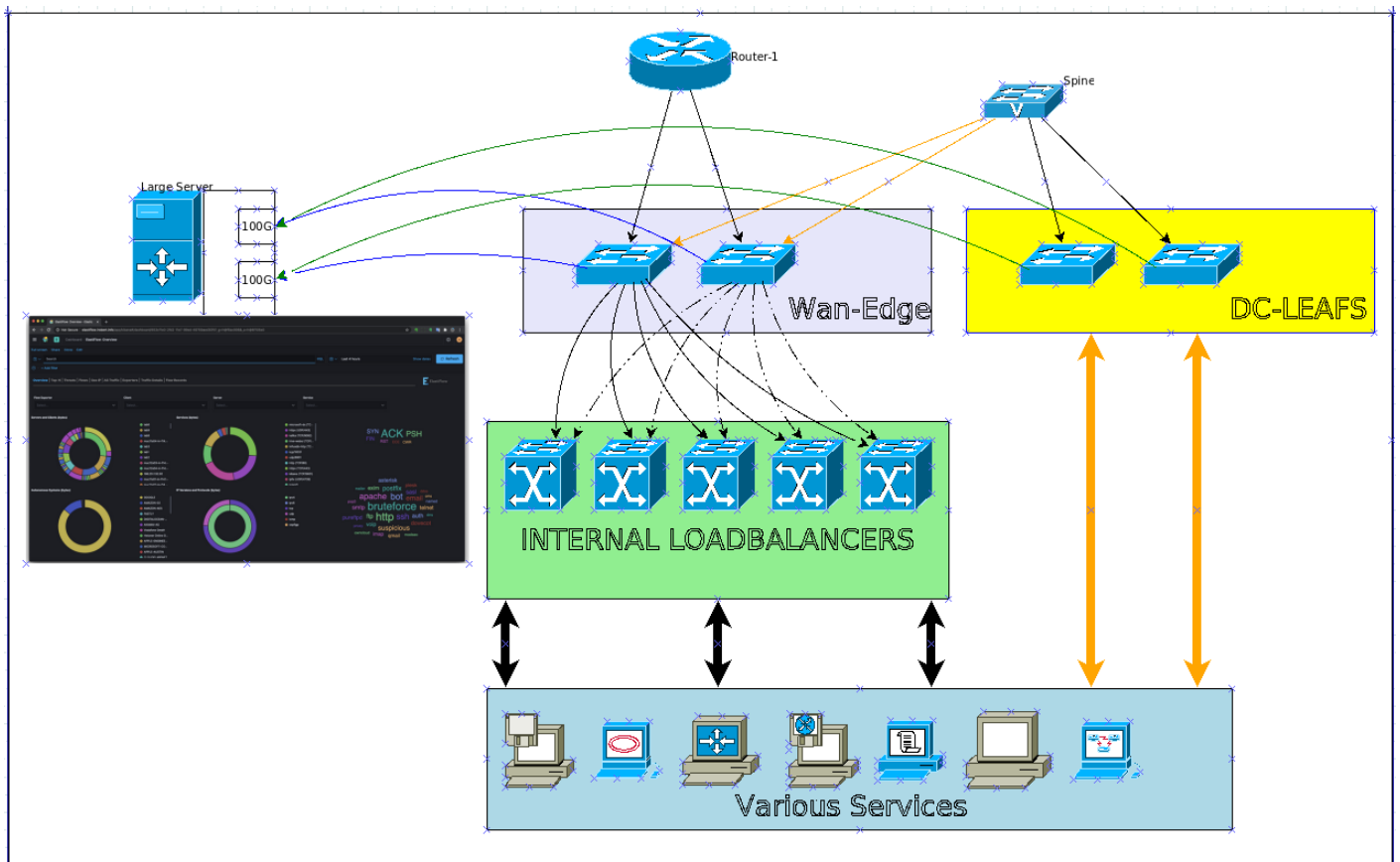


Generating 1:1 ipfix from 10g pipeline - Getting the data - Part 1



Goal : generate lossless ipfix flow's from distributed pipe to monitor application or network performance, identify bottlenecks and generate alerts if possible.

