

Towards 100 Gbit Flow-Based Network Monitoring

Luca Deri <deri@ntop.org>

Alfredo Cardigliano <cardigliano@ntop.org>

Outlook

1. Motivation: Towards 100 Gbit Traffic Monitoring
2. Our Heritage: ntop Tools
3. Cento: 100 Gb Flow Monitoring

Towards 100 Gbit Networking

- In the past few years the market has moved from 10 Gbit to 40 Gbit, and recently to 100 Gbit. 100 Gbit is still not very popular but it is becoming more and more popular as per-port/optical adapters price drop.
- As it happened years ago with the transition from 1 to 10 Gbit, 100 Gbit has been initially available only on switches and routers, but there are now 100 Gbit host network adapters on the market.

100 Gbit Packet Capture

- FPGA-based NICs (e.g. Accolade and Napatech) have been out since more than a year now, with prices (not including optics) starting in the sub-10K USD range.
- Recently Intel has introduced the FM10000 10/25/40/100 Gbit Ethernet controller (Red Rock Canyon) that offloads to the controller selected features (e.g. packet switching /distribution/drop). The first products have been announced and will be available in 1Q16 for < 1.5k USD (dual 100 Gbit).

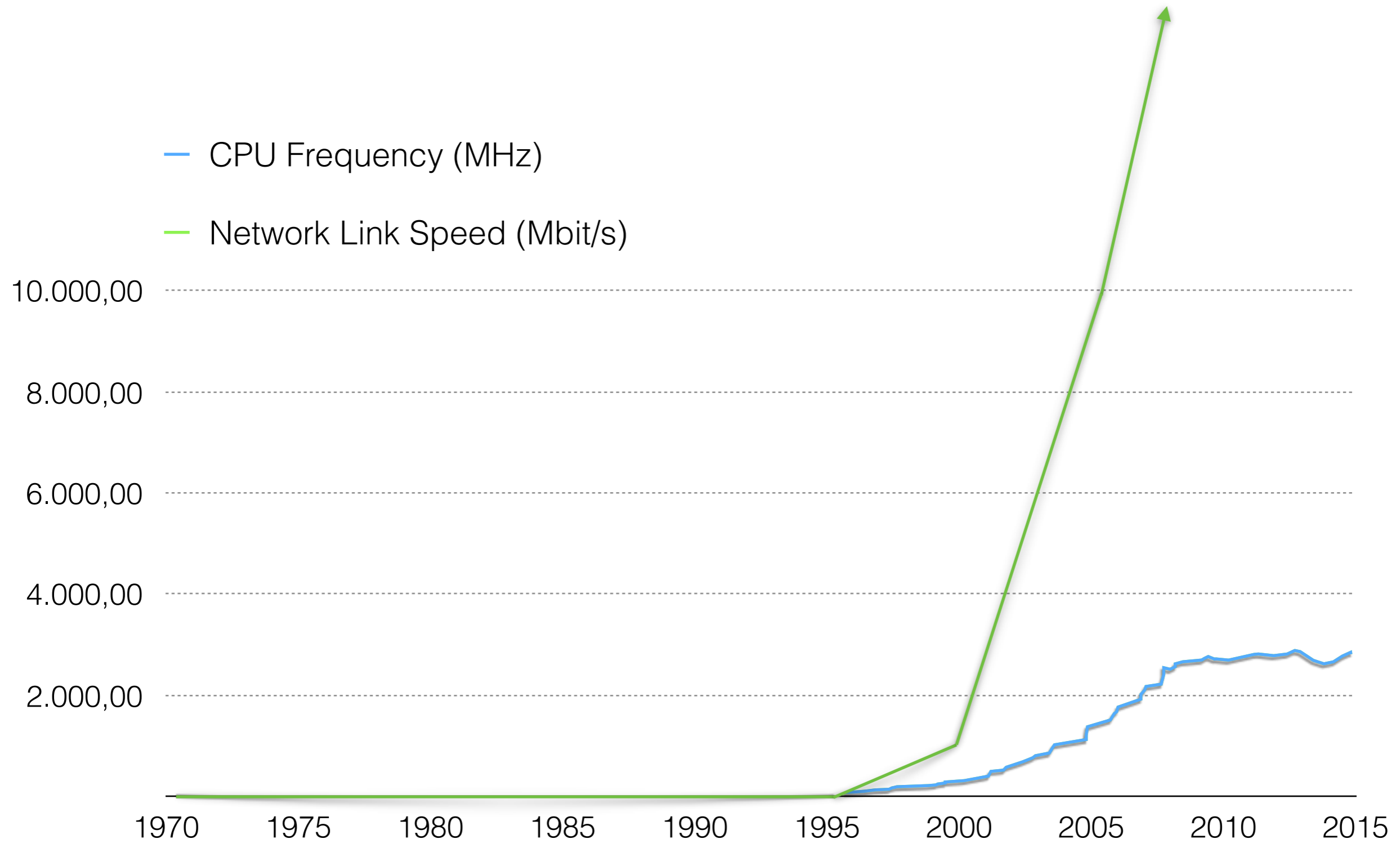


From 10 Gbit to 100 Gbit

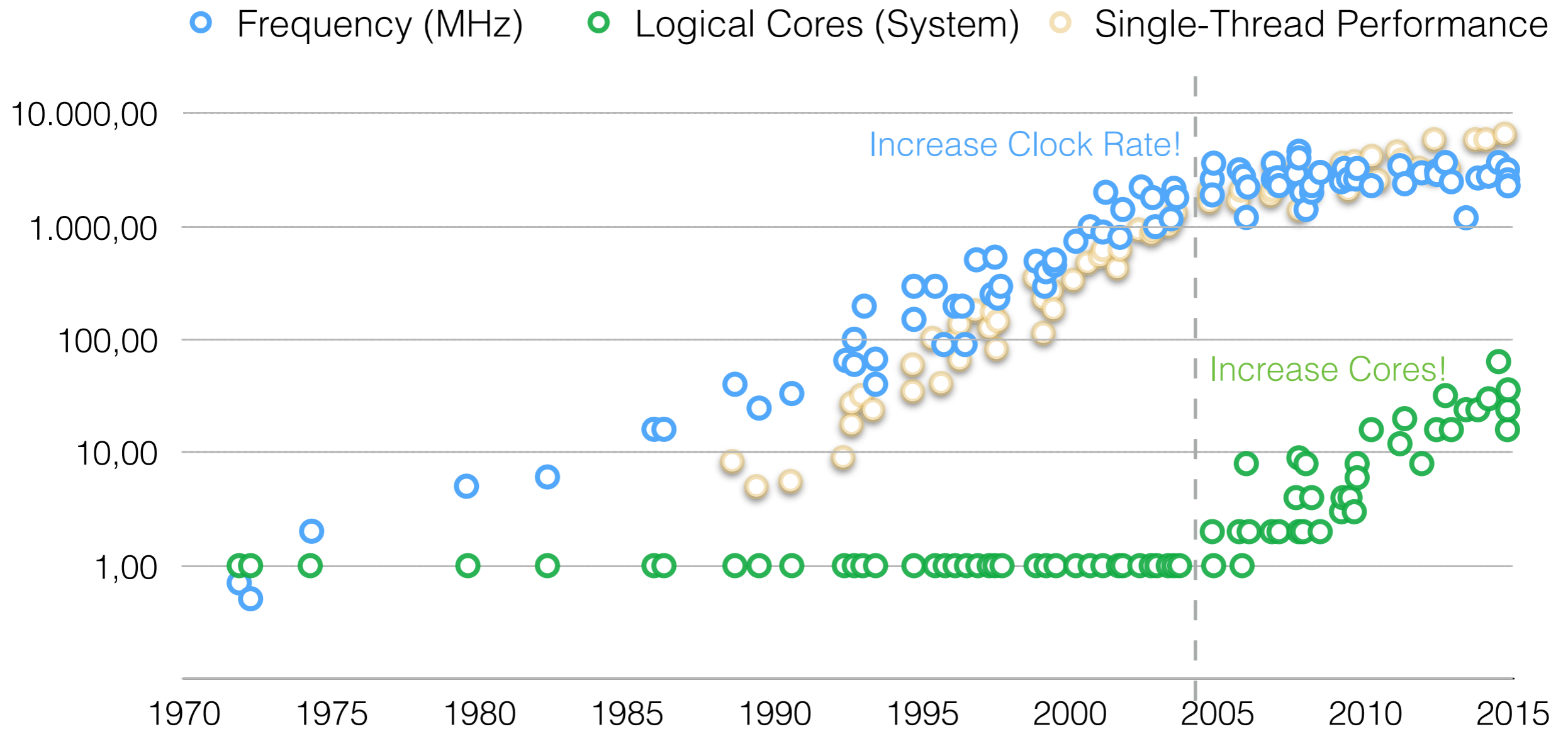
- As RSS (Resource Side Scaling) allows network adapters to distribute traffic across cores on modern network adapters, it is possible to do the same using network switches.
- Some products (e.g. Dell Z9100-ON) allow traffic to be statically (no RSS-like features) distributed across multiple 10/25/40 Gbit ports for reducing 100 Gbit monitoring to multi 10-Gbit monitoring.



Trends: Network vs CPU



Trends: CPU Performance



