Packet Capture Benchmark on 1 GE

Tomáš Mrázek, Jan Vykořal

Received 12.12.2008

Abstract

This report describes packet capture benchmark test and its results. We compared packet capture throughput of the standard Linux TCP/IP stack and PF_RING with common Intel NIC and szedata2 with COMBO cards developed by the Librouter project. The throughput is measured according RFC 2544 and by simpler but widespread “iperf-like” technique on 1 Gigabit Ethernet.

Keywords: packet capture, benchmark, PF_RING, szedata2, COMBO6X, Intel, Linux

1 Introduction

There are many applications requiring reliable packet capture at gigabit speeds, for example network monitoring tools or intrusion detection and prevention systems. The most common way of packet capture is via the standard Linux TCP/IP stack. The next method, PF_RING [1], bypasses the TCP/IP stack and thus speeds up packet capture. The last mentioned method is szedata2 [2]. This hardware-independent method is believed to provide the highest performance in connection with hardware-accelerated COMBO\(^1\) cards. A common multiport NIC (Network Interface Card) is one of applications of COMBO cards.

Figure 1 depicts these three approaches in the host system architecture.

Figure 1. All three tested approaches and host system architecture.

\(^1\) http://www.liberouter.org/hardware.php

© CESNET, 2009
If we use the first method, the incoming packets are forwarded to the operating system. The kernel handles the traffic and the frames go through the TCP/IP stack to the running application via libpcap\(^2\) library. This library is widely used, e.g., in\ tcpdump, Wireshark, Snort, fprobe and many others applications are based on this library. Unfortunately, the performance is slowed down by large amount of CPU interrupts during handling the traffic.

\textit{PF\_RING} copies all incoming packets into the circle buffer and discards the original ones. If the buffer is full, the packets are discarded immediately without being copied. The copied packets are not forwarded to the higher layers of the TCP/IP stack but they stay in the buffer. If the application need to read the packets, it has to call function \textit{mmap()} to get a pointer to the circle buffer. Then it can read them. There are two independent pointers for reading and writing data. Newly incoming packets cause rewriting the old values.

The last method is called \textit{szedata2}. It derives benefit from special card of the \textit{COMBO} family. Similarly to \textit{PF\_RING}, the main advantage of this approach is bypassing the TCP/IP stack. \textit{Szedata2} should provide high throughput thanks to aggregation of short packets into bigger blocks. This approach lowers the overhead for DMA transmissions. The principle of transmitting packets is based on swapping of these packets between two circle buffers. The first one is in RAM and the second one in the network interface card. DMA controllers handles the proper usage of pointers for flawless communication of buffers. The size of buffer can vary (hundreds of MB by default). The buffer is divided into 4kB blocks that corresponds to the size of a page in the most of recent operating systems.

\section{Test Description}

We tested packet capture performed by the mentioned techniques to benchmark their performance, namely throughput. To ensure test objectivity we put a lot of effort on a configuration of a test application and a \textit{device under test} (DUT in short). DUT stands for the network cards in a host computer and the operating system. Following sections of the report focus on every part of preparations and testing itself in detail.

There are two main methods how to measure throughput. The first one is defined in RFC 2544 \cite{RFC2544} as “the fastest rate at which the count of test frames transmitted by the DUT is equal to the number of test frames sent to it by the test equipment.”

The test application sends traffic from the hardware test tool Spirent TestCentre 2000\(^3\) (TC2000 in short) at wire speed on 1 Gigabit Ethernet. The RFC 2544 defines test frame lengths contained in test traffic as follows: 64, 128, 256, 512, 1024, 1280 and 1518 B. In addition, we also randomly chose these frame lengths: 73, 207, 313, 726, 1174 and 1319 B.

According the RFC 2544, the time of sending can vary but the default value is 60 seconds. The test application uses \textit{binary searching algorithm} that means halving intervals corresponding to the result of the last iteration. The test starts at wire

\footnotesize

\(^2\) http://www.tcpdump.org/

\(^3\) http://www.spirent.com/portal/index.cfm?ws=356&media=8
speed. If there is a packet loss, it lowers the speed by the defined step. If not, it raises the speed. The packet loss is the difference between sent packets from TC2000 and received packets by the test application. The step depends on the difference between the two last speeds. The end of the test is set by the precision of this step. For example, the value of 1 % means that if the step is lower than 1 % of bandwidth the application ends and prints the results.

