P4 based Programmable Data Plane

Haitao Kang
CUSTOMER ENGINEER
P4 and Data Plane Programming
A network is a means to forward packets from one place to another, and modify them along the way. If we cannot control how packets are forwarded and processed, we are not really in control.
“This is *how I know* to process packets” 
(i.e. the ASIC datasheet makes the rules)
Top-down network element design

“This is how I want the network to behave and how to switch packets…”
(the user / controller makes the rules)

“P4: Programming Protocol-Independent Packet Processors”
P4 Community – Growing Momentum

Independent Consortium
Free to join
Apache 2.0 License

~1500 developers
~5000 commits
~1500 followers
~800 forks

~200 contributors
~30 Repositories
~12 teams
~Multiple targets

~100 Industry and Academia Members
~4 Working Groups
~4 Bi-weekly face-to-face meetings
~8 Mailing Lists
Programmable Network Devices

- PISA: Flexible Match+Action ASICs
  - Barefoot Tofino, Intel Flexpipe, Cisco Doppler, Cavium (Xpliant), …
- NPU
  - EZchip, Netronome, …
- CPU
  - Open Vswitch, eBPF, DPDK, VPP…
- FPGA
  - Xilinx, Altera, …

These devices let us tell them how to process packets.
Benefits of Data Plane Programmability

"Think programming rather than protocols..."

- **Data Plane Telemetry**: Monitor with Scalable, Programmable, Real-Time, Inband Network Telemetry (INT)
- **Simplicity**: Focus just on your requirements
- **Future Proof**: Continuously deliver features with no rip and replace
- **Scale**: Maximize data-plane resources to your needs
- **Speed & Agility**: Innovate at speed of software

Think programming rather than protocols...
What can you do with P4?

- In-band Network Telemetry – INT[1]
- Low Latency Congestion Control – NDP[2]
- Layer 4 Load Balancer – SilkRoad[3]
- Fast In-Network cache for key-value stores – NetCache[4]
- Consensus at network speed – NetPaxos[5]
- Aggregation for MapReduce Applications [6]

PISA: Protocol-Independent Switch Architecture

- **Programmable Parser**
  - Programmer declares the headers that should be recognized and their order in the packet

- **Programmable Match-Action Pipeline**
  - Programmer defines the tables and the exact processing algorithm
  - Programmer declares how the output packet will look on the wire

- **Programmable Deparser**
PISA in Action

- Packet is parsed into individual headers (parsed representation)
- Headers and intermediate results can be used for matching and actions
- Headers can be modified, added or removed
- Packet is deparsed (serialized)
Mapping a Simple L3 Data Plane Program on PISA

Programmable Parser

Programmable Match-Action Pipeline

Programmable Deparser

- L2
- IPv4
- IPv6
- ACL
- Ethernet MAC Address Table
- IPv4 Address Table
- IPv6 Address Table
- ACL Rules
Example P4 Program

Parser Program

```
parser parse_ethernet {
    extract(ethernet);
    return switch(ethernet.ethertype) {
        0x8100 : parse_vlan_tag;
        0x0800 : parse_ipv4;
        0x8847 : parse_mpls;
        default: ingress;
    }
}
```

Header and Data Declarations

```
header_type ethernet_t { ... }
header_type l2_metadata_t { ... }

header ethernet_t ethernet;
header vlan_tag_t vlan_tag[2];
metadata l2_metadata_t l2_meta;
```

Tables and Control Flow

```
table port_table {
}

control ingress {
    apply(port_table);
    if (l2_meta.vlan_tags == 0) {
        process_assign_vlan();
    }
}
```
Barefoot Tofino and Applications
“Programmable switches are 10-100x slower than fixed-function switches. They cost more and consume more power.”

No longer true!

CONVENTIONAL WISDOM IN NETWORKING
Barefoot Tofino – Programmability & Performance

The world’s **fastest** and most **programmable** Ethernet switch ASIC family.

- Open Source **P4** Programming Language
- Open **PISA** Target Architecture
Arista 7170 Series of Multi-function Programmable Platforms

1U 7170-32C
32 ports of 100G
2.5B pkts/s. Multiple profiles.

2U 7170-64C
64 ports of 100G
5B pkts/s. Multiple profiles.

Cisco Nexus® 34180YC programmable switch

1U Cisco Nexus® 34180YC programmable switch. High-speed, low-power, high-density data center switch. 48 ports of 10/25G and 6 ports of 100G. Multiple profiles.
Data-Plane Telemetry

THE NETWORK SHOULD ANSWER THESE QUESTIONS

1. “Which path did my packet take?”
2. “Which rules did my packet follow?”
3. “How long did it queue at each switch?”
4. “Who did it share the queues with?”

Tofino + Deep Insight can answer all four questions. At full line rate. Without generating any additional packets!
1 “Which path did my packet take?”

