100G In-Band Network Telemetry with P4 and FPGA

Viktor Puš, Pavel Benáček
pus,benacek@cesnet.cz
CESNET, a.l.e.
Zikova 4
160 00 Prague 6
Czech Republic

Pavel Minařík, Jan Pazdera
minarik,pazdera@flowmon.com
Flowmon Networks a.s.
Sochorova 3232
616 00 Brno
Czech Republic

Lukáš Richter
richter@netcope.com
Netcope Technologies a.s.
Sochorova 3232
616 00 Brno
Czech Republic

In-Band Network Telemetry (INT) is a relatively new concept of network performance monitoring, utilizing modern programmable data path elements. INT allows to trace the exact path that a packet took, together with the internal status of data path elements it has visited. This results in unprecedented visibility into the dynamics of network behavior, allowing detailed network debugging and performance analysis. INT is in contrast to more traditional flow-based monitoring, which loses fine-grained time resolution in the process of flow aggregation, and has no information about the internal state of switches due to its fully passive nature. But since flow-based monitoring focuses mostly at L3-L4 (and sometimes L7), both methods can be seen as complementary and mutually beneficial. Since INT is not a standardized protocol, programming it in P4 offers good flexibility for future modifications.

Our live demo shows a proof of concept implementation of a line rate 100 Gbps P4-programmed INT traffic sink - the point where INT marked traffic is received and the INT headers are processed. The NFB-100G2 PCI Express card with Xilinx Virtex-7 FPGA was chosen as hardware target. The Netcope P4 to VHDL compiler is used to generate the packet processing pipeline that strips INT headers away from the packets. The cleaned traffic is forwarded by the card, so that the final traffic destination is not aware that INT was in progress and suffers no INT processing overhead.

We also demonstrate how INT statistics can be combined with and integrated into production-quality flow-based monitoring environment. The extracted INT headers are transferred from the card to host CPU for further processing - P4's `generate_digest` action is used for this. The software running at CPU generates NetFlow messages with the information about delays within individual switches that the packets have visited. The NetFlow messages are then forwarded to the Flowmon Networks Collector, which stores the statistics and provides visualization and query interface for the measured switch delay values. This L2 information is combined in a unified web GUI with traditional flow-based L3-L4 statistics to provide very complete view of the network.

Presenter's name:
Lukáš Richter

Presenter's organization:
Netcope Technologies