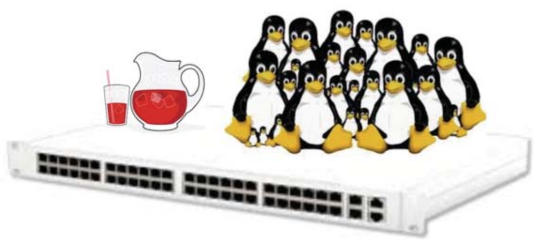
Performance SDN

Yatish Kumar
CTO Corsa Technology

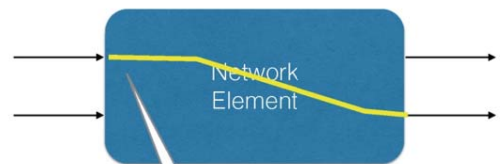
yatish@corsa.com

All you need is white box hardware and SDN to solve any networking problem !



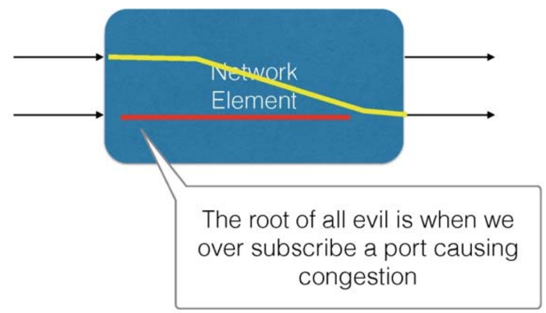
It's a switch
It's a router
It's a load balancer
It's a scalable fabric
It's powered by Linux !

Performance Fundamentals

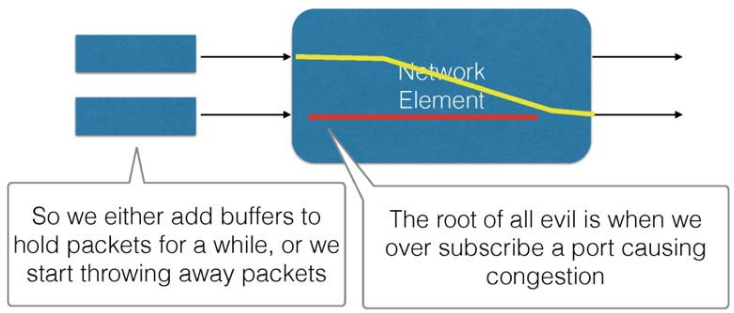


The purpose of all networking boxes is to inspect packets and forward them to an output port

