The difference in network equipment in the 25/100/400G era and how to test/break them

> Tristan Suerink
IT Architect

Nikhef
Nationaal instituut voor subatomaire fysica

V The past

Aspen
FOUNDRY NETWORKS
BigIron 4000
WS-C6509-E
BERKELEY COMMUNICATIONS

What happened?

- Custom silicon development is expensive
- Using merchant silicon can save cost
- Arista was one of the first on the market
- Example from 2009:
  - Brocade MLX16: 64x10Gbit/s
  - Arista 7148SX: 48x10Gbit/s
- But you could buy 10x 7148SX vs 1x MLX16
- Ok, there is a big feature difference

They are all the “same”

- Broadcom Trident3:
  - Cisco Nexus 3100-Z
  - Arista 7050X3
  - Dell S5248F
  - Juniper QFX5120
  - Edgecore AS7326-56X
- Broadcom Tomahawk:
  - Dell Z9100
  - Juniper QFX5200
  - Extreme X870
  - Arista 7060CX
  - Quanta T7032-IX1B

But what's different?

- Software stack
- Command line interface
- Supporting certain hardware features
- Hardware build quality
- Stacking options
- Prices

Open networking

- Decoupling hardware and software
- SAI/SONiC
- Switchdev
- Commercial network OS-es
  - Big Switch Networks
  - Cumulus Linux
  - IP Infusion
  - Pluribus Networks

Switch Abstraction Interface (SAI)

- Being used by Microsoft SONiC
- Support for multiple ASICs
- Part of the Open Compute Project
- Started by Microsoft
- Uses binary blobs for controlling ASICs
- Open API on top

Switchdev

- Switch ASIC SDK included in the Linux kernel
  - So you can install whatever distribution you want
- Using standard linux tools
- Complete freedom!
- And no binary blobs needed!

- Only Mellanox Spectrum is supported
- Ask your vendor for Switchdev support!

V Commercial network OS-es

- Capable of running the same software on multiple hardware platforms
- Specialized in certain type of networks
- Most of them are scale ups/start ups
- Potential support issues between HW and SW
- Possible vendor locking by the software vendor