The second method of measuring the throughput is simpler than the previous one. We call it "iperf-like" method. The same traffic is continuously sent at wire speed and the test application captures this traffic and it prints instant throughput value every second. After defined time traffic and the application is stopped. The application counts the throughput by dividing the received packets by total time of sending. We chose this method because it corresponds to the networking best effort policy. It is similar to the frame loss rate defined in RFC 1242 [4] as “percentage of frames that should have been forwarded by a network device under steady state (constant) load that were not forwarded due to lack of resources.”

These two methods are different so the results can also be different. The measuring method defined in RFC is more strict to packet loss.

Another problem that could occur during testing is a saturation. It can happen after sending the traffic at the speed when the DUT is unable to receive all packets. In that case, following behaviour of DUT is unpredictable and the test results are distorted.

3 Test Bed

We chose TC2000 as a lossless generator of synthetic traffic at wire speed (1 Gbps). It is capable of sending various types of traffic: frames with fixed, random or incremental length and IMIX which can correspond to common Internet traffic.

Technical parameters of the host computer are described in Appendix A.
We tested Intel 82546GB⁴ and COMBO6X/4SFPRO⁵ network interface cards. Both cards were plugged into PCI-X slot in the host machine and connected to TC2000 by one port. The first two packet capture methods, the standard Linux TCP/IP stack and PF_RING, were tested with Intel NIC and the third method, szedata2, with COMBO6X card.

We believe the following settings should help to increase the packet capture throughput:

— Enabled kernel compile parameters RX polling (NAPI) and mmapped IO.
— Disabled CRC check in case of the packet capture via standard TCP/IP stack.
— Tuning e1000⁶ driver parameters such as InterruptThrottleRate and RxIntDelay.
— Disable useless services in operating system.
— Balance load among several CPU cores by taskset.

4 Test Application

We developed a test application counter for measuring the throughput. There are three “versions”. The only difference between them was a compilation process. We used standard libpcap v0.9.7 for TCP/IP stack testing, patched libpcap v0.9.7 with PF_RING and patched libpcap v0.9.8 with szedata2. The differences between version 0.9.7 and version 0.9.8 are irrelevant to our testing and thus it should not influence the performance.

The test application works as follows. From the beginning it waits in poll. After the first burst of received packets, the test application starts to print instant throughput to the output. In case of measuring the throughput according to RFC 2544, the continuous output is suppressed. For capturing a packet we can use two functions: pcap_loop() or pcap_next(). We decided for pcap_next() because it should provide higher performance or at least the same as pcap_loop().

Next, received packets are discarded at libpcap level to provide more CPU time for handling interrupts and thus to increase the throughput. We can stop the test application at any time by sending SIGINT signal.

For more details see the source code in Appendix B.

5 Settings and Test Results

First, we measured iperf-like throughput. We also monitored the CPU load caused by the test application by Linux function time and kernel from top output (the value was only approximate and differed slightly during the tests). Table 1 depicts the results for szedata2.

---

⁵ http://www.liberouter.org/hardware.php
Table 1. Iperf-like throughput of szedata2 with COMBO6X NIC.

<table>
<thead>
<tr>
<th>Frame Length [B]</th>
<th>Throughput [pps / Gbps]</th>
<th>App./kernel load [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1518</td>
<td>81275 / 1.0</td>
<td>11.5 / 0</td>
</tr>
<tr>
<td>1319</td>
<td>93353 / 1.0</td>
<td>12.3 / 0</td>
</tr>
<tr>
<td>1280</td>
<td>96154 / 1.0</td>
<td>12.4 / 0</td>
</tr>
<tr>
<td>1174</td>
<td>104690 / 1.0</td>
<td>12.4 / 0</td>
</tr>
<tr>
<td>1024</td>
<td>119732 / 1.0</td>
<td>12.8 / 0</td>
</tr>
<tr>
<td>726</td>
<td>167560 / 1.0</td>
<td>14.2 / 0</td>
</tr>
<tr>
<td>512</td>
<td>234963 / 1.0</td>
<td>17.4 / 0</td>
</tr>
<tr>
<td>313</td>
<td>375375 / 1.0</td>
<td>22.1 / 0</td>
</tr>
<tr>
<td>256</td>
<td>452899 / 1.0</td>
<td>25.8 / 0</td>
</tr>
<tr>
<td>207</td>
<td>550066 / 1.0</td>
<td>29.4 / 0</td>
</tr>
<tr>
<td>128</td>
<td>844595 / 1.0</td>
<td>41.7 / 0</td>
</tr>
<tr>
<td>73</td>
<td>1344086 / 1.0</td>
<td>57.1 / 0</td>
</tr>
<tr>
<td>64</td>
<td>1488095 / 1.0</td>
<td>63.6 / 0</td>
</tr>
</tbody>
</table>