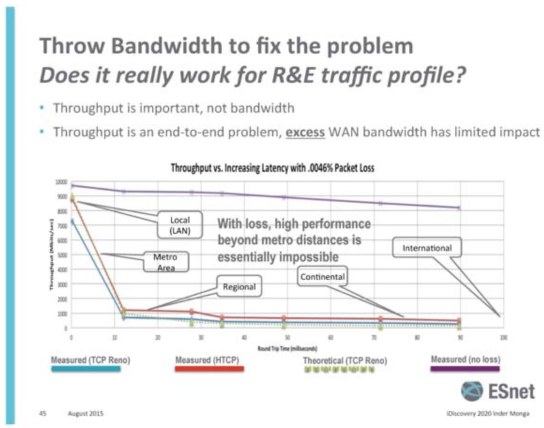
Performance Fundamentals

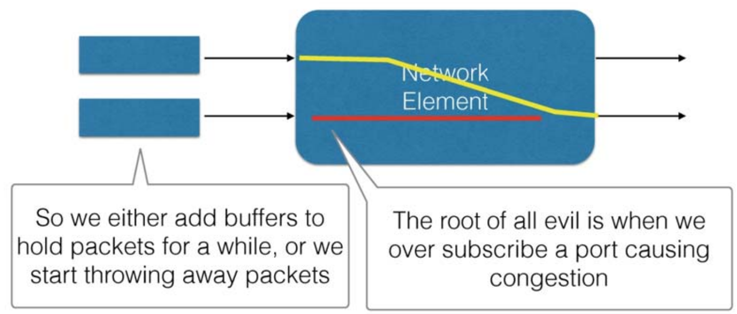


Performance Fundamentals



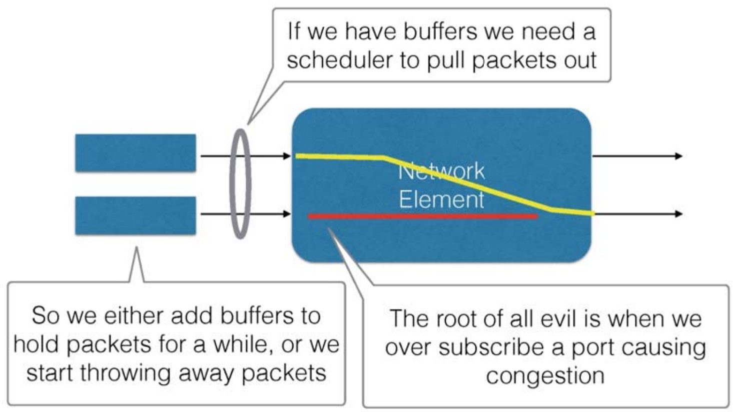
Performance Fundamentals



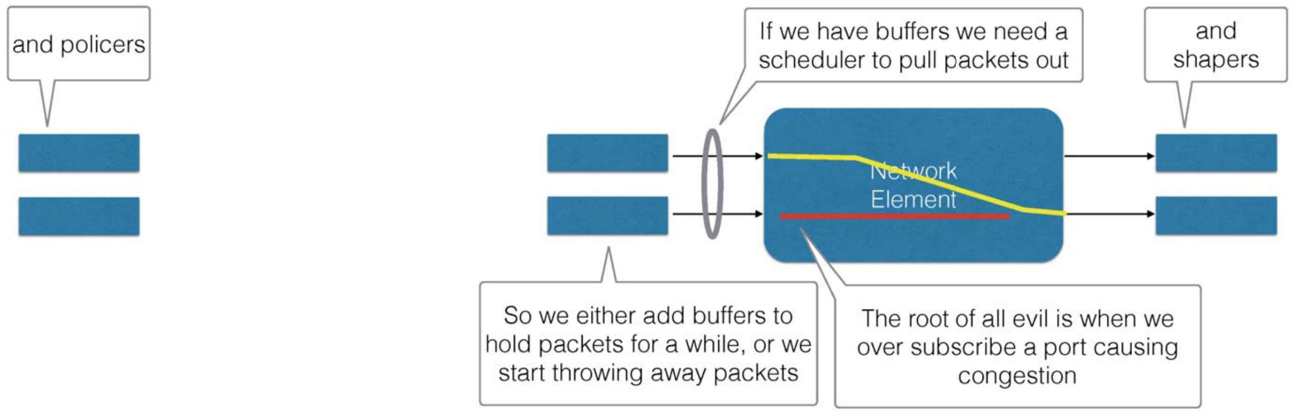


What would these guys do ?
Keep packets or throw them away ?