Trends: User Requirements [1/2]

- For years the industry and community focused mainly on packet capture/filtering.
- Applications were relatively simple and self-contained: network security, traffic monitoring, high-frequency trading, packet-to-disk....
- With the advent of big-data systems (and not only that) and reduction of data center space, people would like to collapse multi-apps on a single box.

Trends: User Requirements [2/2]

- 100 Gbit (and partially 40 Gbit) are raising the bar once more, and (FPGA-based) NICs are solving “just” the packet capture problem.
- Unfortunately “packet capture acceleration” is no longer enough as we often need to combine it with multi-app traffic distribution, balancing, and pipelining in order to collapse on one box at high speed, functionalities that were previously implemented onto multiple boxes.

Trends: Application Performance

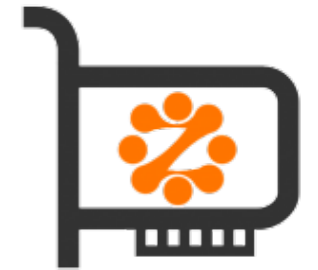
- IDS/IPS Applications (e.g. Snort, Bro and Suricata) are CPU bound. Suricata on a E5-2690v2 3GHz (10 cores+HT all in use) and a FPGA NIC can do ~14.1 Gbps with real traffic (~512 bytes or more in average).
- ntop's NetFlow probe (nProbe) can process ~3.5 Mpps (so line rate with real traffic) per core on a Intel E3-1230v3.

Problem Statement

- Is it possible to implement line-rate 100 Gbit (or 10x10 Gbit) flow based monitoring on a x86 box ?
- Can we instrument a flow-monitoring tool to egress/drop selected traffic to other applications that can further process the traffic ?
- Can we combine flow visibility with other functionality such as IDS or packet to disk on the same box ?
- Can we build multi 1/10 Gbit flow sensors using low-cost hardware?



PF_RING ZC [1/2]

