

Ladění propustnosti TCP přenosů programem Tbwtools

Sven Ubik, CESNET
Olomouc, 29.-30. května 2007

1. Úvod

Páteřní linky sítí mají v současné době instalovanou kapacitu řádu jednotek až desítek gigabitů za sekundu. Přesto bývá často propustnost přenosů dat přes síť nižší než požadovaná a očekávaná. Příčina může být v kterémkoliv komponentu komunikační trasy nebo v jejich interakci:

- Nedostatečná volná kapacita některé linky v trase
- Porucha nebo limitace hardwarového komponentu v trase (např. špatný transceiver, neshoda full a half duplex, přetížení procesoru směrovače)
- Limitace hardware koncové stanice (např. přetížení procesoru, PCI sběrnice, paměťové sběrnice)
- Limitace operačního systému koncové stanice (např. malé soketové vyrovnávací paměti, velká režie obsluhy přerušení, neoptimální konfigurace řízení zahlcení v TCP)
- Limitace aplikace (např. neoptimální využití kapacity procesoru pro zpracování dat)

Při lokalizaci limitujícího komponentu obvykle používáme kombinaci technik a nástrojů, například:

- Test propustnosti programem iperf
- Odchycení paketů programem tcpdump a následná analýza programy tcptrace a xplot
- Sledování runtime proměnných TCP rozšířením jádra web100
- Sledování charakteristik na koncové stanici knihovnou PAPI (Performance Application Programmable Interface)
- Sledování a konfigurace soketových vyrovnávacích pamětí nástrojem sysctl
- Sledování aktivity přerušení v souboru /proc/interrupts
- Sledování aktivity volání jádra nástrojem strace

Poslední čtyři úlohy stačí obvykle provést pouze jednou. Naproti tomu analýza paketů získaných z nástroje tcpdump a runtime proměnných TCP obvykle vyžaduje několik iterací pro různé případy provozu v síti a manuální srovnání údajů z různých zdrojů, jde tedy o pracnou a časově náročnou úlohu.

Pro snazší a rychlejší práci s údaji z různých zdrojů jsme proto navrhli sadu nástrojů a framework nazvaný tbwtools. Tento framework používá kombinaci aktivního monitorování, pasivního monitorování a monitorování koncové stanice a je celý instalovaný na koncových stanicích, nevyžaduje žádný přístup k prvkům síťové infrastruktury.

2. Požadavky

Na základě našich zkušeností s nástroji pro ladění propustnosti TCP přenosů jsme definovali následující požadavky pro vyvíjený framework:

- Zobrazení charakteristik z nástroje tcpdump (např. průběh propustnosti, frekvence opakování paketů)
- Zobrazení runtime proměnných TCP z web100 (např. vnitřní stav konečného automatu TCP, průběh RTT)
- Zobrazení soketových vlastností z nástroje bulk (např. průběh okénka zahlcení - cwnd a limitu fází slow start a prevence zahlcení - ssthresh)
- Všechna charakteristiky musí být korelovány v časové ose a prezentovány v jednotném uživatelském rozhraní
- Uživatel musí stačit standardní webový prohlížeč bez instalace dalších programů
- Uživatel musí mít možnost testovat trasu buď mezi svým PC a vzdáleným serverem nebo mezi dvěma vzdálenými servery
- Framework musí komunikovat se systémem perfSONAR vyvíjeným pro monitorování v síti GN2 v rámci aktivity JRA1 GN2

Předpokládáme následující způsoby použití:

- Zkušený uživatel provede testovací spojení a použije uživatelské rozhraní tbwtools pro prohlédnutí výsledků a provede sám analýzu a predikci příčin nízké propustnosti
- Uživatel provede testovací spojení, získá data ID, pošle ho síťovému odborníkovi, který použije data ID pro získání výsledků a provede analýzu
- Uživatel poskytne do tbwtools soubor s odchycenými pakety získaný externě a požádá a výpočet a zobrazení charakteristik (pochopitelně tato varianta umožňuje získat pouze charakteristiky z odchycených paketů a ne z jiných zdrojů)

3. Architektura

Komunikace mezi komponenty frameworku tbwtools je znázorněna na obrázku 1. Uživatelské rozhraní je implementováno jako applet tbwApplet do webového prohlížeče v jazyce Java a je nahráno do webového prohlížeče po jeho nasměrování na adresu stránky, kde je applet uložený.

