# Managing path diversity in Layer 2 critical networks by using OpenFlow

Elias Molina, Jon Matias, Armando Astarloa and Eduardo Jacob
University of the Basque Country (UPV/EHU)
Alda. Urquijo s/n, 48013 Bilbao (Spain)
{elias.molina, jon.matias, armando.astarloa, eduardo.jacob}@ehu.es

***Abstract*—Critical environments demand redundant networks to achieve high availability. Also, many industrial applications have stringent latency requirements that must be met by the Ethernet mesh networks in which they are supported. However, redundant resources are usually used as backup solutions, being underutilized most of the time. This paper analyzes the relation between the network meshing and load balancing with the latency. As a representative example, it is shown that network management in modern substation automation systems can be improved through the Software-Defined Networking (SDN) paradigm. In contrast to spanning tree-based networks, this paper describes how the OpenFlow technology is used to control Local Area Networks (LANs) to make the most of redundant topologies. Thus, an external controller provides flow-aware load balancing that impacts directly latency reduction, meeting the IEC 61850 requirements. Through emulation it is studied how data flows in IEC 61850-based substation communication systems are balanced as required.**

## I. Introduction

Increasing resilience is an essential objective of industrial applications, making it necessary to deploy redundant nodes and links to reduce the interruption time in case of failure. It is common for these safety critical environments, such as transportation or automation and control systems, to be based on the Ethernet technology. The importance of determinism in such networks can be understood with emerging efforts such as the IEEE 802.1 Time Sensitive Networks (TSN) Task Group [1], which is focused on providing low and guaranteed latency, as well as active redundancy to ensure seamless communication. Nevertheless, nowadays Layer 2 networks are usually based on traditional spanning tree protocols that prevent loops at the expense of disabling redundant paths. Namely, they are configured in active-passive mode, so that redundant links are used as spare paths. This means that traffic flows cannot be balanced or forwarded along the shortest paths, so that network resources are not efficiently utilized.

In automation networks, the design has to meet the requirements of flows with different priorities, such as maximum latency or minimum throughput. In particular, the IEC 61850 is a standard for the design of "communication networks and systems for power utility automation". Among other topics, this standard defines protocols to transmit monitoring and control data, which pose stringent performance requirements on the underlying LANs. For example, in a publisher-subscriber model where a device continuously sends a sequence of control signals and status measurements, IEC 61850 defines the Sampled Values (SV, IEC 61850-9-2) and the Generic Object Oriented Substation Event (GOOSE, IEC 61850-8-1) services. Both of them are carried as payload of link layer frames, without a TCP/IP stack. Hence, SV and GOOSE are suitable to communicate Merging Units (MU) and Intelligent Electronic Devices (IEDs: sensors, relays, meters, etc.) as fast as possible. Otherwise, IEC 61850-8-1 also includes the Manufacturing Message Specification (MMS) to exchange general information. These messages are not as low-latency demanding as the previous ones, being transported over TCP connections.

The network engineering guidelines for these infrastructures are detailed in IEC TR 61850-90-4 [2], which analyzes the behavior of robust systems, including redundant trees, rings or mesh topologies. Regarding latency, on the one hand, maximum transmission times required for different services are indicated in IEC 61850-5 [3]. For example, the most stringent services, such as SV raw data or trip messages to open a breaker, demand 3 ms as transmission limit; whereas MMS may require up to 100 ms for "medium speed" messages. On the other hand, according to IEC 61850-10, 20% of the total transmission time is reserved for network latency, being the remaining 80% related to the processing times at the sender and receiver.

The network latency is affected largely by the Layer 2 control technologies, including redundancy and multihoming ones, which are summarized in Section II. In this paper, a combination of load balancing and multipath techniques takes advantage of partial mesh topologies to improve communication performance. Specifically, the SDN approach allows us to dynamically control network resources and set several data paths between source and destination. Thus, lower latencies are achieved, which is an essential requirement in critical, time-sensitive applications.

