EPJ Web of Conferences **214**, 08013 (2019) https://doi.org/10.1051/epjconf/201921408013
CHEP 2018

Equal-cost multi-pathing in high power systems with TRILL

Andrey Baginyan[1,*], *Vladimir* Korenkov[1], *Andrey* Dolbilov[1], *Ivan* Kashunin[1]

[1]JINR, Laboratory of Information Technologies, 141980 Dubna, Russia

Abstract. The article presents a hierarchical diagram of the network farm and a model of the network architecture levels. Protocols for disposal full mesh network topologies are considered. Modern data transfer protocol TRILL is presented. Its advantages are analysed in comparison with other possible protocols that may be used in the full-mesh topology. Empirical calculations of data routing based on a Dijkstra© algorithm and a patent formula of the TRILL protocol are given. Two monitoring systems of downloading data channels are described. The data obtained from 40G interfaces through each monitoring systems is presented, and their behaviour is analysed. The main result is that the discrepancy of experimental data with theoretical predictions to be equal to the weight balancing of the traffic when transmitting the batch information over the equivalent edges of the graph. It is shown that the distribution of the traffic over such routes is of arbitrary and inconsistent with the patent formula character. The conclusion analyses issues of the traffic behaviour under extreme conditions.

1 Data processing center in the Joint Institute for Nuclear Research

The Joint Institute for Nuclear Research (JINR) participates in the multipurpose experiment named Compact Muon Solenoid (CMS) [1], carried out on the bunches of the accelerator facilities Large Hadron Collider (LHC) in the European Organization for Nuclear Research (CERN). The CMS test bench allows carrying out various experiments with a considerable data flow.

One of new data processing and storage centers was commissioned in the JINR before the second launch (RUN 2) of the Large Hadron Collider. According to the technical collaboration task, the CMS network segment was to ensure an uninterrupted interaction between 160 disk servers, 25 blade servers, 100 infrastructure servers and tape robot. The first module required 80 disk servers (160 10G-ports in bonding mode), 15 blade servers (30 10G-ports in bonding mode), 60 servers infrastructure (40 10G-ports and 40 1G-ports in bonding mode). Finally, the data center network segment should provided with 230 10G-ports and 40 1G-ports. A similar situation should be observed in the second phase of the project planned to be launched in the end of 2019 [2].

* Corresponding author: bag@jinr.ru

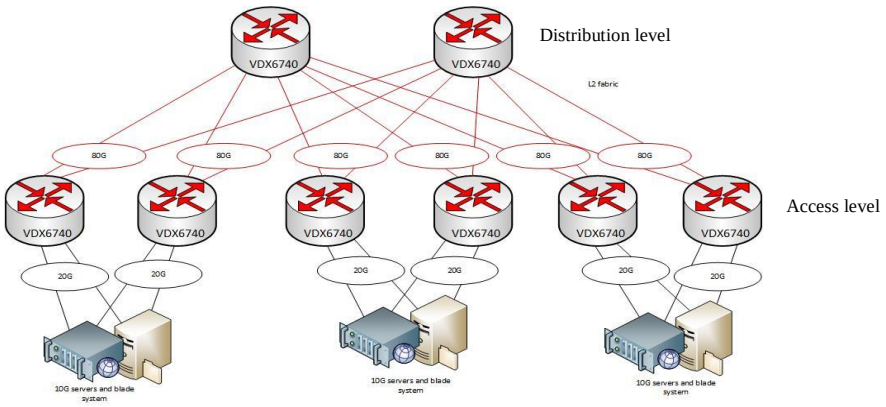

Fig. 1. A three-tier model of network segment Tier1 in the JINR.

Figure 1 shows the project documentation of the network architecture of the first module of Center Tier1 at JINR. Full redundancy of links are provided at all levels. As a result of such an architecture, failure of one switch should lead to the reduction in the total traffic capacity of the network segment by only 25%. In such case, all servers will have access to the external network [2].