Why this way ? : It was expensive to do it with proprietary solutions. Plus we needed to have a flexible, open source option to work on. The closest solution cost was \$1M

Challenges ;

- We are going to do it using cpu, so lot's of flows require lots of processing power,
- short flows create large ipfix messages than themselves, any dns flow is is mostly 100 bytes long but a ipfix message for that flow costs us 1000 bytes
- many packets traversing the kernel would create losses so we needed to bypass that
- we need to create custom dashboards, or create alerts.
- deduplication

Plan :

- get mirrors from devices
- produce ipfix from mirrors
- send ipfix data to elastiflow
- design custom dashboards
- design monitor and alerters

What happened along the way;

- Obviously there will be lot's of storage requirement and cpu requirements how ever the current write rate wasn't going to be deadly so we decided to go with using only one server with many ssd drives. We lost the array drives because of a failure in array controller
- We were going to aggregate mirrors from fabric using a Mellanox switch by utilizing it's bridge functions, the asic couldn't manage packet replication this step failed, we bought additional cards and skipped aggregation layer.
- The intel x810-CAQxx cards failed to go in Zero Copy mode, we needed to wait for 2 months for a new firmware from intel
- The tool that we used created problems with intel cards firmware and kernel module. We had to reflash them twice till we found a working state
- The tool had a bug with combining mirrors so we waited for a bugfix.

Current status

We have a working setup, using this version of intel module

```
filename:      /lib/modules/5.4.0-128-generic/kernel/drivers/net/ethernet/intel/ice/ice.ko
firmware:     intel/ice/ddp/ice.pkg
version:      0.8.1-k
license:      GPL v2
description:   Intel(R) Ethernet Connection E800 Series Linux Driver
```

With this version of firmware

```
root@TTPOP-PNPR-LNTW001:/etc/apt# ethtool -i ens2f1
driver: ice_zc
version: 1.9.11
firmware-version: 4.01 0x80013c9a 1.3256.0
expansion-rom-version:
bus-info: 0000:5f:00.1
supports-statistics: yes
supports-test: yes
supports-eeprom-access: yes
supports-register-dump: yes
supports-priv-flags: yes
```