## II. Redundancy and multihoming

*Spanning tree protocols:*
As noted above, spanning tree is the most common technique for handling the redundancy in Layer 2 networks. These protocols avoid the appearance of loops in Ethernet networks by making certain links into passive resources that are activated in the event of a network element failure. This prevents the packets from being always forwarded through the shortest path, which directly affects the latency.

There are several spanning tree algorithms, either standard or proprietary, such as for example the Rapid Spanning Tree Protocol (RSTP, IEEE 802.1D), MSTP (IEEE 802.1s), which enables multiple instances of Spanning Tree Protocol

978-3-901882-77-7 

per Virtual LAN, or Rapid Per-VLAN Spanning-Tree Plus (RPVST+) developed by Cisco.

*Link aggregation and multihoming:*
Since network design must ensure high availability of mission-critical applications, multihomed devices are appropriate. That is, end hosts with multiple interfaces, one or more of which may be active.

On the one hand, Ethernet channel bonding allows multiple physical interfaces to bundle into one logical link. Therefore, bonding provides resilience between ports in case of a link failure, as well as load balancing that increases bandwidth. Several technologies allow nodes to use multiple links jointly, such as, for example, the Link Aggregation Control Protocol (LACP, IEEE 802.3ad) or proprietary Multi-Chassis Link Aggregation (MC-LAG) implementations. The latter ones avoid a single point of failure by aggregating the capacity of multiple switches, thereby requiring a synchronization protocol between switches. On the other hand, the Parallel Redundancy Protocol (PRP, IEC 62439-3) is a redundancy active method that, implemented in end nodes, achieves seamless communication by duplicating data in two networks simultaneously. It has been selected as suitable [2] redundant technique for protecting SV and GOOSE services, since it guarantees zero recovery time in case of single network failure.

*TRILL and IEEE 802.1aq protocols:*
The new protocols, Transparent Interconnection of Lots of Links (TRILL) and IEEE 802.1aq, rely on a Layer 2 link state protocol to improve the Ethernet control plane. Both technologies enable shortest path forwarding in a mesh topology by calculating a hash (based on e.g. Ethernet addresses, IP addresses and TCP/UDP port numbers) of the packets. Also, using IEEE 802.1aq has been recently proposed in industrial networks [4]. Nevertheless, with IEEE 802.1aq and TRILL protocols multipathing can only be carried out through equal cost paths. Moreover, according to [5], another disadvantage of the hashing technique is that "usually all links get the same percentage of the hash values and therefore all the paths need to have the same capacity". SDN meets the lack of programmability in these networking architectures.

## III. OpenFlow as a multipath flow-based substrate

In line with the SDN paradigm, the OpenFlow technology enables data and control plane decoupling. An controller establishes, via OpenFlow protocol, the forwarding rules for flows arriving at a switch. These rules are based on a priority value, packet headers and instructions that define the data path. In this way, network programmability can be achieved reactively or proactively. In the former the OpenFlow controller set forwarding rules in response to requests from switches, whereas in the latter rules are preinstalled.

Regarding the use of OpenFlow as control plane of critical infrastructures, challenges and benefits for a SDN-enabled Smart Grid communication network have been discussed qualitatively in [6]. For instance, in previous work [7], an SDN approach has been proposed to implement several network features in IEC 61850-based systems, such as traffic filtering or Quality of Service (QoS). Also, in [8] OpenFlow is used together PRP to establish multiple active paths, resisting multiple failures without interruption. Furthermore, the authors of [9] propose to allocate bandwidth dynamically, and based on the propagation time, in industrial networks via OpenFlow.

As network conditions change, a proper critical infrastructure design must be well adaptable to provide the required performance, as long as they meet specific application requirements. For example, in contrast to proposals based on MSTP static configurations [10], SDN provide flexibility so that non-critical flows can be forwarded reactively with the aim of using underutilized routes. Latency and jitter are affected by the number of hops and amount of traffic in Layer 2 networks. Taking into consideration the importance of latency in time-sensitive environments, its reduction due to this approach is discussed below.