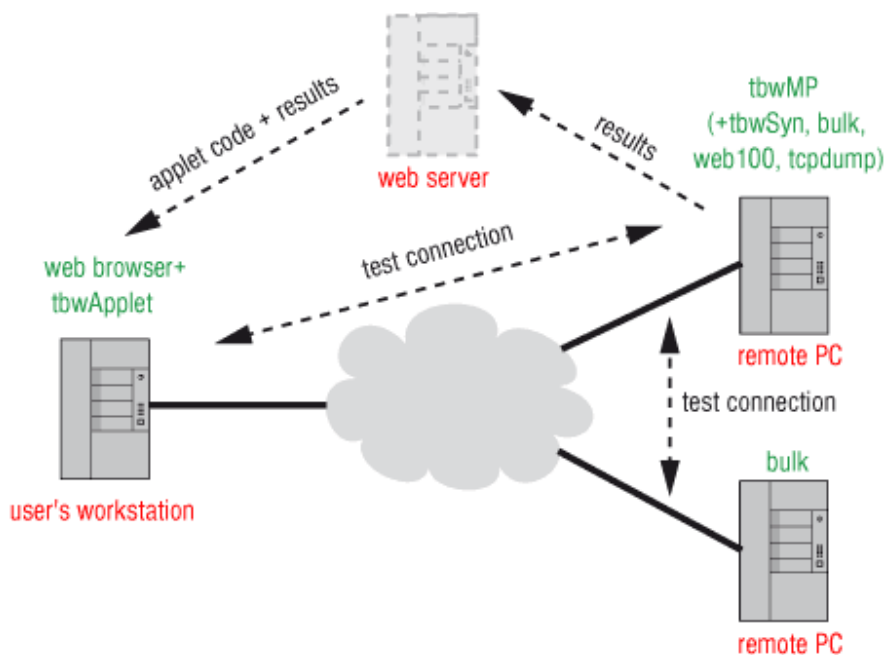
Ladění propustnosti probíhá ve dvou fázích. V první fázi je provedeno testovací spojení a jsou sbírána data o tomto spojení. V druhé fázi jsou získaná data zpracována a prezentována.

Fáze 1 - měření

Testovací spojení může být provedeno a) mezi PC uživatele a vzdáleným serverem nebo b) mezi dvěma vzdálenými servery a to ve zvoleném směru. Spojení je provedeno nástrojem bulk, který jsme vytvořili. Jde o samostatně použitelný program pro stresový test propustnosti trasy obdobný

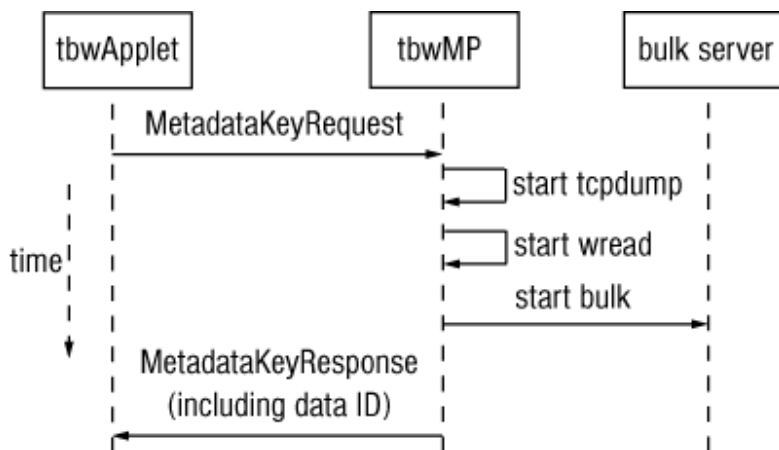
nástroji iperf. Oproti nástroji iperf provádí nástroj bulk navíc sběr hodnot zadaných soketových proměnných během spojení. Pro naše účely je nejzajímavější vlastnost TCP_INFO.

TbwApplet má zabudovanou testovací část programu bulk. Proto je možné provést test z nebo na PC uživatele. Monitorovací část programu bulk není v tbwApplet přítomna, protože tbwApplet je obvykle provozován v operačním systému Windows, který neposkytuje přístup k soketovým vlastnostem.



