Test Your Limits With TRex Traffic Generator

Hanoch Haim
Cisco systems

***Abstract*—Performance measurement tools are an integral part of network testing. There is no shortage of open source tools for network performance testing in the Linux world. To enumerate a few popular tools in the Linux world: Netperf, iperf, Linux kernel based pktgen. These tools tend to fall into two categories:**

- **Stateless packet shooting such as the linux kernel pktgen traffic generator**
- **Stateful client-server tools such as netperf and iperf.**

When very high performance network performance testing is required (quantified as many 10s of Gigabits per second/100MPPS and/or hundreds of thousands of flows) or more advanced functionality (e.g. realistic) is required the linux classical tools are insufficient. Most organizations will opt for very expensive commercial tools such as Ixia, Spirent. In this paper we will introduce TRex a high performance realistic traffic generator and illustrate sample stateless and stateful use cases that apply to testing Linux networking. We will also discuss its design and tricks that help us achieve such high performance.

***Index Terms*—tcp, performance, scale,realistic traffic generation**

I. Introduction

TRex [**?**] is an advanced traffic generator, it has the following interesting features:

- It leverages COTS x86/ARM servers and Physical NICs (Intel, Mellanox etc) for high scale
- Can support linux driver or paravirtual (e.g. virtio) for low scale advance feature
- It can serve both Stateless and Stateful traffic generation. tcp stack for stateful traffic and emulation layer to simulate L7 applications
- It outperforms all of iperf, netperf and pktgen
  - It can generate upto 200gbps/100mpps advance traffic pattern and millions of real world tcp/udp flows.
  - High connection rate - order of Millions of Connection/s (CPS)
- It is extensible
  - Emulate L7 application (on top of TCP/IP), e.g. HTTP/HTTPS/Citrix using a programable language
  - Ability to change fields in the L7 application - for example, change HTTP User-Agent field
- Support routing protocols like BGP/OSPF/RIP using BIRD [**?**]
- Support high scale client simulation protocols (arp.ipv6-nd,dhcp,802.1x,icmp)

Although TRex is implemented on top of DPDK, a lot of the issues we had to deal with when writing the tool apply equally to scaling Linux networking; we share our experiences in that regard and hope to inspire some of the techniques to be adopted in Linux.

II. Software design high level

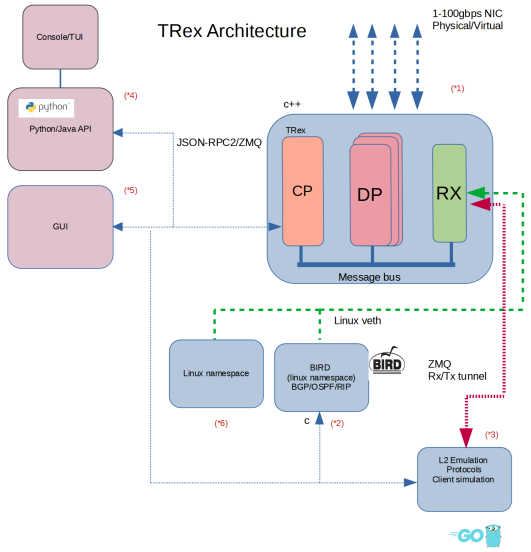

Fig. 1: architecture

Figure **??** presents the main processes. TRex server (*1) is a multi-threaded process, each thread is pinned to a core and works in event driven fashion using a user-space scheduler with a few hierarchy. There is one control-plane (CP) thread that handle the RPC over ZMQ requests and maintenance tasks. The RX thread responsible to handle the low latency traffic for accurate latency measurement. This thread is usually in very low CPU utilization. The DP threads are generating the traffic using and interact with DPDK to transmit/receive traffic via PMD queues. The number of DP threads can be scale up as the number of Tx/Rx queues. There is almost no sharing data structure and no locks to get to best performance. Any information is moved between the threads using a messaging bus which is a shared rings (DP→CP, CP→DP, Rx→DP, RX→CP). There is minimum system call to the kernel (only when required for example with PF_PACKET/AF_XDP driver) (*4) is a Python wrapper to the JSON-RPC2 API over ZMQ so it would be easy automation (e.g. load a profile, get statistics etc) on top of Python API there is a Console that can simplify the way to work with the API. (*5) is the GUI process that written in Java that works directly with the JSON-RPC2 and supports only the Stateless mode. (*2) is a BIRD [**?**] process that works inside a linux namespace and

connected to TRex via a programmable veths. Inside TRex RX thread there is a Switch that forward packets to/from the veth related to the linux namespace. BIRD is used to simulate routing protocols like BGP/OSPF/RIP. (*6) and (*3) is used for simulating clients slow-path protocols like ARP/IPV6-ND/IGMP/MLD/802.1x/DHCP/DHCPv6 while TRex server is for the fast-path high speed TCP/UDP