### A. Ethernet network latency

There are several sources of latency of an Ethernet network. Although other parameters, such as QoS policies (priority queuing, classification of traffic, etc.), also affects overall latency, the main ones for a switch are the following:

- Physical paths (either copper, fiber or radio links): elapsed time frame to traverse the physical medium.
- Store-and-forward latency, defined as *last bit in first bit out*. Otherwise, according to [2], using cut-through (bit forwarding) "reduces average latency but does not improve its worst case".
- Switch port-to-port latency: delay incurred by frames traversing the switch fabric.
- Queuing latency, which depends on the traffic pattern, and the output scheduling policy.

These latencies are calculated for SV, GOOSE and MMS frames in [11], whereas an analytical calculation of the worst case latency for Ethernet frames is detailed in [12]. In the case of an OpenFlow switch, these latencies would be the same when flow rules are already installed. It should be kept in mind that latency-critical data forwarding must be carried out proactively.

*Number of hops and connectivity:*
Obviously, these factors are multiplied by the number of switches that a data has to traverse to reach the destination. That is to say, latency increases with the number of switches in series. Therefore, performance and real time response depends on the number of hops per path. Concretely, given a connected, undirected graph ($G$), the average shortest path length is

$$a = \sum_{s,t \epsilon V} \frac{d(s,t)}{n(n-1)}$$

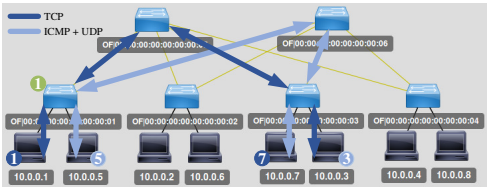
where $V$ is the set of nodes in $G$, $d(s,t)$ is the shortest path from $s$ to $t$, and $n$ is the number of nodes in $G$.

An OpenFlow controller usually discovers the network topology by means of the Link Layer Discovery Protocol (LLDP). Therefore, a controller can set up shortest paths in contrast to spanning tree-based topologies, which implies a latency improvement associated with the number of hops. Thus, Table I contains a comparison of average path length ($a$) for different topologies: rings, two-tier and three-tier designs, grids and full mesh topologies. All except the last two of these networks can be found in data center, campus infrastructures and industrial automation networks. Grids and full mesh topologies have been included to complement the results of [4],

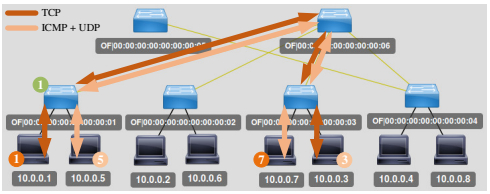
| Topology | Size | Spanning tree | | Entire redundant network | |
|---|---|---|---|---|---|
| | | $a$ | $k$ | $a$ | $k$ |
| Ring | 5 | 2 | 1:2.0, 2:1.67 | 1.5 | 2:2 |
| | 10 | 3.67 | 1:2.0, 2:1.88 | 2.78 | 2:2 |
| | 15 | 5.33 | 1:2.0, 2:1.92 | 4 | 2:2 |
| Two-tier [core, edge] | 2,2 | 1.67 | 1:2, 2:1.5 | 1.33 | 2:2 |
| | 2,4 | 1.87 | 1:3.5, 2:2.5, 4:1.25 | 1.47 | 2:4, 4:2 |
| | 4,8 | 2.15 | 8:1.38, 1:6.8, 4:2.75 | 1.52 | 8:4, 4:8 |
| Three-tier [core, aggregation, edge] | 2,2,4 | 1.93 | 1:5.33, 2:3.5, 6:1.16 | 1.86 | 1: 6.0, 2:3.32, 6:1.33 |
| | 2,4,8 | 2.24 | 8:1.63, 1:7.167, 6:2.17 | 2.09 | 8:2.5, 1:6.0, 2:8.0, 6:3.33 |
| | 2,4,16 | 2.23 | 16:1.31, 1:13.5, 6:3.5 | 2.10 | 16:2.25, 1:6, 2:16, 6:6 |
| Grid | 3x3 | 2.61 | 1:2.25, 2:2, 3:2.17 | 2 | 2:3, 3:2.67, 4:3 |
| | 4x4 | 4.1 | 1:2.33, 2:2.33, 3:2 | 2.67 | 2:3, 3:3, 4:3.5 |
| | 5x5 | 4.85 | 1:2.91, 2:2.81, 3:2, 4:1.92 | 3.33 | 2:3, 3:3.11, 4:3.67 |
| Mesh | 5 | 1.6 | 1:4, 4:1 | 1 | 4:4 |
| | 10 | 1.8 | 1:9, 9:1 | 1 | 9:9 |
| | 15 | 1.87 | 1:14, 14:1 | 1 | 14:14 |

