BISDN GmbH

Open Source vBNG Architecture and Performance Evaluation

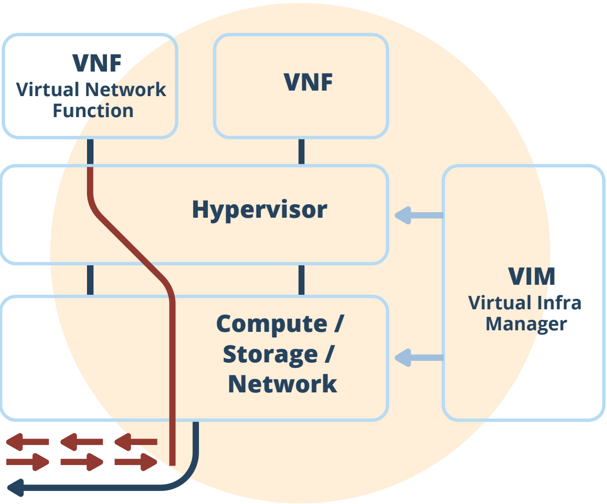
Hagen Woesner

ONF Broadband Spotlight Presentation, June 2020

Challenge: Traditional clouds are unsuited for telco

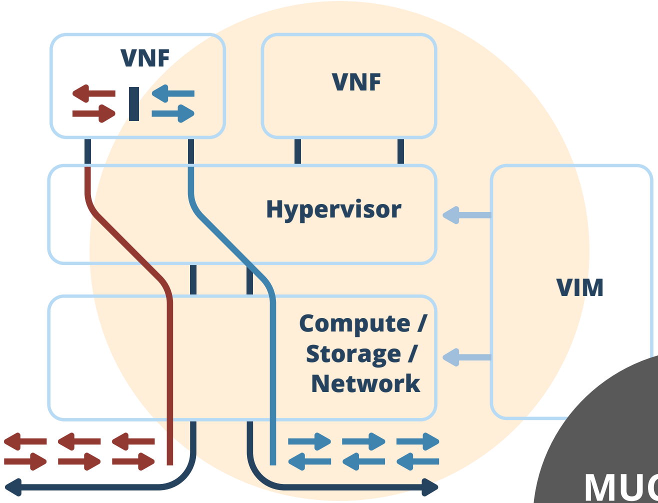
At least twice the bandwidth required for telco services

Traditional cloud: **End of the wire**



Most cloud applications are web services

Telco cloud: **Bump in the wire**

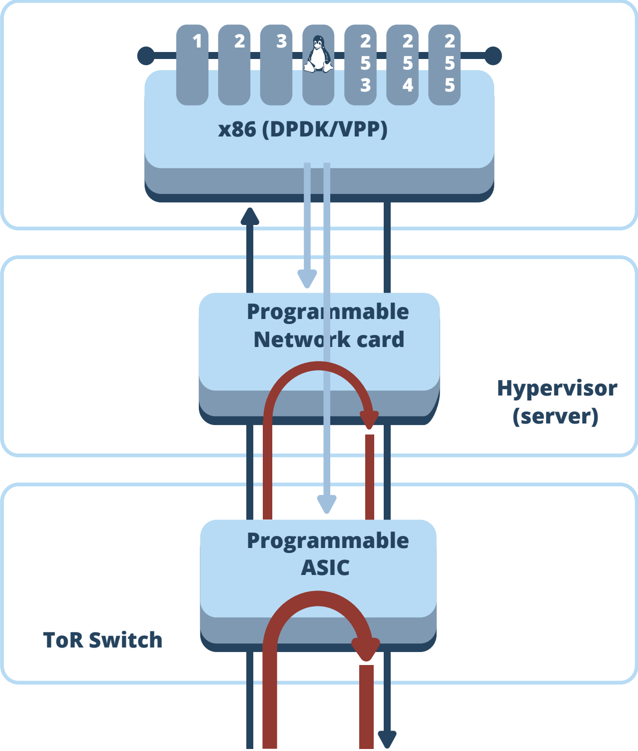


Most telco applications are traffic filters or gateways

MUCH higher throughput required!

Solution: Offload packet handling to infrastructure

SDN required to ‘remote control’ ASICs and FPGAs



Control/User Plane Separation (CUPS)

- Decompose network functions to become ‚cloud native‘

Common user plane abstraction: Linux pipeline

- baseboxd SDN controller translates Linux pipeline to whitebox
- No separate north-bound interface to controller – no SDN “apps“
- Re-use of millions of lines of Linux code

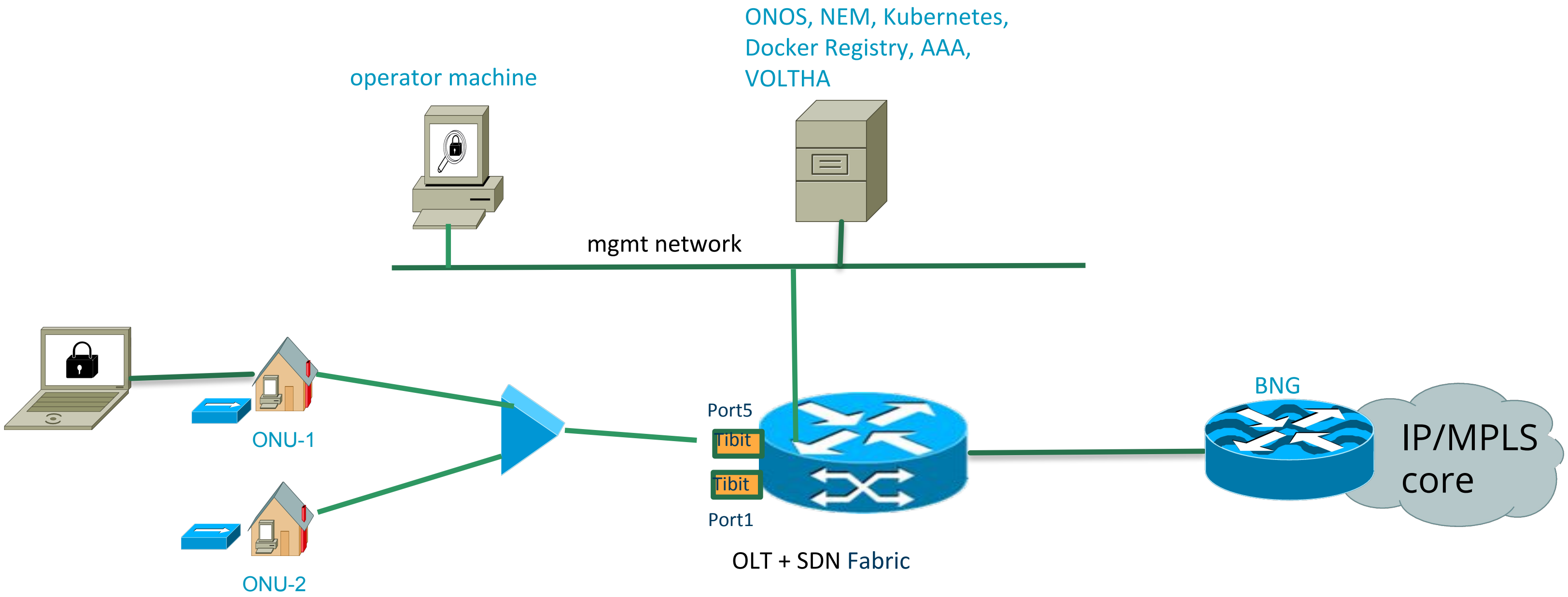
Hardware acceleration (up to factor 1.000)

- Step 1: DPDK* inside VNF – factor 10
- Step 2: Programmable NIC on server – factor 10
- Step 3: Whitebox switch – factor 10