2 SFP protocol and TRILL

To provide data transfer in multilink networks, there several solutions exists, each of which have certain advantages and weaknesses. One of the most well-known protocols – Shortest Path First or SPF, based on Dijkstra's algorithm [3], finds the shortest distance from one curve peak to all the others. It works only for graphs with negative weigh ribs, which is suitable to use in data communication networks. Transparent Interconnection of Lots of Links (TRILL) [4] protocol is developed based on this algorithm.

The TRILL protocol processes paths as implemented in the Intermediate System to Intermediate System (IS-IS) protocol on the second tier of the OSI model, enabling building solutions for campus networks and data centers.

Cloud services and data centers use a distributed network architecture for data storage, queries and search engines. A cluster with such architecture creates an immense east-west traffic. Currently we have information relating to traffic in some communication lines between disks and counting servers achieving 18G. Presently, clusters use more and more virtualization technologies, and therefore each server begins to run much more tasks than earlier and this, in its turn, results in the increase of traffic at input/output interfaces. Virtualization which is used to improve reliability, reduce services cost and increase flexibility in deploying services is also used to provide for the ability to migrate from one physical server to another.

Given all the above, the traditional approach which provides STPs (spanning tree protocols) [5] running at the access level, and three tier [6] protocols operating at the aggregation/core level, cannot fully complete the tasks relating to east-west traffic, because a significant part of cable infrastructure remains uninvolved.

The TRILL protocol allows building an unblocked network architecture which fully provides for the total use of the network imperceptible for users (Figure 2), and helps to involve new servers, because all active channels stay redundant.

The network architecture of the Tier1 data center in the Joint Institute for Nuclear Research is planned with two-way path between the access level and the server level on the

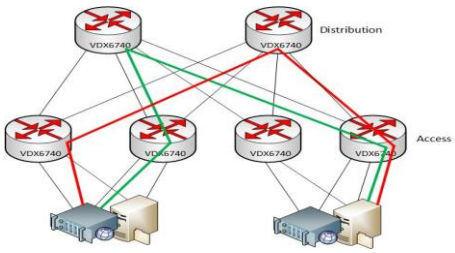

Fig. 2. Unblocked network architecture (two of the eight possible paths are shown).

equipment manufactured by Brocade. Each server will have access to the network segment via two links 10G each with total traffic capacity 20G. In addition to virtualization tasks, in Tier1 data center in JINR an important load on network is applied by the tasks relating to the simulation of test events by Monte-Carlo methods.

The communication between the access and the distribution levels will have four 40G paths, which will enable 160G data transmission.

Each 40G communication line is the sum of four 10G communication lines. That means that the 40G SFP+ module contains 4 transmitters of optic signal, each of which insures a two-way 10G data communication. The scheme of data communication between two data center nodes is shown in Figure 3. Looking at Figure 3, we may affirm that path is a-1-1'-a' ~ a-2-2'-a' ~ 1-3-3'-a' ~ a-4-4'-a'.

Basing on graph calculation according to Dijkstra's algorithm and on the patent claims for calculation of the shortest path "TRILL optimal forwarding and traffic engineered multi pathing in cloud switching", we determine the equal cost of path for all communication lines between the access level and the distribution level. Thus, we come to the conclusion that the load on data channel via path a-1-1'-a' is equal to the traffic via path a-2-2'-a', 1-3-3'-a' and a-4-4'-a'.

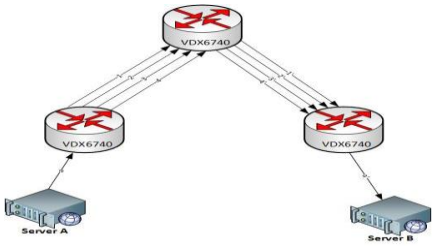

Fig. 3. Similar data communication paths between two networks nodes.

3 CACTI monitoring system

Cacti is a web application with an open source code. Cacti enables building graphs using ring database tools. Cacti acquires statistic data relating to the specified time periods and allows their graphical display. The standard patterns for graphical display are used for CPU start, RAM allocation, number of running processes, incoming/outgoing traffic.