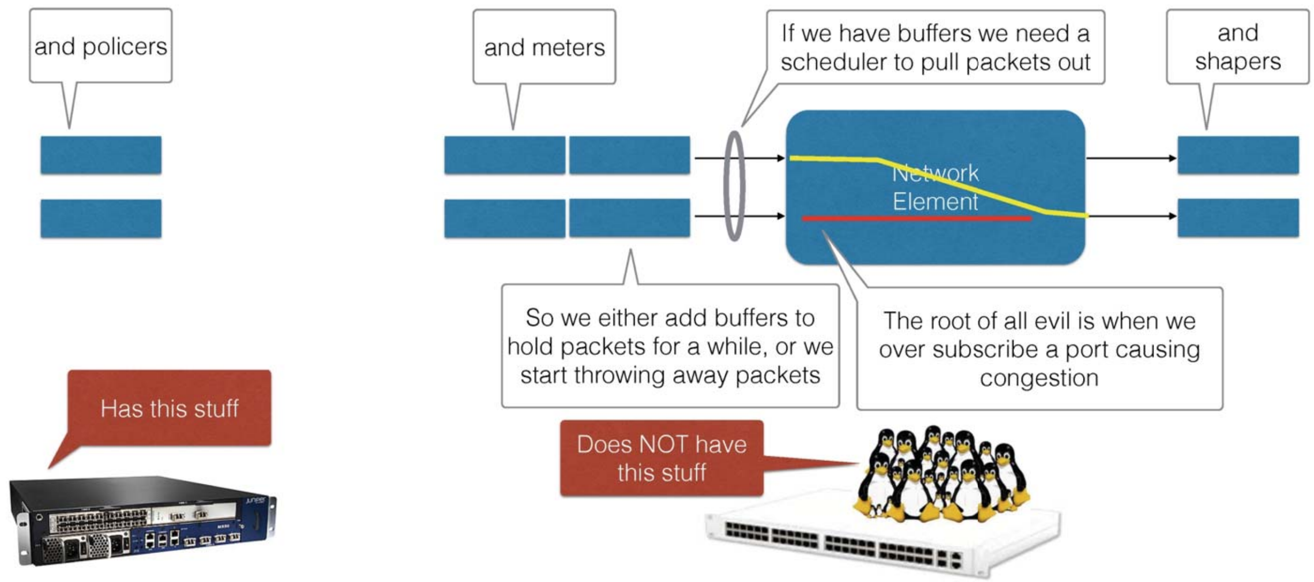
Performance Fundamentals



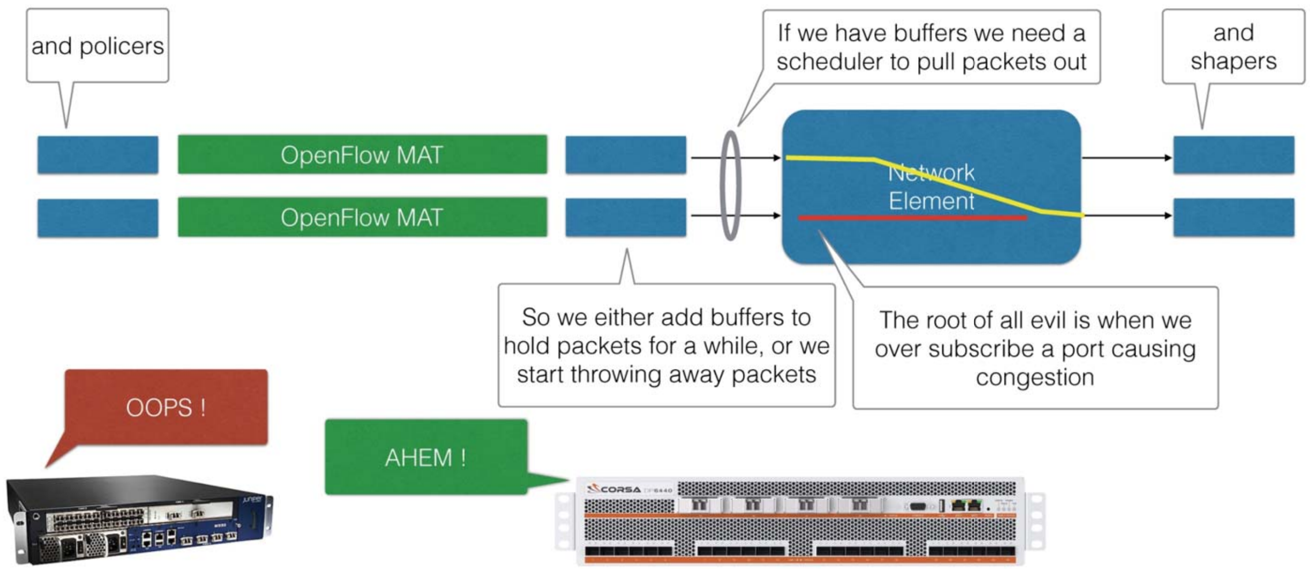
Performance Fundamentals



Performance Fundamentals



Performance Fundamentals

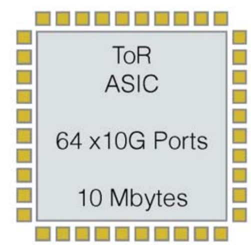


Just one more thing !!!

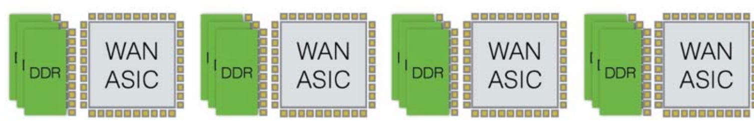
At 100 Gbits per second
According to Inder Monga we need 100 ms of buffering
Which means we need 10Gbits (Roughly 1 GigaByte) of buffering
Per 100G port !