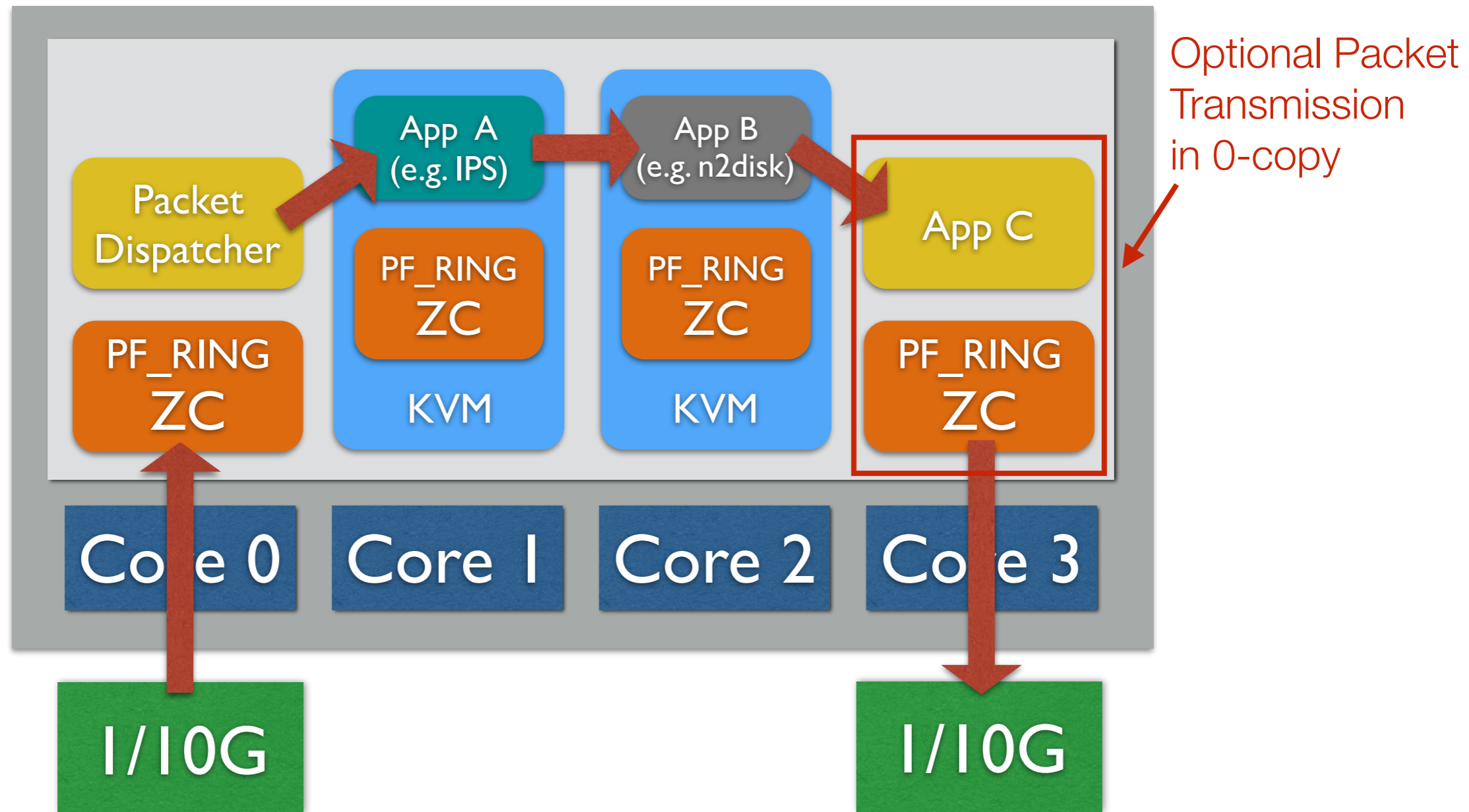
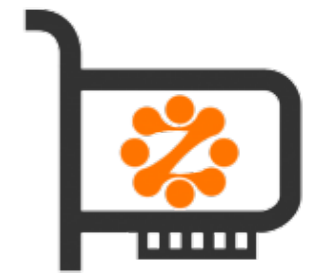


- PF_RING is a home-grown open source packet processing framework for Linux.
- Support of legacy pcap-based applications as well FPGA NICs.



- ZC has simple yet powerful components (no complex patterns, queue/consumer/balancer).
- KVM/Docker/OpenStack support: ability to setup Inter-VM clustering.
- Native PF_RING ZC support in many open-source applications such as Snort, Suricata, Bro, Wireshark.
- Ability to operate on top of sysdig.org for dispatching system events to PF_RING applications.

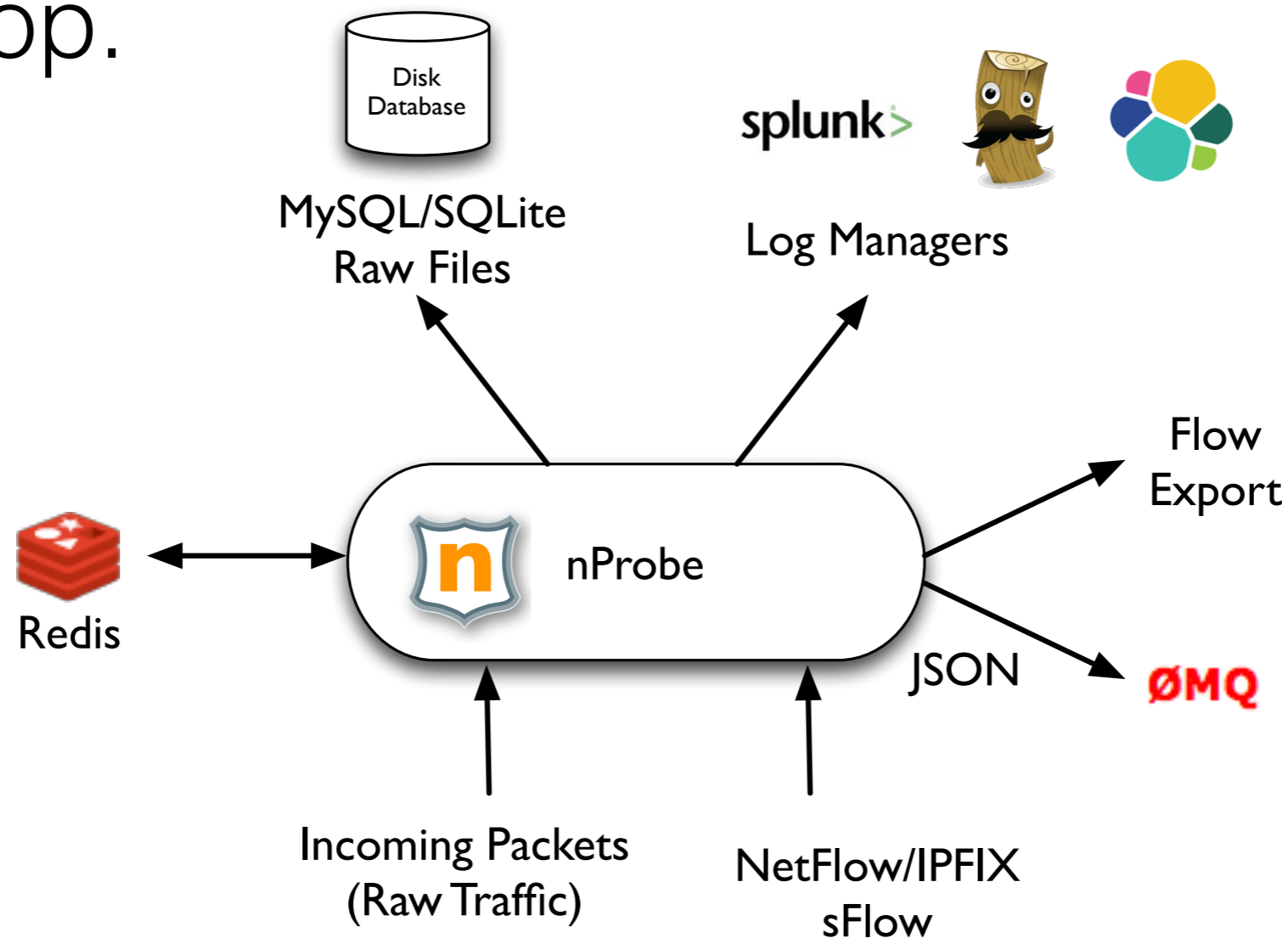
PF_RING ZC [2/2]



```
(Host) $ ./zpipeline_ipc -i zc:eth2,0 -o zc:eth3,1 -n 2 -c 99 -r 1 -t 2 -Q /tmp/qmp0
(VM)   $ ./zbounce_ipc -c 99 -i 0 -o 1 -g 3
```

nProbe [1/2]

