SC23 Network Research Exhibition

PolKA routing approach to support traffic engineering for data-intensive science

Magnos Martinello, Rafael Guimarães, Everson Scherrer Borges, Cristina Dominicini,

Diego Maffioletti, Jordi Ros-Giralt, Edgar Pontes, Moisés Ribeiro and Harvey Newman

UFES {magnos.martinello, moises.ribeiro, edgar.pontes,}@ufes.br

IFES {everson,cristina.dominicini,rafaelg,maffioletti}@ifes.edu.br

Qualcomm Europe, Inc. {jros@qti.qualcomm.com}

Caltech {newman@hep.caltech.edu}

Abstract

To manage terabit/sec competing data flows across complex intercontinental networks for Data Intensive Science (DIS) research programs, we need new Traffic Engineering (TE) approaches that can balance innovative functionality, performance, reliability, and cost. These requirements are driven by the global distribution of DIS workflows and the demanding projections for the future. In this NRE, we present how PolKA (Polynomial Key-based Architecture) Source Routing (SR) offers Better, Faster, and Stronger functionalities to address the TE challenges for DIS networks.

Introduction

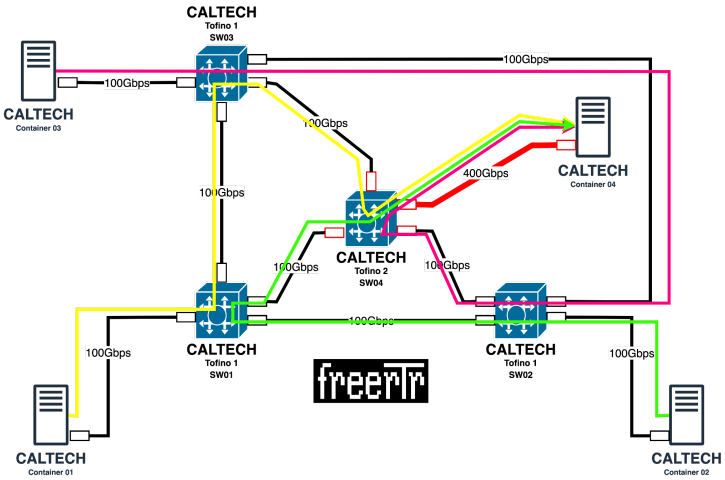
Enhancing the current requirements of globally distributed workflows involves optimizing and ensuring efficient and seamless operations across different locations. For instance, LHC and other data-intensive science programs require new approaches for data transport networks to address the balance between innovative functionality, performance, cost, and other policy-driven factors.

The High Luminosity LHC (HL-LHC) experiments alone, for example, require terabit/sec data flows across regional, national, and intercontinental network paths by around 2028. In addition, the Vera Rubin Observatory, DUNE at LBNF, the Square Kilometer Array Telescope, and others will bring not only extra traffic volumes to be handled across such paths but also new Traffic Engineering (TE) challenges. Challenges include satisfying contrasting requirements of competing flows across a complex intercontinental network topology.

This NRE proposes to demonstrate PolKA functionality to support the TE challenges for data-intensive science. PolKA is a novel source routing approach [1] that explores the Residue Number System (RNS) and Chinese Remainder Theorem (CRT) by performing the forwarding as an arithmetic operation: the remainder of division. PolKA encodes the path in a routeID using the RNS in contrast to the conventional list-based representation, which transports the path information “in clear” inside the packet header. Then, PolKA core nodes use this encoded route label to discover the output ports.

We plan to divide the demonstration of PolKA capabilities into two scenarios: i) in a data-intensive transfer over 100G and 400G network by using PolKA underlay tunnels; ii) in a comparison between segment routing and PolKA over a high-speed intercontinental testbed composed of P4-enabled programmable switches that interconnect DIS research facilities (Europe, South, and North America) employing its controllability of flows to support TE. The flows can be classified, balanced, and steered at the edge using a Policy-Based Routing (PBR) so that TE decisions can be guided by the Quantitative Theory of Bottleneck Structures (QTBS) and GradientGraph (G2)[3] for optimization.