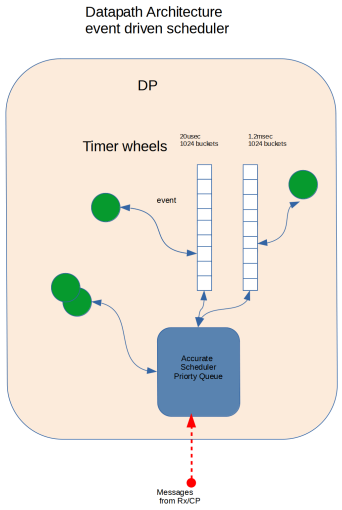

Fig. 2: DP Scheduler

Figure **??** presents the schedulers in each DP thread. The priority queue is the low level scheduler that can schedule events in nsec resolution. On top of that there are two timer wheels for lower resolution events. The first has resolution of 20usec with 1024 buckets for maximum of 2msec time. the second timer wheel has resolution of 1msec with 1024 buckets maximum of 1sec. Each event in the second level is spread each 20usec tick to reduce processing spikes. DP transmit/receive messages from the share rings using events. This design gives linear scale with of performance about 4-20MPPS/core and 200gbps for one COGS server.

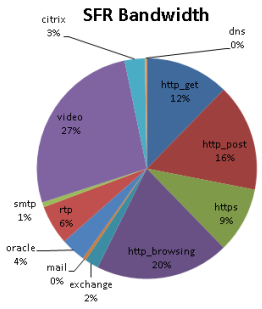

Fig. 3: emix

From functionality point of view TRex has two main operation modes. Stateful and Stateless. Stateful is for testing L7 services that care about clients/flows/L7 application like DPI/NAT/Firewall, Figure **??** is an example of mix of traffic that can be generated using this mode. Stateless is for testing Switch/Filters/ACL/Qos services and has no flow/client state

III. Stateless mode

Stateless traffic model is for testing device under test (DUT) that is relatively simple and does not care about flow/client context.

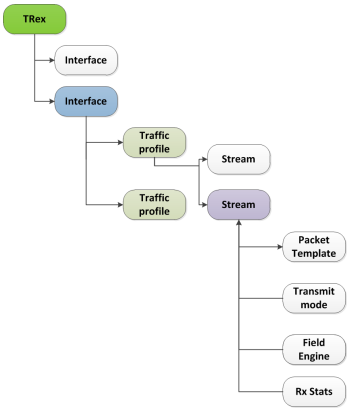

Fig. 4: Stateless objects

Figure **??** shows the model of a profile. Each interface supports one or more traffic profiles in parallel. Each traffic profile supports one or more streams. Each stream includes

- **Packet**: Packet template up to 9 KB
- **Field Engine**: A program that determines which field to change and how to change
- **Mode**: Specifies how to send packets [Continuous,Burst,Multi-burst]
- **Rx Stats**: Which statistics to collect for each stream
- **Rate**: Rate (pps or bps)
- **Action**: Specifies the next stream to follow when the current stream is complete (valid for Continuous or Burst modes)

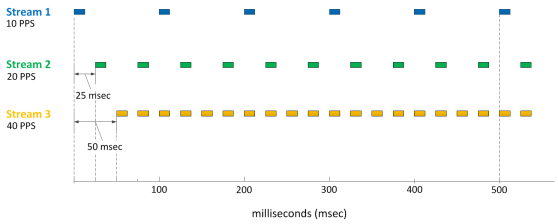

Fig. 5: Stateless profile example

Figure **??** shows an example of a profile with three streams. Stream 1 is in rate of 10pps stream 2 has a rate of 20 pps and streams 3 has rate of 40 pps. All are continues mode.

List **??** shows a profile with one UDP stream. The mode is Continues. It uses Scapy for building the template packet.

