použít dříve uložená data do proměnné `skb->tcindex`, například hodnotu ToS či DSCP. Nepodporuje žádné akce kromě **police**, současná verze jádra spadne na validaci akcí. Plně nahraditelný např. filtrem **basic** s akcí **skbedit**.

u32: Klíči jsou 32bitové hodnoty z klasifikovaného paketu. Kombinací klíčů lze vytvořit víceúrovňovou hashovací tabulku, ve které filtr najde požadované číslo třídy. Možnost nepřímého adresování.

Akce k filtrům

Generické akce (`gact`) jako **drop**/**pass**/**reclassify**/**pipe**/**continue** se umějí pohybovat po subsystému, akce **ipt** umí spustit na právě klasifikovaný paket daný cíl systému Netfilter, akce **mirred** umí paket duplikovat či přesměrovat na jiné rozhraní, což se hodí třeba při omezování rychlosti příchozích dat (*ingress shaping*). Akce **nat** přepisuje bezstavově IPv4 adresy. **pedit**, **skbedit** a **csum** umějí upravit paket, změnit metadata či přepočítat kontrolní součet. Původně jediná akce **police** umí omezovat rychlost parametry `rate` a `burst`; jelikož se jedná o policer, nevyhovující provoz se nezdrží, ale aplikuje se na něj jedna ze specifikovaných akcí.

Našli jste chybu? Pošlete ji na pelikan@storkhole.cz. http://qviz.storkhole.cz/