- nProbe is a high-speed (1/10 Gbit) open source traffic probe/collector developed by ntop.



nProbe [2/2]

- It has an open architecture extensible by means of plugins that include:
 - GTP (v0, v1, v2) plugins.
 - VoIP (SIP and RTP) plugins for analysing voice signalling (who's calling who/when) and voice quality (Jitter and pseudo-MOS/R-Factor).
 - HTTP(S), Email (SMTP, IMAP, POP3), Radius, Database (Oracle and MySQL), FTP, DHCP, and BGP, SSDP, NetBIOS, DNS/MDNS.
 - JSON export (Text File, Splunk, ElasticSearch, Kafka)

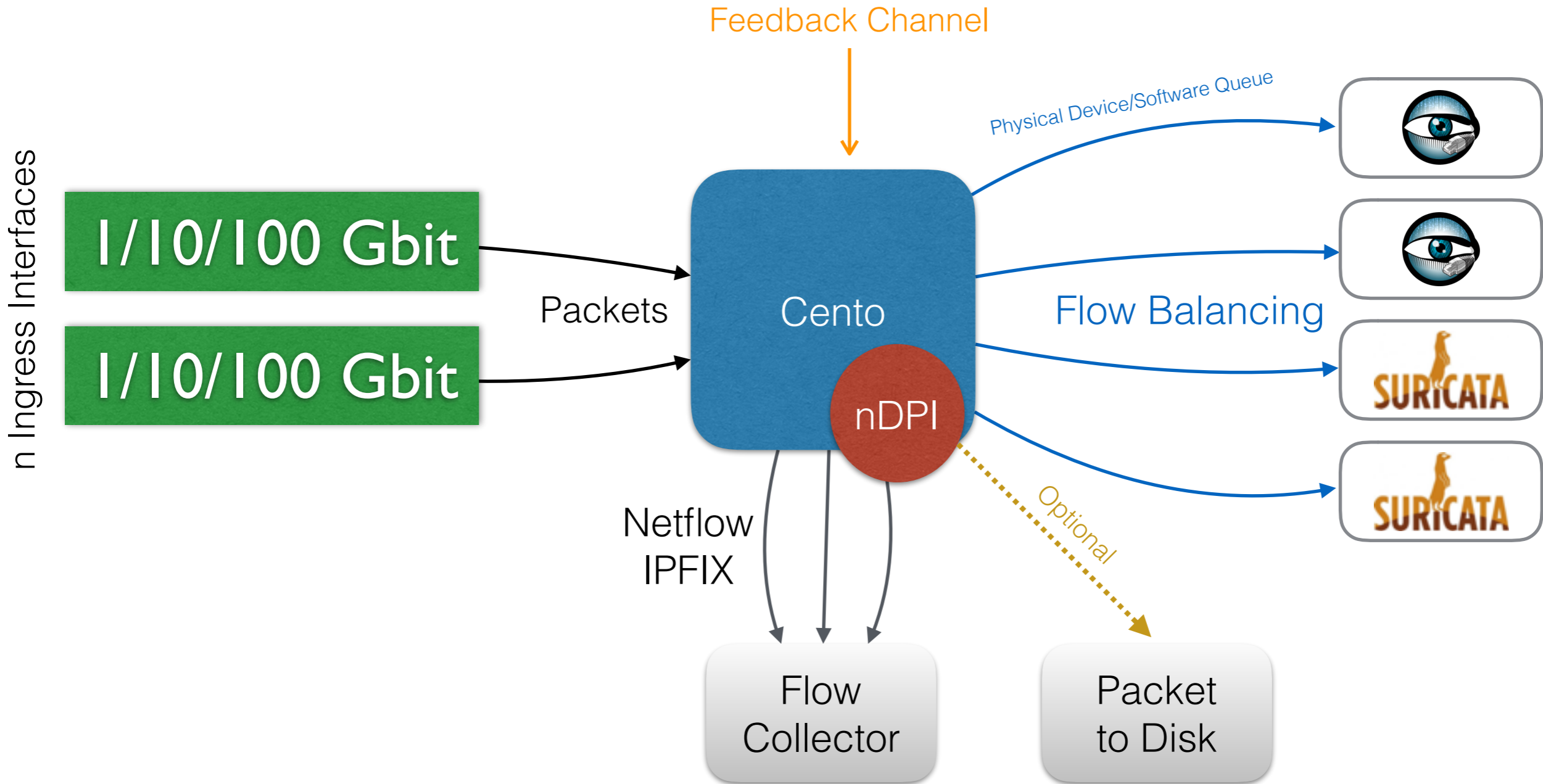
What is Still Missing?

- PF_RING ZC allows to be efficiently received/transmitted (line rate, 64 bytes packet) from/to ethernet devices or in-host/VM queues.
- nProbe is an advanced NetFlow probe featuring many plugins and information elements but it has been designed in the 1/10 Gbit days, so fast but not 100 Gbit-ready (too in depth packet analysis).
- Can we design (from scratch) a new 100 Gbit probe able to exploit on ZC and perform other tasks such as distribute traffic to other applications in order to fully exploit modern multi-core Intel boxes ?

Welcome to Cento (/ 5)

- Leveraging on more than 15 years of high-speed packet processing and NetFlow monitoring we have decided to code from scratch a new flow-based sensor.
- Cento (“one hundred” in Italian) has been designed as the first component of a monitoring system, the one that captures ingress packets, classify them via DPI (Deep Packet Inspection), and performs optional actions on selected packets.