The figure highlights the approach to achieve line rate transfer over a set of tunnels (paths in colors) created in the existing topology at Caltech testbed. Additionally, the demonstration will leverage the P4 World Lab testbed running FreeRtr as the routing platform that supports PolKA, along with other proposed features by this NRE (Network Research Education).



RARE (the Router for Academia, Research & Education) is a GÉANT project which is developing and deploying an open-source routing software platform solution that provides innovative functionality, interoperability, and high-performance data planes. Network solutions are sought out from the perspective of National Research and Education Networks (NRENs), and not from regular service operators', which makes RARE's outcomes very relevant to GNA-G's goals.

RARE subscribes to the new network paradigm of software-defined networking, which is based on a separation between the control plane and data plane to address the innovative functionality issue. Along with a Free and Open Source Software (FOSS) system on the control plane called FreeRouter(freeRtr), RARE manages to re-conciliate open innovation with interoperability with legacy protocols and networks, and adapt to and interwork with key technologies emerging in the network industry. For performance, RARE/freeRtr relies on the P4 open-source network programming language in order to expedite packet processing by exploiting cutting-edge programmable ASICs embedded in modern network boxes. It also supports the P4Runtime API, which is a control plane specification for controlling the data plane elements. Last but not least, FOSS will address the pressing cost issues for acquiring fully-featured operational systems versions of high-end network equipment.

Goals

The goal of this proposal is to investigate whether the PolKA approach deployed at RARE/freeRtr meets the needs of DIS networks, working with other software tools and subsystems developed by the DIS-WG for constructing a packet-switched underlay network composed of network paths with bandwidth guarantees, that offers load balancing at the edge, prioritizing and scheduling flows over selected multi-domain paths. As a result, decisions will be taken in a coordinated way through the network, computing and storage resources to help accelerate the science workflows.

In particular, this NRE will focus on providing TE functionalities allowed by FOSS versions of emerging industry-driven Source Routing Techniques (SRT), such as Segment Routing, and compare those with cutting-edge academic developments. In the latter category, **Pol**ynomial **K**ey-based **A**rchitecture (**PolKA**) [1] for Source Routing in Network Fabrics was chosen since it might be a better enabler for orchestrating computing and storage resources globally among DIS facilities. Our purpose is to:

1. Create PolKA underlay tunnels to support the massive transfer of data. We envision data-intensive transfer over 100G and 400G networks by using PolKA underlay tunnels.
2. Design and execution of emulated experiments based on GNA-G topologies guided by Quantitative Theory of Bottleneck Structures (QTBS) and GradientGraph (G2) [3] framework to reduce the flow completion time by selecting the highest throughput route.
3. Implement the emulated topologies (e.g. multipath) and traffic scenarios on the RARE/freeRtr (P4 Global Lab) testbed with the Barefoot Switch in order to select a representative proof-of-principle case.
4. Validate a data-intensive transfer crossing international testbeds with an edge-based control to prevent congestion and overload of international links capacity.

Controllability enabled by source routing

The most traditional way of executing source routing is to represent the path as a list of output ports and the forwarding operation as a pop. Although the most disseminated source routing protocol is Segment Routing, it has some limitations: i) it depends on expensive equipment and proprietary implementations; ii) its MPLS version still depends on tables in the core nodes; iii) it depends on variable-length headers, which limits the number of maximum hops implementable in hardware switches; and iv) there is no direct multipath/multicast support.

PolKA is a novel source routing approach that explores the RNS and CRT by performing the forwarding as an arithmetic operation: the remainder of division. PolKA can be deployed in high-performance P4-enabled programmable switches with the reuse of the CRC hardware, and its performance is equivalent to traditional approaches. Thus, this special path encoding and routing mechanism allows PolKA to offer the following advantages: i) it does not keep any table in the core network; ii) the packet header has a fixed length and the route label does not change throughout the path; iii) it can represent multipaths [2] in the network layer for any topology.