It is obvious that szedata2 has the throughput equal to 1 Gbps (theoretical maximum throughput) at all measured lengths. The CPU load caused by the test application was higher for the shortest packets than for the longest ones. It is not caused by szedata2 itself but by the test application which reads continuously timestamps from each incoming packet to be able to print instant throughput. Without reading this information the application load is about 6% for all frame lengths. We also measured the load caused by the kernel and for szedata2 it was 0%.

Packet capture via next two methods, `PF_RING` and standard stack, achieved different throughput depending on the CPU on which the test application was running. `Throughput #1` denotes result of a measurement when the test application was running on the same CPU as the CPU which handles the incoming traffic. But in case of `Throughput #2` we assigned a different CPU to the test application. The load #1 corresponds with Throughput #1 and load #2 with Throughput #2 of course.

We were also able to change the parameters of the network driver `e1000`. We tried many combinations of `InterruptThrottleRate` and `RxIntDelay` parameters but none of them had a significant influence (more than several percent) on throughput. The reported results were measured with default settings of `e1000` driver (`Interrupt-ThrottleRate = 8000, RxIntDelay = 0`).

Tables 2 and 3 show the results for `PF_RING` and the standard Linux TCP/IP stack, respectively.

`PF_RING` provide slightly better results and depends on chosen CPUs. Both methods achieved throughput of 1 Gbps at the frame length equal to and greater than 512 B. At length 313 B `PF_RING` achieved only 0.9 Gbps in case we used different CPUs. However, it captured all the packets if only one CPU was utilized. The same behaviour for `PF_RING` was observed for the lengths of 256 and 207 B.

The standard stack was unable to capture all the packets with using the same CPU nor with different ones for handling incoming traffic. `PF_RING` was also unable to capture all the packets for lengths shorter than 207 B where the better performance was measured by using different CPUs.
Table 2. Iperf-like throughput of \textit{PF\_RING} with Intel NIC.

<table>
<thead>
<tr>
<th>Frame Length [B]</th>
<th>Throughput</th>
<th>App./kernel load [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[pps / Gbps]</td>
<td>#1</td>
</tr>
<tr>
<td>1518</td>
<td>81274 / 1.0</td>
<td>81279 / 1.0</td>
</tr>
<tr>
<td>1319</td>
<td>93359 / 1.0</td>
<td>93359 / 1.0</td>
</tr>
<tr>
<td>1280</td>
<td>96160 / 1.0</td>
<td>96160 / 1.0</td>
</tr>
<tr>
<td>1174</td>
<td>104697 / 1.0</td>
<td>104697 / 1.0</td>
</tr>
<tr>
<td>1024</td>
<td>119739 / 1.0</td>
<td>119739 / 1.0</td>
</tr>
<tr>
<td>726</td>
<td>167571 / 1.0</td>
<td>167571 / 1.0</td>
</tr>
<tr>
<td>512</td>
<td>234978 / 1.0</td>
<td>234978 / 1.0</td>
</tr>
<tr>
<td>313</td>
<td>375375 / 1.0</td>
<td>338143 / 0.9</td>
</tr>
<tr>
<td>256</td>
<td>452928 / 1.0</td>
<td>343014 / 0.757</td>
</tr>
<tr>
<td>207</td>
<td>550691 / 1.0</td>
<td>360587 / 0.655</td>
</tr>
<tr>
<td>128</td>
<td>288621 / 0.342</td>
<td>379419 / 0.449</td>
</tr>
<tr>
<td>73</td>
<td>142171 / 0.106</td>
<td>401708 / 0.3</td>
</tr>
<tr>
<td>64</td>
<td>150915 / 0.101</td>
<td>404517 / 0.272</td>
</tr>
</tbody>
</table>

Table 3. Iperf-like throughput of the Linux TCP/IP stack with Intel NIC.