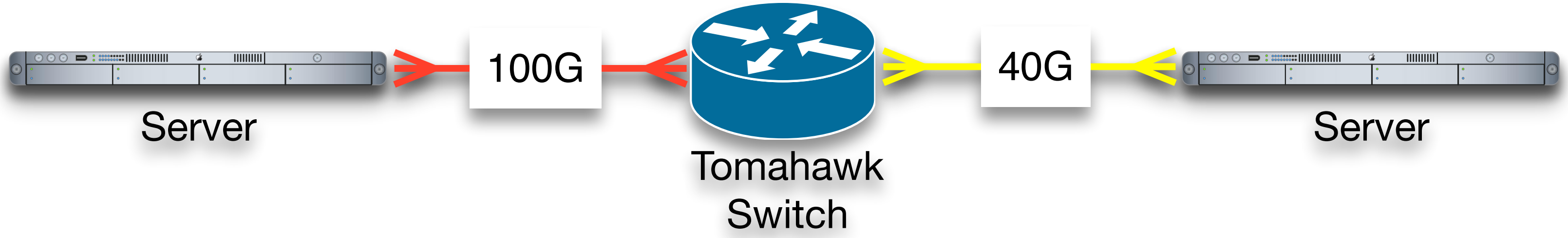
Chapter 2

- Let’s talk about traffic and challenges with it

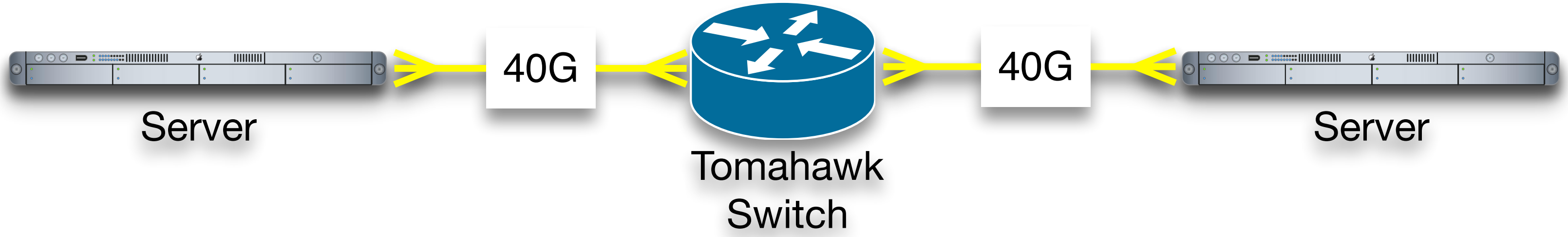
Gbit/s vs packets per second

- Most ASICs can handle line rate with big packets
- But small packets become a big problem
- Firewalls can’t handle a lot of small packets
- Small packets are smaller than 1000 Bytes
- small packets are used:
  - DAQ networks
  - Advanced DDoS attacks (UDP amplification)
  - Certain UDP based protocols (NTP, DNS, DHCP)
  - Realtime streaming protocols (sFlow, telemetry)

Tomahawk speed problem example



Capable doing:
Iperf TCP <30Gbit/s



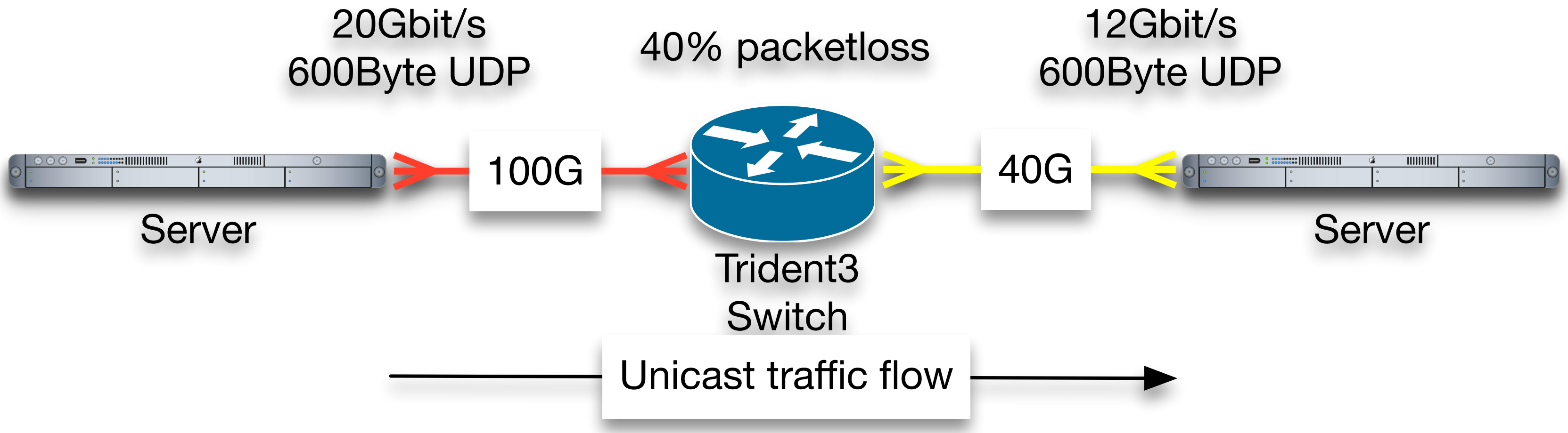
Capable doing:
Iperf TCP 39,4Gbit/s

What happened over there?