Listing 1: Profile with one continues UDP stream

A. Stateless Features

- Large scale - Supports about 10-22 million packets per second (mpps) per core, scalable with the number of cores
- Support for 1, 10, 25, 40, and 100 Gb/sec interfaces with Linux driver vs PF_PACKET
- Support for multiple traffic profiles per interface
- Programable Field Engine to change any field inside the packet (e.g.)
- Ability to change the packet size
- API,Console, GUI
- Statistics, per interface,per stream
- Latency and jitter per stream
- Multi-user support

B. Multi stream example

Figure **??** shows two streams the Stream 0 is a burst stream that activate a multi-burst stream 1 (With 5 burst of 4 packets). List **??** shows the python profile to create this profile

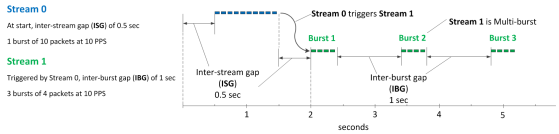

Fig. 6: Multi stream profile

Listing 2: Multi profile example

C. Field Engine

The field engine (FE) is a programable engine that can change any field in the packet and is part of the profile and compiled to bytecode in the TRex server. The challenge was to provide an engine that change packet fields on number of core in parallel but as a black-box behave like it runs on a single thread (hardware like) Let gives an example of syn-attack and explain it

Listing 3: FE syn attack 48.0.0.1 server

List **??** shows a FE program that generate a syn attack using one stream. every stream has a context of a FE. in this context in this example there are two variables **ip_src** for the range of the source ipv4 ips and the **source_port** for the range of the source protocols. Those variables are written to the right offset in the packet and fix the checksum (using hardware assist if exists)

D. Automation using Python API

List **??** shows a simple script to automate TRex. it self explained

Listing 4: Stateless automation example

Table 1. Traffic Profile = Field Engine Cached 255

Packet size	Line Utilization (%)	Total L1 (Gb/s)	Total L2 (Gb/s)	CPU Util (%)	Total MPPS	BW per core (Gb/s) ❶	MPPS per core ❷	Multiplier
Imix	100.04	80.03	76.03	2.7	25.03	89.74	28.07	100%
1514	100.12	80.1	79.05	1.33	6.53	430.18	35.07	100%
590	99.36	79.49	76.89	3.2	16.29	177.43	36.36	99.5%
128	99.56	79.65	68.89	15.4	67.27	36.94	31.2	99.5%
64	52.8	42.3	32.23	14.1	62.95	21.43	31.89	31.5mpps

Table 2. Traffic Profile = Field Engine with 1 variable

Packet size	Line Utilization (%)	Total L1 (Gb/s)	Total L2 (Gb/s)	CPU Util (%)	Total MPPS	BW per core (Gb/s) ❶	MPPS per core ❷	Multiplier
Imix	100.04	80.03	76.03	12.6	25.03	45.37	14.19	100%
1514	100.12	80.1	79.05	2.6	6.53	220.05	17.94	100%
590	99.36	79.49	76.89	5.6	16.29	101.39	20.78	99.5%
128	99.56	79.65	68.89	33.1	67.27	17.19	14.52	99.5%
64	52.8	42.3	32.23	31.3	63.06	9.65	14.37	31.5mpps

❶ Extrapolated L1 bandwidth per 1 core @ 100% CPU utilization.
❷ Extrapolated amount of MPPS per 1 core @ 100% CPU utilization.

Fig. 7: Stateless performance with Intel XL710

E. Stateles Performance

Figure **??** [**?**] shows the measured performance on a Cisco UCS server with dual socket and two Intel XL710 NICS

IV. Stateful mode

Stateful model objective is to simulate realistic L7 applications on top of TCP/UDP stack (based on BSD source code) at high scale. The scale could reach millions of flows and 100k clients/servers up to 200gbps for one server. It is important to test Stateful features using realistic traffic scenarios because this is the only way to estimate more accurate performance metrics and identify bottlenecks in the design. Figure **??** present the traffic generation model.

Each profile includes:

- **Client pool**: Range of clients with a distribution model e.g. random,seq. a profile can include a few pool
- **Server pool**: Range of servers, profile can include a few pools
- **Template**: A model that describe an application on top of TCP/UDP. A pcap for L7 data can be taken
  - **CPS**: How many connection per second for this template
  - **flowLimit**: to limit the flows

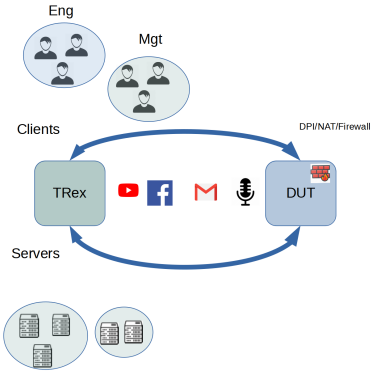

Fig. 8: Stateful model

In this model each core has its own context of TCP/UDP stack with no sharing and no locks. It works on top of the scheduler hierarchy in Figure **??** The changes to the BSD stack were

- Each stack has a context per thread
- Tx work in pool mode (build the packets only when required) and save reference to the template data. This saves three order of magnitude of memory resource

