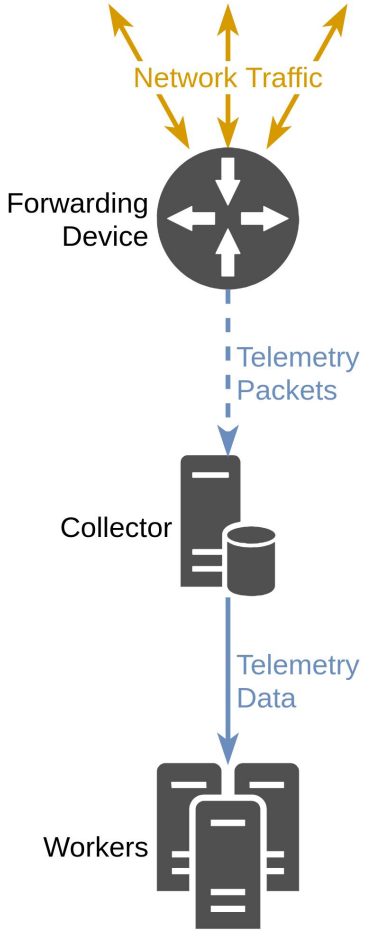
University of Amsterdam
Multiscale Networked Systems

Using P4 and RDMA to Collect Telemetry Data

Rutger Beltman, Silke Knossen,
Joseph Hill, Paola Grosso

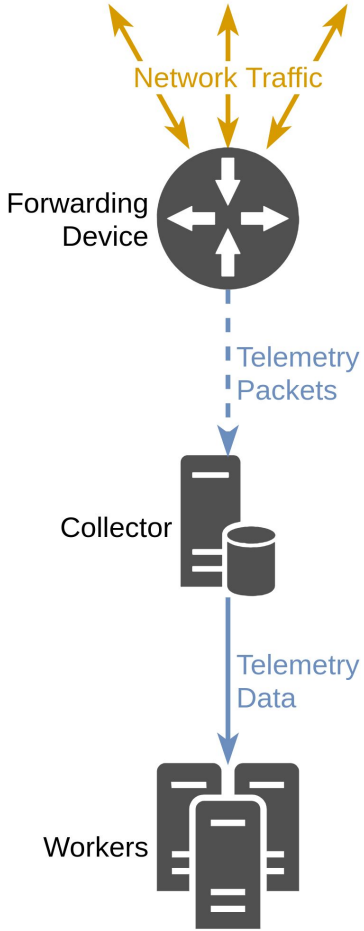
Network Telemetry

- Importance of network telemetry
- Traditional collection methods
  - SNMP
  - netFlow
  - sFlow
- Per packet telemetry data
- Performance considerations
  - Source
  - Collector



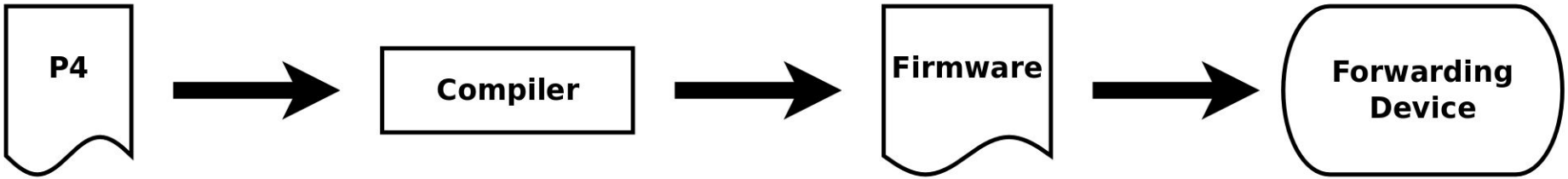
Telemetry Workflow

- General Workflow
  - Generate Telemetry Data
  - Identify Target Traffic
  - Gather Data
  - Send to Collector
  - Store Data
  - Provide Data to Analyzers
- For the Purposes of this Research
  - TCP/IP headers are used as the Telemetry Data
  - No Analysis of Telemetry Data is Performed



P4

- Programming Protocol-Independent Packet Processors
- Domain Specific Language that describes the behavior of the data plane
- Allows for the implementation of custom protocols
- Control plane is outside of scope of the specification



Remote Direct Memory Access (RDMA)

- Direct Access to Memory without involving CPU
- Hardware offloading
- Often used for storage
- Protocols
  - Infiniband
  - RDMA over Converged Ethernet (RoCE)
  - iWARP
- Security Concerns
  - Boundary checking
  - Isolated networks

Related Work

- In-band Network Telemetry (INT) framework
- ESnet High Touch services
- Daehyeok Kim et al. “Generic External Memory for Switch Data Planes.”