* DPDK = Data Plane Developer Kit – Intel’s SDK for forwarding acceleration

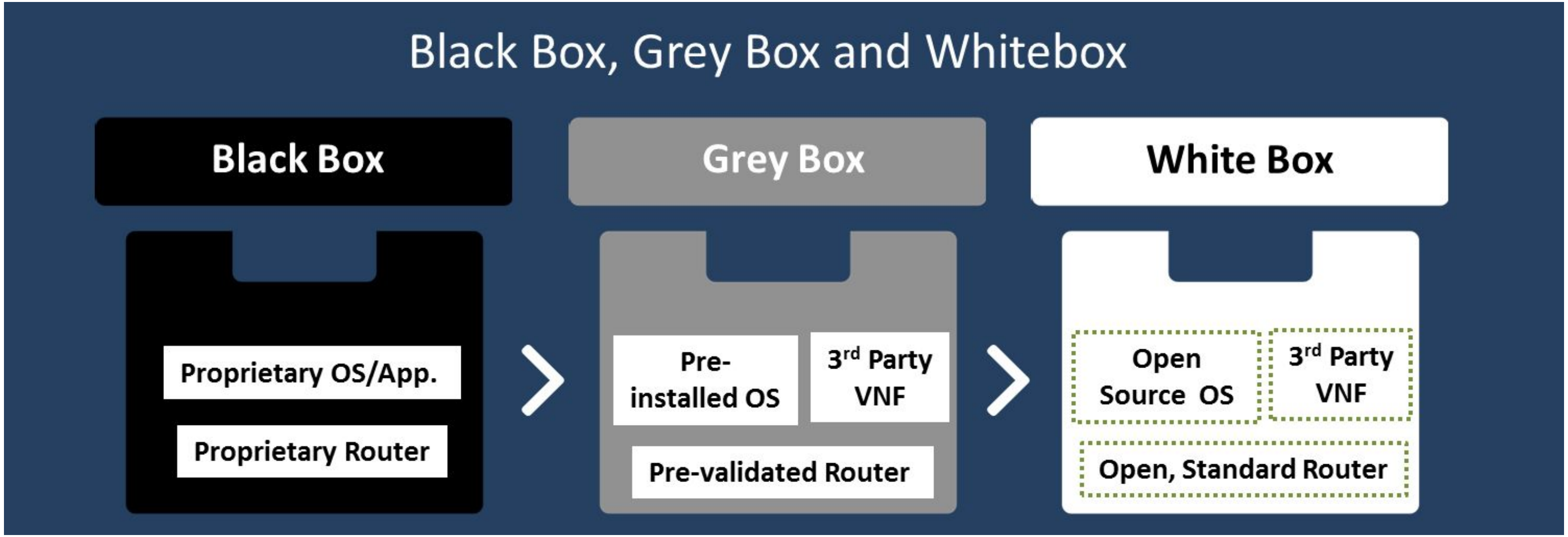
benefit from NFV *and* run high-speed

SEBA POD Topology



BISDN

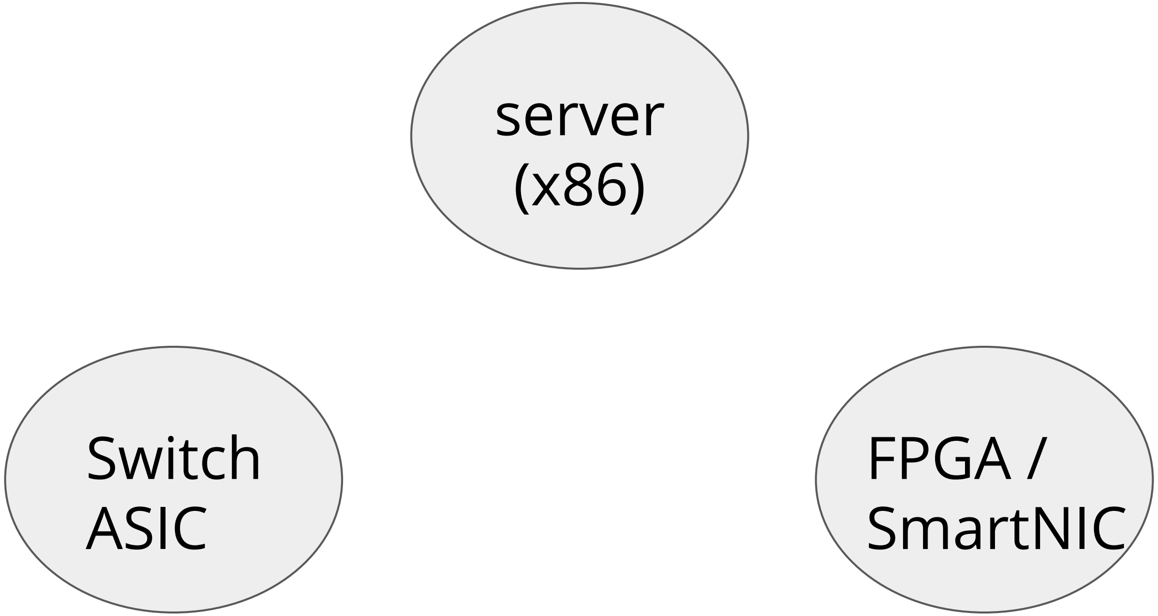
COTS brings cost down



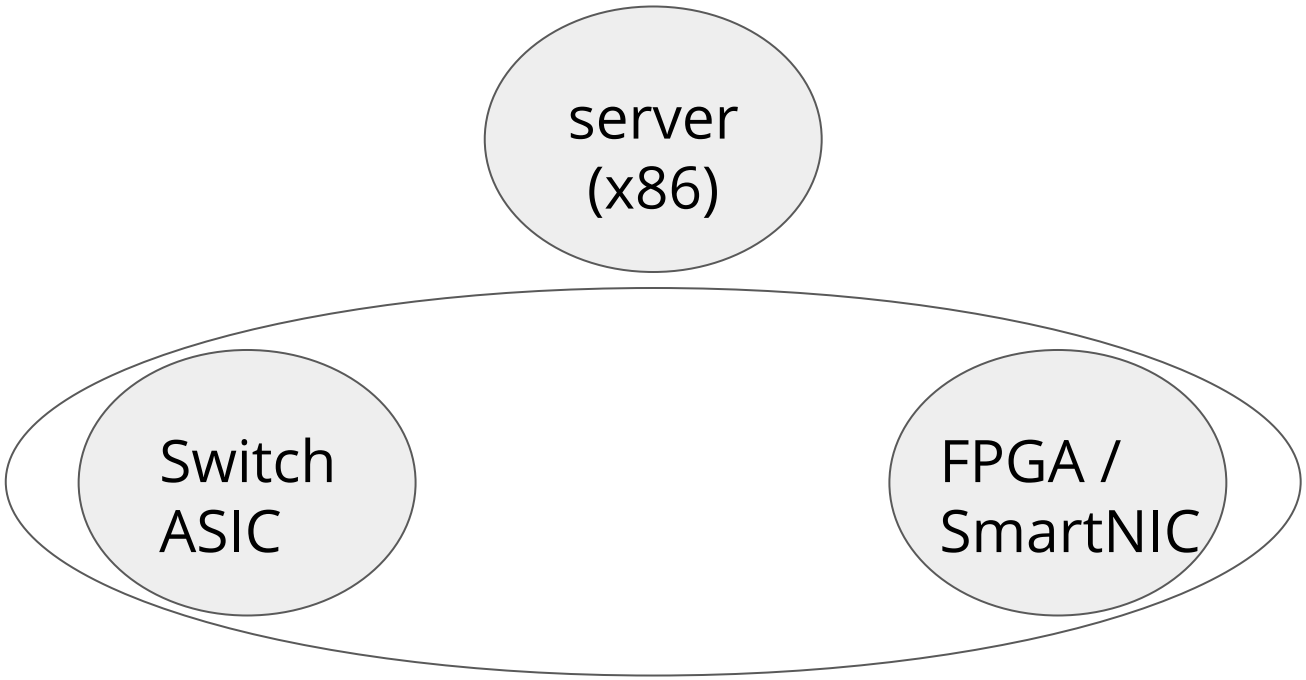
source: https://www.lanner-america.com/latest-news/network-disaggregation/

BISDN

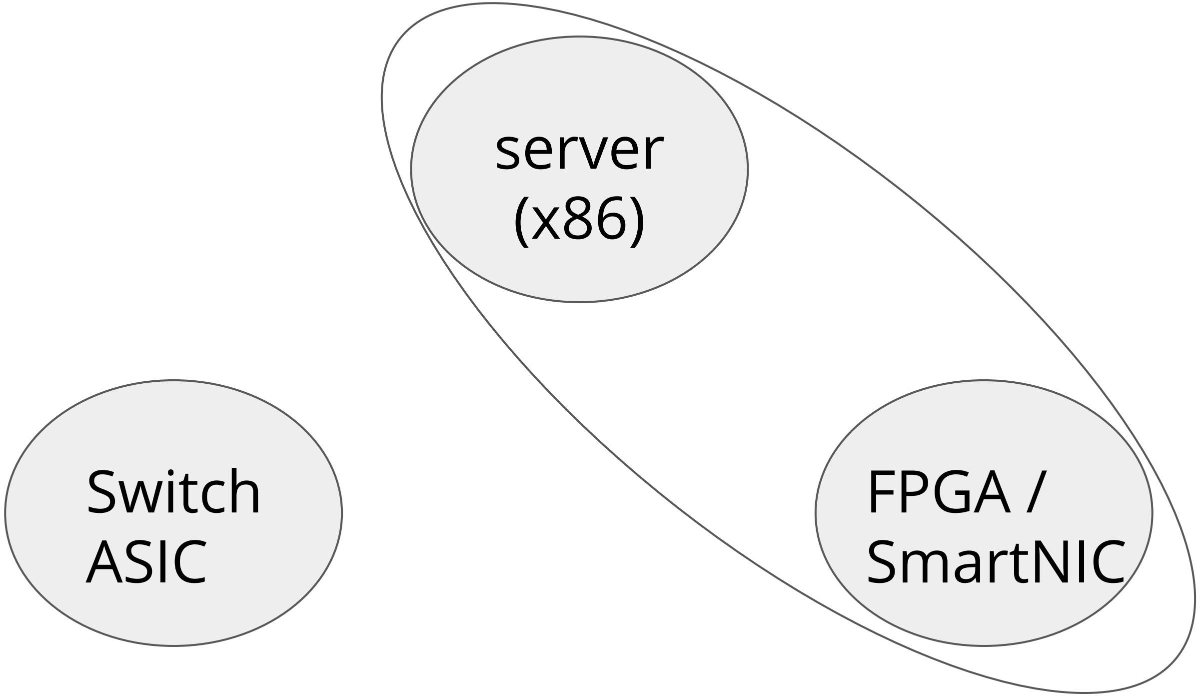
architectural options for BNG



architectural options for BNG

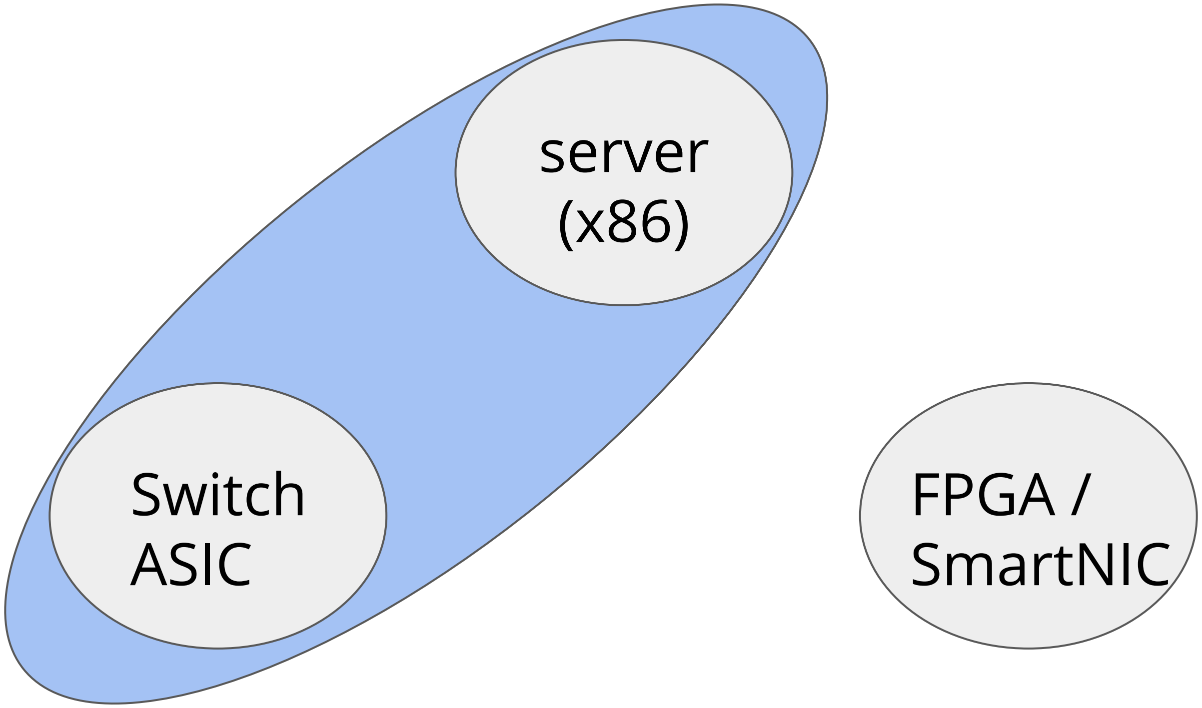


architectural options for BNG



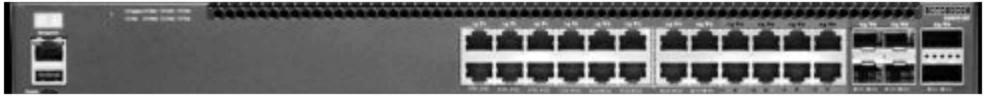
BISDN

architectural options for BNG



BISDN supports low-cost, high-volume platforms

- 1G/10G
  - AS 4610
  - Broadcom Helix4



- 10G/40G
  - AG7648
  - Broadcom Trident2

- 25G/100G
  - AG5648
  - Broadcom Tomahawk+

BISDN

not every ‘white box’ is COTS

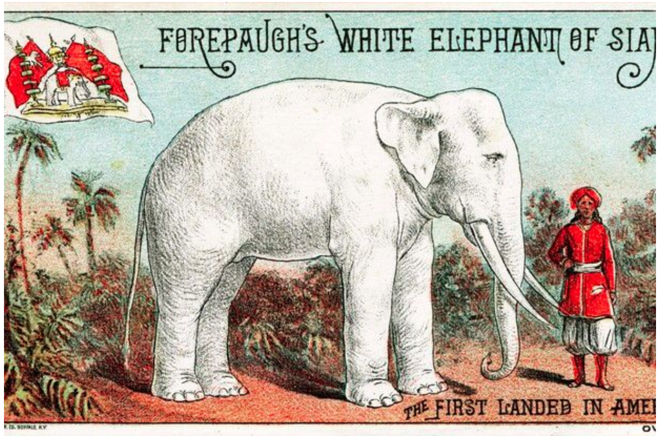
Deutsche Telekom and Delta announced Open BNG platform at OCP Summit 2019

http://www.delta-emea.com/news/pressDetail.aspx?secID=3&pID=1&typeID=1;2;8&itemID=9565&tid=0&hl=en-GB