<table>
<thead>
<tr>
<th>Frame Length [B]</th>
<th>Throughput</th>
<th>App./kernel load [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[pps / Gbps]</td>
<td>#1</td>
</tr>
<tr>
<td>1518</td>
<td>81276 / 1.0</td>
<td>81276 / 1.0</td>
</tr>
<tr>
<td>1319</td>
<td>93356 / 1.0</td>
<td>93357 / 1.0</td>
</tr>
<tr>
<td>1280</td>
<td>96158 / 1.0</td>
<td>96158 / 1.0</td>
</tr>
<tr>
<td>1174</td>
<td>104695 / 1.0</td>
<td>104695 / 1.0</td>
</tr>
<tr>
<td>1024</td>
<td>119736 / 1.0</td>
<td>119737 / 1.0</td>
</tr>
<tr>
<td>726</td>
<td>167571 / 1.0</td>
<td>167571 / 1.0</td>
</tr>
<tr>
<td>512</td>
<td>234964 / 1.0</td>
<td>234971 / 1.0</td>
</tr>
<tr>
<td>313</td>
<td>375389 / 1.0</td>
<td>375370 / 1.0</td>
</tr>
<tr>
<td>256</td>
<td>401845 / 0.887</td>
<td>370045 / 0.817</td>
</tr>
<tr>
<td>207</td>
<td>386925 / 0.703</td>
<td>386306 / 0.702</td>
</tr>
<tr>
<td>128</td>
<td>299121 / 0.354</td>
<td>363735 / 0.431</td>
</tr>
<tr>
<td>73</td>
<td>136142 / 0.101</td>
<td>383673 / 0.285</td>
</tr>
<tr>
<td>64</td>
<td>86989 / 0.058</td>
<td>354233 / 0.238</td>
</tr>
</tbody>
</table>

Generally, the CPU load corresponds with measured throughput. Note that there is an interesting anomaly in case of load generated by \textit{PF\_RING} at length 128 B. We inspected neighbouring frame lengths and found significant difference between lengths of 104 B and 103 B where the load generated by the test application dropped from 89 % to 25 %.

For a better comparison we plotted Figures 3, 4, 5 and 6. Figure 7 and 8 represent kernel and test application load as a function of frame lengths.

Next, we performed the throughput tests defined by \textit{RFC 2544}. The time of each iteration was set to 60 seconds. The results are shown in Tables 4, 5 and 6.

Again, only szedata2 proved to be absolutely lossless at all tested packet lengths. \textit{PF\_RING} also proved to be more efficient than standard TCP/IP stack. From
Figure 3. Iperf-like packet capture throughput #1 of szedata2, PF_RING and the standard Linux TCP/IP stack.

Figure 4. Iperf-like packet capture throughput #2 of szedata2, PF_RING and the standard Linux TCP/IP stack.
Figure 5. Iperf-like packet capture throughput #1 of szedata2, PF_RING and the standard Linux TCP/IP stack.

Figure 6. Iperf-like packet capture throughput #2 of szedata2, PF_RING and the standard Linux TCP/IP stack.
Packet Capture Benchmark on 1 GE

Figure 7. Dependency of kernel load on frame lengths for `szedata2`, `PF_RING` and standard Linux TCP/IP stack.

Figure 8. Dependency of load caused by the test application on frame lengths for `szedata2`, `PF_RING` and standard Linux TCP/IP stack.
Table 4. RFC 2544 throughput of szedata2 with COMBO6X NIC.

<table>
<thead>
<tr>
<th>Frame Length [B]</th>
<th>Throughput [pps / Gbps] #1</th>
<th>Throughput [pps / Gbps] #2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1518</td>
<td>81274 / 1.0</td>
<td>81274 / 1.0</td>
</tr>
<tr>
<td>1319</td>
<td>93353 / 1.0</td>
<td>93353 / 1.0</td>
</tr>
<tr>
<td>1280</td>
<td>96154 / 1.0</td>
<td>96154 / 1.0</td>
</tr>
<tr>
<td>1174</td>
<td>104690 / 1.0</td>
<td>104690 / 1.0</td>
</tr>
<tr>
<td>1024</td>
<td>119732 / 1.0</td>
<td>119732 / 1.0</td>
</tr>
<tr>
<td>726</td>
<td>167560 / 1.0</td>
<td>167560 / 1.0</td>
</tr>
<tr>
<td>512</td>
<td>234963 / 1.0</td>
<td>234962 / 1.0</td>
</tr>
<tr>
<td>313</td>
<td>375375 / 1.0</td>
<td>331753 / 0.884</td>
</tr>
<tr>
<td>256</td>
<td>452899 / 1.0</td>
<td>331713 / 0.732</td>
</tr>
<tr>
<td>207</td>
<td>550066 / 1.0</td>
<td>349541 / 0.635</td>
</tr>
<tr>
<td>128</td>
<td>844595 / 1.0</td>
<td>362912 / 0.430</td>
</tr>
<tr>
<td>73</td>
<td>1344086 / 1.0</td>
<td>387212 / 0.288</td>
</tr>
<tr>
<td>64</td>
<td>1488095 / 1.0</td>
<td>377837 / 0.254</td>
</tr>
</tbody>
</table>