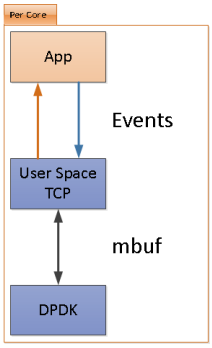

Fig. 9: Stateful stack

Figure **??** shows the stack of the programable application emulation layer. this module responsible to simulate application on top of the TCP/UDP stack. Example for commands are:

- Start write buffer
- Continue write
- End Write
- Wait for buffer/timeout
- OnConnect/OnReset/OnClose

A. *Stateful Profile*

Listing 5: Stateful profile

List **??** shows a simple profile with one pool of client (16.0.0.1-16.0.0.254) and one pool of servers (48.0.0.1-48.0.255.254) The pcap file is parsed and the L7 data is converted to instructions on top of the TCP stack

B. *Emulation layer instructions*

List **??** shows a simple example of a low level instructions of the emulation layer. This example the client send request and wait for response while the server wait for request and send a response.

Listing 6: Emulation layer instructions

Listing 7: Client pseudo code

Listing 8: Server pseudo code

C. *Features*

- high scale (flows, bandwidth, connection per second)
- measure latency/jitter/drop in high rate
- Emulate L7 application, e.g. HTTP/HTTPS/Citrix- there is no need to implement the exact application.
- TCP implementation based on BSD
- automation Python API
- TCP/UDP/Application statistics (per client side/per template)

D. *Automation example*

Listing **??** shows an example of python script to automate load a stateful profile and read port and TCP statistics

Listing 9: Stateless automation example

E. *Memory saving*

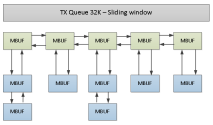

Fig. 10: One Flow Tx Ring

Most TCP stacks have an API that allow the user to provide a buffer (write operation). The TCP module saves the buffer until the data is acknowledged by the remote side. With big windows (required with high RTT) and many flows this could create a memory scale issue. Figure **??** shows one TCP flow TX queue. For 1M active flows with 64K TX buffer the worst case memory requirement is 1M x 64K * mbuf-factor (let's assume 2) = 128GB. The mbuf resource is expensive and needs to be allocated ahead of time. The chosen solution for this problem is to change the API to be a poll API, meaning TCP Tx queue will just save a reference to the constant traffic and offset into this ring the mbufs are allocated only when packets need to be sent (lazy) but virtually have a tx queue only for management of queue (two pointers). Now because most of the traffic is almost constant in traffic generation case (per template) and known ahead of time it was possible to implement and save most of the memory. The same idea happen in the Rx side with reassembly [1]

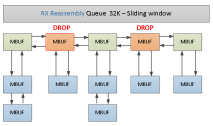

Fig. 11: One Flow Rx Ring

[1]This will not work for TLS streams

F. Benchmark TRex vs Linux kernel

To evaluate the performance and memory scale of TRex and compare it against standard linux tools the following was done. Linux **curl** as a client and **nginx** as a server were compared to TRex for stressing a device under test.

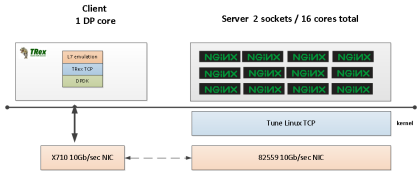

Fig. 12: TRex vs NGINX

The benchmark setup was designed to take a good event-driven Linux server application and to test a TRex client against it. TRex is the client requesting the pages. Figure **??** shows the topology in this case. TRex generates requests using **one** DP core/thread and exercises the **whole** 16 cores of the NGINX/Linux server. The server is installed with the NGINX process on all 16 cores. After some trial and error, it was determined that is is more difficult to separate Linux kernel/IRQ contexts events from user space process CPU, so it was chosen to give the NGINX all server resources, and monitor to determine the limiting factor. The objective is to simulate HTTP sessions as it was defined in our benchmark (e.g. new session for each REQ/RES, initwnd=2 etc.) and not to make the fastest most efficient TCP configuration. This might be the main **difference** between NGINX benchmark configuration and this document configuration. In both cases (client/server), the same type of x86 server was used (2 sockets,CPU E5-2667, Intel X710)

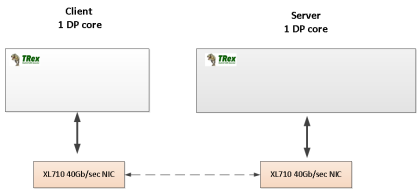

Fig. 13: TRex vs TRex

To compare apples to apples, the nginx server was replaced with a TRex server with **one** DP core, using an XL710 NIC (40Gb). see Figure **??**

G. Benchmark traffic profile

Typically, web servers are tested with a constant number of active flows that are opened ahead of time. In the nginx benchmark blog, only 50 TCP constant connections are used with many requests/responses for each TCP connection see here [**?**]. In our traffic generation use case, each HTTP request/response (for each new TCP connection) requires opening a **new** TCP connection. A simple HTTP flow with a request of 100B and a response of 32KB (about 32 packets/flow with initwnd=2) was used.