AGCX422S
22*QSFP28 and 4*QSFP-DD

AGCVA48S
4*SFP+, 48*SFP28 and 10*QSFP28

- ➢ Dual Q2C,
- ➢ additional counter processor
- ➢ Xeon-D inside

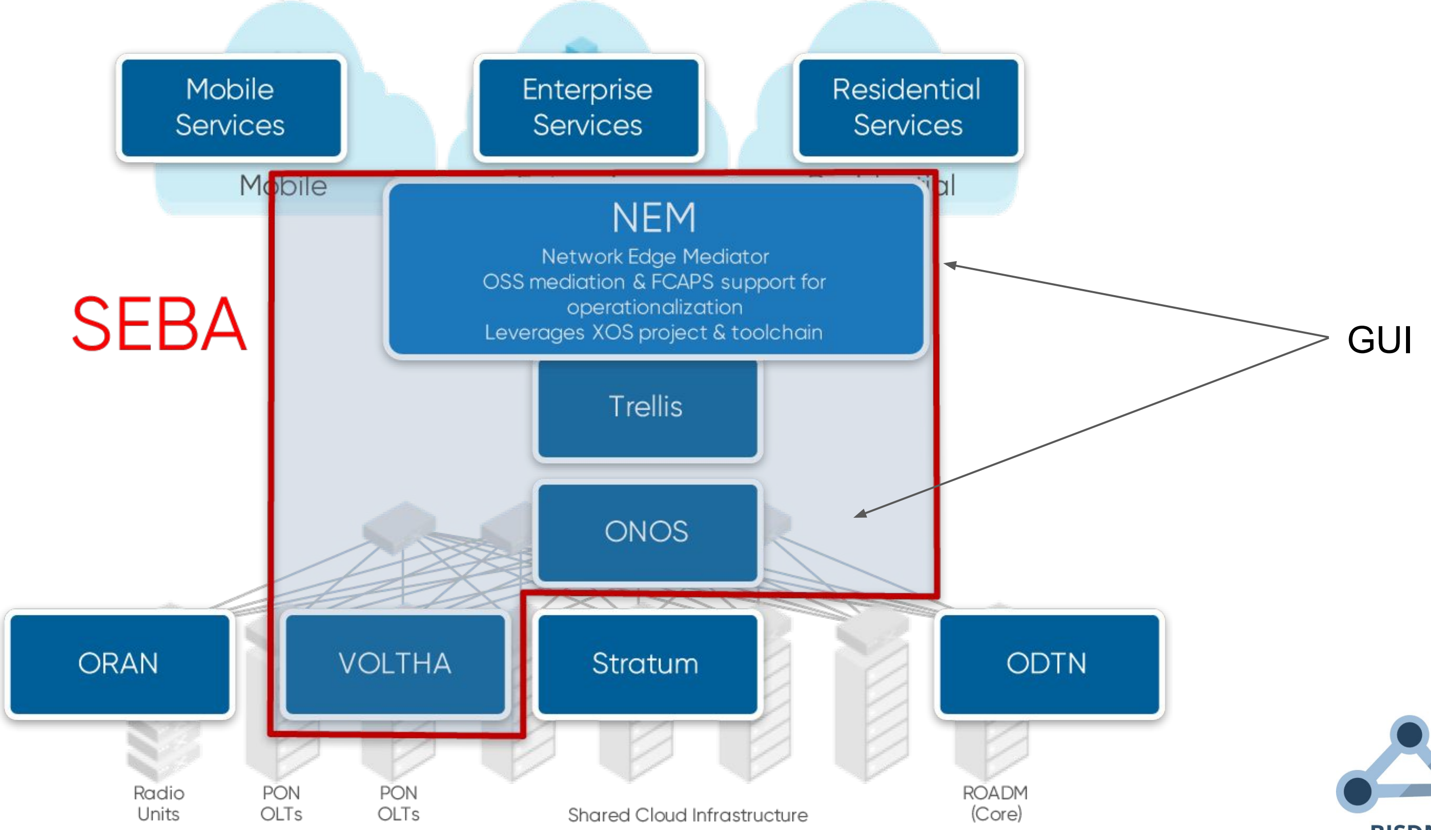


custom designs won’t fly **unless usable for mass market**.

(personal opinion of Dr.-Ing. Hagen Woesner, Berlin)

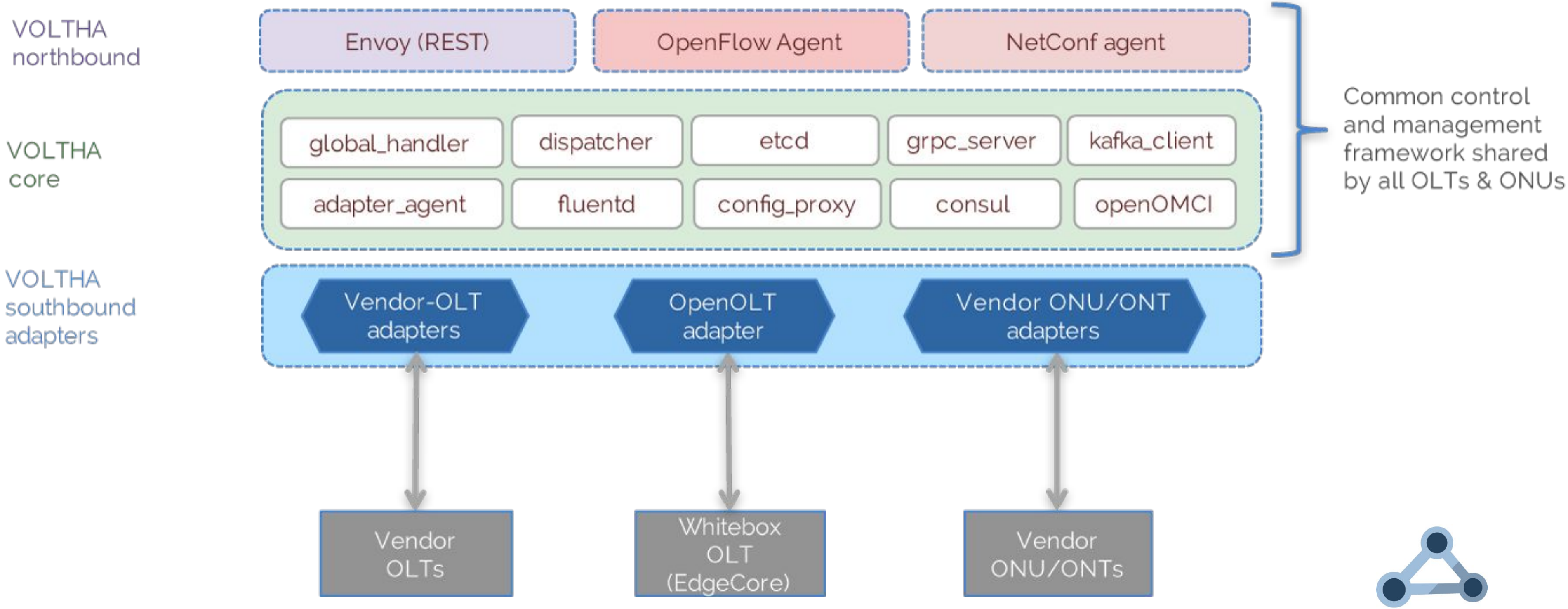
BISDN

SEBA Overview - need to integrate into OSS/BSS



BISDN

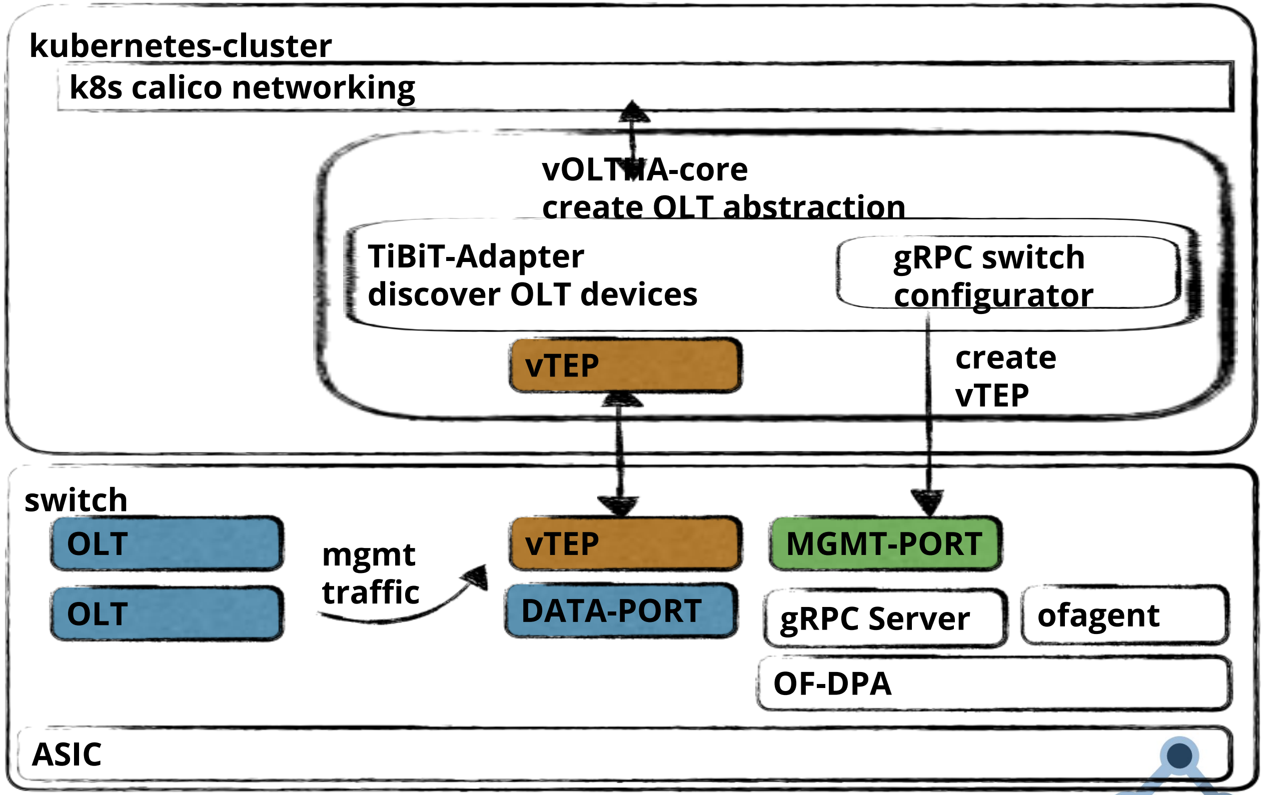
vOLTHA Architecture



BISDN

Schematic of grpc/OF coexistence in k8s

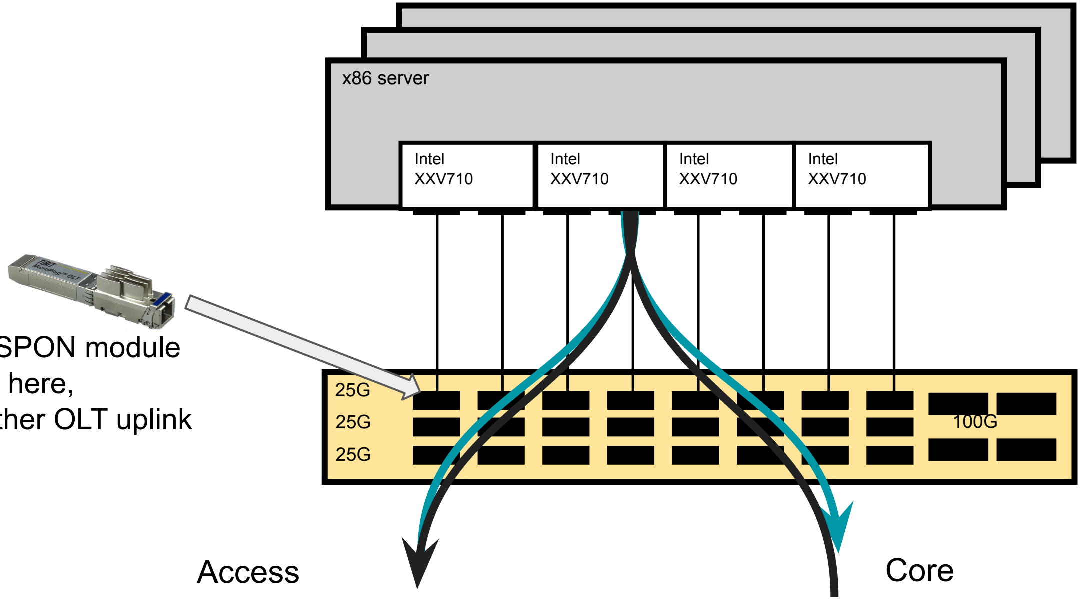
- currently working on vOLTHA 1.6
- control plane extracted into VXLAN port
- transport via data port (not the internal CPU)