Table 5. RFC 2544 throughput of PF_RING with Intel NIC.

<table>
<thead>
<tr>
<th>Frame Length [B]</th>
<th>Throughput [pps / Gbps] #1</th>
<th>Throughput [pps / Gbps] #2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1518</td>
<td>81274 / 1.0</td>
<td>81274 / 1.0</td>
</tr>
<tr>
<td>1319</td>
<td>93353 / 1.0</td>
<td>93353 / 1.0</td>
</tr>
<tr>
<td>1280</td>
<td>96154 / 1.0</td>
<td>96154 / 1.0</td>
</tr>
<tr>
<td>1174</td>
<td>104690 / 1.0</td>
<td>104690 / 1.0</td>
</tr>
<tr>
<td>1024</td>
<td>119732 / 1.0</td>
<td>119732 / 1.0</td>
</tr>
<tr>
<td>726</td>
<td>167560 / 1.0</td>
<td>167560 / 1.0</td>
</tr>
<tr>
<td>512</td>
<td>234963 / 1.0</td>
<td>234962 / 1.0</td>
</tr>
<tr>
<td>313</td>
<td>355947 / 0.948</td>
<td>331753 / 0.884</td>
</tr>
<tr>
<td>256</td>
<td>387441 / 0.855</td>
<td>331713 / 0.732</td>
</tr>
<tr>
<td>207</td>
<td>426439 / 0.774</td>
<td>349541 / 0.635</td>
</tr>
<tr>
<td>128</td>
<td>593031 / 0.702</td>
<td>362912 / 0.430</td>
</tr>
<tr>
<td>73</td>
<td>624790 / 0.465</td>
<td>387212 / 0.288</td>
</tr>
<tr>
<td>64</td>
<td>632150 / 0.425</td>
<td>377837 / 0.254</td>
</tr>
</tbody>
</table>

lengths equal to or greater than 512 B it reached maximum throughput of 1 Gbps. Consequently, the RFC 2544 throughput goes down till approximately 400 Mbps in case of using the same CPU for test application and kernel operations or 250 Mbps in case of using different CPUs. RFC 2544 throughput of standard TCP/IP stack is very low, particularly if we used the same CPU for test application and kernel operations. Figures 9, 10, 11 and 12 depict these results.

The reason of poor throughput of TCP/IP stack is a losing of few packets (for example about 350 packets from several millions) in almost every iteration of binary search. Standard stack was almost unable to capture all the packets except the longest ones.

On the contrary, PF_RING profited from this behaviour. It was unable to capture all the packets shorter than 512 B at speed of 1 Gbps. But when we slowed
Figure 9. RFC 2544 packet capture throughput #1 of szedata2, PF_RING and the standard Linux TCP/IP stack.

Table 6. RFC 2544 throughput of the Linux TCP/IP stack with Intel NIC.

<table>
<thead>
<tr>
<th>Frame Length [B]</th>
<th>Throughput #1 [pps / Gbps]</th>
<th>Throughput #2 [pps / Gbps]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1518</td>
<td>45717 / 0.562</td>
<td>81274 / 1.0</td>
</tr>
<tr>
<td>1319</td>
<td>40842 / 0.438</td>
<td>93353 / 1.0</td>
</tr>
<tr>
<td>1280</td>
<td>36245 / 0.377</td>
<td>96154 / 1.0</td>
</tr>
<tr>
<td>1174</td>
<td>31898 / 0.305</td>
<td>104690 / 1.0</td>
</tr>
<tr>
<td>1024</td>
<td>24087 / 0.201</td>
<td>102778 / 0.858</td>
</tr>
<tr>
<td>726</td>
<td>20618 / 0.123</td>
<td>78380 / 0.468</td>
</tr>
<tr>
<td>512</td>
<td>24093 / 0.103</td>
<td>80768 / 0.344</td>
</tr>
<tr>
<td>313</td>
<td>23461 / 0.062</td>
<td>57919 / 0.154</td>
</tr>
<tr>
<td>256</td>
<td>59266 / 0.131</td>
<td>119417 / 0.264</td>
</tr>
<tr>
<td>207</td>
<td>68833 / 0.125</td>
<td>119382 / 0.217</td>
</tr>
<tr>
<td>128</td>
<td>65159 / 0.077</td>
<td>150113 / 0.178</td>
</tr>
<tr>
<td>73</td>
<td>273017 / 0.203</td>
<td>165386 / 0.123</td>
</tr>
<tr>
<td>64</td>
<td>354585 / 0.238</td>
<td>232515 / 0.156</td>
</tr>
</tbody>
</table>