Obr. 1: Architektura tbwtools

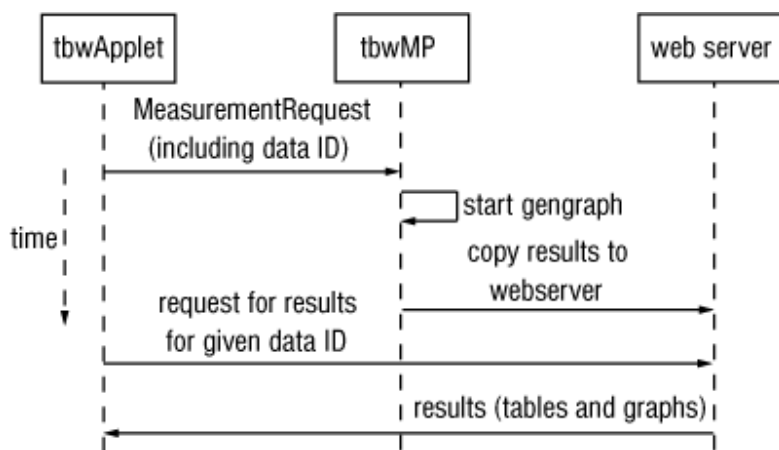
Vzdálený server musí mít spuštěný měřící bod - Tbwtools Measurement Point (tbwMP). TbwApplet a tbwMP komunikují posíláním zpráv ve formátu XML kompatibilním se systémem perfSONAR. Použití tbwApplet není nutné. Uživatel může sám sestavit požadavek na tbwMP ve formátu XML a pomocí klienta systému perfSONAR komunikovat s tbwMP. Zprávy používané v 1. fázi jsou znázorněny na obrázku 2. TbwMP spustí na základě požadavku nástroje pro sběr dat - tcpdump a wread pro runtime TCP proměnné a spustí nástroj bulk se zabudovaným monitorováním soketových vlastností pro testovací spojení.



Obr. 2: Testovací spojení

Fáze 2 - zpracování a prezentace výsledků

Zprávy použité v 2. fázi jsou znázorněny na obrázku 3. TbwApplet prezentuje uživateli shrnutí spojení - objem přenesených dat, podíl času kdy bylo spojení omezeno vysílačem, příjemcem a sítí a atd. TbwApplet dále prezentuje formulář pro výběr charakteristik, které mají být určeny ze získaných dat a prezentovány. K výpočtu charakteristik slouží program gengraph a několik dalších programů z něho volaných. Výsledky ve formě grafů jsou následně přeneseny na web server.



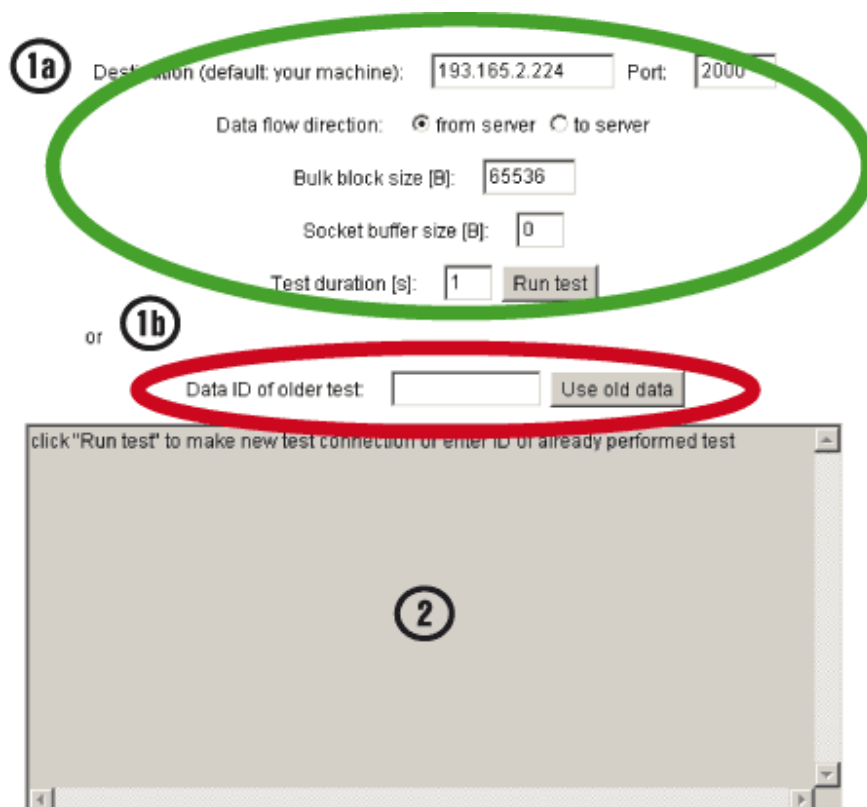
Obr. 3: Zpracování a prezentace výsledků

4. Příklad použití

Uživatelské rozhraní ve formě apletu tbwApplet je znázorněno na obrázku 4.