Figure 4 shows data relating to the load capacity of input/output interfaces between the distribution and access levels for 10G four communication lines.

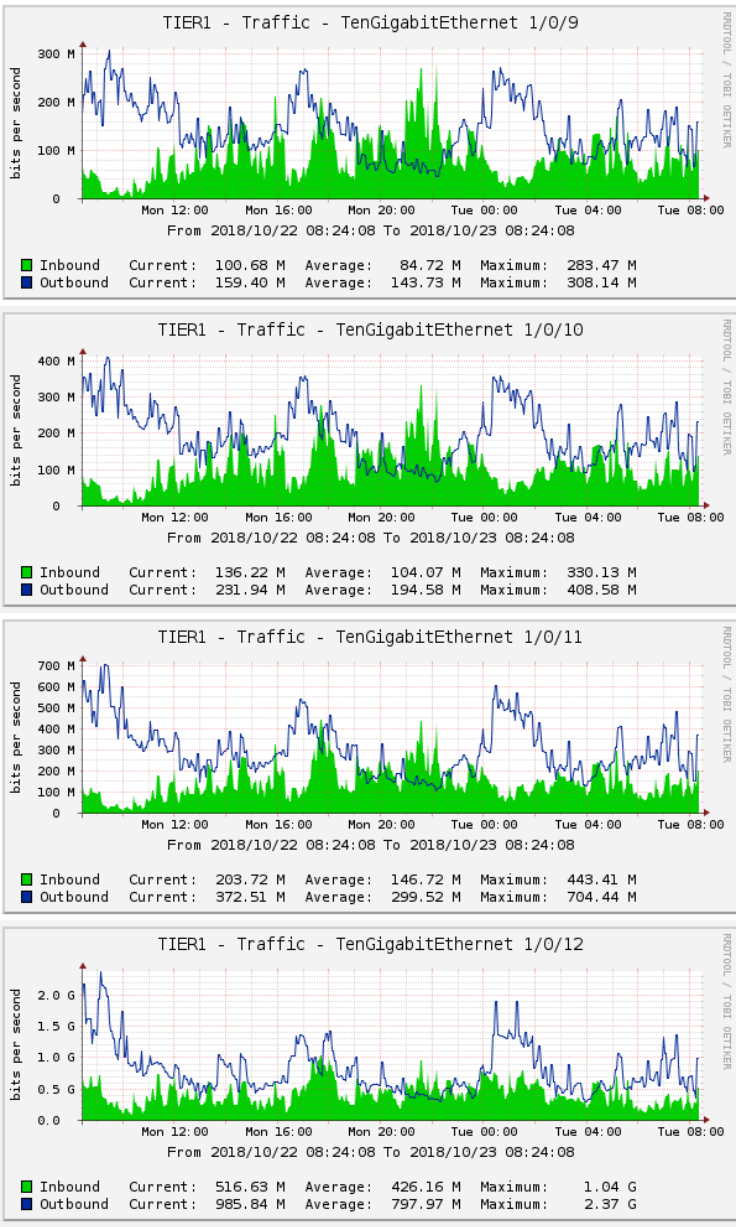

Fig. 4. Example of data loading capacity in a random similar four communication lines.

The curves analysis allows to pay attention to the similar nature of behaviour of sent and received traffic via each communication line. Only vertical axis values differ. However, this difference is not systematic. The ratio is 10%×10%×30%×50%.

Figure 5 shows other four communication lines. This shows a similar nature of curves' behaviour.