down the speed of incoming traffic it did not lose any packet. So the speed could be increased and the final throughput is higher than in case of the iperf-like testing.

Finally, we measured packet capture throughput of Internet mix (IMIX) traffic. We chose the following distribution of four packet lengths: 57% of 64B packets, 7 % of 570B packets, 16 % of 594B packets and 20 % of 1518B packets. The measured throughput was 1 Gbps for all three methods. Only the load caused by the test
Figure 10. RFC 2544 packet capture throughput #2 of szedata2, **PF_RING** and the standard Linux TCP/IP stack.

Figure 11. RFC 2544 packet capture throughput #1 of szedata2, **PF_RING** and the standard Linux TCP/IP stack.
Figure 12. RFC 2544 packet capture throughput #2 of szedata2, PF_RING and the standard Linux TCP/IP stack.

Table 7. Load caused by all tested methods during throughput test of IMIX packet distribution.

<table>
<thead>
<tr>
<th>Method</th>
<th>App./kernel load #1 [%]</th>
<th>App./kernel load #2 [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard Stack</td>
<td>54 / 10</td>
<td>75 / 4</td>
</tr>
<tr>
<td>PF_RING</td>
<td>22 / 5</td>
<td>53 / 1</td>
</tr>
<tr>
<td>szedata2</td>
<td>19 / 0</td>
<td>20 / 0</td>
</tr>
</tbody>
</table>

application and the kernel differed. Table 7 shows actual values.

6 Conclusion and Future Work

We conclude szedata2 is the fastest method in the benchmark. Its throughput does not depend on packet length and is always 1 Gbps (theoretical maximum throughput). On the contrary, the other two methods suffers by significant loss at the shortest packet lengths. PF_RING achieve better performance for packet capture than the standard Linux TCP/IP stack, as we suppose. The described saturation appeared so we excluded anomalous results caused by this effect.

Also the CPU load was monitored during the testing. As opposed to PF_RING and standard stack, szedata2 did not cause any kernel load at all tested frame lengths. We found out the performance depends whether all relevant processes are running on one CPU or are distributed among two processors.

Next, it was expected that the specific settings of parameters of network device
driver e1000 should increase the performance. But we did not notice any significant influence on the throughput during the testing.

Finally, we can confirm that the results may vary due to chosen testing procedure. RFC 2544 testing methodology is rigorous whereas iperf-like testing corresponds to the best effort policy. For example, the former should be used when packet loss is mission critical, such as in security applications.

Our future work will be aimed at packet capture benchmark on 10 Gigabit Ethernet. We would like to benchmark a new generation of COMBO cards, namely COMBOv2, with szedata2 and Myricom network interface card with PF_RING and the standard Linux TCP/IP stack.

Appendix A. Used SW/HW equipment

Testing device: Spirent TestCenter 2000 Testing software: Tcl library version 2.20

Host computer:
- Number of CPU: 2 (quad core)
- CPU: Xeon(R) CPU E5335 @ 2.00GHz 16 KB L1 + 8 MB L2 cache
- RAM: 4 GB
- MB: Supermicro X7DB8
- OS: Redhat Enterprise Linux ES release 4 (Nahant update 1)
- Kernel: Linux 2.6.26.3_32bit
- PCI bridge: c610.05.0c (2008/06/09 19:36) NetCOPE Bridge
- MAC: 00:30:48:35:d1:90
- COMBO cards: COMBO6X / COMBO-4SFPRO/1G
- COMBO6X S/N: CAM7000047
- COMBO-4SFPRO S/N: CAM70000468
- Intel card: 82546GB Gigabit Ethernet Controller
- Driver: e1000
- Driver version: 7.3.20-k2-NAPI
- PCI: PCI-X
- libpcap: version 0.9.7 and 0.9.8
- PF_RING: version 3.8.6
- ethtool: version 1.8