První možností (1a) je provedení testovacího spojení. Zadáme cílovou IP adresu, port, směr spojení, velikost bloku dat (poslaného jedním voláním send()), velikost soketové vyrovnávací paměti a trvání spojení v sekundách. Cílová IP adresa je předvyplněna jako IP adresa uživatele PC (kde běží tbwApplet) a můžeme ji změnit na jinou adresu některého tbwMP. Druhou možností (1b) je zadat data ID získané dříve provedeným testem.

XML zprávy mezi tbwApplet a tbwMP jsou pro informaci vypsány v okénku v dolní části uživatelského rozhraní (2).



Obr. 4: Uživatelské rozhraní - tbwApplet

Po provedení testovacího spojení nebo zpracování data ID může uživatel vybrat charakteristiky, které mají být určeny a prezentovány. Uživatel může zvolit kombinaci charakteristik získaných z odchycených paketů (3a), soketové vlastnosti (3b) a TCP runtime proměnné (3c). Nakonec uživatel zvolí časové období spojení, které má být zpracováno (4) a časový interval (krok), pro který mají být charakteristiky vypočteny.

3a

Tcpdump variables

seq ack window flightsize
 throughput pktdups ackdups rtt

3b

Bulk variables

state ca_state retransmits probes
 backoff options wscalen rto
 ato snd_mss rcv_mss unacked
 sacked lost retrans packets
 last_data_sent last_ack_sent last_data_rcv last_ack_rcv
 pmtu rcv_ssthresh rtt rttvar
 snd_ssthresh snd_cwnd advmss reorderir

3c

Web100 variables

CurRwinRcvd CurAppRQueue CurReasmQueue DupAcksOut
 LimRwin CurRwinSent CurAppWQueue CurRebQueue
 CurMSS CurRTO CountRTT SumRTT
 RTTVar SmoothedRTT SampleRTT DSACKDups
 AckAfterFR NonRecovDA RetranThresh QuenchRcvd
 SendStall ECERcvd PostCongCountRTT PostCongSumRTT
 PreCongSumRTT PreCongSumCwnd SACKBlocksRcvd SACKsRcvd
 DupAcksIn BytesRetrans PktsRetrans AbruptTimeouts
 CurTimeoutCount SubsequentTimeouts Timeouts FastRetran
 LimCwnd CurSsthresh CurCwnd CongestionOverCoun
 OtherReductions CongestionSignals CongAvoid SlowStart
 SndLimTimeRwin SndLimBytesRwin SndLimTransRwin SndLimTimeCwnd
 SndLimBytesCwnd SndLimTransCwnd SndLimTimeSender SndLimBytesSender
 SndLimTransSender ThruBytesReceived RcvNxt ThruBytesAcked
 SndMax SndNxt SndUna DataBytesIn
 DataPktsIn PktsIn DataBytesOut DataPktsOut
 PktsOut MSSRcvd ActiveOpen State

Select

4

Timeline zoom: from ms to ms (leave empty for whole time)

Base period for tcpdump-computed variables [ms]:

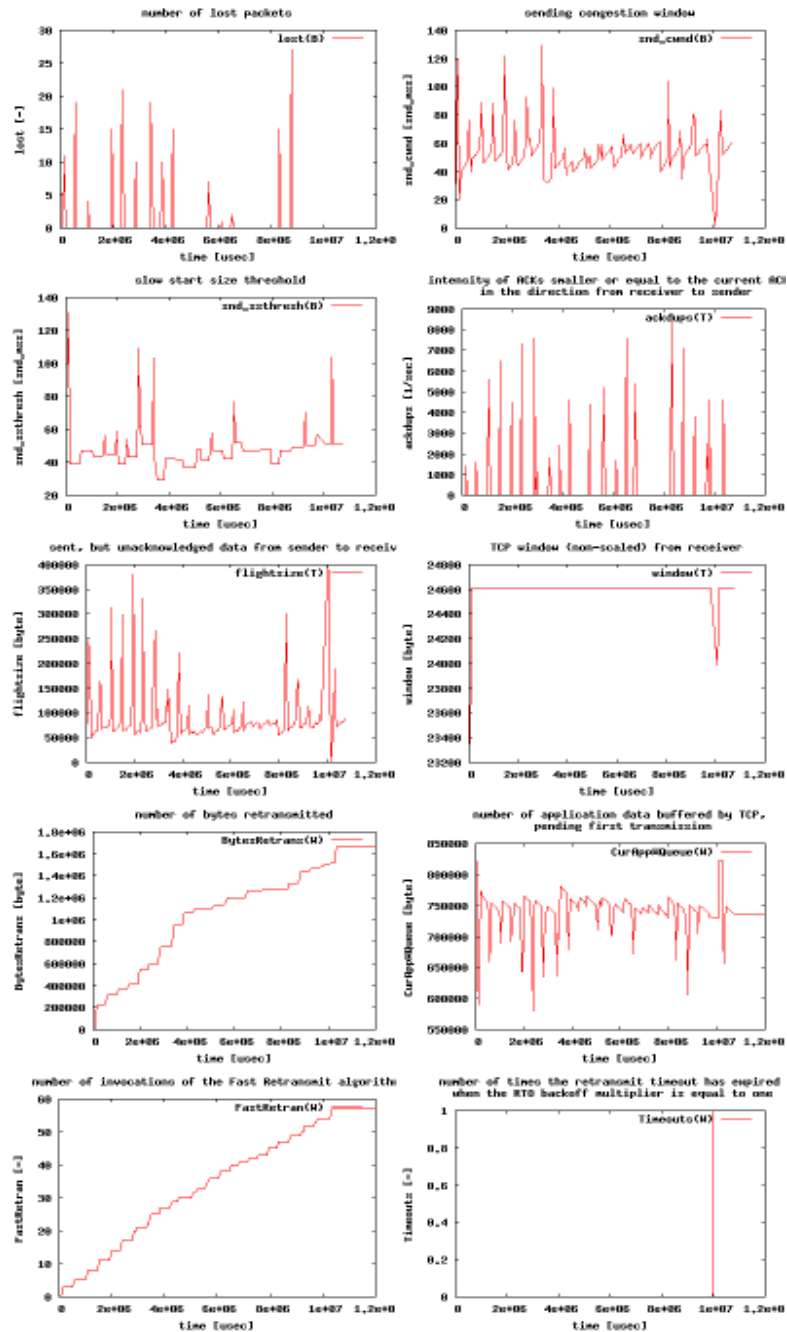
Sample rate (average data over this period) [ms]:

Obr. 5: Výběr charakteristik v *tbwApplet*

Příklad určených charakteristik je znázorněn na obrázku 6. Zleva doprava a shora dolů jde o následující charakteristiky:

- Frekvence ztrát paketů, z nástroje bulk
- Okénko zahlcení (cwnd) z nástroje bulk
- Slow start threshold (sssthresh) z nástroje bulk
- Frekvence duplikací potvrzení z odhycených paketů
- Objem nepotvrzených dat (flight size) z odhycených paketů
- Okénko přijímače (rwin) z odhycených paketů
- Objem přenesených bajtů z web100
- Dosud neodeslaná data v soketové vyrovnávací paměti z web100
- Počet aktivací mechanismu Fast Retransmit z web100
- Frekvence vypršení časovače opakování paketů (retransmission timeout) z web100

V případě tohoto spojení můžeme vidět vysoký počet ztracených paketů, ale jen malý počet vypršení časovače opakování paketů. Většina opakování paketů proběhla v mechanismu Fast Retransmit. Ploché okénko přijímače ukazuje, že přijímač nebyl přetížen a stále přítomný objem dosud neodeslaných dat v soketové vyrovnávací paměti vysílače ukazuje, že vysílač byl schopen trvale poskytovat data k odeslání. Celkově šlo o normálně probíhající spojení, které ale soupeřilo o volnou kapacitu trasy po většinu doby spojení.



Obr. 6: Příklad charakteristik získaných pomocí *tbwtools*

Tbwttools také prezentuje následující shrnutí průběhu spojení:

```
Transported 91226112 bytes in 10.796 s, connection speed 67.600 Mbps
There were 1146 packets retransmitted, 5097 duplicate ACKs received
and 0 SACK blocks received.
Segment size was 1448 B, average round-trip time 8.804 ms,
packet loss rate 0.090413%
The connection was receiver limited 1.68% of time
The connection was network limited 98.23% of time
Maximal window size of your machine was 384 kB which limits the throughput
to 357.899 Mbps
```

5. Shrnutí

Vytvořili jsme sadu nástrojů a framework Tbwttools, který aktivně testuje propustnost trasy se souběžným monitorováním testovacího spojení různými metodami. Charakteristiky získané z naměřených dat z různých zdrojů jsou prezentovány v jednotném uživatelském rozhraní.

Plánujeme upravit známý program iperf tak, aby obsahoval monitorovací funkce nástroje bulk. Předpokládáme, že tím usnadníme přijetí systému uživateli, kteří jsou zvyklí používat program iperf.