H. Benchmark Limitations

The comparison is not perfect, as TRex merely emulates HTTP traffic. It is not a real-world web server or HTTP client. For example, currently the TRex client does not parse the HTTP response for the Length field. TRex simply waits for a specific data size (32KB) over TCP. However the TCP layer is a full featured TCP (e.g. delay-ack/Retransmission/Reassembly/timers) . The bench objective is to compare traffic generation capabilities for stressing network gears.

I. Benchmark results

Comparing 1 DP core running TRex to NGINX running on 16 cores with a kernel that can interrupt any NGINX process with IRQ. Figure 3 shows the performance of one DP TRex. It can scale to about 25Gb/sec of download of HTTP (total of 3MPPS for one core).

	TRex one DP core							
m	CPU (1DP) %%	cps	rps	rx (mb/sec)	pps(tx+rx)	active flows	drop	Memory TCP/MB
1000	0.6	1000	1000	265	34444	600	0	0.3
5000	2.7	5000	5000	1320	172222	3000	0	1.4
10000	5.7	10000	10000	2210	344444	6000	0	2.9
20000	13.5	20000	20000	5410	688889	12000	0	5.7
50000	40.8	50000	50000	13480	1722222	29870	0	14.2
87500	79.0	87500	87500	23670	3013889	55329	0	26.4
90000	86.1	90000	90000	24271	3100000	56217	0.10%	26.8

Fig. 14: TRex 1 DP core

					Linux 16 cores			
m	cps	rps	rx (mb/sec)	active flows	16xCPU	drop	Kernel memory SLAB (MB)	Total Memory used (free-h) MB
1000	1000	1000	265	600		no		21000
5000	5000	5000	1320	3000		no		22000
10000	10000	10000	2210	6000		no		25000
20000	20000	20000	5410	12000	20%/25% 8x cores IRQ break	yes	800	31000
50000	50000	50000	13480	29870				
87500	87500	87500	23670	55329				
90000	90000	90000	24271	56217				

Fig. 15: NGINX 16 cores

nginx cannot handle more than 20K new flows/sec, due to kernel TCP software interrupt and thread processing. The limitation is the kernel and not NGINX user space process. With more NICs and optimized distribution, the number of flows could be increased X2, but not more than that. The total number of packets was approximately 600KPPS (Tx+Rx). The number of active flows were 12K.

TRex with one core could scale to about 25Gb/sec, 3MPPS of the same HTTP profile. The main issue with nginx and Linux setup is the tunning. It is very hard to let the hardware utilizing the full server resource (half of the server was idle in this case and still experience a lot of drop). TRex is not perfect too, we couldn't reach 100% CPU utilization without a drop (CPU was 84%). To achieve 100gbps with this profile on the server side requires 4 cores for TRex, vs. 20x16 cores for NGINX servers. TRex is faster by a factor of **80**. In this implementation, each flow requires approximately 1K bytes of memory (Regardless of Tx/Rx rings because of TRex

architecture). In the kernel, with a real-world server, TRex optimization can't be applied and each TCP connection must save memory in Tx/Rx rings. For about 5Gb/sec traffic with this profile, approximately 10GB of memory was required (both NGINX and Kernel). For 100Gb/sec traffic, approximately 200GB is required (If we will do the extrapolation) With a TRex optimized implementation, approximately 100MB is required. TRex thus provides an improvement by a factor or **2000** in the use of memory resources.

V. Emulation layers

To simulate clients there is a need for many a slow path

References

[1] Cisco systems: "TRex realistic traffic generator",https://trex-tgn.cisco.com/trex/doc/

[2] Cisco systems, "TRex Stateless",,https://trex-tgn.cisco.com/trex/doc/trex_stateless.html

[3] Cisco systems, "TRex Stateful ASTF",,https://trex-tgn.cisco.com/trex/doc/trex_astf.html

[4] BIRD, Faculty of Math and Physics, Charles University Prague https://bird.network.cz/

[5] Cisco systems, "TRex STL benchmark",,https://trex-tgn.cisco.com/trex/doc/trex_stateless_bench.html

[6] NGINX performance https://www.nginx.com/blog/testing-the-performance-of-nginx-and-nginx-plus-web-servers/