WAN Buffers have to use DDR3

Each DDR3 roughly supports 50Gbits of read+write traffic
So we need 2 per 100G port in order to have sustained bandwidth
A 600 Gig WAN switch needs 12 DDR3 modules
12 Modules x 240 pins = 2880
Is a LOTTA PINS !

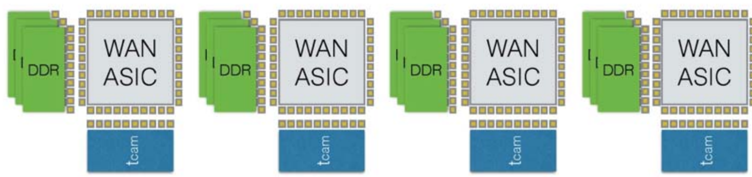


WAN Buffers have to use distributed processing



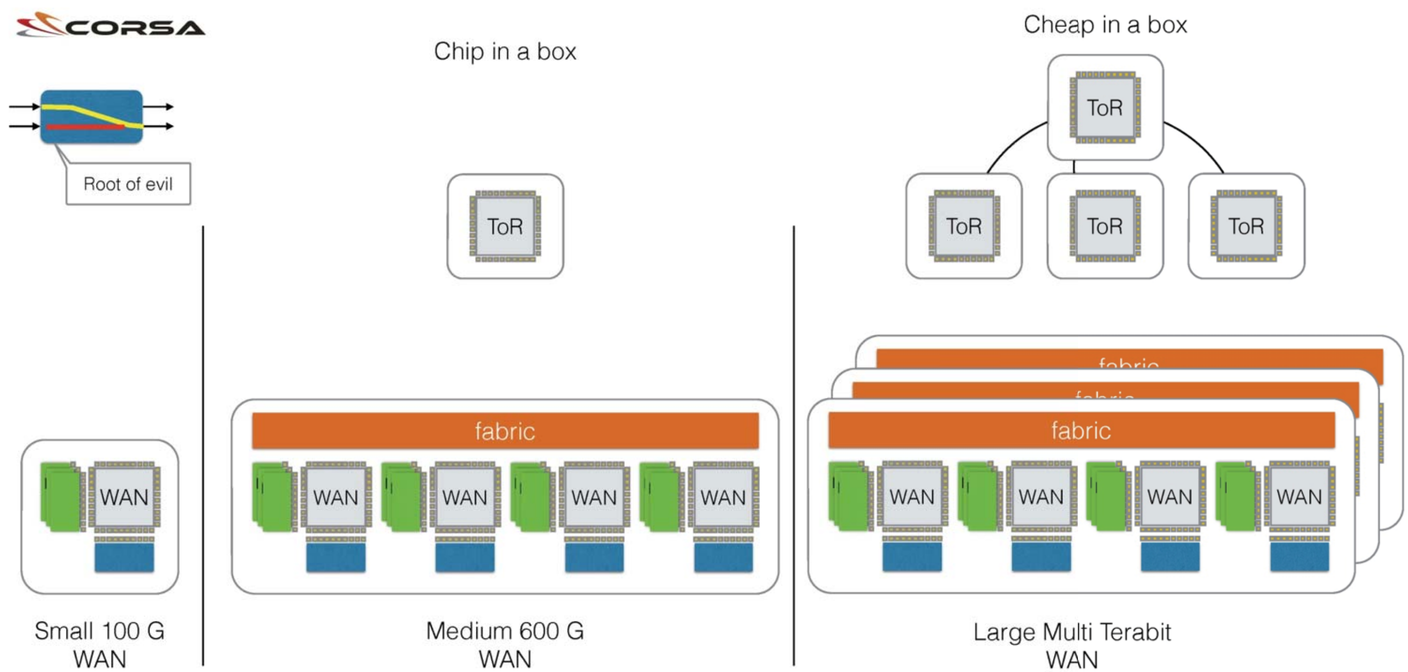
What about TCAMs ?
ToR White Box switches support roughly 10k TCAM entries
WAN routers need 1M IPv4 routes
Now we need external TCAMs too !