BISDN

COTS BNG architecture - switch+server

Extend switch pipeline into server - balance the tasks to where they can be handled best



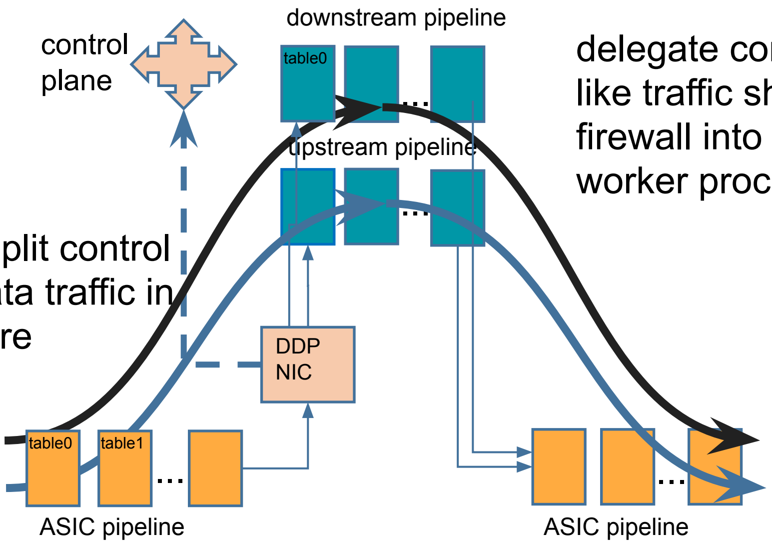
Tibit XGSPON module could go here, or any other OLT uplink

SuperMicro X11DPI-N, 2 x CPU Intel Xeon Gold 5120 (Skylake)
w/ 4 Intel XXV710DA2 cards

Delta AG5648v1 (BRCM Tomahawk+)
BISDN Linux, baseboxd controller for L2 access and L3 core routing to spine

BISDN

Packets do not leave hardware pipeline



delegate complex tasks like traffic shaping, NAT, firewall into DPDK worker processes

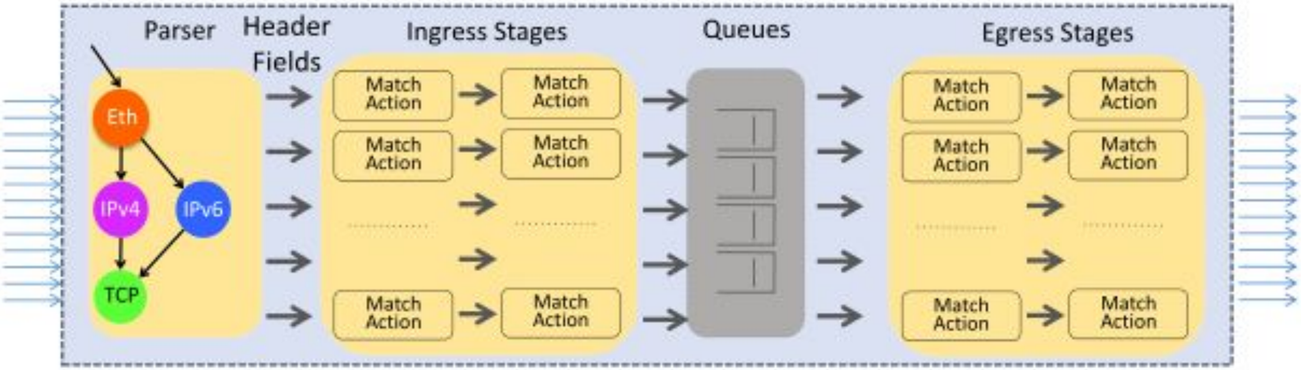
DDP: Split control from data traffic in hardware

keep simple tasks like fan-out, routing, MPLS, in whitebox ASIC

BISDN

P4 pipeline / folded

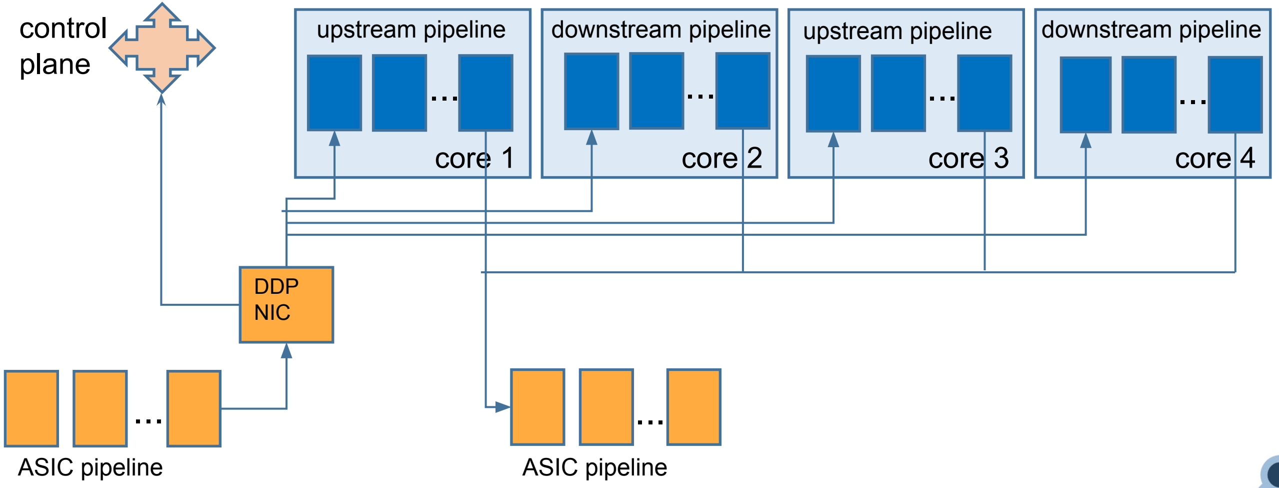
- similarity in the pipeline structure
- both ingress and egress stage are handled in the same switch
- control interface is PFCP (NOT p4runtime), but similar paradigm
- use local OpenFlow controller for switch (github.com/bisdn/basebox)



BISDN

horizontal scaling

- scale to 16 parallel BNG instances on a single server
  - 2 instances per 25G port