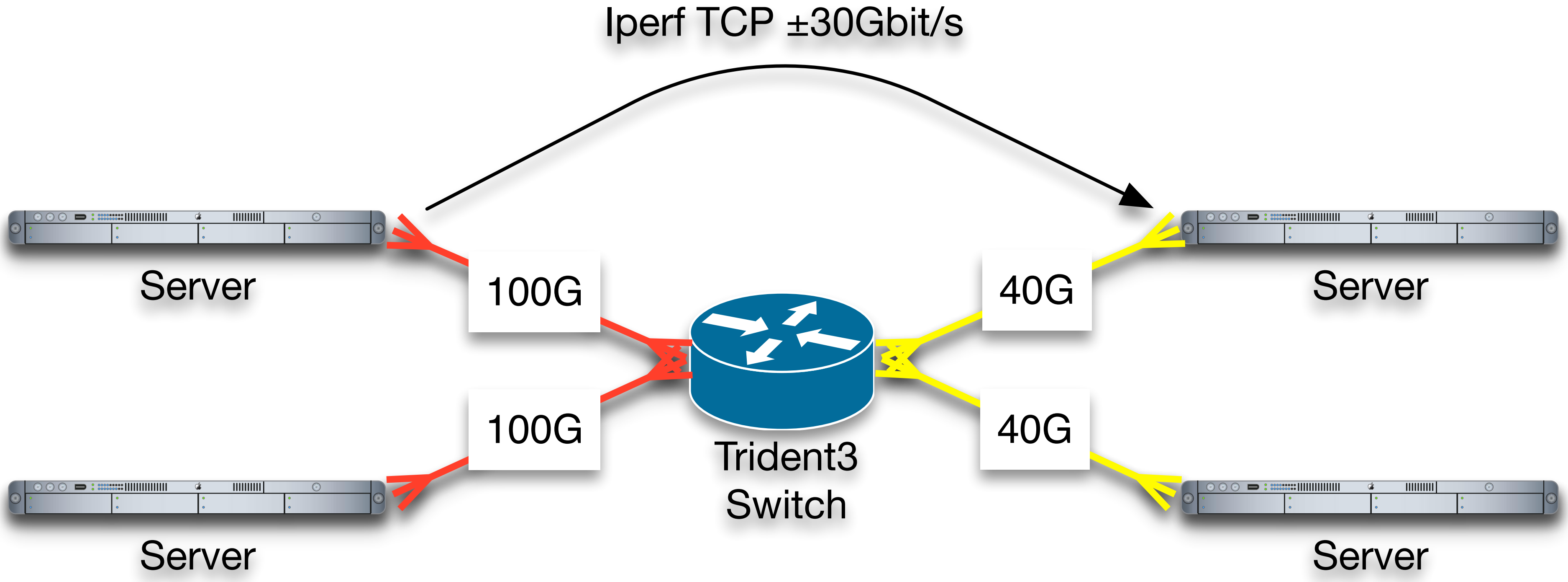
- 100G is 4x25G under water
- 40G is 4x10G under water
- The ASIC uses the buffer when changing speeds
- The buffer is too small and too slow to handle it
- TCP saves the day by lowering the Gbit/s

- By changing the cable from 100G to 40G
  - Buffer doesn’t need to be used because of the same speed on both ends