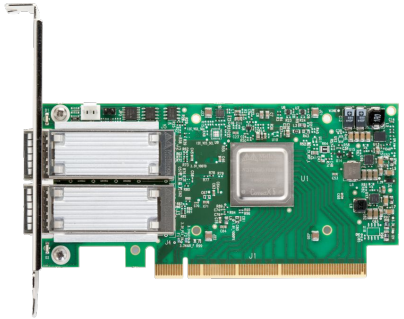
RoCE v1 Packet Structure

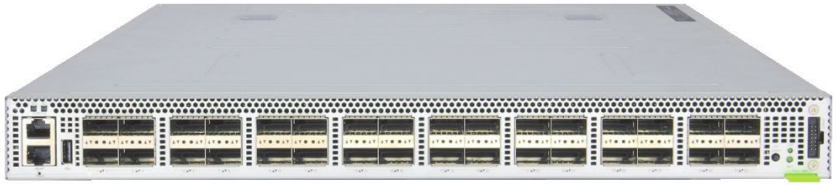
- Specific to RDMA write only
- Key Fields
  - Virtual Address (VA)
  - Remote Key (R Key)
  - Destination Queue Pair (QP)
- Derived Fields
  - Packet Sequence Number (PSN)
  - Invariant CRC

Headers	Fields
Ethernet	...
	EtherType: 0x8915
Global Routing Header (GRH)	...
Base Transport Header (BTH)	...
	Destination QP
	...
	PSN
RDMA Extended Transport Header (RETH)	Virtual Address
	R Key
	...
Payload	
Invariant CRC	

Hardware

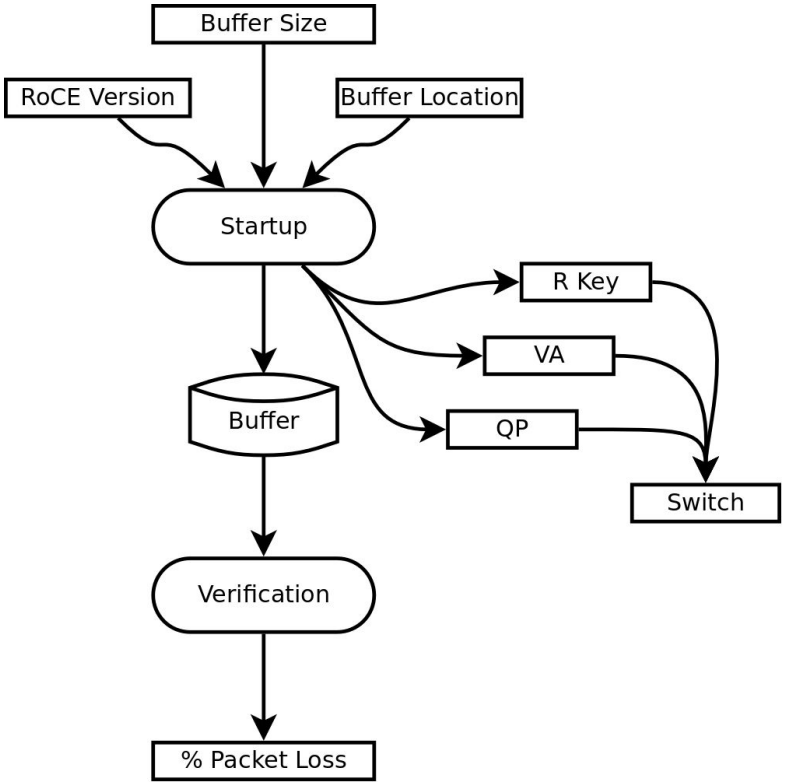
- Mellanox Connect-X5 EN NIC
  - RDMA with Infiniband, RoCE v1/2
  - Two 100 GbE Interfaces
  - Up to 200 million packets per second
- EdgeCore Wedge 100BF-32X switch
  - Packet Generator
  - P4 Forwarding Device
  - Thirty-two 100 GbE Interfaces
  - Up to 4.7 billion packets per second





