PROCEEDINGS OF SCIENCE

The Software Defined Networks implementation for the KM3NeT networking infrastructure

Tommaso Chiarusi *
INFN-Sezione di Bologna, Viale Berti-Pichat 6/2, 40127, Bologna, Italy
E-mail: tommaso.chiarusi@bo.infn.it

Lorenzo Chiarelli
GARR, Via dei Tizii, 6 - 00185 Roma, Italy
INFN-CNAF, Viale Berti-Pichat 6/2, 40127, Bologna, Italy
E-mail: chiarelli@garr.it

Emidio Giorgio
INFN-Laboratori Nazionali del Sud, via S. Sofia 62, 95123 Catania, Italy
E-mail: emidio.giorgio@infn.it

Stefano Zani
INFN-CNAF, Viale Berti-Pichat 6/2, 40127 Bologna, Italy
E-mail: stefano.zani@cnaf.infn.it

Silvia Celli†
Gran Sasso Science Institute, Viale Francesco Crispi 7, 67100 L'Aquila, Italy
INFN-Sezione di Roma, P.le Aldo Moro 2, 00185 Roma, Italy
E-mail: silvia.celli@roma1.infn.it

On behalf of the KM3NeT Collaboration

The Software Defined Networks technology is used to configure and operate the core of the switch fabric of the networking system of the KM3NeT shore infrastructures. It is organised according to a star-center layout, interconnecting all the on-shore resources. The used switches are DELL Series S devices, compatible with the OpenFlow 1.3 protocol and managed by dedicated OpenDaylight 4.2 (Beryllium) controller servers. With a limited number of Layer 2 forwarding rules, expressly developed for the KM3NeT use-case, which keeps independent from any scaling of the off-shore detector, the SDN technology allows the best handling of the KM3NeT asymmetric network topology, optimising the layout of connections on shore, preventing loops and enhancing the data taking stability.

35th International Cosmic Ray Conference — ICRC2017
10–20 July, 2017
Bexco, Busan, Korea

*Corresponding author
†Speaker.

PoS(ICRC2017)940

1. Introduction

KM3NeT is a distributed neutrino observatory in abyssal sites of the Mediterranean Sea [1] with two installations, ARCA and ORCA, 80 km off the Sicily coast (Italy) at the dept of 3500 m and 40 km off Toulon (France) at the depth of 2500 m, respectively. Both of them are based on a grid of thousands Digital Optical Modules (DOMs), interconnected to shore via an electro-optical seafloor infrastructure. The DOMs are organised in vertical structures, each one provided with a Base-module for controlling the power supply and optical amplification of the attached devices. Exploiting a custom FPGA-based White Rabbit kernel with Ethernet connectivity, the DOMs and Base-modules are submarine nodes of the global Layer 2 optical networking infrastructure [2], with two significant characteristics that make it unique with respect to the other DAQ networking installations: the *asymmetry* of the connections between the station and the detector and a custom implementation of the White Rabbit (WR) protocol [3] for synchronisng the detector DOMs and Base-modules, which exploits a *hybrid* layout of the onshore switching infrastructure.