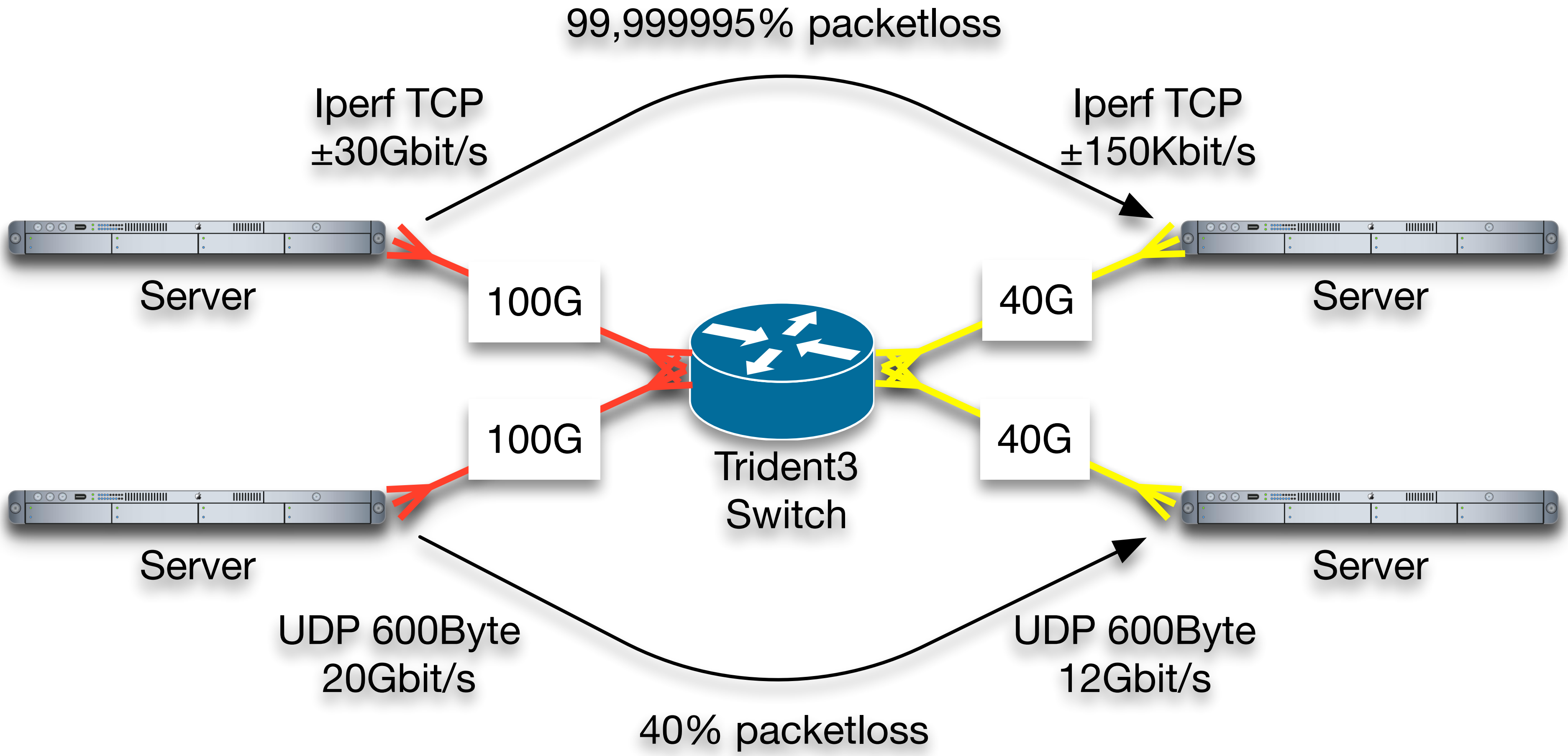
Trident3 buffer problem example



Trident3 buffer problem impact



Trident3 buffer problem impact



The dark side of Broadcom ASICs

- Modern ASICs can’t handle speed differences that well
  - Losing ±20% traffic capacity per port
- Poor use of buffer in the switch
  - Maximum capacity reached with only 1% traffic usage
- Proprietary SDK
  - If there is a software bug, you can’t fix it!
- Combining features can be tricky
  - We’re getting more demanding of our network boxes

Let’s get the IXIA!

- Easy to use
- Generate the type of packets that you want
- But hold on…
- Commercial network testers are too perfect
- Lacks a bit of chaos to correctly simulate the real world
- Are very expensive…

- Now what?

Software alternatives

- Iperf
  - TCP and UDP user space network tester
- kernel pktgen module
  - UDP kernel mode packet generator
- pkt-gen netmap
  - UDP packet generator and PCAP replay tool
- dpdk-pktgen
  - “Any” type packet generator using DPDK

Network card options

- Intel X710
  - 10, 25, 40Gbit/s support
  - It’s less stable compared with Intel X520
- Chelsio T6
  - 10, 25, 40, 50, 100Gbit/s support
  - Extensive storage offloading
  - Works well with FreeBSD
- Mellanox Connect X5
  - 10, 25, 40, 50, 100Gbit/s support
  - Stable and has a lot of features

40Gbit/s test machine

- hardware specs:
  - Chassis: Fujitsu RX1330 or Dell R340 or something else
  - CPU: Intel E series CPU’s (more GHz is better!)
  - NIC: 25/40/50Gbit/s network card
- Capable of generating 42Mpps using 40Gbit NIC
- Cost: <€2500,-

V 100Gbit/s test machine

> hardware specs:
  > Chassis: Dell R7415 or Gigabyte R271-Z00
  > CPU: AMD EPYC 7371
  > NIC: 25/40/50/100Gbit/s network card