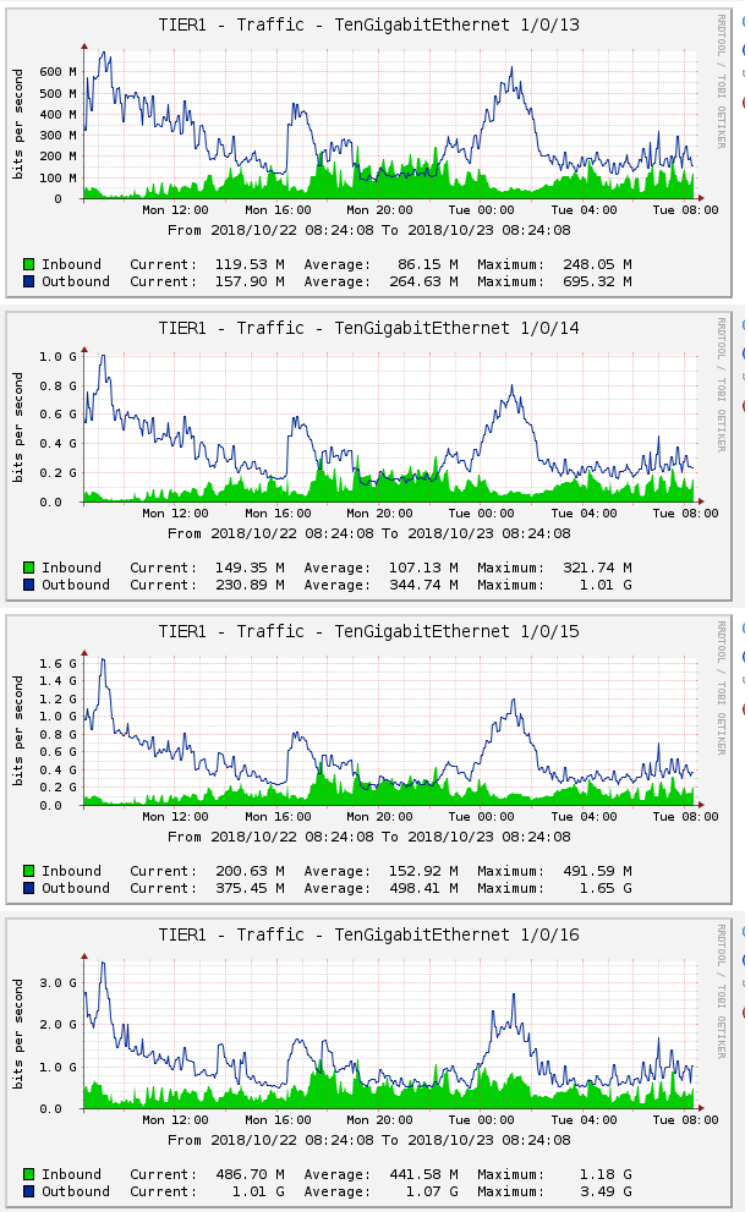

Fig. 5. Similar example of other similar four communication lines.

Thus, it turns out that our assumption about equal weigh categories for this path is incorrect. To verify this assumption and to exclude third party software products, a new monitoring system for communication lines was developed.

4 LITMON monitoring system and algorithm of calculation traffic

To build the incoming and outgoing traffic calculation algorithm we need to introduce the definitions of incoming and outgoing traffic.

Incoming traffic (D) is the relation of incoming bytes (I) par a time unit (t) (formula 1):

$$D=\Delta I/\Delta t \quad (1)$$

Outgoing traffic (U) is the relation of outgoing bytes (O) par a time unit (t) (formula 2):

$$U=\Delta O/\Delta t \quad (2)$$

To calculate the incoming and the outgoing traffic on the switch, the following steps are required:

1. obtain incoming/outgoing octets;
2. monitor the data collection time;
3. wait for a time off;
4. make next measurement;
5. determine the value as a relation;
6. convert the obtained value into megabits.

Incoming/outgoing octets could be obtained using SNMP [7] protocol. The data collection time unit shall be chosen according to the required accuracy. In Litmon it is 10 seconds.

It is possible to calculate incoming and outgoing traffic using the algorithm. According to the acquired data, it is possible to compare the results of different monitoring systems. For this purpose, the following parameters similarly specified for Cacti monitoring system are included:

- Measurement date from October 23, 2018;
- TenGigabit 09-012 network interfaces.