TABLE I: Average shortest path length and connectivity for different networks[a].

[a]The number of end hosts has not been considered in the calculation, therefore only the distribution nodes are taken into account.




(a) Topology with all links enabled and load balancing.




(b) RSTP configuration (distribution switch 6 is root).

Fig. 1: Two-tier network topology discovered by the OpenDayLight controller and flow paths.

where the authors obtain the minimum and maximum number of hops in IEEE 802.1aq mesh networks without comparing them with spanning tree-based ones.

Since the spanning-tree construction allows each host interface to communicate with another one along a single path, multipath connections and load balancing are infeasible. Table I also includes the average degree connectivity, which is the average nearest neighbor degree of nodes with degree $k$. Each connection can be weighted by network characteristics (ie, bandwidth). Thus, for a node $i$,

$$k^w_{nn,i} = \frac{1}{s_i} \sum_{j \epsilon N(i)} w_{ij} k_j$$

where $s_i$ is the weighted degree of node $i$, $w_{ij}$ is the weight of the edge that links $i$ and $j$, and $N(i)$ are the neighbors of node $i$. Table I collects each degree $k$ with the value of average connectivity for unweighted networks. For example, as can be seen, a full mesh topology becomes a star one when it is configured by a spanning tree protocol.

*Network load:*
With our proposal, traffic can be spread among multiple paths of any length, including the shortest one. The partial network load is reduced as a consequence. Paying attention to the queuing latency, in [11] and [12] the average latency due to queuing ($L_Q$) "is assumed to be" directly proportional to the network load ($Load_{Network}$), as indicated by the following formula:

$$L_Q = Load_{Network} * L_{SF(max)}$$

Where $L_{SF(max)}$ is the store and forward latency, which is the ratio of the size of the frames and the bit-rate capacity.

## IV. Tests and Results

Experimentally, the traffic performance of a sampled value process bus is analyzed in [13], focusing on the maximum number of connected devices sending SVs without packet loss. The authors state that "once sampled value frames are queued by an Ethernet switch, the additional delay incurred by subsequent switches is minimal". However, the mentioned paper does not take into consideration a possible variation in the network load or load balancing techniques. The following validation tests compare the effect of traffic interactions when frames are balanced dynamically.

### A. Test bed configuration

The test bed uses several open source tools, as listed below:

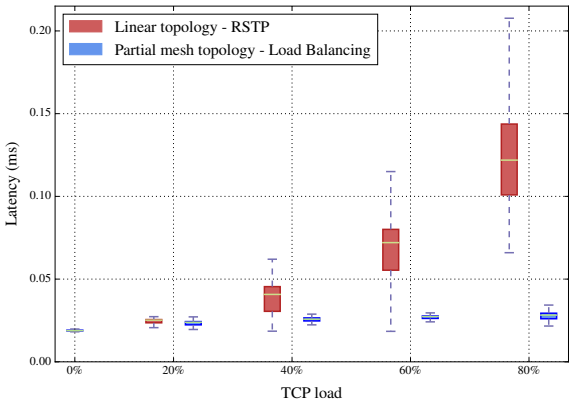
- Mininet software serves to emulate "networks, running real kernel, switch and application code, on a single machine". In addition, NetEm and Linux Traffic Control tools allow us to configure parameters of links and packet queues on interfaces.
- Open vSwitch (OVS) is used as OpenFlow software switch.
- The OpenDaylight project is used as OpenFlow controller. Specifically, the *MultipathODL* fork [14] enables Link-layer MultiPath Switching. It calculates multiple link disjoint paths per flow that are exposed as a path-finder service. Moreover, *MultipathODL* includes multipathing reactive flow handling that pushes, with each new flow, the forwarding rules to all switches on the paths. In addition, it provides different path calculators and selectors, for example: "shortest path, random path, round robin path, maximum available bandwidth path, path with fewest flows or path with highest capacity". These selectors can be chosen via REST API.

