COIN
Internet-Draft
Intended status: Informational
Expires: June 10, 2021

H. Singh
MNK Labs and Consulting

December 7, 2020

Requirements for P4 Program Splitting for Heterogeneous Network Nodes
draft-hsingh-coinrg-reqs-p4comp-00

Abstract

The P4 research community has published a paper to show how to split a P4 program into sub-programs which run on heterogeneous network nodes in a network. Examples for nodes are a network switch, a smartNIC, or a host machine. The paper has developed artifacts to split program based on latency, data rate, cost, etc. However, the paper does not mention any requirements. To provide guidance, this document covers requirements for splitting P4 programs for heterogeneous network nodes.

Table of Contents

1. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

2. Introduction

The research paper [FLY] covers splitting a P4 program into sub-programs to run the sub-programs on heterogeneous network nodes. The requirements are:

1. If the heterogeneous network includes a switch, the ARP [RFC0826] and IPv6 ND [RFC4861] data plane P4 code cannot be split. This code replicates packets on switch ports to issue broadcast ARP or IPv6 ND multicast messages. If this code moves outside the switch, then another node has to send each packet to the switch to issue broadcast or multicast messages, causing delay with address resolution.

2. Likewise ARP or IPv6 ND Proxy data plane code cannot be split to run outside the switch.

3. BGP table cannot move outside the switch to another node. Distributed BGP is a research topic.

4. A switch likely includes TCAM (ternary content-addressable memory) and thus the P4 program may use P4 ternary table match kind. If such a table is moved to another node due to program split, the node the code moves to is important. A FPGA (field-programmable gate array) does not use TCAM and a host machine may

not either. The FPGA and host use hash-based table lookup. Depending on the table key size, an appropriate hash is required. Either the splitting tool prompts the user for what hash to use or deduces what hash - user input is desirable. For example, for a 6-tuple IPv4 key, a 128 bit key is used and for the same 6-tuple, the IPv6 key uses 320 bits. Appropriate hashes are required for such keys.

5. Splitting algorithms should not develop High Availability. Network deployments already use dual switches, or CLOS (leaf and spine switch redundant network) topology for redundancy. BFD [RFC5880] is recommended for use with liveliness detection.

3. Discussion

The two largest public cloud operators are Amazon AWS and Microsft Azure [NIC]. Both operators run Software Defined Networking (SDN) in the smartNIC (smart Network Interface Card). The reason is running SDN stack in software on the host requires additional CPU cycles. Burning CPUs for SDN services takes away from the processing power available to customer VMs, and increases the overall cost of providing cloud services. Azure uses a FPGA on smartNIC and programs the FPGA in Verilog, not P4. Amazon uses multi-core npu (Graviton uses 64 cores) on smartNIC and does not program Graviton in P4. Both these operators do not use host cpu or network switch for SDN operations. In future, even if both operators program smartNIC in P4, the operators do not have heterogeneous nodes running SDN.

4. Security Considerations

Use IPSec [RFC4301] to secure any control plane communications.

5. IANA Considerations

None.

6. Acknowledgements

Thanks (in alphabetical order by first name) to.

7. References

7.1. Normative References

[RFC0826] Plummer, D., "An Ethernet Address Resolution Protocol: Or Converting Network Protocol Addresses to 48.bit Ethernet Address for Transmission on Ethernet Hardware", STD 37, RFC 826, DOI 10.17487/RFC0826, November 1982, <https://www.rfc-editor.org/info/rfc826>.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC4301] Kent, S. and K. Seo, "Security Architecture for the Internet Protocol", RFC 4301, DOI 10.17487/RFC4301, December 2005, <https://www.rfc-editor.org/info/rfc4301>.

[RFC4861] Narten, T., Nordmark, E., Simpson, W., and H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)", RFC 4861, DOI 10.17487/RFC4861, September 2007, <https://www.rfc-editor.org/info/rfc4861>.

[RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, DOI 10.17487/RFC5880, June 2010, <https://www.rfc-editor.org/info/rfc5880>.

7.2. Informative References

[FLY] Sultana, N., Sonchack, J., Giesen, H., Pedisich, I., Han, Z., Shyamkumar, N., Burad, S., DeHon, A., and B. T. Loo, "Flightplan: Dataplane Disaggregation and Placement for P4 Programs", November 2020, <https://flightplan.cis.upenn.edu/flightplan.pdf>.

[NIC] Firestone, D., "Azure Accelerated Networking: SmartNICs in the Public Cloud", April 2018, <https://www.microsoft.com/en-us/research/uploads/prod/2018/03/Azure_SmartNIC_NSDI_2018.pdf>.

Author's Address

Hemant Singh
MNK Labs and Consulting
7 Caldwell Drive
Westford, MA 01886
USA

Email: hemant@mnkcg.com
URI: https://mnkcg.com/