Cento Architecture



Cento Design Principles [1/3]

- Export in NetFlow v5/v9/IPFIX, JSON, Text (soon also Kentik Flow and Kafka).
- Coded in “simple” C++ to avoid runtime slowdowns.
- Ingress traffic is split across multiple interfaces/memory areas by RSS-like techniques: one thread per interface.
- Avoid locks when possible: one flow exporter/cache per sensor thread.
- Short locks when unavoidable: in case of synchronisation (e.g. during export) locks enclose very short code and are clustered (e.g. one lock every 500k flow export).

Centos Design Principles [2/3]

- All data structures are cache-line aligned (typically 64 bytes).
- Large hash tables (used to host flows) put a lot of pressure on memory and so prefetching is widely used to increase performance.
- As ZC prefetches packets too, prefetching must be carefully used to avoid it to slow down the application instead of accelerating it.

Centro Design Principles [3/3]

- Ingress packet processing has maximum priority over other tasks (e.g. flow export), and thus multiple low-priority actives can be collapsed on the same core, while dedicating a complete core to packet processor threads.
- In case centro has been configured to egress packets, ZC does that by putting the packet reference onto the egress queue/device (no memory copy unless FPGA-adapters are used as they allocate memory on a custom/proprietary fashion).

Hardware Hashing and Prefetching

- Modern network adapters are able to provide both packet payload and metadata. Inside packet metadata many adapters (e.g. Intel) can store a packet hash (e.g. used for RSS) in addition to hardware timestamp.
- We planned to use this hash to both:
 - avoid hashing in software
 - prefetch hash buckets in order to increment the performance.
- Unfortunately when using the 5-tuple hash (e.g. Toeplitz Hash on Intel) the number of hash collision was higher than the software hash we implemented, and thus the performance was worse. Not a good idea, then we stick to the software hash.

DPI Impact on Performance

- ntop develop and maintains an open source DPI toolkit named *nDPI* that supports over 200 protocols (e.g. Skype, SSL, BitTorrent...).
- Cento can optionally use nDPI to identify application protocols. Detection happens at flow start using less than 10 packets.
- Enabling nDPI with all protocols on Cento the performance is reduced of 20/40% (depends on flow duration).
- We have developed a μ -nDPI that contains just a few protocols such as HTTP/SSL/DNS and enabling it, the performance degradation it is almost unnoticeable.

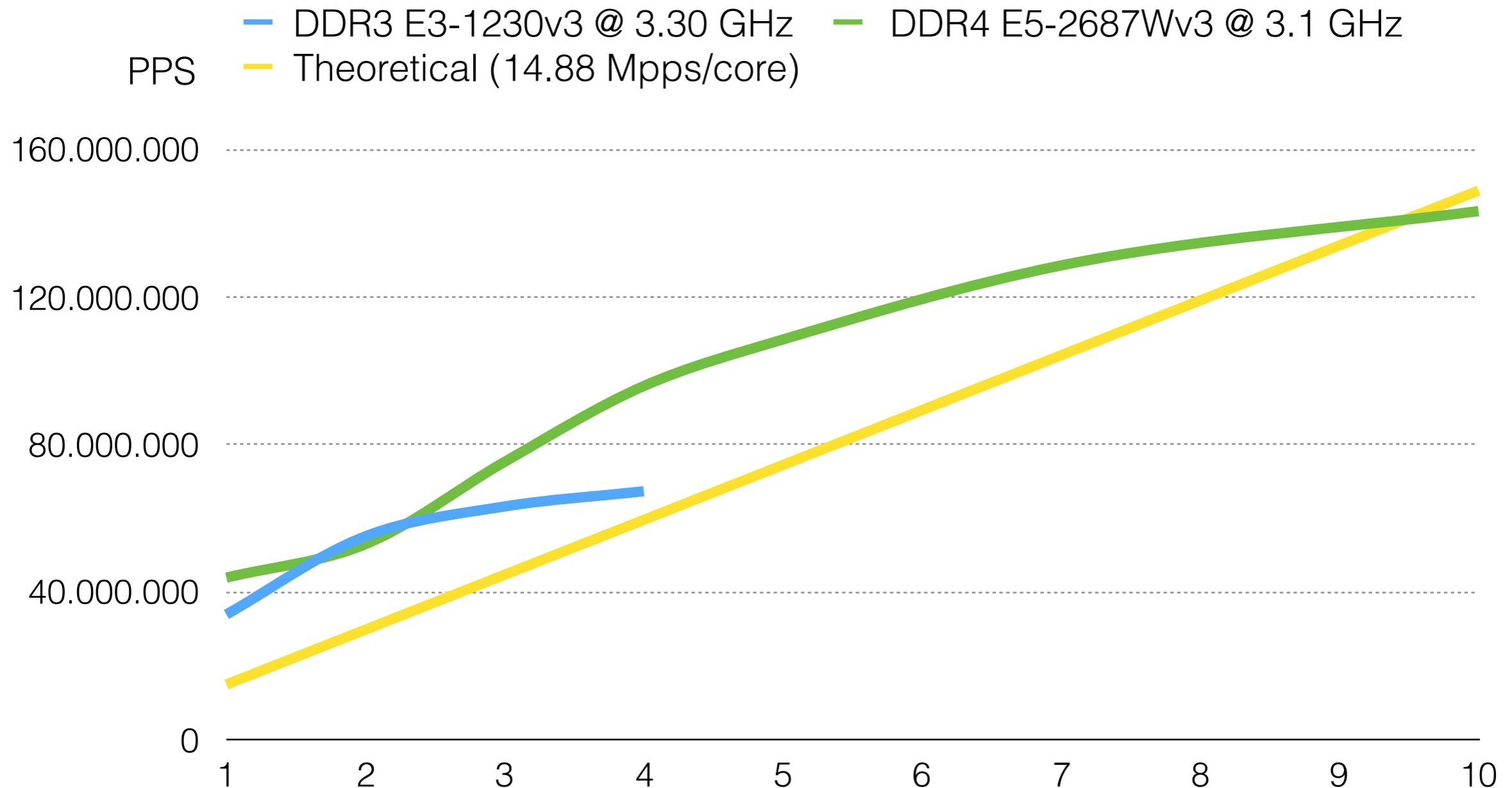


Performance Evaluation

In order to evaluate the performance we have tested cento on:

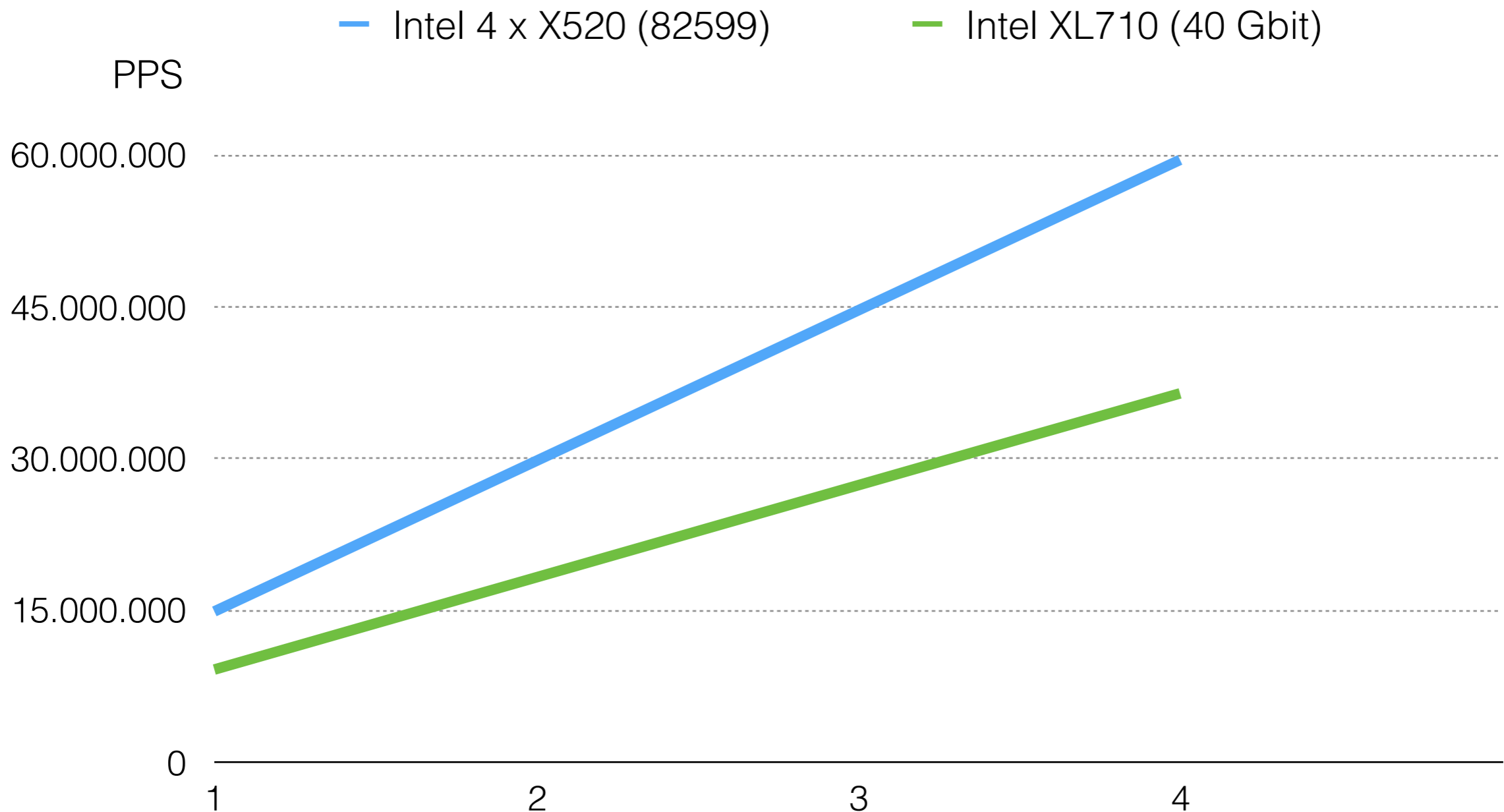
- “Dummy” interface (32k packets/interface in memory to simulate impact on CPU cache).
- PF_RING ZC over Intel 82599 and X710.
- 100 Gbit interfaces (Accolade Technology and Napatech).

Scalability: Dummy Interface



Per-core Processing Performance (500k Flows/Interface, Dummy Interface)

Scalability: Intel multi-10 Gbit



Per-core Processing Performance (500k Flows/Interface, Intel E3-1230v3)