A two-tier network topology, widely adopted in real critical infrastructures, is used in the following tests, where end hosts and switches are connected through 100 Mbit/s full-duplex interfaces. Figure 1 shows two screenshots of the OpenDayLight web interface: the entire setup is given in Figure 1a, whereas Figure 1b displays the same topology, but where the switches run RSTP, proving the reduction in the number of available links.

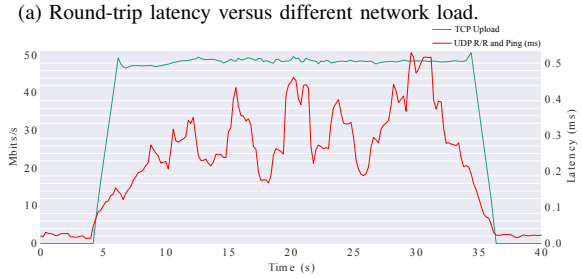
### B. Interaction testing

The latency impact when different data flows are transmitted at the same time is evaluated by generating ICMP, UDP and TCP streams, which may typify the interaction of critical real-time data and non-critical background ones. In this experiment, two incoming ports in switch 1 receive the following traffic:

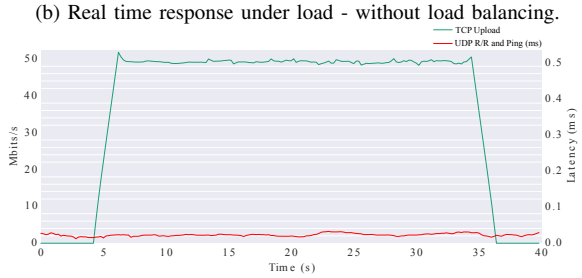
- Flows UDP and ICMP are generated through *Netperf* and *ping* tools, respectively. UDP requests/responses and *ping* allow us to measure the round-trip latency, computed as the average of both services.
- A TCP connection injected with the *iperf* tool at different data rates. The data rates are limited by the output link capacities in the sender, and the edge switch has also installed an ingress policing rule to limit the maximum rate.




(a) Round-trip latency versus different network load.




(b) Real time response under load - without load balancing.




(c) Real time response under load - with load balancing.

Fig. 2: Latency comparison with and without load balancing.

Nodes *1* and *5* communicate with *3* and *7*, respectively, as depicted in Figure 1. The size of all packets transmitted by *1* are fixed to 126 bytes, which is in compliance with the 61850-9 SV-LE specification; while frames sent by *5* are fixed to 300 bytes, a typical value for MMS services [2]. Load balancing is based on destination MAC addresses so that, since ICMP messages and UDP datagrams provide no protection from duplication and no guarantees for delivery, traffic behavior is comparable to SV/GOOSE frames.

Figure 2a shows box plots that summarize the distributions of the round-trip latency for different data rates of TCP background connections (0, 20, 40, 60 and 80 Mbit/s). These statistics have been performed in 30 trials, each 10 seconds long. As a result, it can be said that *1*-to-*3* transmission results in significant lower latency when traffic is balanced compared to the non-balanced case. Moreover, Figures 2b and 2c show two experiments where *5* establishes a TCP connection that tries to consume all the bandwidth, limited in this case to 60 Mbit/s. These graphs compare the TCP goodput versus the *ping* and UDP request/response performance with and without load balancing.

As a summary, detailed experiments outline that latency grows quickly as TCP connection is set up for a linear topology, whereas latency is slightly affected in the balanced case. This load balancing can be configured statically, assigning links to critical and non-critical flows, or dynamically based on parameters such as the bandwidth available on the path.