Resources

RARE/freeRtr has a wide range of legacy protocol FOSS components and subsystems with (operator-class) validated implementations. It can be used in all-in-one installations in order to emulate large networks for testing and validating new functionalities. Moreover, it has a unique seamless portability process (i.e., from emulation setups into testbed experiments) for performance test validation on different physical data planes. Finally, it should be highlighted that freeRtr is one of the few FOSS router operating systems that have segment routing implemented; and has a full implementation of PolKA. Thus, RARE/freeRtr suits the role of being a sandbox in prototyping pre-production networks for DIS-WG applications. As a concrete use case, we should focus on the SRT as it is a prominent alternative to conventional table-based routing for reducing the number of network states and also providing deterministic paths for TE policies.

SRT has recently been made available via Segment Routing by major vendors, but the price tag is still high. This motivates us to use RARE/freeRtr to tackle the issue of cost, but a FOSS solution will be considered valid if only and only if performance metrics behave accordingly under the extremely demanding DIS traffic scenarios. Besides bearing industry-led solutions, recent advances employing a residue number system (RNS) bring a new way of executing fully stateless SRT, in which forwarding decisions rely on a simple modulo operation over a routing label as done in PolKA.

Because much of the state information can be derived from the RNS coding embedded in the packets themselves, this approach could open new pathways to operate and manage networks, making decisions that respond to the progress of data flows, the state of the end sites, and time-dependent constraints associated with priorities, delivery deadlines, and other policies.

Involved Parties

As far as experimental facilities and teamwork are concerned, this proposal also counts on other collaborators, namely, Frédéric Loui at RENATER/GÉANT and Csaba Mate RARE/GÉANT for consulting and granting access to RARE/freeRtr GÉANT's testbed during initial testing for performance validation; and Marcos Schwarz at RNP for advising us on deploying proof-of-principle experiments on the GNA-G's AutoGOLE/SENSE testbed.

- Harvey Newman, Caltech, newman@hep.caltech.edu
- Marcos Schwarz, RNP, marcos.schwarz@rnpbr
- Frederic Loui, Renater, frederic.loui@renater.fr
- Csaba Mate, RARE/FreeRtr, cs@mp.ls
- Frank Slyne, Trinity College Dublin, fslyne@tcd.ie
- Marco Ruffini, Trinity College Dublin, marco.ruffini@scss.tcd.ie
- Eoin Kenny, HEANET, eoin.kenny@heanet.ie
- Qiao Xiang, Xiamen University, xiangq27@gmail.com
- Jeronimo Bezerra, FIU, Amlight, jbezerra@fiu.edu
- Marinho Barcellos, Waikato University, marinho.barcellos@waikato.ac.nz

References

[1] C. Dominicini *et al*., "PolKA: Polynomial Key-based Architecture for Source Routing in Network Fabrics," *2020 6th IEEE Conference on Network Softwarization (NetSoft)*, 2020, pp. 326-334, doi: 10.1109/NetSoft48620.2020.9165501.

[2] R. S. Guimarães et al., "M-PolKA: Multipath Polynomial Key-Based Source Routing for Reliable Communications," in IEEE Transactions on Network and Service Management, pp. 2639-2651, Sept. 2022, doi: 10.1109/TNSM.2022.3160875.

[3] Jordi Ros-45/33667Giralt et al., "On the Bottleneck Structure of Congestion-Controlled Networks". Proc. ACM Meas. Anal. Comput. Syst. (2019). https://doi.org/10.1107

Acknowledgements

This work was supported by the following Brazilian projects: FAPESP/MCTI/CGI.br (PORVIR-5G #20/05182-3, SAWI #20/05174-0), FAPES (#284/2021). M. Martinello is funded by the grant #306225/2020-4, CNPq-Brazil. Also, it was a recipient of the 2021 Google Research Scholar Award, and the 2022 Intel Fast Forward Initiative.