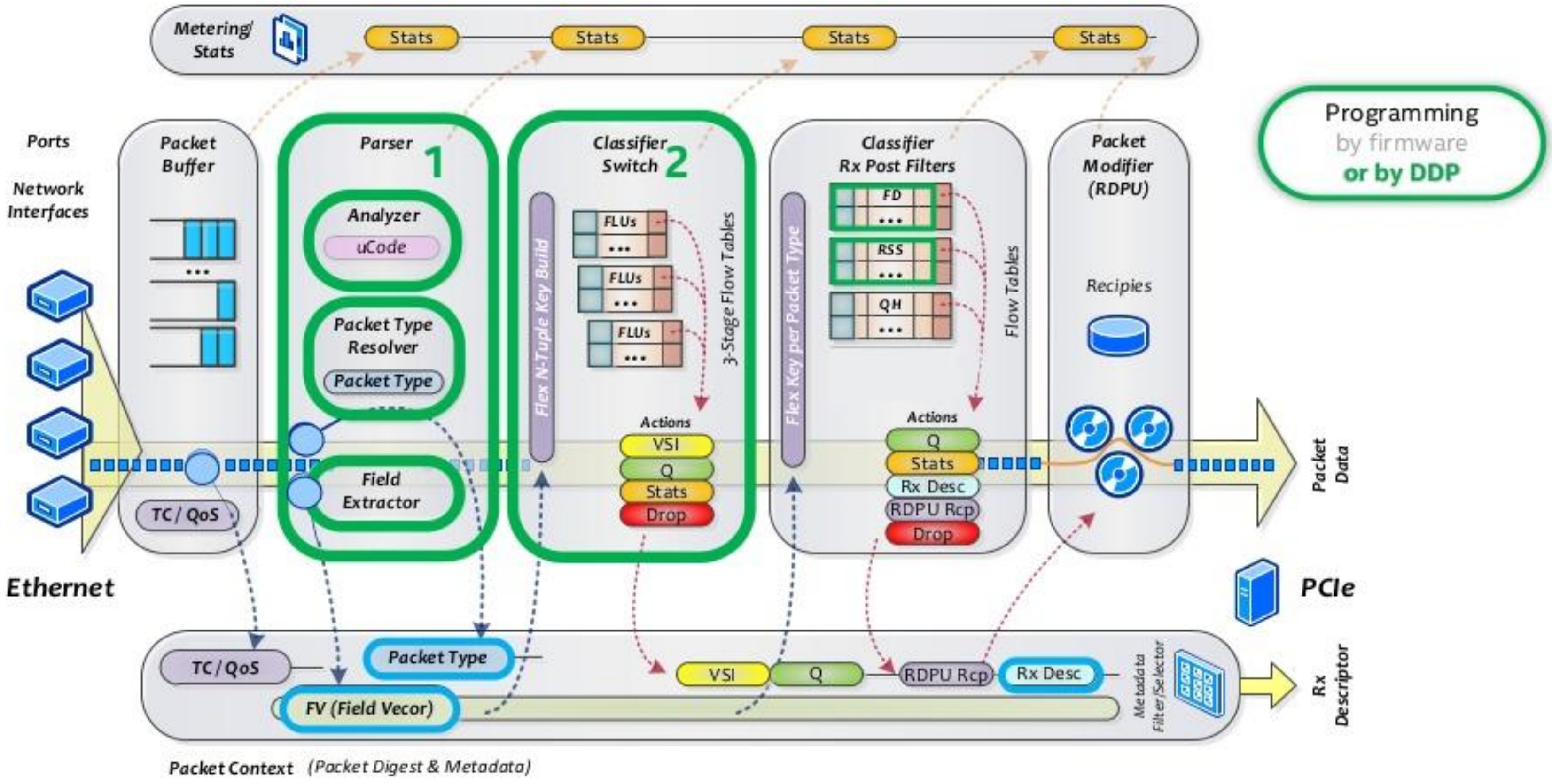


Dynamic Device Personalization - DDP

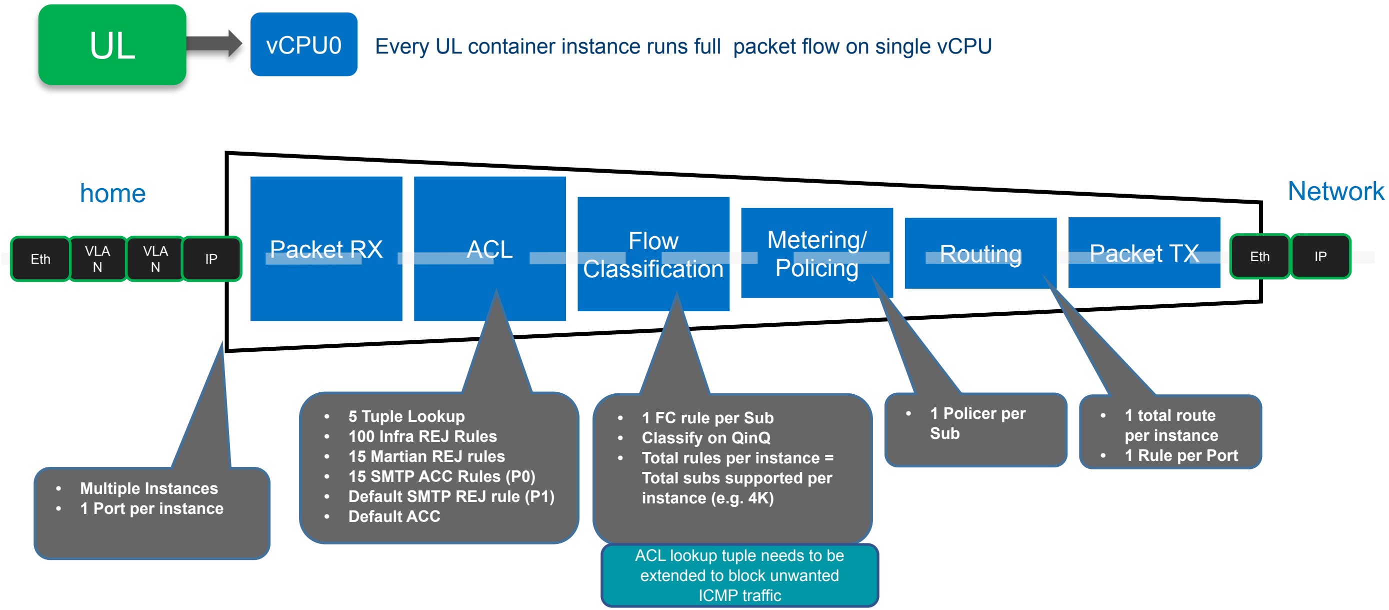
- program filters into NIC pipeline, profiles available for PPPoE, IPSec(?), GTP
  - cloud filters can be set/programmed with ethtool
- Here we use some specific filter config (made by Intel)
  - create 5 VFs
  - VF_0 is 'control VF', bind to kernel driver, on to VXLAN (remote control server)
  - VF_1, VF_3 downlink
  - VF_2, VF_4 uplink
- initialize NIC such that PPPoE control traffic is sent to CP VFs

BISDN

Intel® Ethernet 700 Series Rx Programmable Pipeline

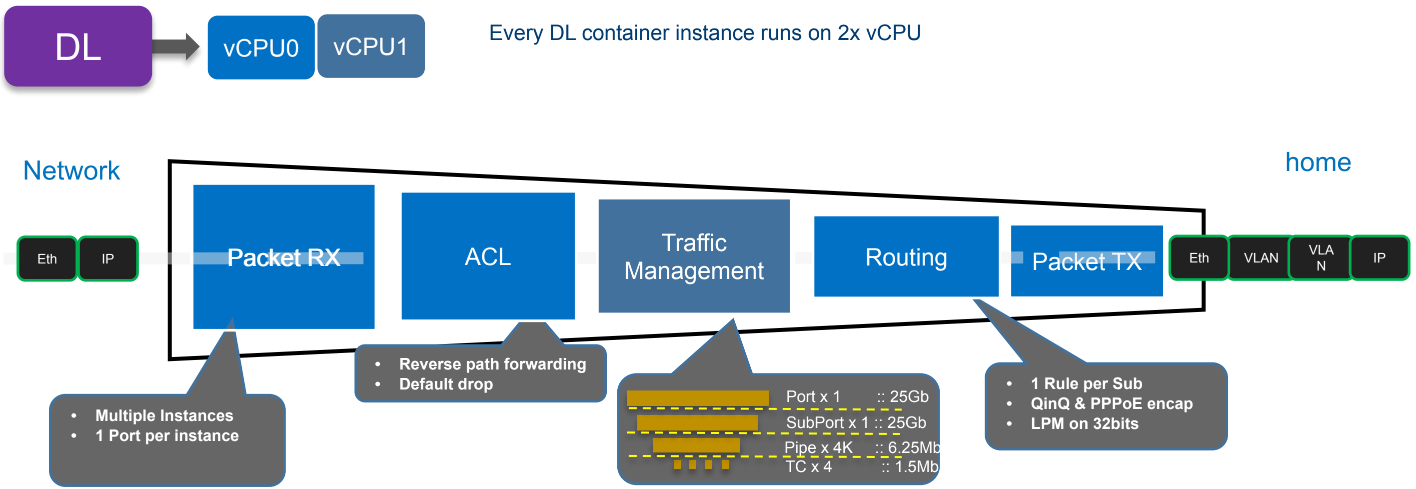


vBNG Base Config - Uplink Packet Flow stages

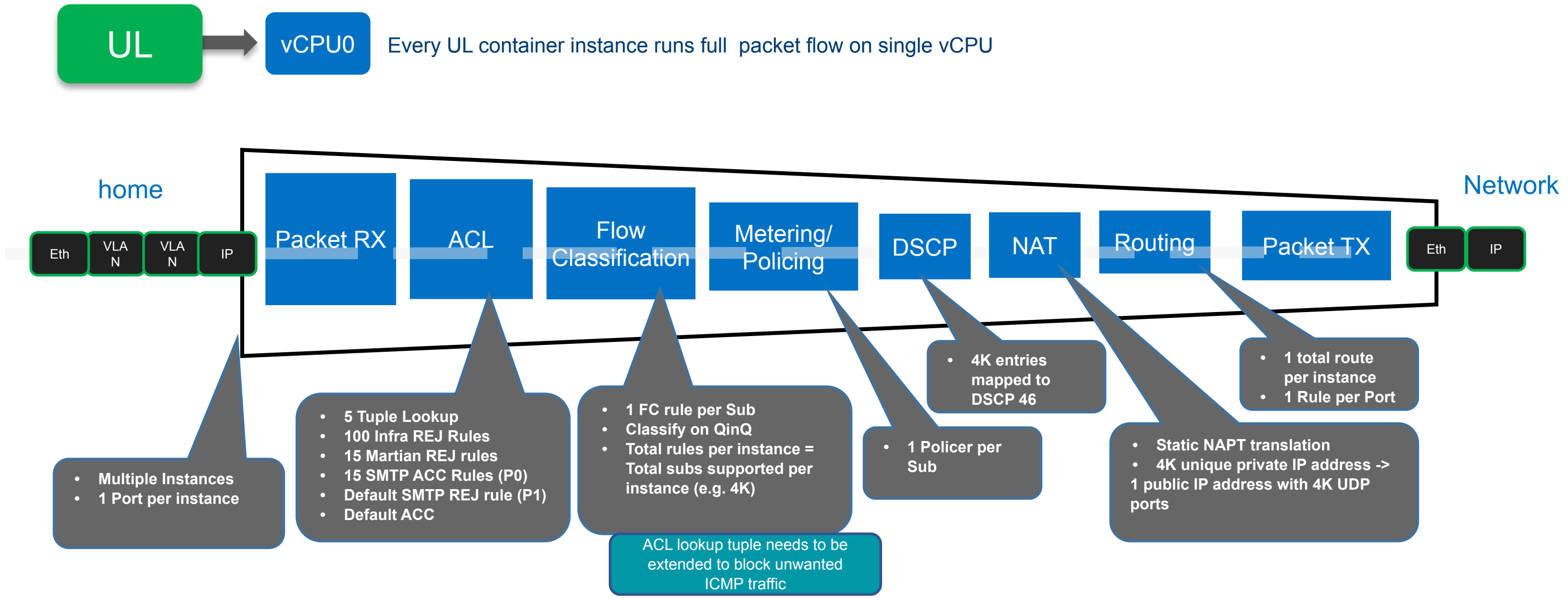


BISDN

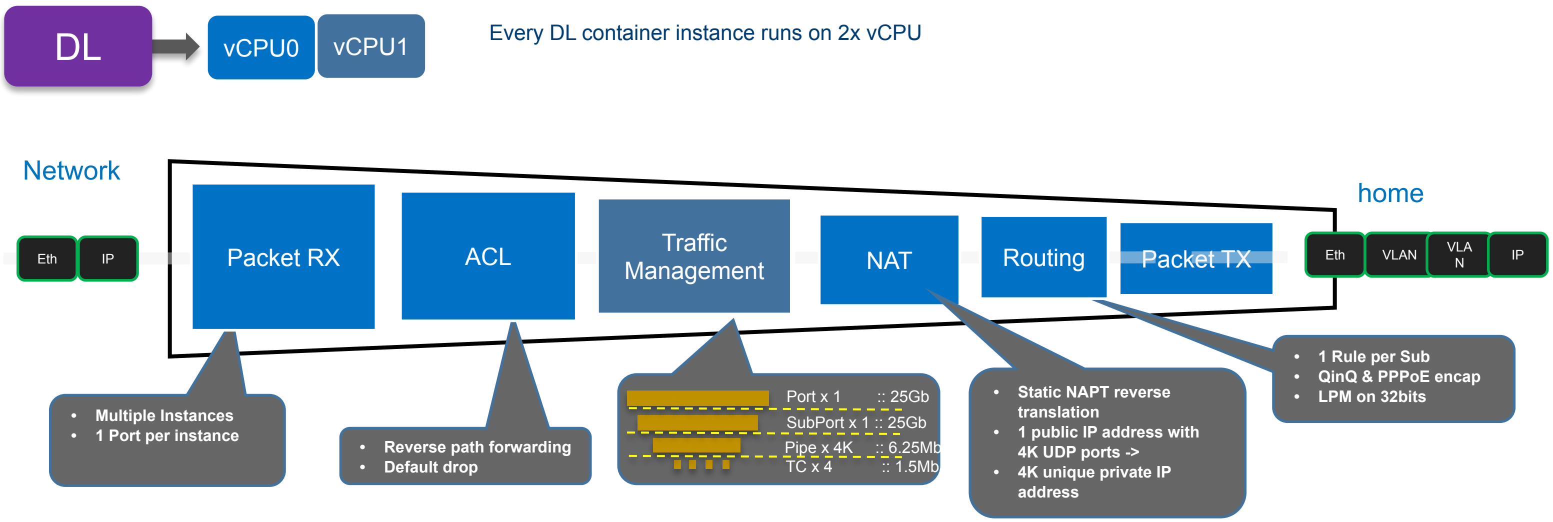
vBNG Base Config- Downlink packet flow stages



vBNG Full Config - Uplink Packet Flow stages



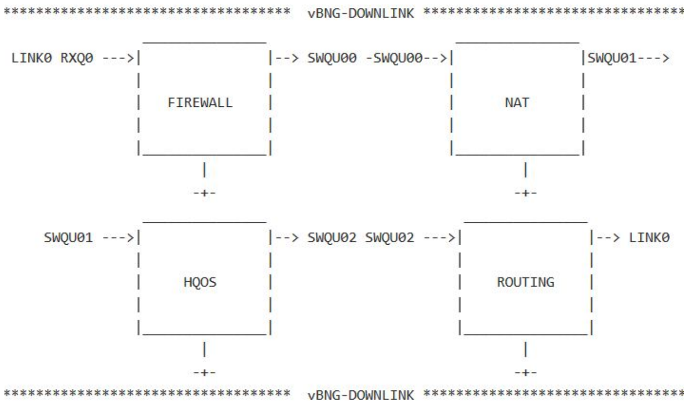
vBNG Full Config- Downlink packet flow stages