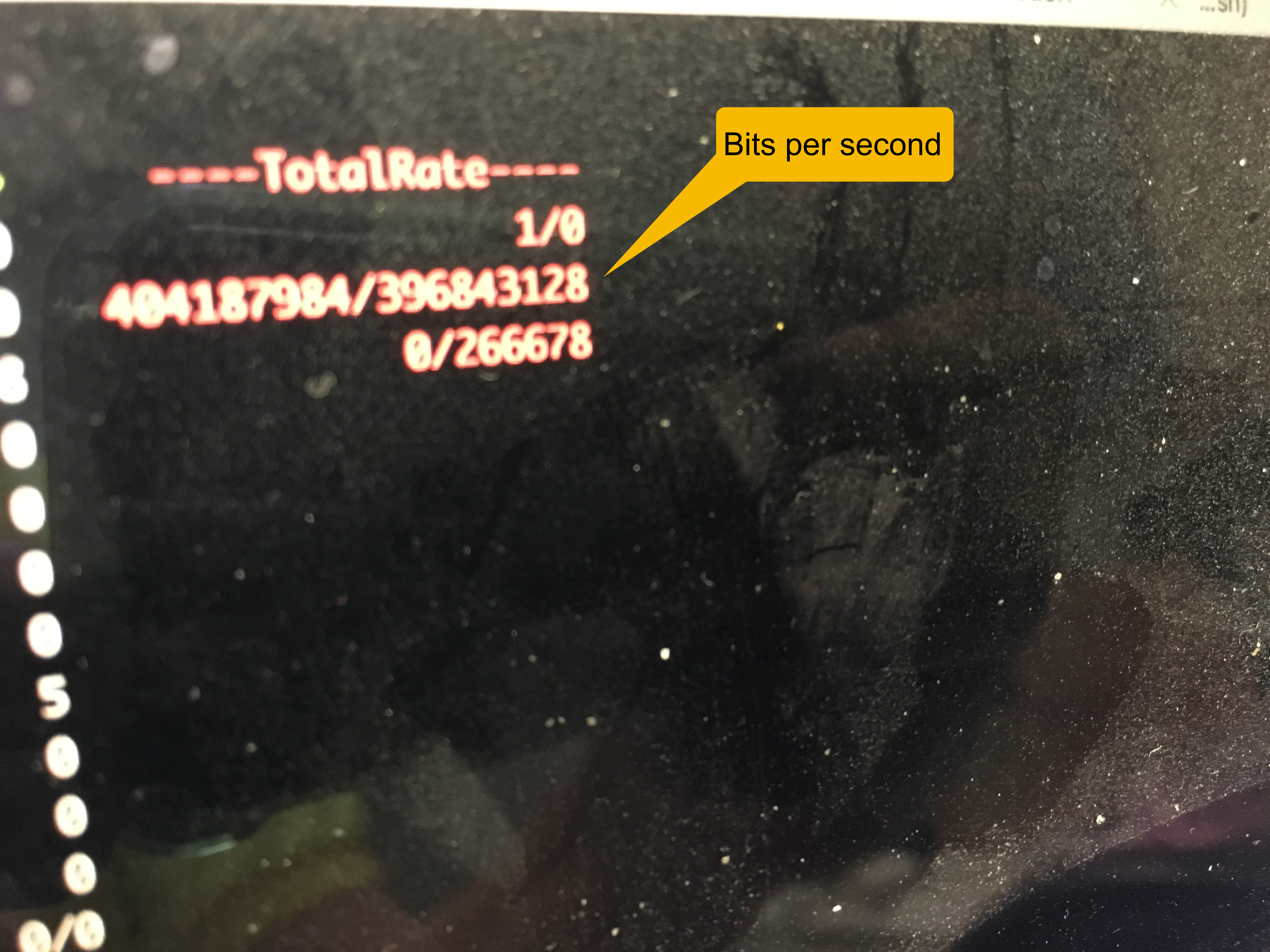
> Capable of generating ±80Mpps per 100Gbit NIC

> Cost: <€5000,-

V Multi 100Gbit/s test machine

- hardware specs:
  - Chassis: IBM S922LC or something else
  - CPU: POWER9
  - NIC: Mellanox Connect-X5 dual port PCI-e Gen-4
- Capable of generating 200Mpps per 100Gbit NIC
- Cost: <€10000,-

- Possible alternative later this year: AMD Rome?



```
  ----TotalRate----
                1/0
404187984/396843128
           0/266678
```

Bits per second

V Why is this important?

- > UDP based protocols are getting more popular
- > Examples:
  - > HTTP3/QUIC
  - > UDP-lite
  - > PTP
  - > Realtime monitoring protocols
- > Side effect is that these protocols will not be friendly for your network compared with TCP

V Conclutions

- Mixing speeds on modern Broadcom ASICs is a **bad** idea
- ASIC bugs aren’t fixable with software
- Most Open Networking solutions aren’t really open
- Building network test boxes isn’t expensive
- More protocols could be UDP-based moving forward
- Hoping on more good ASICs coming to market
- 400G ASICs will be hit even worse? Let’s see!

Questions?

- Ask me during the coffee breaks about Jericho flaws
- Or other ASICs on the market :-)

- Arista 7500R == Broadcom Jericho ;-)

- Brocade/Extreme SLX == Broadcom Jericho ;-)

- Juniper QFX10000 == Juniper custom ASIC

- More info: https://wiki.nikhef.nl/grid/SystemDesign