Scalability: Real 100 Gbit Traffic (64 bytes)

```
30/Nov/2015 14:05:51 [anic:0@0] [6'102'338 pps/4.10 Gbps][500'000/0/0 act/exp/drop flows][0/0 RX/TX pkt drops]
30/Nov/2015 14:05:51 [anic:0@1] [4'638'028 pps/3.12 Gbps][500'000/0/0 act/exp/drop flows][0/0 RX/TX pkt drops]
30/Nov/2015 14:05:51 [anic:0@2] [4'637'369 pps/3.12 Gbps][500'000/0/0 act/exp/drop flows][0/0 RX/TX pkt drops]
30/Nov/2015 14:05:51 [anic:0@3] [3'905'982 pps/2.62 Gbps][500'000/0/0 act/exp/drop flows][0/0 RX/TX pkt drops]
30/Nov/2015 14:05:51 [anic:0@4] [5'125'179 pps/3.44 Gbps][500'000/0/0 act/exp/drop flows][0/0 RX/TX pkt drops]
30/Nov/2015 14:05:51 [anic:0@5] [5'373'763 pps/3.61 Gbps][500'000/0/0 act/exp/drop flows][0/0 RX/TX pkt drops]
30/Nov/2015 14:05:51 [anic:0@6] [6'105'389 pps/4.10 Gbps][500'000/0/0 act/exp/drop flows][0/0 RX/TX pkt drops]
30/Nov/2015 14:05:51 [anic:0@7] [4'637'409 pps/3.12 Gbps][500'000/0/0 act/exp/drop flows][0/0 RX/TX pkt drops]
30/Nov/2015 14:05:51 [anic:0@8] [4'149'903 pps/2.79 Gbps][500'000/0/0 act/exp/drop flows][0/0 RX/TX pkt drops]
30/Nov/2015 14:05:51 [anic:0@9] [6'348'062 pps/4.27 Gbps][500'000/0/0 act/exp/drop flows][0/0 RX/TX pkt drops]
30/Nov/2015 14:05:51 [anic:0@10] [4'393'731 pps/2.95 Gbps][500'000/0/0 act/exp/drop flows][0/0 RX/TX pkt drops]
30/Nov/2015 14:05:51 [anic:0@11] [7'080'337 pps/4.76 Gbps][500'000/0/0 act/exp/drop flows][0/0 RX/TX pkt drops]
30/Nov/2015 14:05:51 [cento.cpp:499] Actual stats: 62'497'490 pps/0 drops
```

```
%Cpu14 : 50.8 us, 0.0 sy, 0.0 ni, 49.2 id, 0.0 wa, 0.0 hi, 0.0 si, 0.0 st
%Cpu15 : 39.5 us, 0.3 sy, 0.0 ni, 60.1 id, 0.0 wa, 0.0 hi, 0.0 si, 0.0 st
%Cpu16 : 39.2 us, 0.7 sy, 0.0 ni, 60.1 id, 0.0 wa, 0.0 hi, 0.0 si, 0.0 st
%Cpu17 : 32.2 us, 0.3 sy, 0.0 ni, 67.4 id, 0.0 wa, 0.0 hi, 0.0 si, 0.0 st
%Cpu18 : 45.5 us, 0.3 sy, 0.0 ni, 54.2 id, 0.0 wa, 0.0 hi, 0.0 si, 0.0 st
%Cpu19 : 45.2 us, 0.3 sy, 0.0 ni, 54.5 id, 0.0 wa, 0.0 hi, 0.0 si, 0.0 st
%Cpu20 : 51.3 us, 0.0 sy, 0.0 ni, 48.7 id, 0.0 wa, 0.0 hi, 0.0 si, 0.0 st
%Cpu21 : 39.3 us, 0.3 sy, 0.0 ni, 60.3 id, 0.0 wa, 0.0 hi, 0.0 si, 0.0 st
%Cpu22 : 35.2 us, 0.3 sy, 0.0 ni, 64.5 id, 0.0 wa, 0.0 hi, 0.0 si, 0.0 st
%Cpu23 : 50.7 us, 0.3 sy, 0.0 ni, 49.0 id, 0.0 wa, 0.0 hi, 0.0 si, 0.0 st
%Cpu24 : 35.9 us, 0.0 sy, 0.0 ni, 64.1 id, 0.0 wa, 0.0 hi, 0.0 si, 0.0 st
%Cpu25 : 56.5 us, 0.7 sy, 0.0 ni, 42.9 id, 0.0 wa, 0.0 hi, 0.0 si, 0.0 st
```



~40% Line Rate

NOTE: Due to traffic generator limitations, testing at line rate has not been possible.

Scalability: Real 100 Gbit Traffic (389 bytes)

```
30/Nov/2015 14:17:42 [anic:0@0] [1'933'581 pps/6.33 Gbps][392'869/18'062'399/0 act/exp/drop flows][0/0 RX/TX pkt drops]
30/Nov/2015 14:17:42 [anic:0@1] [1'933'801 pps/6.33 Gbps][469'480/17'786'458/0 act/exp/drop flows][0/0 RX/TX pkt drops]
30/Nov/2015 14:17:42 [anic:0@2] [3'975'172 pps/13.01 Gbps][500'000/18'026'978/0 act/exp/drop flows][0/0 RX/TX pkt drops]
30/Nov/2015 14:17:42 [anic:0@3] [3'112'961 pps/10.19 Gbps][500'000/17'665'265/0 act/exp/drop flows][0/0 RX/TX pkt drops]
30/Nov/2015 14:17:42 [anic:0@4] [3'920'094 pps/12.83 Gbps][500'000/18'032'939/0 act/exp/drop flows][0/0 RX/TX pkt drops]
30/Nov/2015 14:17:42 [anic:0@5] [1'932'245 pps/6.32 Gbps][394'838/17'996'664/0 act/exp/drop flows][0/0 RX/TX pkt drops]
30/Nov/2015 14:17:42 [anic:0@6] [0 pps/0.00 Gbps][500'000/4'582'548/0 act/exp/drop flows][0/0 RX/TX pkt drops]
30/Nov/2015 14:17:42 [anic:0@7] [3'952'837 pps/12.93 Gbps][500'000/17'909'541/0 act/exp/drop flows][0/0 RX/TX pkt drops]
30/Nov/2015 14:17:42 [anic:0@8] [1'933'887 pps/6.33 Gbps][393'166/17'648'878/0 act/exp/drop flows][0/0 RX/TX pkt drops]
30/Nov/2015 14:17:42 [anic:0@9] [3'909'272 pps/12.79 Gbps][500'000/18'260'938/0 act/exp/drop flows][0/0 RX/TX pkt drops]
30/Nov/2015 14:17:42 [anic:0@10] [3'896'342 pps/12.75 Gbps][500'000/17'854'718/0 act/exp/drop flows][0/0 RX/TX pkt drops]
30/Nov/2015 14:17:42 [anic:0@11] [0 pps/0.00 Gbps][500'000/4'632'836/0 act/exp/drop flows][0/0 RX/TX pkt drops]
30/Nov/2015 14:17:42 [cento.cpp:499] Actual stats: 30'500'192 pps/0 drops
```

```
%Cpu14 : 66.6 us, 0.3 sy, 0.0 ni, 33.1 id, 0.0 wa, 0.0 hi, 0.0 si, 0.0 st
%Cpu15 : 65.2 us, 0.7 sy, 0.0 ni, 34.1 id, 0.0 wa, 0.0 hi, 0.0 si, 0.0 st
%Cpu16 : 98.3 us, 0.3 sy, 0.0 ni, 1.3 id, 0.0 wa, 0.0 hi, 0.0 si, 0.0 st
%Cpu17 : 98.7 us, 0.0 sy, 0.0 ni, 1.3 id, 0.0 wa, 0.0 hi, 0.0 si, 0.0 st
%Cpu18 : 98.7 us, 0.0 sy, 0.0 ni, 1.3 id, 0.0 wa, 0.0 hi, 0.0 si, 0.0 st
%Cpu19 : 66.4 us, 0.3 sy, 0.0 ni, 33.2 id, 0.0 wa, 0.0 hi, 0.0 si, 0.0 st
%Cpu20 : 1.0 us, 0.3 sy, 0.0 ni, 98.7 id, 0.0 wa, 0.0 hi, 0.0 si, 0.0 st
%Cpu21 : 97.7 us, 0.3 sy, 0.0 ni, 2.0 id, 0.0 wa, 0.0 hi, 0.0 si, 0.0 st
%Cpu22 : 66.0 us, 0.7 sy, 0.0 ni, 33.3 id, 0.0 wa, 0.0 hi, 0.0 si, 0.0 st
%Cpu23 : 99.0 us, 0.0 sy, 0.0 ni, 1.0 id, 0.0 wa, 0.0 hi, 0.0 si, 0.0 st
%Cpu24 : 98.3 us, 0.0 sy, 0.0 ni, 1.7 id, 0.0 wa, 0.0 hi, 0.0 si, 0.0 st
```