WAN Lookups have to use distributed processing

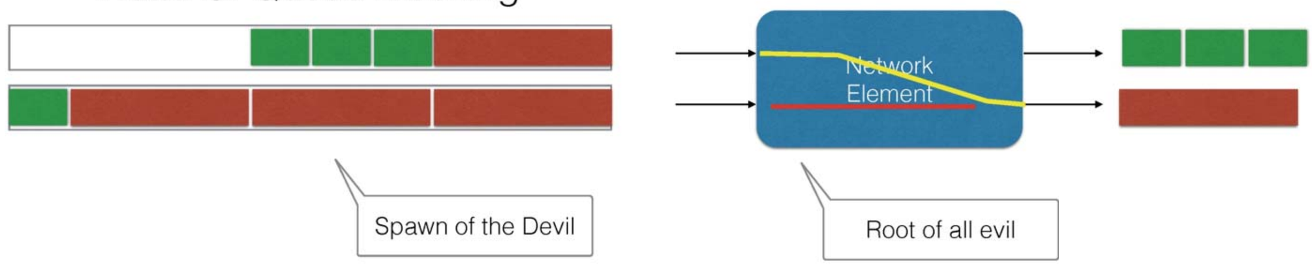


ToR discards packets

small internal TCAM

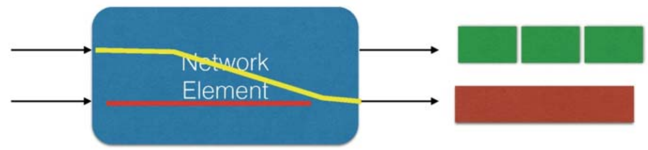

Root of evil
Chip in a box
Cheap in a box
ToR
ToR
ToR
ToR
ToR
fabric
fabric
WAN
WAN
WAN
WAN
WAN
WAN
WAN
WAN
WAN
Small 100 G
WAN
Medium 600 G
WAN
Large Multi Terabit
WAN

Head Of Queue Blocking



CORSA

Head Of Queue Blocking

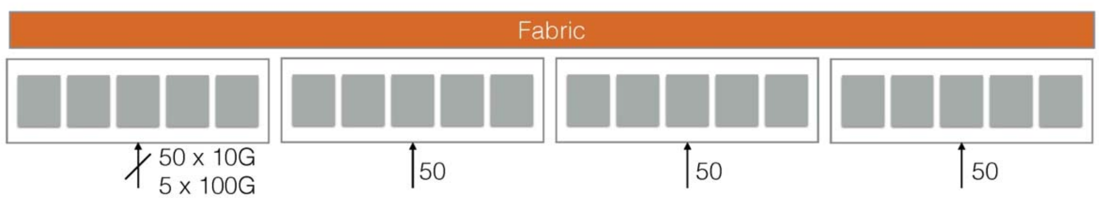


N inputs per module
M Modules
Q QoS

Q x N x M

8 QoS x 4 boxes x 50 egress ports = 1600 Queues

NOT 8 Queues !!



Leading to Papers in Sigcomm 2015

Congestion Control for Large-Scale RDMA Deployments

Yibo Zhu[1,3] Haggai Eran[2] Daniel Firestone[1] Chuanxiong Guo[1] Marina Lipshteyn[1]
Yehonatan Liron[2] Jitendra Padhye[1] Shachar Raindel[2] Mohamad Haj Yahia[2] Ming Zhang[1]

[1]Microsoft [2]Mellanox [3]U. C. Santa Barbara

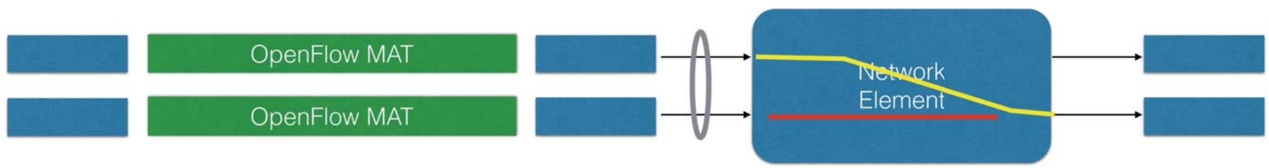
Their solution signals back to the servers for congestion control. Unfortunately this doesn't work in the WAN !!