BISDN

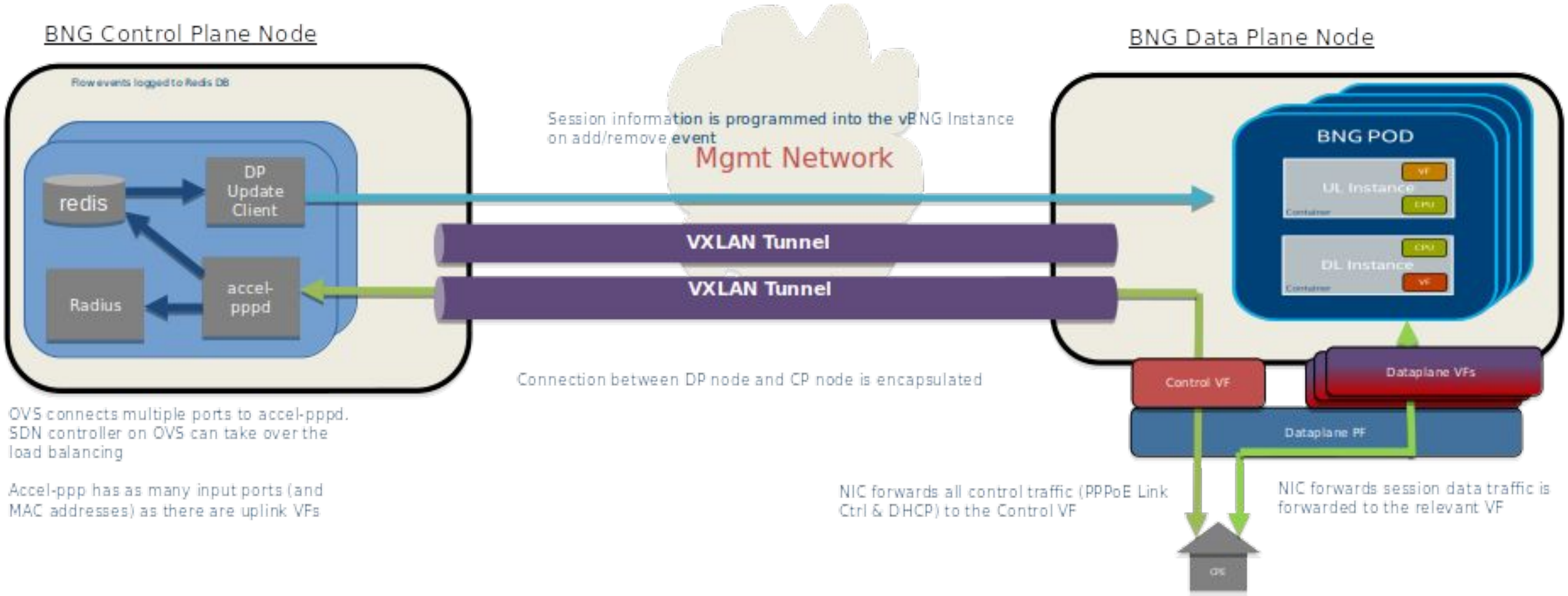
Example IP pipeline code

```
;**************DECLARE ACTION PROFILE********
table action profile AP5 ipv4 offset 270 fwd encap qinq_pppoe
;**************PB'S CREATION****************
pipeline downstream|firewall period 10 offset_port_id 0 cpu 0
;**************PB LINKING*******************
pipeline downstream|routing port in bsz 32 swq SWQU02
;**************TABLE CREATION****************
pipeline downstream|routing table match lpm ipv4 offset 286 size 4K action AP5
;***********TABLE ASSOCIATION****************
pipeline downstream|routing port in 0 table 0
;********************************PIPELINE BLOCKS********************************;
pipeline downstream|routing table 0 rule add match default action fwd drop
pipeline downstream|routing table 0 rule add bulk
./bng_configs/bulk_rules/dl_bng/route_bulk_pppoe_ins_%S5.txt
```



BISDN

vBNG Control Plane attachment



Assume that Dataplane VF MAC addresses are replicated at accel-pppd’s input ports
Static mapping, input port determines the DP instances to be called from python script

Packet Forwarding Control Protocol (PFCP)

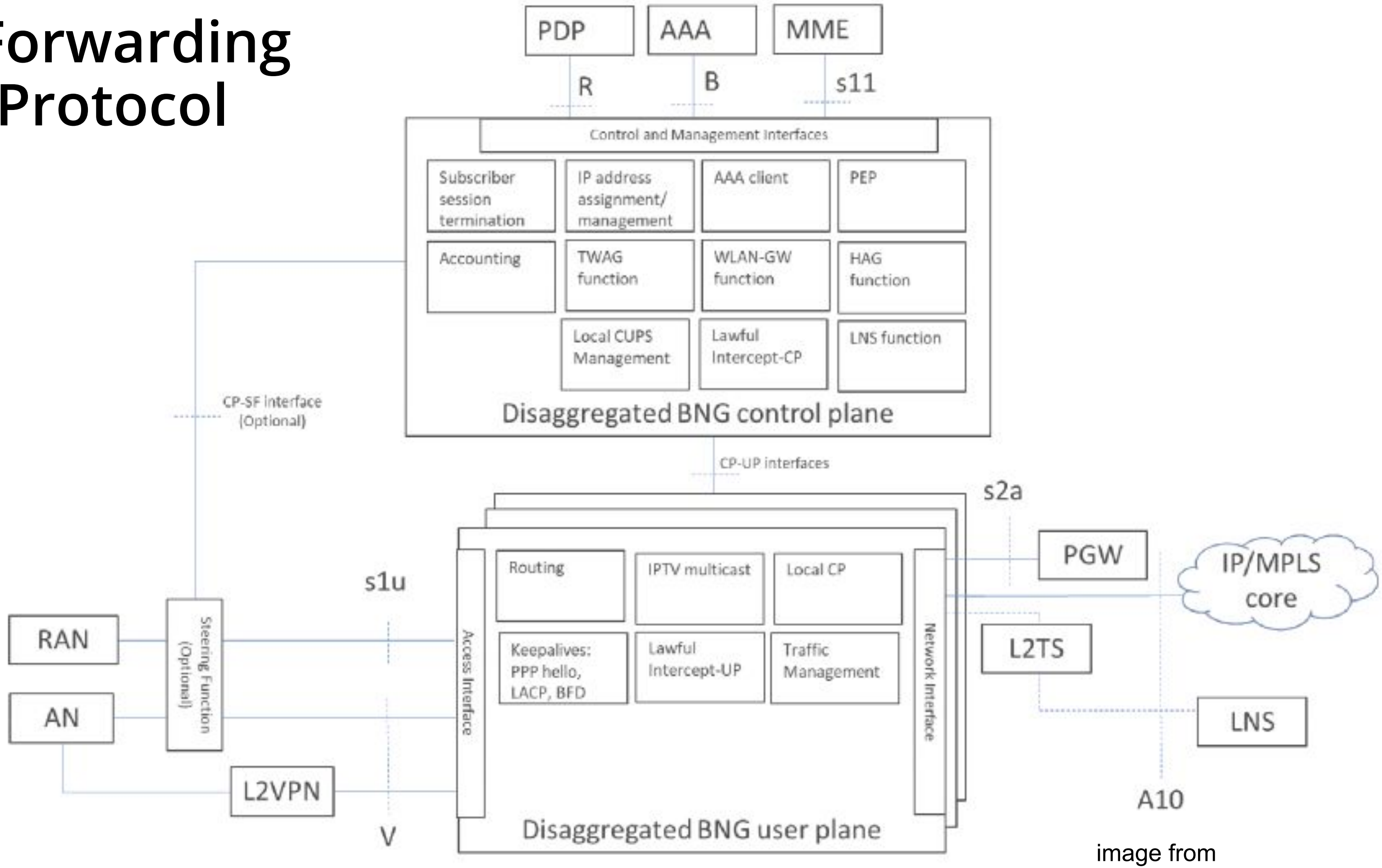
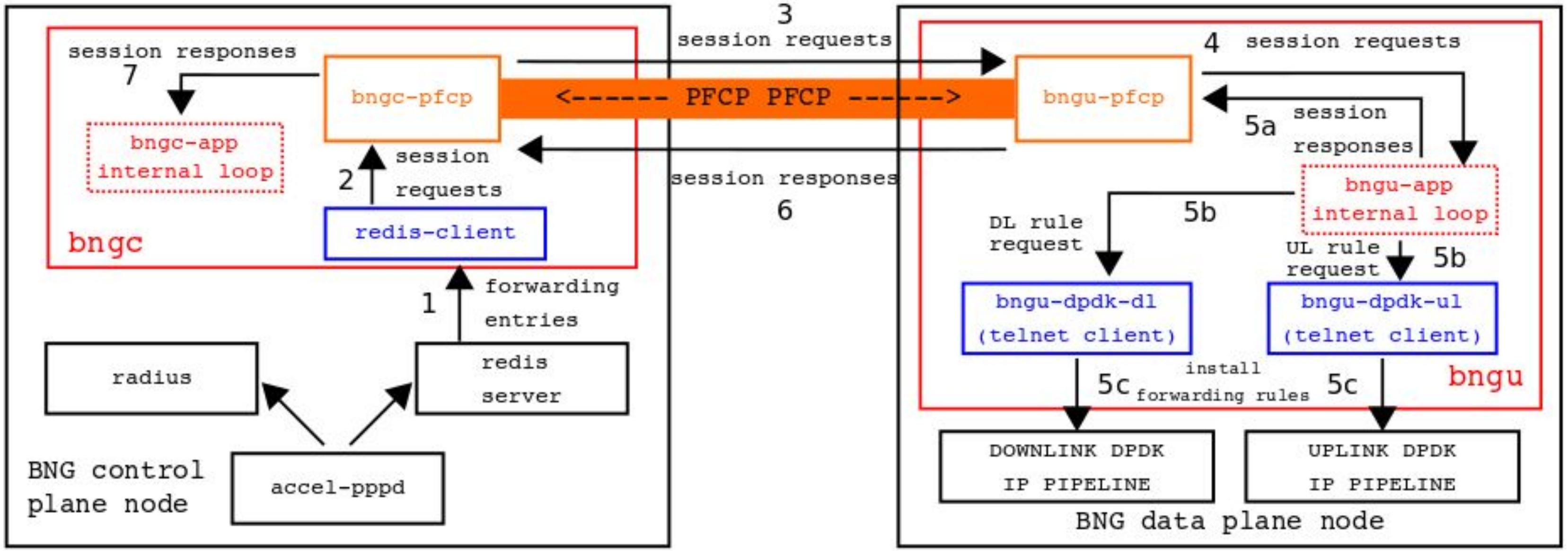


image from
WT-459 -10SB, page 24

BNG with PFCP:



BISDN

Measurement setup
using a Spirent TestCenter 5.01 at DT

- 2*100Gbit/s Spirent ports attached to single server
  - SuperMicro X11DPI-N, 2xCPU Intel Xeon Gold 5120
    - 14 core @ 2.2GHz
  - Delta Agema AG5648 (BRCM Tomahawk) does routing and switching
    - split to 8*25G Intel XXV710DA2 cards
    - baseboxd controller on switch
- bidirectional traffic access+core in one port
- asymmetric upstream to downstream traffic ratio
  - upstream between 10 and 25% of downstream
- session setup speed limited by DPDK pipeline implementation
  - currently ~10 sessions/s per instance, will be fixed

BISDN

Queue dimensioning

identified three bottlenecks:

- DPDK IP Pipeline
- Fortville ASIC on NIC
- and then “something somewhere” in PCI/CPU, practical limitations of PCI bandwidth to CPU.

main focus on downstream pipeline.

upstream traffic in general much lower than downstream, and no HQoS (we did measure upstream, too.)