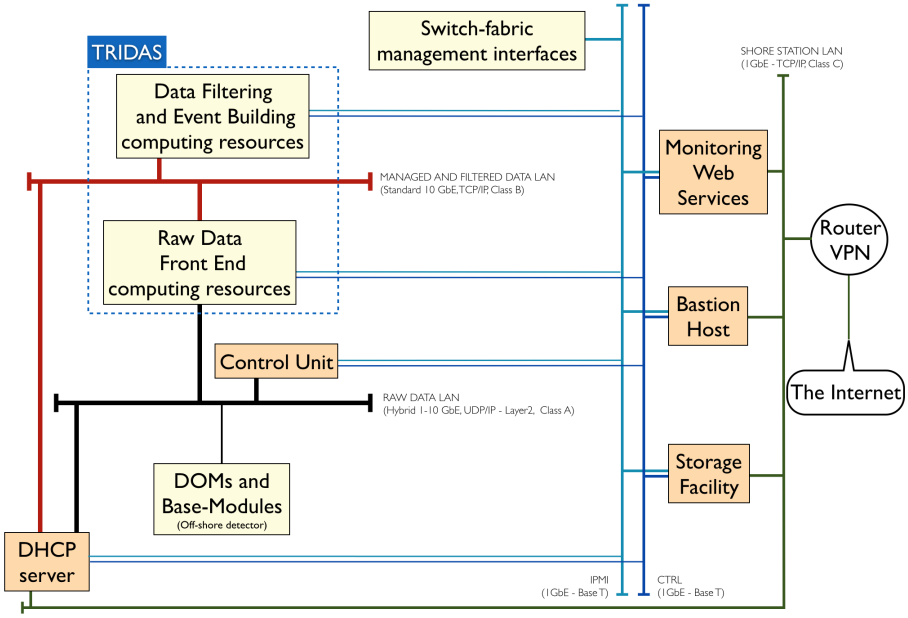

Figure 1: The KM3NeT network segments with the basic elements of the Trigger and Data Acquisition System and the other computing resources available at the shore-stations. See text for details.

2. The KM3NeT networking design: involved elements and constraints

The segments of the KM3NeT network are represented in the scheme sketched in Figure 1, together with all the relevant elements of the shore-station infrastructure. While a detailed description of the KM3NeT DAQ system can be found in [5], it is convenient here to highlight the role of the following components: the *Control Unit* (CU) [6] is the process running on a dedicated server (or on a group of servers) with the task to concertate the off-shore Detector and the on-shore DAQ facilities via a prefixed State Machine, optimising the data taking. The CU is the origin of the Slow

Control commands to the detector. The *Trigger and Data Acquisition System* (TriDAS) is the group of computing resources which manipulate and filter the incoming optical and acoustic data from the DOMs and Base-modules. The shore-station exposes the collected data, metadata and monitoring information through dedicated web-based services or via direct access to the storage facility. All the nodes in the network, including the off-shore elements, are properly configured via static definitons in the DHCP context. The bandwidth of each network segment is properly dimensioned in order to avoid oversubscriptions.

Among the five LAN segments sketched in Figure 1, we focus here on the *RAW DATA LAN*, which is the Layer 2 segment that interconnects the off-shore detector to the CU and TriDAS servers, on-shore. As anticipated, it has to comply with the two following important constraints:

- **Asymmetry of the connections**: the asymmetry directly originates after the so-called *optical broadcast* (or simply *broadcast*) architecture adopted for the global optical infrstructure. It aims at best exploiting the number of optical fibers contained in the many km-long electro-optical cable (EOC), which connects the shore stations with the detectors. In practice, all the information addressed from shore to the DOMs or to the Base-modules (essentially WR-PTP packets for timing and slow control UDP packets) are embedded in a single stream of data, propagating through a single fiber in the EOC. At any next branching step of the line, the optical signal is splitted and routed to all the final endpoints. Thus all the DOMs and Base-modules receive the same information, but only the device which is the actual destination process them. Such a redundant traffic, which is proportional to the number of endpoints is however very low ($\sim$10 kBps with 24 strings, i.e. $\sim$20 Bps/DOM), and the global incoming throughput per DOM or Base-module keeps well below the 1 Gbps bandwidth of the SFP transceiver onboard of the node under the sea .

- **Hybrid switching layout**: the asymmetry described above violates the requirements of the WR protocol. A customization of the White Rabbit Switch (WRS) infrastructure is then necessary. In the standard WR implementation, the master-slave relation is obtained via a point-to-point connection between two communicating devices, which mutually exchange WR-PTP packets for the time-synchronisation; in the KM3NeT context, it has been necessary to split the on-shore master functionalities in two cathegories: the WRS-Broadcast and the WRS-Level1 ones. The WRS-Broadcast and the WRS-Level1 share the mastership versus the off-shore endpoints. This architecture was developed by the Seven Solution Company [7] and released in 2014. It is based on WR version 3.3.1. A possible upgrade to more recent versions (4.2 or 5.05) is currently under evaluation.