You can find some performance outputs as shown below

Server's status

```

1 [|||||] 16.0% 25 [|||||] 32.5% 49 [|||||] 10.5% 73 [|||||] 7.8%
2 [|||||] 9.1% 26 [|||||] 58.7% 50 [|||||] 2.6% 74 [|||||] 1.9%
3 [|||||] 19.6% 27 [|||||] 10.3% 51 [|||||] 9.1% 75 [|||||] 10.8%
4 [|||||] 3.2% 28 [|||||] 3.9% 52 [|||||] 2.6% 76 [|||||] 2.6%
5 [|||||] 7.7% 29 [|||||] 7.3% 53 [|||||] 13.1% 77 [|||||] 7.2%
6 [|||||] 3.1% 30 [|||||] 2.6% 54 [|||||] 2.6% 78 [|||||] 1.9%
7 [|||||] 9.1% 31 [|||||] 9.7% 55 [|||||] 11.6% 79 [|||||] 6.5%
8 [|||||] 3.2% 32 [|||||] 58.7% 56 [|||||] 2.6% 80 [|||||] 1.9%
9 [|||||] 4.6% 33 [|||||] 9.2% 57 [|||||] 10.3% 81 [|||||] 12.3%
10 [|||||] 2.6% 34 [|||||] 57.9% 58 [|||||] 4.5% 82 [|||||] 1.9%
11 [|||||] 10.3% 35 [|||||] 8.4% 59 [|||||] 7.1% 83 [|||||] 8.3%
12 [|||||] 3.6% 36 [|||||] 59.1% 60 [|||||] 2.6% 84 [|||||] 2.6%
13 [|||||] 7.9% 37 [|||||] 8.5% 61 [|||||] 9.1% 85 [|||||] 5.4%
14 [|||||] 2.6% 38 [|||||] 4.6% 62 [|||||] 2.6% 86 [|||||] 1.9%
15 [|||||] 35.5% 39 [|||||] 10.3% 63 [|||||] 9.0% 87 [|||||] 8.5%
16 [|||||] 2.6% 40 [|||||] 2.6% 64 [|||||] 1.9% 88 [|||||] 1.9%
17 [|||||] 36.2% 41 [|||||] 7.2% 65 [|||||] 11.0% 89 [|||||] 36.4%
18 [|||||] 4.2% 42 [|||||] 1.2% 66 [|||||] 4.2% 90 [|||||] 2.0%
19 [|||||] 25.7% 43 [|||||] 9.0% 67 [|||||] 9.0% 91 [|||||] 20.5%
20 [|||||] 1.9% 44 [|||||] 2.6% 68 [|||||] 2.6% 92 [|||||] 4.5%
21 [|||||] 6.4% 45 [|||||] 2.6% 69 [|||||] 12.6% 93 [|||||] 100.0%
22 [|||||] 2.6% 46 [|||||] 2.6% 70 [|||||] 2.6% 94 [|||||] 4.5%
23 [|||||] 6.4% 47 [|||||] 8.5% 71 [|||||] 4.5% 95 [|||||] 5.9%
24 [|||||] 4.5% 48 [|||||] 2.6% 72 [|||||] 4.5% 96 [|||||] 4.5%

Mem [|||||] Tasks: 144, 972 thr: 13 running
Swap [|||||] Load average: 10.20 8.78 6.82
Uptime: 6 days, 6:35:28

PID USER PRI NI VIRT RES SHR S CPU% MEM% TIME+ Command
1 root 20 0 167M 1348K 1856K 0 0 0 0:13:20 /sbin/init
235767 nprobe 20 0 1369M 941M 21992K 75.2 0.2 3850:33 /usr/bin/nprobe /run/nprobe-flow.conf
235819 nprobe 20 0 1369M 941M 21992K 0 0 0:01:02 /usr/bin/nprobe /run/nprobe-flow.conf
235818 nprobe 20 0 1369M 941M 21992K 0 0 0:01:02 /usr/bin/nprobe /run/nprobe-flow.conf
235817 nprobe 20 0 1369M 941M 21992K 2.6 0.2 6:12:61 /usr/bin/nprobe /run/nprobe-flow.conf
215345 root 20 0 2730M 47260 19656 0 0 0:09:52 /usr/lib/snmpd/snmpd
215333 root 20 0 2730M 47260 19656 0 0 0:00:23 /usr/lib/snmpd/snmpd
215332 root 20 0 2730M 47260 19656 0 0 0:00:22 /usr/lib/snmpd/snmpd
215331 root 20 0 2730M 47260 19656 0 0 0:00:06 /usr/lib/snmpd/snmpd
215329 root 20 0 2730M 47260 19656 0 0 0:00:24 /usr/lib/snmpd/snmpd
215328 root 20 0 2730M 47260 19656 0 0 0:00:25 /usr/lib/snmpd/snmpd
215327 root 20 0 2730M 47260 19656 0 0 0:00:25 /usr/lib/snmpd/snmpd
215326 root 20 0 2730M 47260 19656 0 0 0:00:26 /usr/lib/snmpd/snmpd
215325 root 20 0 2730M 47260 19656 0 0 0:00:27 /usr/lib/snmpd/snmpd
215324 root 20 0 2730M 47260 19656 0 0 0:00:22 /usr/lib/snmpd/snmpd
215323 root 20 0 2730M 47260 19656 0 0 0:00:26 /usr/lib/snmpd/snmpd
215322 root 20 0 2730M 47260 19656 0 0 0:00:28 /usr/lib/snmpd/snmpd
215321 root 20 0 2730M 47260 19656 0 0 0:00:21 /usr/lib/snmpd/snmpd
215320 root 20 0 2730M 47260 19656 0 0 0:00:35 /usr/lib/snmpd/snmpd
215319 root 20 0 2730M 47260 19656 0 0 0:00:24 /usr/lib/snmpd/snmpd
215318 root 20 0 2730M 47260 19656 0 0 0:00:28 /usr/lib/snmpd/snmpd
215317 root 20 0 2730M 47260 19656 0 0 0:00:36 /usr/lib/snmpd/snmpd
215316 root 20 0 2730M 47260 19656 0 0 0:00:14 /usr/lib/snmpd/snmpd
215315 root 20 0 2730M 47260 19656 0 0 0:00:35 /usr/lib/snmpd/snmpd
215314 root 20 0 2730M 47260 19656 0 0 0:00:29 /usr/lib/snmpd/snmpd
215313 root 20 0 2730M 47260 19656 0 0 0:00:28 /usr/lib/snmpd/snmpd