It is to be taken into consideration that the Cacti monitoring system counts data in Megabytes, while the Litmon monitoring system counts data in Megabits. Nevertheless, the curves shall not be different, as the difference between megabytes and megabits is a fixed value (Figure 6).

As we can see from the graph, the curves factually simulate those mapped in the Cacti monitoring system. There are some differences; they may be, however, explained by the fact that different monitoring systems acquire data in different time intervals. As the incoming/outgoing traffic is a permanently varying value, it is virtually impossible to obtain ideally identical data. Nevertheless, the main extremums within different curve intervals match, which may evidence the correctness of obtained data both by Cacti and by Litmon monitoring systems.

Let us also consider other network interfaces. (Figure 7).

As in the previous case, here we can observe the similarity of the data acquired by monitoring system Cacti with Litmon. Moreover, the main extremums of the incoming and the outgoing traffic match within intervals. All this proves that the data acquired by monitoring systems are equal, but, as we noticed earlier, the vertical axis values are not the same.

Fig. 6. Incoming/outgoing traffic of TenGigabit 09-012 network interfaces.

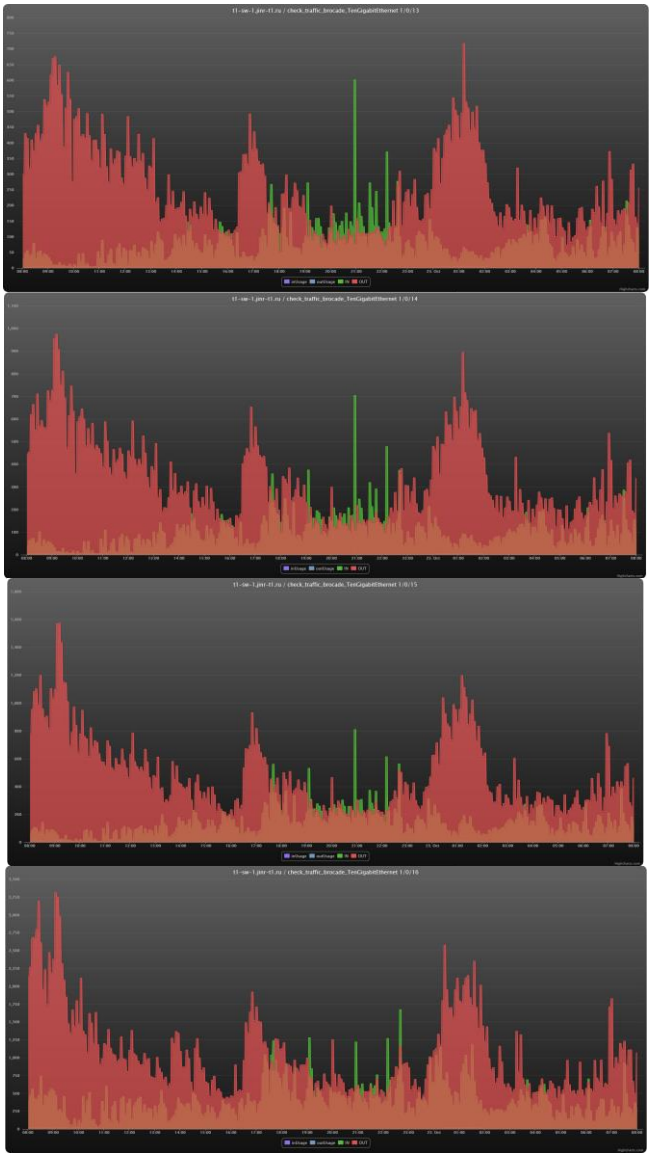

Fig. 7. Incoming/outgoing traffic of TenGigabit 013-016 network interfaces.

And again our research team has got confirmation that our assumption about equal weigh categories for this path is incorrect and the traffic via similar paths is communicated by unequal portions. On all 24 link, that were observe, data acquired ratio is 10%×10%×30%×50%. This exclude impact of source-destination-mac-ip address balance, because data spread in different ways.