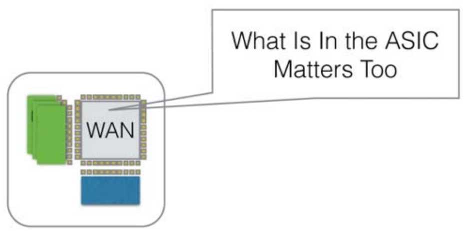
So taking ToR CLOS fabrics and building WAN routers doesn't work very well !

ABSTRACT

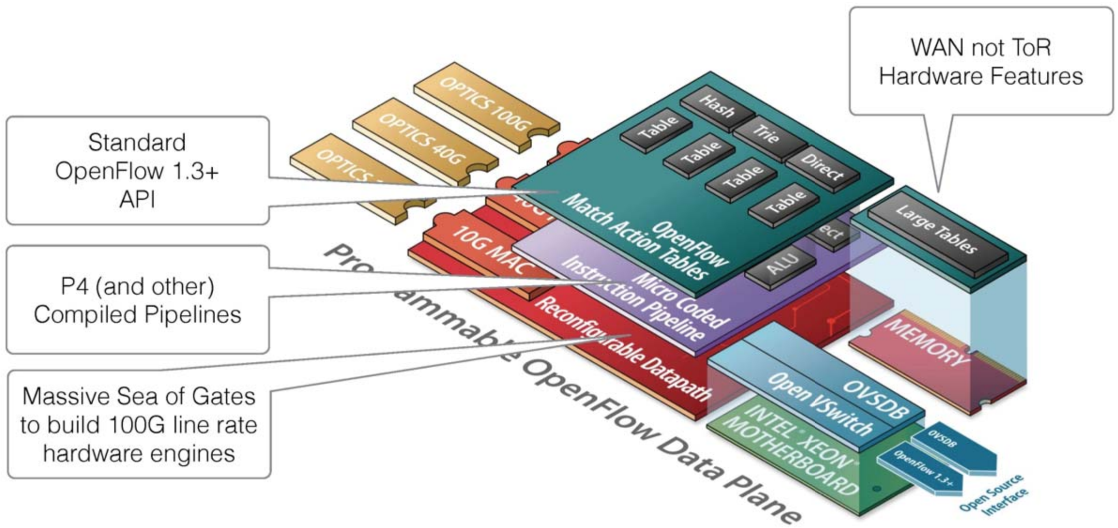
Modern datacenter applications demand high throughput (40Gbps) and ultra-low latency ($< 10\ \mu s$ per hop) from the network, with low CPU overhead. Standard TCP/IP stacks cannot meet these requirements, but Remote Direct Memory Access (RDMA) can. On IP-routed datacenter networks, RDMA is deployed using RoCEv2 protocol, which relies on Priority-based Flow Control (PFC) to enable a drop-free network. However, PFC can lead to poor application performance due to problems like head-of-line blocking and unfairness. To alleviates these problems, we introduce DC-QCN, an end-to-end congestion control scheme for RoCEv2.

Match Action Hardware Implications





Inside The Corsa Dataplane



Key Attributes to Shop For

Multiple match/action tables
Millions of flow entries
Large scale packet buffers
Metering and QoS
100-Gigabit ports with full OpenFlow 1.3
Extremely fast flow modifications per second 60k+ flow entries per second.
(as opposed to 100 flow mods / sec on white box ToR)

Key Bottlenecks To Understand

How big are the match tables ?

How much search bandwidth ?

What DoS Conditions Exist ?

Too many packets ?
Not enough control plane mips ?
Tail drop under load ?

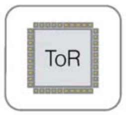
Match Action
MIPS

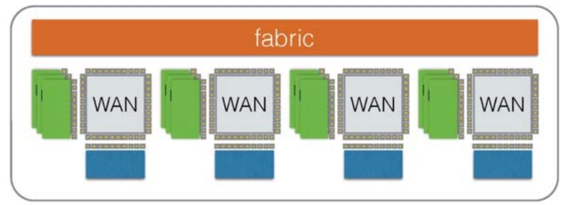
How many Packets Per Second ?

Buffer Memory Size and Bandwidth

How many ms ?
What is the read/write bandwidth ?

How is QoS maintained in the fabric ?





Thank You

Skip the cool aid. Get a stiff drink.