BISDN

Characterization of IP pipeline

single vBNG instance consists of two containers

- **upstream**: one hyperthreaded core
- **downstream**: two hyperthreaded cores
  - siblings on one CPU core, one for HQoS only, the other for all the rest

DDP allows splitting the traffic into more than one instance

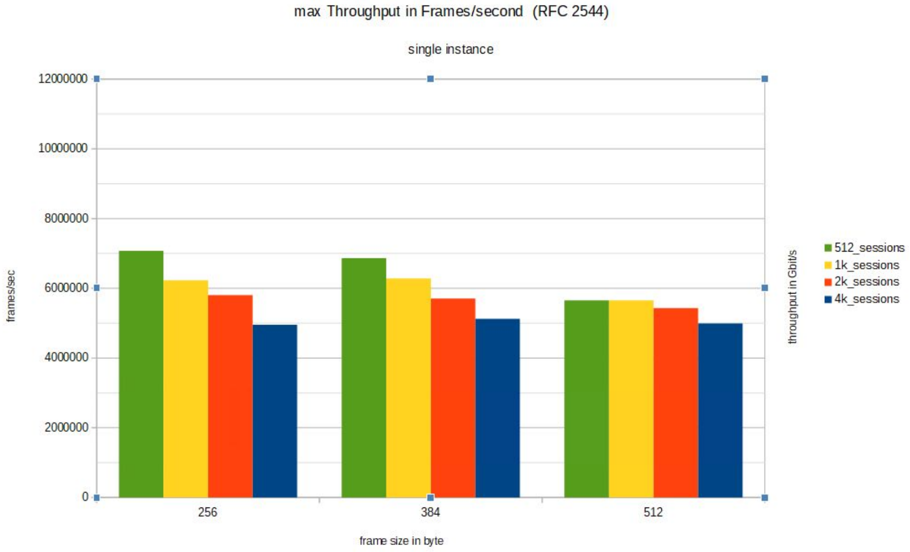
- single instance cannot quite serve a full 25 GbE port
  - this depends on the CPU, we used a 2.2 GHz CPU from 2017

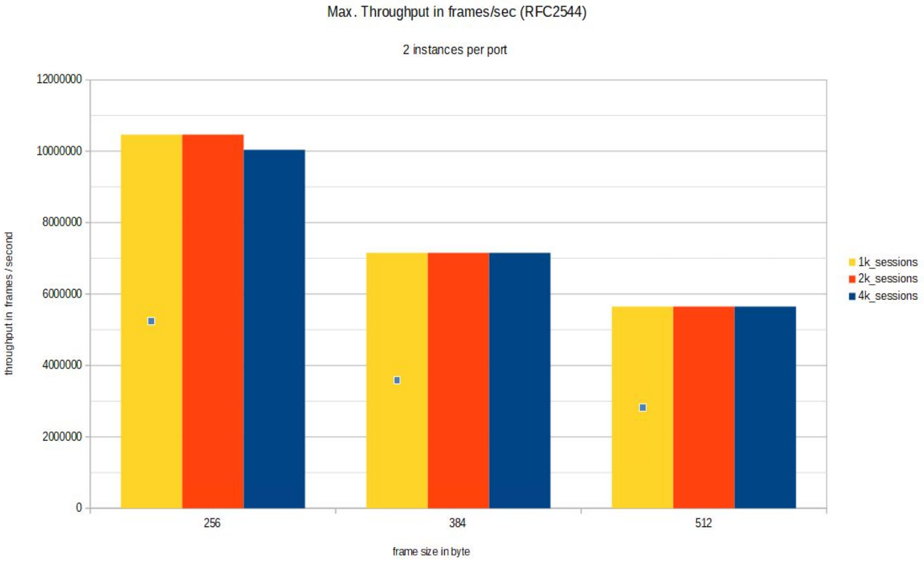
... just take two.

DPDK pipeline measurements

throughput test (RFC 2544) - achievable throughput with zero packet loss

- for 4K sessions, single instance can process 5Mpps. ⇒ 11.1 Gbit/s @256 byte
- two instances in the same port process 10Mpps. ⇒ 22.2 Gbit/s @256 byte
- dominating factor for pipeline processing is number of sessions (lookup time for routing and HQoS)

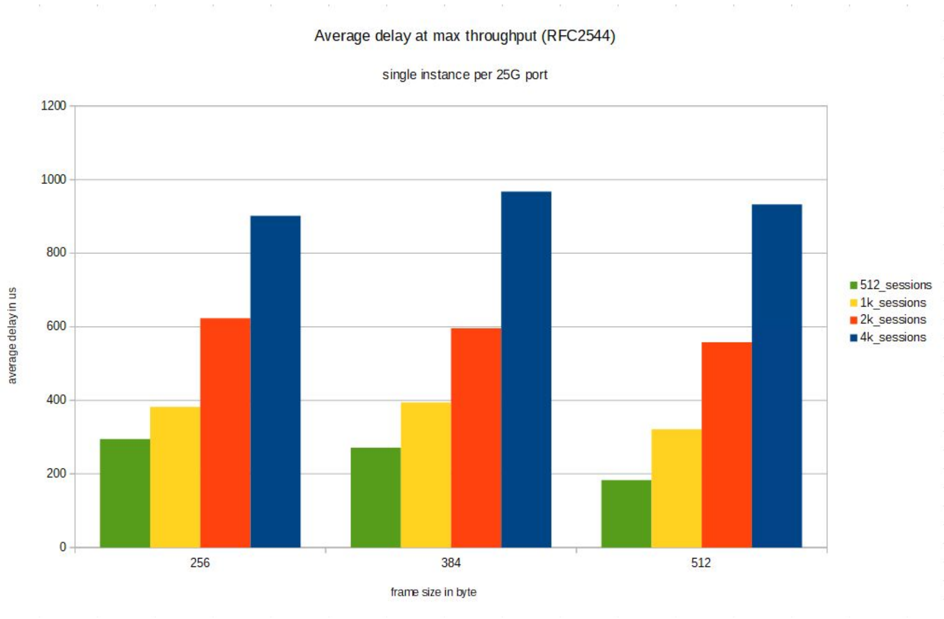


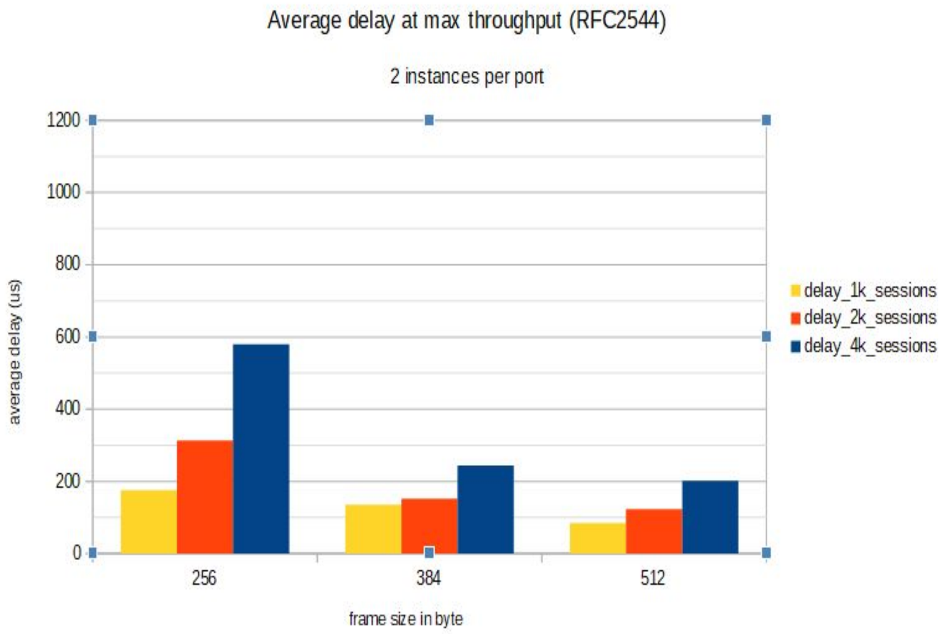


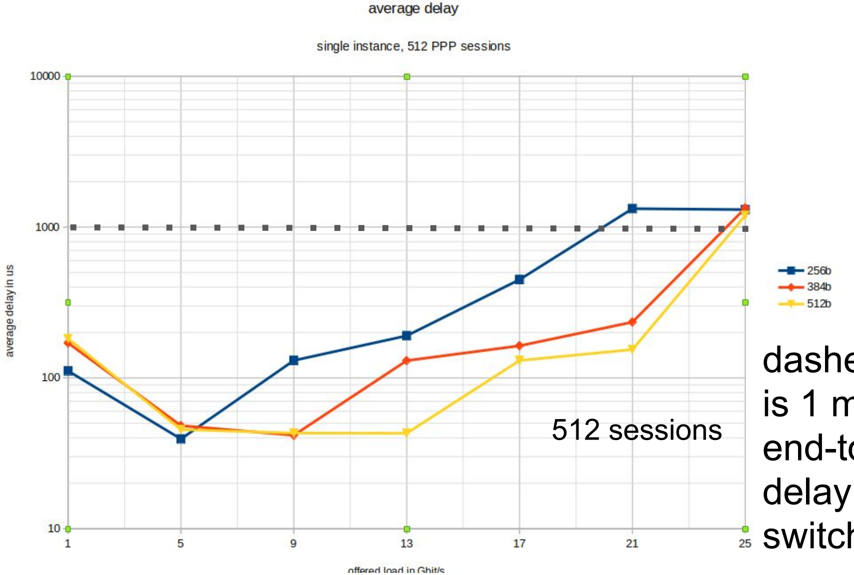
DPDK pipeline measurements
throughput test (RFC 2544), average delay at peak load

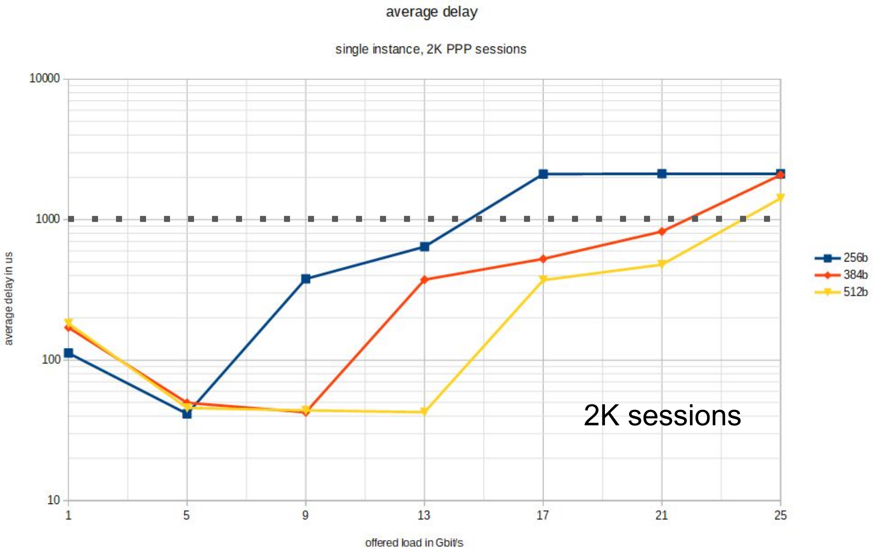
average delay stays below 1ms

max delay is typically twice the average, min delay around 30 microseconds



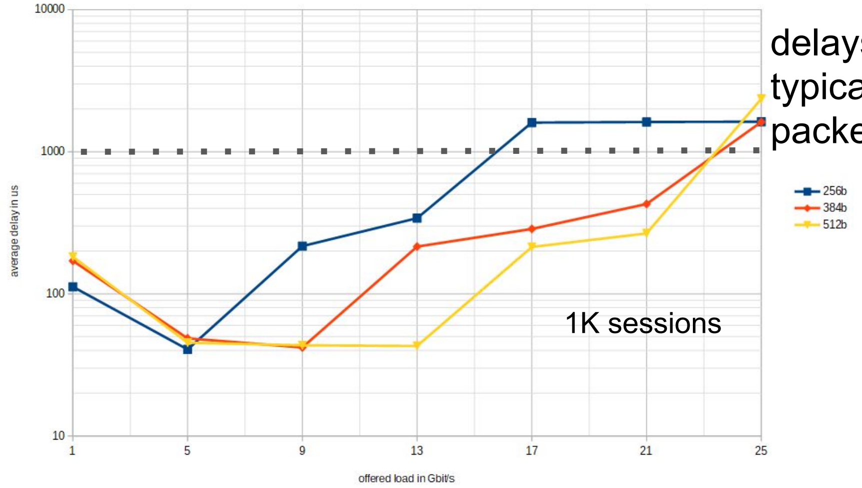


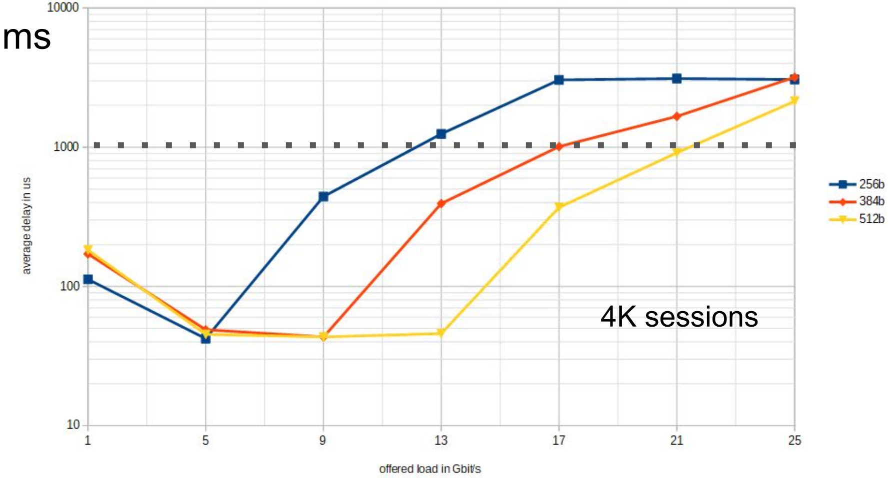




dashed line is 1 ms end-to-end delay (incl. switch)