5 CONCLUSION

As it has been mentioned before, the network architecture of the first module of the Tier1 data center in JINR is built on hardware supplied by Brocade company using the modern multichannel data transfer protocol TRILL. Obtained experimental data direct our

activities for the further research of the nature of traffic distribution in redundant topologies. Nowadays, all collected information is transferred to the manufacturer with several obvious questions.

How is the distribution done while transmitting packed data by four peer paths, provided that the conditions of the patent claims are not met? What will happen if the traffic via one of the communication channels achieves peak values? No answers yet.

Detail analysis of packet error rate on each of the four path did not reveal any violations. Works in this field are continued and we are developing a test bench to carry out a similar experiment using a traffic generator. It is possible that, in order to confirm the collected data, it will be necessary to build a similar network fabrics running via TRILL protocol on a manufacturer's equipment. The proposal for construction of the second data center module is received from Huawei. At present, its implementation possibility is being analyzed. In case of completing the harmonization of this matter, we should carry out a comparative functional analysis of the TRILL protocol in various data center modules by various manufacturers. This work has an immense importance because the collected data will be used for the construction of the Data Processing Center within the NICA (Nuclotron-based Ion Collider fAcility) megaproject [8].

References

1. N.S.Astakhov, A.S.Baginyan, A.G.Dolbilov, N.I.Gromova, I.A.Kashunin, V.V.Korenkov, V.V.Mitsyn, S.V.Shmatov, A.Strizh, V.V.Trofimov, N.N.Voitishin, V.E.Zhiltsov JINR Tier1-level computing system for the CMS experiment at LHC: status and perspectives / Computer studies and simulation, vol. 7, No 3, 2015. pp. 455−462.
2. Baginyan A.S., Dolbilov A.G., Korenkov V.V. Network for data-center Tier1 at JINR for experiment CMS (LHC) / T-Comm, Vol. 10, No.1, 2016. pp. 25-29.
3. E. W. Dijkstra A note on two problems in connexion with graphs/ Numerische Mathematik 1, 269 - 271 (l959)
4. Touch, J. and R. Perlman, "Transparent Interconnection of Lots of Links (TRILL): Problem and Applicability Statement," RFC 5556, May 2009.
5. Perlman, Radia An Algorithm for Distributed Computation of a Spanning Tree in an Extended LAN / ACM SIGCOMM Computer Communication Review. 15 (4): 44–53 1985.
6. Baginyan A.S., Dolbilov A.G. TCAM - from Ipv4 to Ipv6 / T-Comm, No.4, 2013. pp. 24-28.
7. J. Case, K. McCloghrie, M. Rose, S. Waldbusser RFC 1448 – Protocol Operations for version 2 of the Simple Network Management Protocol / April 1993.
8. N.S.Astakhov, A.S.Baginyan, S.D.Belov, A.G.Dolbilov, A.O.Golunov, I.N.Gorbunov, N.I.Gromova, I.S.Kadochnikov, I.A.Kashunin, V.V.Korenkov, V.V.Mitsyn, I.S.Pelevanyuk, S.V.Shmatov, T.A.Strizh, E.A.Tikhonenko, V.V.Trofimov, N.N.Voitishin, V.E.Zhiltsov JINR Tier1 center for the CMS Experiment at LHC / Particles and Nuclei, Letters, vol.13, No 5, 2016. pp. 1103-1107.
9. V.V.Korenkov, A.V.Nechaevskiy, G.A.Ososkov, D.I.Pryahina, V.V.Trofomov, A.V.Uzhinskiy Simulation concept of NICA-MPD-SPD Tier0-Tier1 computing facilities / Particles and Nuclei, Letters, vol.13, No 5, 2016, Pp.1074-1083.G. O. Young, “Synthetic structure of industrial plastics (Book style with paper title and editor),” in *Plastics*, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 15–64.