## V. Conclusion

Despite the fact that critical networks have redundant resources, they are usually underused as active-passive configurations, mainly through spanning tree-based protocols. In order to facilitate compliance with strict time-sensitive requirements, a software-based control plane has been presented. Concretely, OpenFlow is used for efficiently exploiting multiple paths simultaneously, which affects the latency in time-sensitive scenarios. The IEC 61850 standard has been chosen as a validation case, which entails deploying robust topologies that fulfill latency and recovery time requirements.

The latency reduction has been illustrated using both an analytical approximation and emulation tests. As demonstrated, among other actors that affect performance, latency is reduced via load balancing as traffic spreading reduces the traffic interaction and network load.

## Acknowledgment

This research was partly funded by the Spanish Ministry of Economy and Competitiveness under the "Secure deployment of services over SDN and NFV based networks" project S&NSEC TEC2013-47960-C4-3-P. This work was produced within the Training and Research Unit UFI11/16 supported by the UPV/EHU. Furthermore, the authors would like to acknowledge the ZABALDUZ Program.

## References

[1] 802.1 TSN, "Time-Sensitive Networking Task Group," IEEE, 2015.

[2] IEC TC57, *Communication networks and systems in substations Part 90-4: Network engineering guidelines*, 2013, IEC/TR 61850-90-4.

[3] IEC TC57, *Communication networks and systems for power utility automation - Part 5: Communication requirements for functions and device models*, Geneva, January 2013, IEC 61850-5.

[4] P. Ferrari, A. Flammini, S. Rinaldi, G. Prytz, and R. Hussain, "Multipath redundancy for industrial networks using IEEE 802.1aq Shortest Path Bridging," in *Factory Communication Systems (WFCS), 2014 10th IEEE Workshop on*, 2014.

[5] R. van der Pol, M. Bredel, A. Barczyk, B. Overeinder, N. van Adrichem, and F. Kuipers, "Experiences with MPTCP in an intercontinental OpenFlow network," in *29th TERENA Network Conference*, 2013.

[6] X. Dong, H. Lin, R. Tan, R. K. Iyer, and Z. Kalbarczyk, "Software-Defined Networking for Smart Grid Resilience: Opportunities and Challenges," in *1st ACM Workshop on CPS Security*, 2015.

[7] E. Molina, E. Jacob, J. Matias, N. Moreira, and A. Astarloa, "Using Software Defined Networking to manage and control IEC 61850-based systems," *Computers and Electrical Engineering*, vol. 43, pp. 142–154, 2015.

[8] E. Molina, E. Jacob, J. Matias, N. Moreira, and A. Astarloa, "Availability Improvement of Layer 2 Seamless Networks Using OpenFlow," *The Scientific World Journal*, Jan 2015.

[9] H. Miyata, M. Namiki, and M. Sato, "pmqFlow: Design of propagation time measuring QoS system with OpenFlow for process automation," in *Industrial Electronics Society (IECON), 40th IEEE Conference*, 2014.

[10] D. Ingram, P. Schaub, R. Taylor, and D. Campbell, "Network Interactions and Performance of a Multifunction IEC 61850 Process Bus," *Industrial Electronics, IEEE Transactions on*, vol. 60, no. 12, pp. 5933–5942, 2013.

[11] X. Xu and Y. Ni, "Analysis of networking mode caused by GOOSE delay of smart substation," in *Software Engineering and Service Science (ICSESS), 4th IEEE International Conference on*, 2013.

[12] Azarov, Max, *Worst-case Ethernet Network Latency for Shaped Sources*, December 2005, IEEE 802.3 ResE study group.

[13] D. Ingram, P. Schaub, R. Taylor, and D. Campbell, "Performance Analysis of IEC 61850 Sampled Value Process Bus Networks," *Industrial Informatics, IEEE Transactions on*, vol. 9, no. 3, pp. 1445–1454, 2013.

[14] Julian, Bunn, *Experience with the OpenDaylight Controller in a multi-vendor 1 Tbps network*, December 2014, Caltech.