```

Current traffic rate

```

Nov 16 15:42:43 TTPOF-PNPR-LNTW001 nprobe[235767]: 16/Nov/2022 15:42:43 [nprobe.c:4206] -----
Nov 16 15:42:43 TTPOF-PNPR-LNTW001 nprobe[235767]: 16/Nov/2022 15:42:43 [nprobe.c:4209] Average traffic: [316.03 K pps][All Traffic 2.32 Gb/sec][IP Traffic 2.21 Gb/sec][ratio 0.96]
Nov 16 15:42:43 TTPOF-PNPR-LNTW001 nprobe[235767]: 16/Nov/2022 15:42:43 [nprobe.c:4217] Current traffic: [476.94 K pps][3.53 Gb/sec]
Nov 16 15:42:43 TTPOF-PNPR-LNTW001 nprobe[235767]: 16/Nov/2022 15:42:43 [nprobe.c:4258] Flows exports (including drops) [32504914 flows][avg: 541748.6 flows/sec][latest 60 sec avg: 2445.4 flows/sec]
Nov 16 15:42:43 TTPOF-PNPR-LNTW001 nprobe[235767]: 16/Nov/2022 15:42:43 [nprobe.c:4266] Flow drops (export queue full: 0)
Nov 16 15:42:43 TTPOF-PNPR-LNTW001 nprobe[235767]: 16/Nov/2022 15:42:43 [nprobe.c:4269] Packet Drops (too many flow buckets: 0)
Nov 16 15:42:43 TTPOF-PNPR-LNTW001 nprobe[235767]: 16/Nov/2022 15:42:43 [nprobe.c:4272] Flow Buckets (active: 374649)[allocated: 374649][toBeExported: 0]
Nov 16 15:42:43 TTPOF-PNPR-LNTW001 nprobe[235767]: 16/Nov/2022 15:42:43 [nprobe.c:4276] Export Queue (current: 0)[max: 10485760][fill level: 0.0%]
Nov 16 15:42:43 TTPOF-PNPR-LNTW001 nprobe[235767]: 16/Nov/2022 15:42:43 [nprobe.c:4010] Processed packets: 6448290598 (max bucket search: 7)
Nov 16 15:42:43 TTPOF-PNPR-LNTW001 nprobe[235767]: 16/Nov/2022 15:42:43 [nprobe.c:3993] Fragment queue length: 281
Nov 16 15:42:43 TTPOF-PNPR-LNTW001 nprobe[235767]: 16/Nov/2022 15:42:43 [nprobe.c:4041] Flow export stats: [1704116834 bytes/2056566499 pkts][32504914 flows/9521355 pkts sent]
Nov 16 15:42:43 TTPOF-PNPR-LNTW001 nprobe[235767]: 16/Nov/2022 15:42:43 [nprobe.c:4047] Flow export drop stats: [0 bytes/0 pkts][0 flows/0.0%]
Nov 16 15:42:43 TTPOF-PNPR-LNTW001 nprobe[235767]: 16/Nov/2022 15:42:43 [nprobe.c:4055] Total flow stats: [1704116834 bytes/2056566499 pkts][32504914 flows/9521355 pkts sent]