  There is an important difference concerning how the Base-modules and the DOMs are handled. The WRS-Broadcast sends continuously the WR-PTP packets via the *broadcast* channel to all the nodes underwater, but only the Base-modules stream back all their data (i.e. WR-PTP, SC feedbacks and acoustic packets) to WRS-Level1 switches. It was proved [8] that the time synchronization of a DOM keeps stable by simply latching to the WR-PTP packets emitted from the WRS-Broadcast. It was then decided to modify the WR implementation on the DOM's FPGAs so that the DOMs do not reply with any WR-PTP packets. For this reason, the DOM returning connections can be set directly on standard switches.

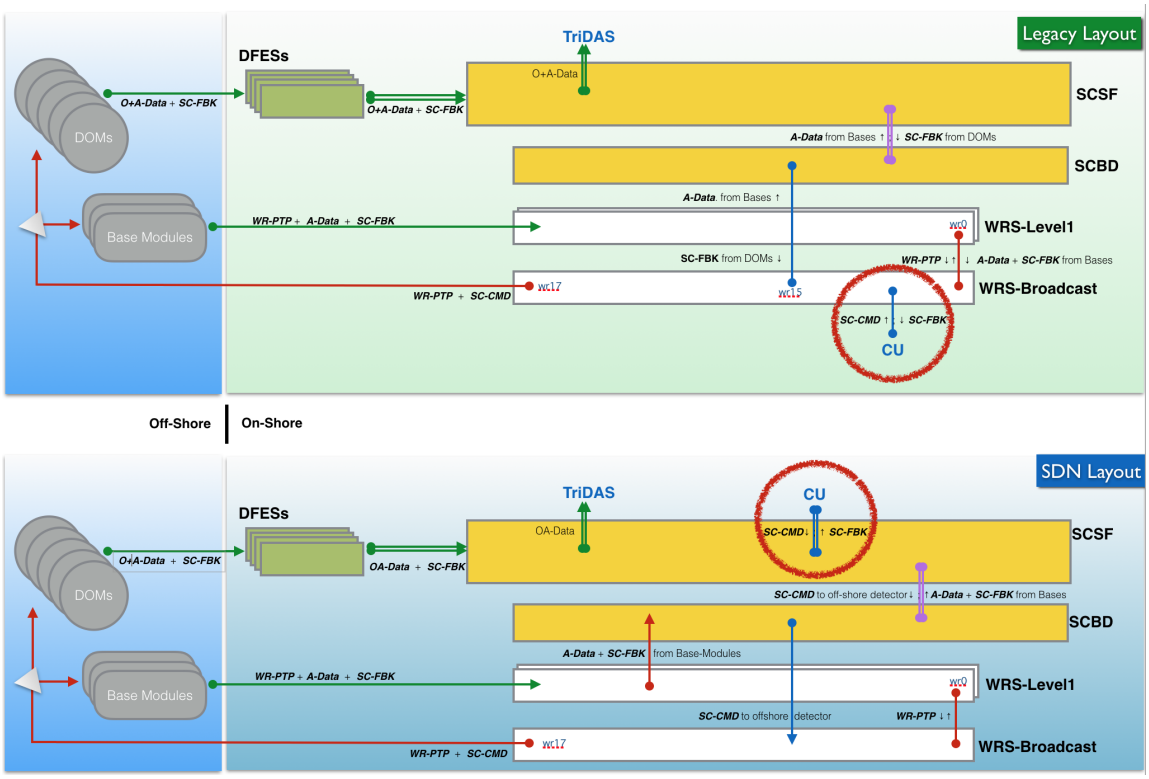

Figure 2: The two possible layouts for the network connections. Top: the *Legacy* layout with the CU connected directly on the WRS-Broadcast. Bottom: the SDN layout, with the CU connected on the SCSF. Single or double lines represent 1 GbE and 10 GbE connections, respectively. Single directional data-flows are indicated with arrows; bi-directional connections have only dots. The data flow notation is: *WR-PTP*: the WR implementation of the Precision Time Protocol (PTP); *SC-CMD, SC-FBK*: the slow control commands and feedbacks exchanged between the CU and the detector; *O + A-Data*: the optical and acoustic data, respectively. Note that the Base-modules don't produce any optical data.