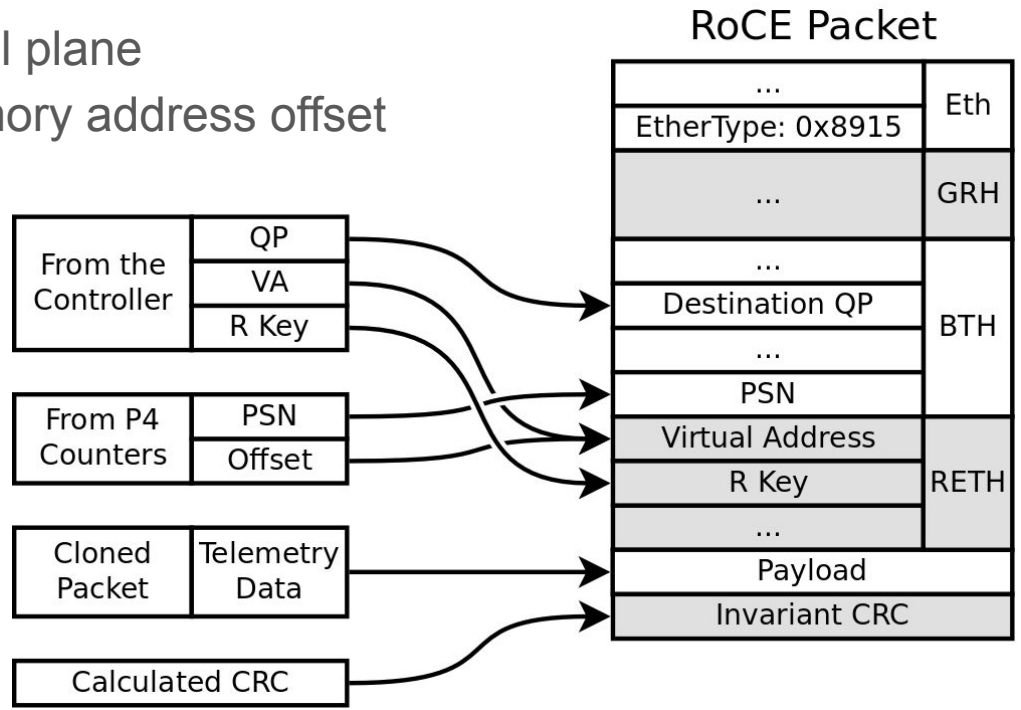
Collector Implementation

- Creates RoCE session
  - Sets RoCE version
  - Buffer size
  - Memory vs Disk
- Provides RoCE parameters
  - Queue Pair (QP)
  - Remote Key (R Key)
  - Virtual Address (VA)
- Verifies received data
  - Reports packet loss



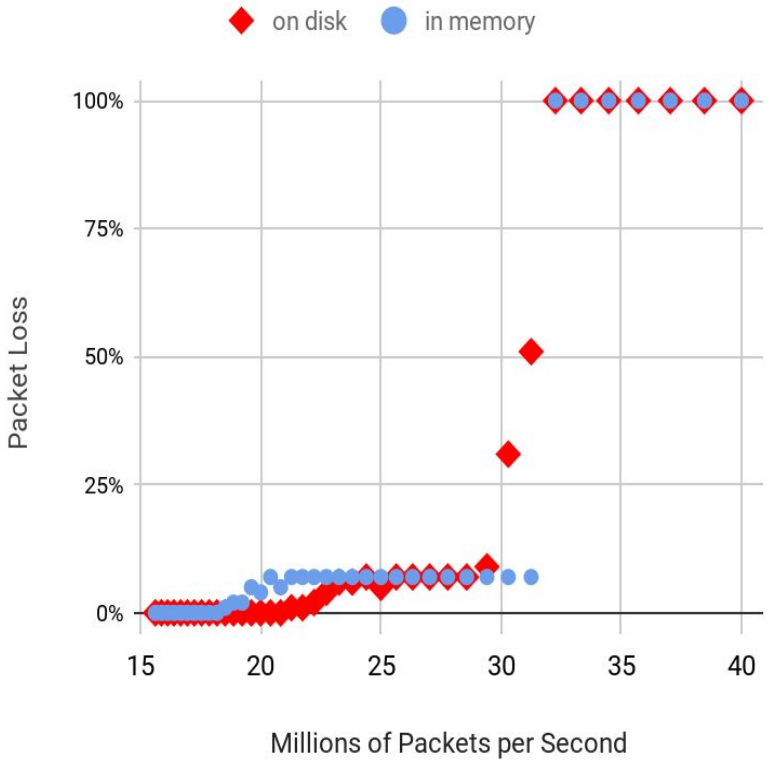
Switch Implementation

- Parameters loaded via control plane
- Keeps track of PSN and memory address offset
- Calculates Invariant CRC
- Builds RoCE packet



Results

- Packet loss versus packet rate
- Packet rates of 15 to 40 million pps
- 10 measurements at each packet rate
- Separate measurements for on disk and in memory buffers
- No packet loss at 15 million pps
- Less than 10% packet loss up to 30 million pps
- Packet loss increases sharply above 30 million pps



Challenges

- Not using Converged Ethernet
- Switch is not an RoCE endpoint
- Packet loss and the PSN
- Invariant CRC
  - Trailers in P4
  - Hash / Checksum functions in P4

Future Work

- Performance optimizations
- Use RoCE Unreliable Connection
- Compare with other methods
  - DPDK
  - tcpdump
  - UDP socket
- Implement using other RDMA protocols

Acknowledgments

- SURFnet - Research on Networks (RoN)

- Security, Stability and Transparency in inter-network Communication (2STiC)

- Energy Sciences Network (ESnet) High Touch Services