3 sec Flow Lifetime
(Real NetFlow Export)

Accolade
Technology

~100% Line Rate


Traffic Generator Limitation

Scalability: Real 100 Gbit Traffic

More Streams (70 bytes)

26 Streams

```
01/Dec/2015 16:00:12 [nt:stream0] [5'077'577 pps/3.49 Gbps][100/100/0 act/exp/drop flows][0/0 RX/TX pkt drops]
01/Dec/2015 16:00:13 [nt:stream1] [5'076'595 pps/3.49 Gbps][100/100/0 act/exp/drop flows][0/0 RX/TX pkt drops]
01/Dec/2015 16:00:13 [nt:stream2] [5'078'730 pps/3.49 Gbps][100/100/0 act/exp/drop flows][0/0 RX/TX pkt drops]
01/Dec/2015 16:00:14 [nt:stream3] [5'079'094 pps/3.49 Gbps][100/100/0 act/exp/drop flows][0/0 RX/TX pkt drops]
01/Dec/2015 16:00:14 [nt:stream4] [5'079'561 pps/3.49 Gbps][100/100/0 act/exp/drop flows][0/0 RX/TX pkt drops]
01/Dec/2015 16:00:15 [nt:stream5] [5'079'930 pps/3.49 Gbps][100/100/0 act/exp/drop flows][0/0 RX/TX pkt drops]
01/Dec/2015 16:00:15 [nt:stream6] [5'079'448 pps/3.49 Gbps][100/100/0 act/exp/drop flows][0/0 RX/TX pkt drops]
01/Dec/2015 16:00:16 [nt:stream7] [5'080'528 pps/3.50 Gbps][100/100/0 act/exp/drop flows][0/0 RX/TX pkt drops]
01/Dec/2015 16:00:16 [nt:stream8] [5'080'557 pps/3.50 Gbps][100/100/0 act/exp/drop flows][0/0 RX/TX pkt drops]
01/Dec/2015 16:00:17 [nt:stream9] [5'081'058 pps/3.50 Gbps][100/100/0 act/exp/drop flows][0/0 RX/TX pkt drops]
01/Dec/2015 16:00:17 [nt:stream10] [5'080'978 pps/3.50 Gbps][100/100/0 act/exp/drop flows][0/0 RX/TX pkt drops]
01/Dec/2015 16:00:18 [nt:stream11] [5'081'977 pps/3.50 Gbps][100/100/0 act/exp/drop flows][0/0 RX/TX pkt drops]
01/Dec/2015 16:00:18 [nt:stream12] [5'082'670 pps/3.50 Gbps][100/200/0 act/exp/drop flows][0/0 RX/TX put drops]
01/Dec/2015 16:00:19 [nt:stream13] [5'082'088 pps/3.50 Gbps][100/200/0 act/exp/drop flows][0/0 RX/TX pkt drops]
01/Dec/2015 16:00:19 [nt:stream14] [5'082'256 pps/3.50 Gbps][100/100/0 act/exp/drop flows][0/0 RX/TX pkt drops]
01/Dec/2015 16:00:20 [nt:stream15] [5'082'632 pps/3.50 Gbps][100/100/0 act/exp/drop flows][0/0 RX/TX pkt drops]
01/Dec/2015 16:00:20 [nt:stream16] [5'083'798 pps/3.50 Gbps][100/100/0 act/exp/drop flows][0/0 RX/TX pkt drops]
01/Dec/2015 16:00:21 [nt:stream17] [5'084'046 pps/3.50 Gbps][100/100/0 act/exp/drop flows][0/0 RX/TX pkt drops]
01/Dec/2015 16:00:21 [nt:stream18] [5'083'934 pps/3.50 Gbps][100/100/0 act/exp/drop flows][0/0 RX/TX pkt drops]
01/Dec/2015 16:00:22 [nt:stream19] [5'084'371 pps/3.50 Gbps][100/100/0 act/exp/drop flows][0/0 RX/TX pkt drops]
01/Dec/2015 16:00:22 [nt:stream20] [5'084'487 pps/3.50 Gbps][100/100/0 act/exp/drop flows][0/0 RX/TX pkt drops]
01/Dec/2015 16:00:23 [nt:stream21] [5'084'625 pps/3.50 Gbps][100/100/0 act/exp/drop flows][0/0 RX/TX pkt drops]
01/Dec/2015 16:00:23 [nt:stream22] [5'084'374 pps/3.50 Gbps][100/100/0 act/exp/drop flows][0/0 RX/TX pkt drops]
01/Dec/2015 16:00:24 [nt:stream23] [5'083'037 pps/3.50 Gbps][100/100/0 act/exp/drop flows][0/0 RX/TX pkt drops]
01/Dec/2015 16:00:24 [nt:stream24] [5'083'832 pps/3.50 Gbps][100/100/0 act/exp/drop flows][0/0 RX/TX pkt drops]
01/Dec/2015 16:00:25 [nt:stream25] [5'083'563 pps/3.50 Gbps][100/100/0 act/exp/drop flows][0/0 RX/TX pkt drops]
01/Dec/2015 16:00:25 [cento.cpp:513] Actual stats: 132'125'746 pps/0 drops
```

napatech  ~100% Line Rate

Final Remarks

- Cento has demonstrated that 100 Gbit traffic monitoring is feasible using an x86 server and a 100 Gbit interface as well multiple 10/40 Gbit interfaces.
- It is not uncommon to have 10/14 physical CPUs cores, and Cento can take advantage of them, so you can effectively combine 100 Gbit flow-monitoring with other activities.
- You can use Cento at 10 Gbit line rate to generate flows with as few as 2 cores (1 processing + 1 export) while leaving the remaining cores for Bro, Suricata, n2disk, Wireshark, Snort...
- We plan to release Cento by 2Q16 (beta is already available from <http://packages.ntop.org/>, more info at <http://www.ntop.org>).
- Many thanks to Accolade Technology and Napatech for providing support throughout our tests.