2.1 Legacy and SDN connections layouts

According to such asymmetric and hybrid context, the global adopted switching infrastructure is composed of the following elements:

- the White Rabbit switch fabric, which includes a WRS-Broadcast plus a number of WRS-Level1 elements, dimensioned to the total number of Base-modules;
- the Standard Switch fabric, including: a number of *DOM Front End Switch* (DFES) elements, dimensioned to host the total number of DOMs; one *Star Center Switch Fabric* (SCSF) element, which interconnects the TriDAS resources and forward to them the optical and acoustic data flows; one *Slow Control and Base Data* (SCBD) element, which is responsible to connect the standard switch fabric with the White Rabbit one.

Figure 2 sketches the two connection layouts experienced with KM3NeT networking system: the top configuration is referred to as the *Legacy* layout, and it is based on standard working principles of the switches (i.e. no SDN). The bottom configuration refers to the SDN case. Refer to the caption for the notation used to classify the various data-flows.

In the Legacy scenario, holding the MAC-learning procedure common to all the switches, and willing to avoid any data flooding, the only possible location for the CU is on the WRS Broadcast. For similar reasons, it is not possible to set a connection from the WRS-Level 1 which takes care of the A-Data packets directly to the SCSF or to the SCBD. In fact it would generate a *network loop*. With the Legacy layout, the SC-FBK and A-Data data must pass through the WRS-Broadcast. As the detector scales to a larger number of DUs, the global SC and A-Data traffic generated by the Base-modules increases significanly, with the risk to compromise the WRS-Broadcast performance.

In order to preserve the stability of the WR infrastructure and consequently grant a reliable time-sinchronisation mechanism, it was mandatory to minimise the traffic passing through the WRS-Broadcast. This is now possible with the use of the Software Defined Networks, a frontieer technology applicable by modern high-level switches. By means of the specific OpenFlow protocol, it is possible to fix the relation between a particular data flow and the used port of the switch, via a simple Layer 2 routing. In this way it is possible to choose the most convenient connections topology, evolving it from the Legacy layout to the SDN scenario.

3. The Software Defined Networks implementation

The core of the SDN implementation is the OpenFlow protocol. It gives access to the forwarding plane of a network switch. In practice, OpenFlow allows the user to define the needed forwarding rules (called flows) and implement them into the SDN switch. A flow defines how the transient packets are forwarded from the input ports to the output ports. The flow/rules are essentially based on the source and destination MAC addresses contained in the packets. If a packet does not fulfill the rules, the packet is dropped. In the SDN scenario, the automatic MAC learning is not active. OpenFlow 1.3 was chosen in order to exploit some of its important features, like the masking for handling multiple MAC addresses within 1 single rule. Furthermore, OpenFlow 1.3 has the advantage that the injected flows are made persistent in the memory of the SDN switch. Not all the switches can implement the OpenFlow protocol. The KM3NeT Collaboration selected the DELL Series S which allows to fully exploit the OpenFlow features keeping the possibility to configure the speed and the autonegotiation of the in-band interfaces. This is important for correctly interfacing the DELL switches with the WR devices.

The choice of the DELL models follows up from the complexity of the experimental context: in the shore stations, dealing with a large number of off- and on-shore nodes, we selected the DELL S6000 (32 x 40 GbE QSFP interfaces) as SCSF and a DELL S3124F (24x 1 GbE SFP + 2x 10 GbE SFP+ interfaces) as SCBD, the latter working also as a media converter between the SCSF and the WR infrastructure. In test-stations, where the number of connected nodes is limited, we use the DELL S4048-ON (48x 10 GbE + 6x 40 GbE QSFP interfaces) for implementing both the SCSF and SCBD through two independent OpenFlow instances.

3.1 The SDN rules and the OpenDayLight Controller

To handle all the data-flows implied with the KM3NeT Data Acquisition, only a restricted number of SDN rules are required for both SCSF and SCBD instance. The rules developed for

KM3NeT-Phase1 (24 ARCA DUs and 7 ORCA DUs) are summarised in Table 1, where it must be assumed that the DHCP server for the RAW DATA LAN is implemented in the CU.

It is important to note that the number of rules does not depend on the size of the detector, i.e. on the number of DOMs and Base-modules, thanks to the fact that the MAC addresses of the off-shore nodes always have the *08:00:30* prefix. It increases, instead, with the number of TriDAS Front End resources, which is however limited. The maximum number of rules applicable with the full ARCA and ORCA detector sizes is expected not to be larger than 15.