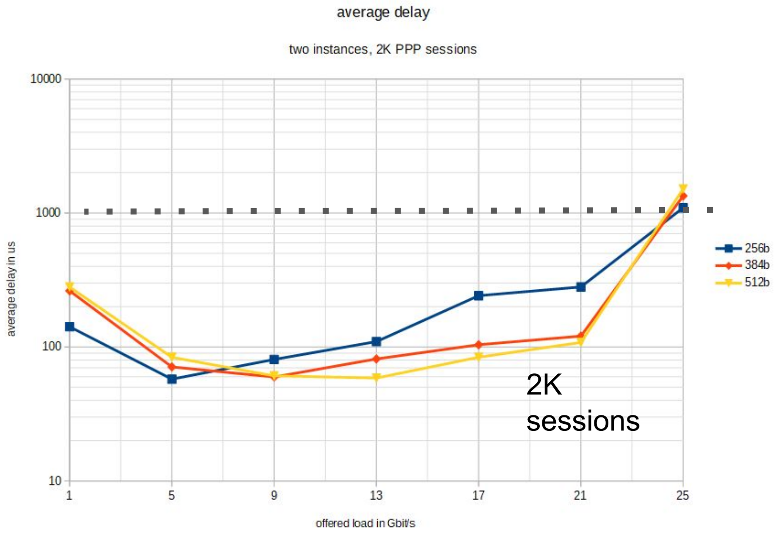
delays above 1ms typically mean packet loss

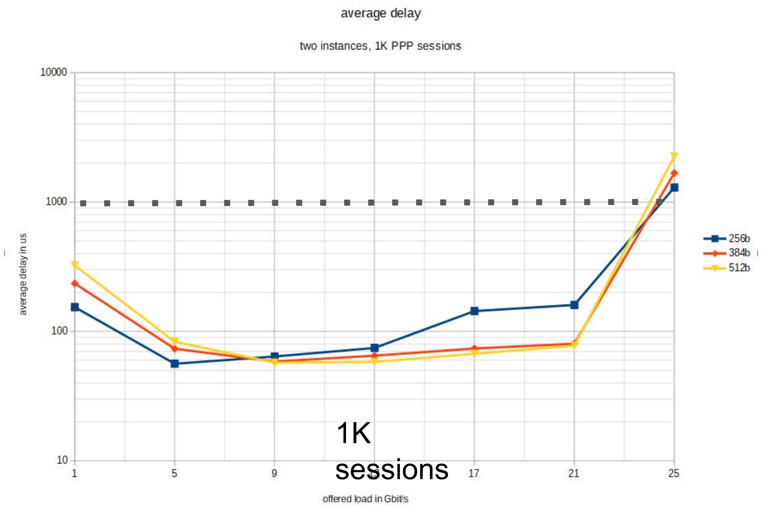


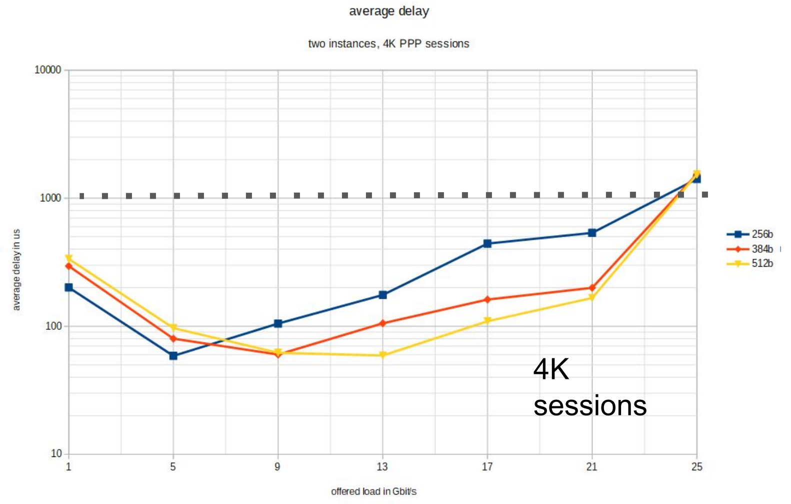


For 2 instances
… typical delay around 70 μs

- delay curves are flat until very high load is applied
- switch traversal is 3.8 μs, i.e. a total of 7.6 μs is included that is caused by the switch
    - similar number for Mellanox and Broadcom







port vs. card limitations

single port can serve full line rate of a dual port XXV710 card

dual-port card is limited to ~20 Mpps

both ports can be filled up to 45-46 Gbit/s in total

above that, packet losses “before” the pipeline

BISDN

single card to multiple cards

multiple cards on the same socket are not completely independent **in our server (!)**

loss-free operation until 80% offered load.

multiple sockets are indeed completely independent

total achievable loss-free throughput is 160Gbit/s

combined up+downstream traffic

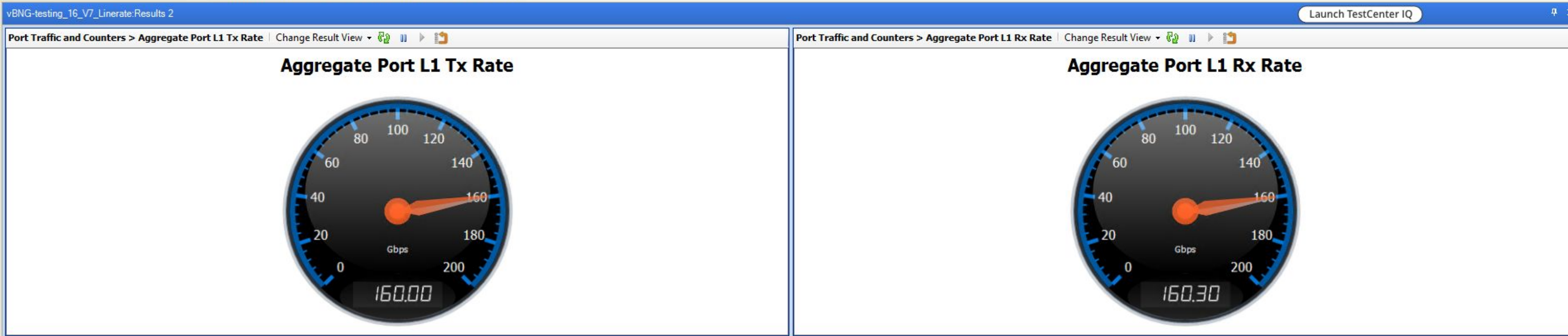
BISDN

160 Gbit/s total throughput, lossless

8 ports, 2 instances per port, 16 G down/ 4G upstream traffic

Down 1a	Click to ad...	generator	Device core 1a (211.1.0.1...	Device 1a (192.51.5.184.../1...	Pair	Port	Port //11/1	Port //11/1,	Ready	512	8.000	Mbps	Fixed	512
Down 1b	Click to ad...	generator	Device core 1b (211.2.0.1...	Device 1b (192.51.22.205.../...	Pair	Port	Port //11/1	Port //11/1,	Ready	512	8.000	Mbps	Fixed	512
Down 2a	Click to ad...	generator	Device core 2a (212.1.0.1...	Device 2a (192.51.45.237.../...	Pair	Port	Port //11/1	Port //11/1,	Ready	512	8.000	Mbps	Fixed	512
Down 2b	Click to ad...	generator	Device core 2b (212.2.0.1...	Device 2b (192.51.50.228.../...	Pair	Port	Port //11/1	Port //11/1,	Ready	512	8.000	Mbps	Fixed	512
Down 3a	Click to ad...	generator	Device core 3a (213.1.0.1...	Device 3a (192.51.78.30.../1...	Pair	Port	Port //11/1	Port //11/1,	Ready	512	8.000	Mbps	Fixed	512
Down 3b	Click to ad...	generator	Device core 3b (213.2.0.1...	Device 3b (192.51.90.216.../...	Pair	Port	Port //11/1	Port //11/1,	Ready	512	8.000	Mbps	Fixed	512
Down 4a	Click to ad...	generator	Device core 4a (214.1.0.1...	Device 4a (192.51.110.235....	Pair	Port	Port //11/1	Port //11/1,	Ready	512	8.000	Mbps	Fixed	512
Down 4b	Click to ad...	generator	Device core 4b (214.2.0.1...	Device 4b (192.51.122.85.../...	Pair	Port	Port //11/1	Port //11/1,	Ready	512	8.000	Mbps	Fixed	512
Up 1a	Click to ad...	generator	Device 1a (192.51.5.184....	Device core 1a (211.1.0.1/16)	Pair	Port	Port //11/1	Port //11/1,	Ready	512	2.000	Mbps	Fixed	128
Up 1b	Click to ad...	generator	Device 1b (192.51.22.205...	Device core 1b (211.2.0.1/16)	Pair	Port	Port //11/1	Port //11/1,	Ready	512	2.000	Mbps	Fixed	128
Up 2a	Click to ad...	generator	Device 2a (192.51.45.237....	Device core 2a (212.1.0.1/16)	Pair	Port	Port //11/1	Port //11/1,	Ready	512	2.000	Mbps	Fixed	128
Up 2b	Click to ad...	generator	Device 2b (192.51.50.228...	Device core 2b (212.2.0.1/16)	Pair	Port	Port //11/1	Port //11/1,	Ready	512	2.000	Mbps	Fixed	128

Displaying Stream Blocks 1 - 16 | Total Stream Blocks: 16 | Selected 1 of 16



vBNG x86+whitebox implementation
Lessons learned

DDP allows forwarding of traffic into separate vBNG instances

- no hashing, no cache pollution, 4K entries fit into cache

delay is generally low, but moreover: **manageable**

- if delay is too high ⇒ add a pipeline

processing of 160 Gbit/s of traffic is possible on a commodity (low end) server

there are bottlenecks outside of the pipeline

- increasing the Rx/Tx queue size to 4K and mempool to 320K helped a lot

BISDN

Conclusions

- COTS is bringing cost of access equipment down
- Linux as common API over switches and servers
- stateful functions like OLTs can be separated from forwarding
- stateful functions like PPPoE session termination, HQoS, NAT should be separated from forwarding
- whiteboxes + pluggable OLTs + programmable NICs allow this

BISDN