```

Current io rate

```

Total DISK READ:      0.00 B/s | Total DISK WRITE:      14.25 M/s
Current DISK READ:    0.00 B/s | Current DISK WRITE:    28.14 M/s

  TID  PRIO  USER      DISK READ  DISK WRITE  SWAPIN     IO>    COMMAND
 2006  be/3  root       0.00 B/s   13.03 K/s   0.00 %    2.19 % [jbd2/sda1-8]
 5756  be/4  elastics   0.00 B/s   172.71 K/s  0.00 %    0.13 % java -Xshare:auto -Des.networkaddress.cache.ttl=60
 5876  be/4  elastics   0.00 B/s   192.26 K/s  0.00 %    0.11 % java -Xshare:auto -Des.networkaddress.cache.ttl=60
 5877  be/4  elastics   0.00 B/s   192.26 K/s  0.00 %    0.11 % java -Xshare:auto -Des.networkaddress.cache.ttl=60
 5878  be/4  elastics   0.00 B/s   175.97 K/s  0.00 %    0.11 % java -Xshare:auto -Des.networkaddress.cache.ttl=60
 6326  be/4  elastics   0.00 B/s   172.71 K/s  0.00 %    0.10 % java -Xshare:auto -Des.networkaddress.cache.ttl=60
247066  be/4  root       0.00 B/s    0.00 B/s   0.00 %    0.08 % [kworker/u193:1-events_power_efficient]
 6425  be/4  elastics   0.00 B/s   143.38 K/s  0.00 %    0.08 % java -Xshare:auto -Des.networkaddress.cache.ttl=60
 6013  be/4  elastics   0.00 B/s   101.02 K/s  0.00 %    0.06 % java -Xshare:auto -Des.networkaddress.cache.ttl=60
 6466  be/4  elastics   0.00 B/s   101.02 K/s  0.00 %    0.04 % java -Xshare:auto -Des.networkaddress.cache.ttl=60
 6464  be/4  elastics   0.00 B/s   140.12 K/s  0.00 %    0.04 % java -Xshare:auto -Des.networkaddress.cache.ttl=60
 6420  be/4  elastics   0.00 B/s    97.76 K/s  0.00 %    0.04 % java -Xshare:auto -Des.networkaddress.cache.ttl=60
 6419  be/4  elastics   0.00 B/s    87.98 K/s  0.00 %    0.04 % java -Xshare:auto -Des.networkaddress.cache.ttl=60

```

Disk config

```

root@TTPOP-PNPR-LNTW001:/etc/apt# df -h
Filesystem                Size      Used Avail Use% Mounted on
udev                      252G         0   252G   0% /dev
tmpfs                     51G         2.9M    51G   1% /run
/dev/mapper/ubuntu--vg-ubuntu--lv 109G        23G    81G  22% /
tmpfs                     252G         0   252G   0% /dev/shm
tmpfs                     5.0M         0    5.0M   0% /run/lock
tmpfs                     252G         0   252G   0% /sys/fs/cgroup
/dev/sdb2                 974M       206M   701M  23% /boot
/dev/sdb1                 511M        5.3M   506M   2% /boot/efi
/dev/sda1                 11T         5.0T   4.7T  52% /opt
/dev/loop1                56M         56M     0 100% /snap/core18/2566
/dev/loop3                64M         64M     0 100% /snap/core20/1623
/dev/loop4                68M         68M     0 100% /snap/lxd/22526
/dev/loop7                68M         68M     0 100% /snap/lxd/22753
/dev/loop8                48M         48M     0 100% /snap/snapd/17336
tmpfs                     51G         0    51G   0% /run/user/1001
/dev/loop5                56M         56M     0 100% /snap/core18/2620
/dev/loop9                64M         64M     0 100% /snap/core20/1695
/dev/loop0                50M         50M     0 100% /snap/snapd/17576
root@TTPOP-PNPR-LNTW001:/etc/apt# lsblk
NAME                MAJ:MIN RM   SIZE RO TYPE MOUNTPOINT
loop0                7:0      0  49.7M  1 loop /snap/snapd/17576
loop1                7:1      0  55.6M  1 loop /snap/core18/2566
loop3                7:3      0  63.2M  1 loop /snap/core20/1623
loop4                7:4      0  67.9M  1 loop /snap/lxd/22526
loop5                7:5      0  55.6M  1 loop /snap/core18/2620
loop6                7:6      0   48M   1 loop
loop7                7:7      0  67.8M  1 loop /snap/lxd/22753
loop8                7:8      0   48M   1 loop /snap/snapd/17336
loop9                7:9      0  63.2M  1 loop /snap/core20/1695
sda                  8:0      0  10.2T  0 disk
└─sda1                8:1      0  10.2T  0 part /opt
sdb                  8:16     0  223.5G  0 disk
├─sdb1                8:17     0   512M  0 part /boot/efi
├─sdb2                8:18     0    1G   0 part /boot
└─sdb3                8:19     0  222G   0 part
   └─ubuntu--vg-ubuntu--lv 253:0    0  111G   0 lvm /

```

What's Next:

Well, we started to deep dive into traffic and analyze, create widgets and all the necessary stuff to have a Management dashboard. We saw some interesting stuff too which will need a lot of troubleshooting and investigation.

Revision #9

Created 2022-11-16 15:27:33 UTC by Mesut Bayrak

Updated 2022-11-17 16:50:33 UTC by Mesut Bayrak