Finally, failover procedures can require additional rules. The most effective one implements some drop actions, in order to temporarily remove unwanted data-flows.

All the rules are handled by a *Controller* service, which is an additional new element in the network. The Controlller has essentially the role to inject or remove the SDN-flows (i.e. the forwarding rules) into the SDN switches to which it is connected through the out-of-band ports. The rules management is done via a dedicated software called OpenDaylight (ODL) [9]. The currently used version is the ODL 4.0, *Beryllium*. ODL implements a REST API interface for manipulating the information related to the SDN. In particular, ODL implements the standards of the YANG data modelling language and, via a dedicated web-GUI, it provides the user with an intuitive way to define the various rules (in JSON format) and then send them to the SDN switches. One Controller suffices to drive both the OpenFlow-Instances. In case of accidental or intentional reload of the SDN instances, the Controller restores all the SDN flows corresponding to the predefined rules.

Auxiliary services, expressely developed for KM3NeT and interfaced to the Controller, allow to access via RESTCONF queries to the statistics information about the application of the various rules. This is relevant for monitoring the SDN system and possibly intervene, even automatically, with failover strategies.

Rule #	**Source MAC**	**Destination MAC**	**Action**
SCSF-1	any	broadcast (ff:ff:ff:ff:ff:ff)	to DHCP server
SCSF-2	08:00:30:00:00:00/ff:ff:ff:00:00:00	Control Unit	SC-FBK to CU
SCSF-3	08:00:30:00:00:00/ff:ff:ff:00:00:00	TriDAS Front End	O+A Data to TriDAS
SCSF-4	CU	08:00:30:00:00:00/ff:ff:ff:00:00:00	SC-CMD to SCBD
SCBD-1	08:00:30:00:00:00/ff:ff:ff:00:00:00	any	SC-FBK + A-Data to SCSF
SCBD-2	Any from uplink to SCSF	08:00:30:00:00:00/ff:ff:ff:00:00:00	SC-CMD to WRS-Broadcas
SCBD-3	08:00:30:00:00:00/ff:ff:ff:00:00:00	broadcast (ff:ff:ff:ff:ff:ff)	to SCSF

Table 1: The SDN rules for both SCSF and SCBD instances. The MAC address prefix *08:00:30* is always the same for DOMs and Base Modules, which differentiate for the next 3 bytes digits. The *Action* column summarises the flow handling as sketched in Figure 2, bottom.

4. Overview of the SDN performances in the ARCA shore station

There are different KM3NeT experimental contexts which make use of the official DAQ implementation: the Bologna Common Infrastructure (BCI) for the development and testing of all the aspects of the DAQ, the integration sites where the new DUs are validated before the deployment, and finally the production sites, i.e. the ARCA and ORCA shore stations. Every such context implements the SDN layout, with the SCSF and SCBD OpenFlow instances as presented in the previous sections. They all offer the advantage of multiple and independent validation of the DAQ system. The ARCA shore station, active since December 2015, actually motivated the adoption of the SDN for the first time. The networking infrastructure was initially compliant with the Legacy layout; annoying instabilities of the WR switch fabric limited the live-time of the experiment with the loss of time synchronisation, which required an almost daily reboot of the WRS-Broadcast. Figure 3 shows two periods of about 20 days each of monitored traffic passing through the SCBD

