Data Center ASICs Considerations
IEEE 802 / IETF Data Center Workshop – Bangkok
Barak Gafni
Nov 2018
Why are we here?
The End of Moore’s Law – The Networking Aspect

▪ Capacity doubles every 1.5y-2y, but...

▪ ASICs are getting bigger
  ▪ Higher ASIC cost: area is growing, yield is decreasing
  ▪ Force to consider new ASICs and packaging architectures

▪ The IO problem
  ▪ Analog doesn’t scale with process
  ▪ More, not only faster, IO is embedded in the networking ASICs

▪ Power density is rising
  ▪ New cooling technologies
  ▪ Higher systems’ cost
The Data Center Networks are Changing

- **Big Data / Machine Learning**
  - Distributed computation, Incast/Microbursts are common
  - GPUs are driving high BW

- **Cloud**
  - East-west traffic
  - Multi-tenancy

- **Storage**
  - Higher throughput with SSDs and NVMeoF
  - Hyperconverged architectures are adopted
RoCE – RDMA over Converged Ethernet

- RoCE becomes the data centers standard
  - Higher Efficiency and throughput in the network
  - Accelerate AI, ML, storage applications

RoCEv2 packet format:

- Data
- BTH (InfiniBand Transport)
- UDP
- IP
- MAC
Agenda

▪ The Buffers Challenge
▪ The Forwarding Resources Challenge
▪ The Telemetry Challenge
The Buffers Challenge - Requirements

- **Fairness**
  - Invisible infrastructure
  - SLA

- **Efficient burst absorption**
  - Incast / Microburst mitigation
  - Linear expansion of buffer size (?)

- **Lossy and lossless**
  - At the same device, at the same time

- **High throughput, low latency**
  - Network shouldn’t limit applications
The Buffers Challenge - Fully Shared Buffer!

- Provides bandwidth and buffering Fairness
  - Across ports, queues
  - Across applications, tenants
  - Simplifies buffer management

- Provides efficient burst absorption
  - All the memory is accessible by all the ports
  - ASIC dynamically adopts queues’ thresholds

- Enables ingress AND egress buffer admission
  - Lossy AND lossless co-exist

- Flexible PFC
  - Support various cable length
The Forwarding Resources Challenge

- Data center scale requirements are growing...
  - `draft-dt-rtgwgdcrouting-requirements`
  - Need for overlay networks support

- Table per pipe stage / network application doesn’t scale

- The Solution – Fully Shared Forwarding Resources
  - All the Forwarding Resources are accessible by all networking functions
  - Dynamic (re)allocation of resources

The Telemetry Challenge

- CPU doesn’t keep up with packet rate
  - At 100G port – 150 million packets per second!
  - Polling queue status is like shooting in the dark
  - Watermarks are not enough granular

- Application-aware telemetry is needed

- New data-plane solutions are needed
  - ASIC generated histograms
  - ASIC assisted event driven telemetry
  - ASIC support for iOAM

Data Fields for In-situ OAM
draft-ietf-ippm-ioam-data-04

Histogram - Queue Utilization
Thank You