Appendix B. Source Code of the Test Application

```c
#include <pcap.h>
#include <stdio.h>
#include <signal.h>
#include <stdlib.h>
#include <unistd.h>
#include <sys/poll.h>
#include <time.h>
#include <stdint.h>
#include <sys/time.h>
#include <sys/socket.h>
```
```c
#define INTERVAL 1000    /* 1 second by default */
int ctrl_c = 0;
/* to catch an interrupt signal (ctrl+c) */
static void sig_usr(int signo)
{
    if (signo == SIGINT) {
        ctrl_c = 1;
    }
}

int main(int argc, char *argv[])
{
    uint64_t p_count = 0;    /* Total packet count */
    uint64_t p_last = 0;    /* Packets in last interval */
    uint64_t next;        /* Next time check */
    uint64_t now;         /* Current time */
    uint64_t start;       /* Start packet arriving time */
    uint64_t last;        /* Last packet incoming time */
    struct timeval t_now;
    pcap_t *handle;        /* Pcap handle */
    int pcap_socket;
    /* Pcap error string buffer */
    char errbuf[PCAP_ERRBUF_SIZE];
    //catch SIGINT
    signal(SIGINT, sig_usr);
    if (argc < 2) {
        fprintf(stderr, "Need argument where to listen
                     (%s eth0)\n", argv[0]);
        exit(-1);
    }
    char *dev = argv[1];    /* The device to sniff on */
    printf("Pcap library version:%s \n",
            pcap_lib_version());
    /* Open the session in promiscuous mode */
    handle = pcap_open_live(dev, BUFSIZ, 1, 1000, errbuf);
    if (handle == NULL) {
        fprintf(stderr, "Couldn't open device %s: %s\n",
                dev, errbuf);
        return (-2);
    }
    pcap_socket = pcap_fileno(handle);
    /* set first next interval */
    gettimeofday(&t_now, NULL);
    /* accuracy in ms */
    start = next = (t_now.tv_sec * INTERVAL) +
                (t_now.tv_usec / INTERVAL);
    next += INTERVAL;
```
/* using polling */
while (!ctrl_c) {
    /* The header that pcap gives us */
    struct pcap_pkthdr header;
    /* do not use poll() function with szedata2 */
    #ifdef POLL
    struct pollfd pfd;
    int ret;
    pfd.fd = pcap_socket;
    pfd.events = POLLIN;
    ret = poll(&pfd, 1, INTERVAL);
    /* packet may not come */
    if (ret > 0 && pfd.revents & POLLIN
        && pcap_next(handle, &header)) {
    } else {
        if (pcap_next(handle, &header)) {
        }
    } #endif
    last = now = header.ts.tv_sec * INTERVAL
        + header.ts.tv_usec / INTERVAL;
    if(!p_count) {
        start = now;
    }
    p_count++;
} else {
    gettimeofday(&t_now, NULL);
    /* accuracy in ms */
    now = (t_now.tv_sec * INTERVAL) +
        (t_now.tv_usec / INTERVAL);
}
if (now >= next) {
    uint64_t delta = now - next + INTERVAL;
    #ifdef OUTPUT
    /* print instant throughput, not in rfc2544 */
    if (p_count) {
        printf("iteration: packets %llu, pps %llu\n", p_count - p_last,
            ((p_count - p_last) * INTERVAL) / delta);
    } #endif
    p_last = p_count;
    next = now + INTERVAL;
}
} if (p_count) {
    /* different output */
    #ifdef OUTPUT
/*print real packet rate from first to end packet*/
printf ("total: packets %llu, time %llu.%llu, 
pps %llu\n", p_count, (last-start) / INTERVAL, 
(last-start) % INTERVAL, 
(p_count * INTERVAL) / (last - start));
#else
printf("%llu\n", p_count);
#endif
/* close handle */
pcap_close(handle);
return (0);
}

References


\(^7\) https://svn.ntop.org/trac/browser/trunk/PF_RING/doc/
\(^8\) http://is.muni.cz/th/98734/fi_m/
\(^9\) http://tools.ietf.org/html/rfc2544
\(^10\) http://tools.ietf.org/html/rfc1242