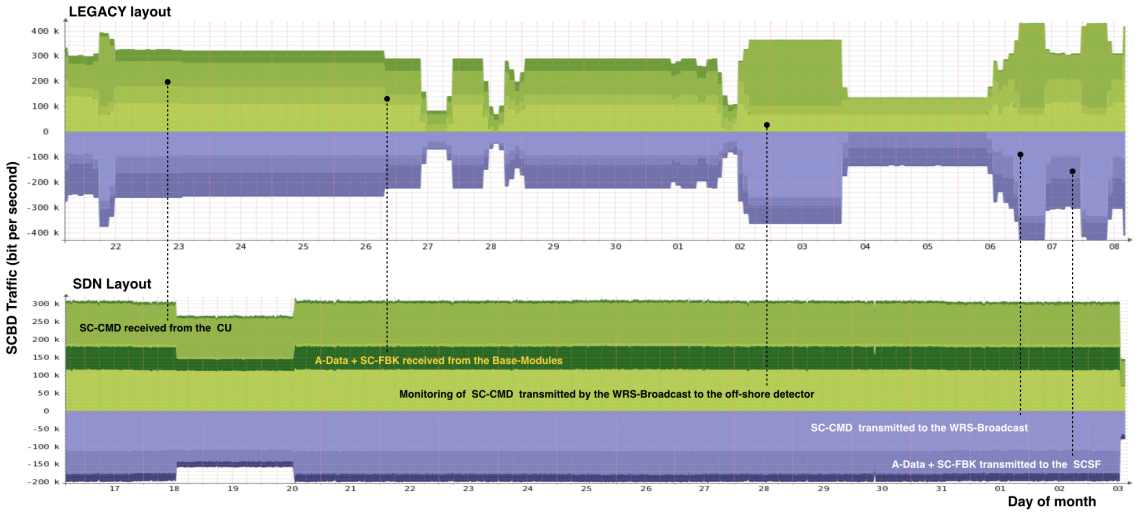

Figure 3: The comparison of the traffic through the SCBD during two different data-taking periods: top, with the Legacy layout; bottom: exploiting the SDN layout. Diagrams with shades of green and violet stand for incoming and outgoing throughputs in the SCBD, respectively.

instance, which is directly related to the WRS-Broadcast traffic. The top panel is referring to when the Legacy layout was in use, while the bottom panel concerns an example of the traffic with the SDN layout. The gain in stability with SDN is noticeable. It is worth to mention that the two temporary drops of traffic apparent in the SDN case are not due to network instabilites but to different running conditions.

In order to enhance the robustness of the full hybrid switch fabric, some optimisation was performed also on the White Rabbit devices. Essentially, most of the inerventions relate to static entries in the forwarding tables of both WRS-Broadcast and WRS-Level 1 together with some new characterisation of the interested interfaces for matching the data flows allowed by the SDN rules.

5. Conclusions and outlooks

The Software Defined Networks technology was succesfully introduced in the KM3NeT DAQ system. If not unique, it represents one of the few real use-cases of SDN implementation for experiments in the Hight Energy Physics field. The use of SDN highly improved the data taking stability and fixed important issues of the network, due to matching standard switches with White Rabbit devices.

Beside the extremely important positive results, SDN is still a frontier technology and lacks open-source application which allow an user-friendly configuration and maintenance. The current goal of the KM3NeT Collaboration is to consolidate the due expertise necessary to realise a custom framework, compliant with the OpenDaylight software, for better controlling and maintaining the networking infrastructure exploited by an experiment with decades of lifetime.

Acknowledgments

The authors want to thank Fabio Bellini and Marco Pinna from DELL for the discussions and the valuable technical support provided during the tests of SDN with DELL Series *S* and *N* models. An important contribution for better accomplishing the optimisation of the White Rabbit fabric with the SDN layout was given by Emilio Marin Lopez and David Martin from Seven Solutions.

References

[1] A. Heijboer, *KM3NeT, Physics and status* in these proceedings

[2] D. Real, D. Calvo, *Digital optical module electronics of KM3NeT*, *Phys. Part. Nuclei* **47**:918 (2016)

[3] The Open Hardware White Rabbit Group, *http://www.ohwr.org*

[4] The Open Networking Foundation, *https://www.opennetworking.org*

[5] C. Pellegrino and T. Chiarusi et *al*., *The Trigger and Data Acquisition System for the KM3NeT neutrino telescope* , *EPJ Web of Conferences*, **116** (2016) 05005

[6] C. Bozza et *al.*, *The Control Unit of KM3NeT data acquisition*, *EPJ Web of Conferences*, **116** (2016) 05001

[7] The Seven Solution Company, *http://sevensols.com/*

[8] M. Bouwhuis, *Time synchronization and time calibration in KM3NeT* in proceeding of *ICRC 2015*, PoS ICRC2015 (2016)

[9] The OpenDayLight Project, *https://www.opendaylight.org*