“I visited Switch 1 @780ns, Switch 9 @1.3µs, Switch 12 @2.4µs”

2 “Which rules did my packet follow?”

“In Switch 1, I followed rules 75 and 250. In Switch 9, I followed rules 3 and 80.”

<table>
<thead>
<tr>
<th>#</th>
<th>Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>75</td>
<td>192.168.0/24</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>
3 “How long did my packet queue at each switch?”

4 “Who did my packet share resources with?”

“Delay: 100ns, 200ns, 19740ns”
3. “How long did my packet queue at each switch?”

4. “Who did my packet share resources with?”

“Delay: 100ns, 200ns, 19740ns”
How it works and how we use the data

Leverages In-Band Network Telemetry (INT)
https://github.com/p4lang/p4-applications/tree/master/telemetry/specs

Add: SwitchID, Arrival Time, Queue Delay, Matched Rules, ...

INT Source
Instruments packets for Telemetry

INT Transit
Adds metadata based on INT instructions

INT Sink
Removes metadata

Original Packet

Log, Analyze and Visualize
**SPRINT: A Fully Featured, High-Performance INT**

**FULLY COMPATIBLE SUPERSET OF A VANILLA INT IMPLEMENTATION**

- **S**mart
- **P**rogrammable
- **R**eal Time

- What to Observe
- What to Collect
- Intelligent Triggers
- Built-in Load Balancing
- Adapt to customers requirements
- Flexible encapsulation through P4
- Open specifications and ecosystem
- Data-plane Streaming
- Packet-by-packet Anomaly detection
- Real time Analytics with Deep Insight
Barefoot Deep Insight Analytics

Barefoot SPRINT Data-Plane Telemetry
- In-Band Network Telemetry (INT.P4)
- Intelligent Deduplication and Triggers
- Hardware Primitives
- Line Rate Monitoring

Open Telemetry Report Format defined by the P4.org Applications Working Group

3rd Party Network Management Solutions

Deep Insight Analytics Software
- Real-time Anomaly Detection
- Rich Analytics
- Dashboards and Drill down workflows
- Modular Architecture for scale-out
- Runs on Commodity Servers

Answer for Every Packet...
1. How did it get here?
2. Why is it here?
3. How long was it delayed?
4. Why was it delayed?
Financial and Enterprise – Congestion analysis

TCP INCAST PROBLEMS (MICRO-BURSTS)

Tofino Switch (6.5 Tb/s)

Aggressor (Sender)

Victim (Sender)

Queue

INT is enabled

QUEUE is FULL!

Applications

• **Financial** and **Trading** Apps
• **Distributed storage Apps** (Hadoop, MapReduce, HDFS, Cassandra)
• **Cloud data centers Apps** (web search, maps, social networks, data warehousing and analytics)

Deep Insight

Copyright 2018 - Barefoot Networks
In-Network and Smart Appliance Model for the Cloud Era

Problem:
• Customers and Operators devoting thousands of x86 servers for network and security functions
• High CAPEX, Poor efficiency (sized for worst case) and Application performance (Latency)

Accelerate Network, Security and Apps efficiency
• 100x better performance
• 1000x lower latency (nanoseconds)
• 100x lower power
L4 Load-balancing at every ToR (1 of 2)

Problem in Today’s Networks

- Centralized Hardware or Software Load Balancing with high Infrastructure Cost and Latency
- Hair-pinning of E-W and N-S Traffic to Virtual or Physical Load-Balancers
- Hard to scale to Millions of Connections at Line Rate
L4 Load-balancing at every ToR (2 of 2)

Problem in Today’s Networks

- Centralized Hardware or Software Load Balancing with high Infrastructure **Cost and Latency**
- **Hair-pinning** of E-W and N-S Traffic to Virtual or Physical Load-Balancers
- **Hard to scale** to Millions of Connections at Line Rate

P4 Tofino Solution

- **High Scale** (Millions of Flows) with frequent DIP Pool updates
- Per-connection **Consistency**
- **Optimized** Traffic Flow and Latency (Ideal for E-W LB)
- **Consolidation** of Middle-boxes or x86 to Lower Cost
- **Robustness** and DDOS protection built into Tofino

Flexible Tofino Deployment Model

- **Service Appliance:** L4 LB appliances using Tofino
- **Distributed:** Embed L4 LB capabilities into regular switches

Copyright 2018 - Barefoot Networks
In summary…

1. **SDN is about who is in control!**
   - **Part 1:** Network owners decided how their networks are controlled.
   - **Part 2:** With P4, they could decide how packets are processed.

2. **Chip technology:** Programmable switch now has the same power, performance and cost as fixed function.

3. **Line-rate telemetry now possible.** Per-packet, flexible. 100% in data plane at line rate.

4. **New ideas:** Beautiful new ideas are now owned by the programmer, not the chip designer.
Thank You!

Haitao Kang
BAREFOOTNETWORKS.COM