The Software Defined Networks in KM3NeT

Tommaso Chiarusi —— INFN Sezione di Bologna

on behalf of the **KM3NeT Collaboration**

- Collaboration of 51 institutions in 15 countries

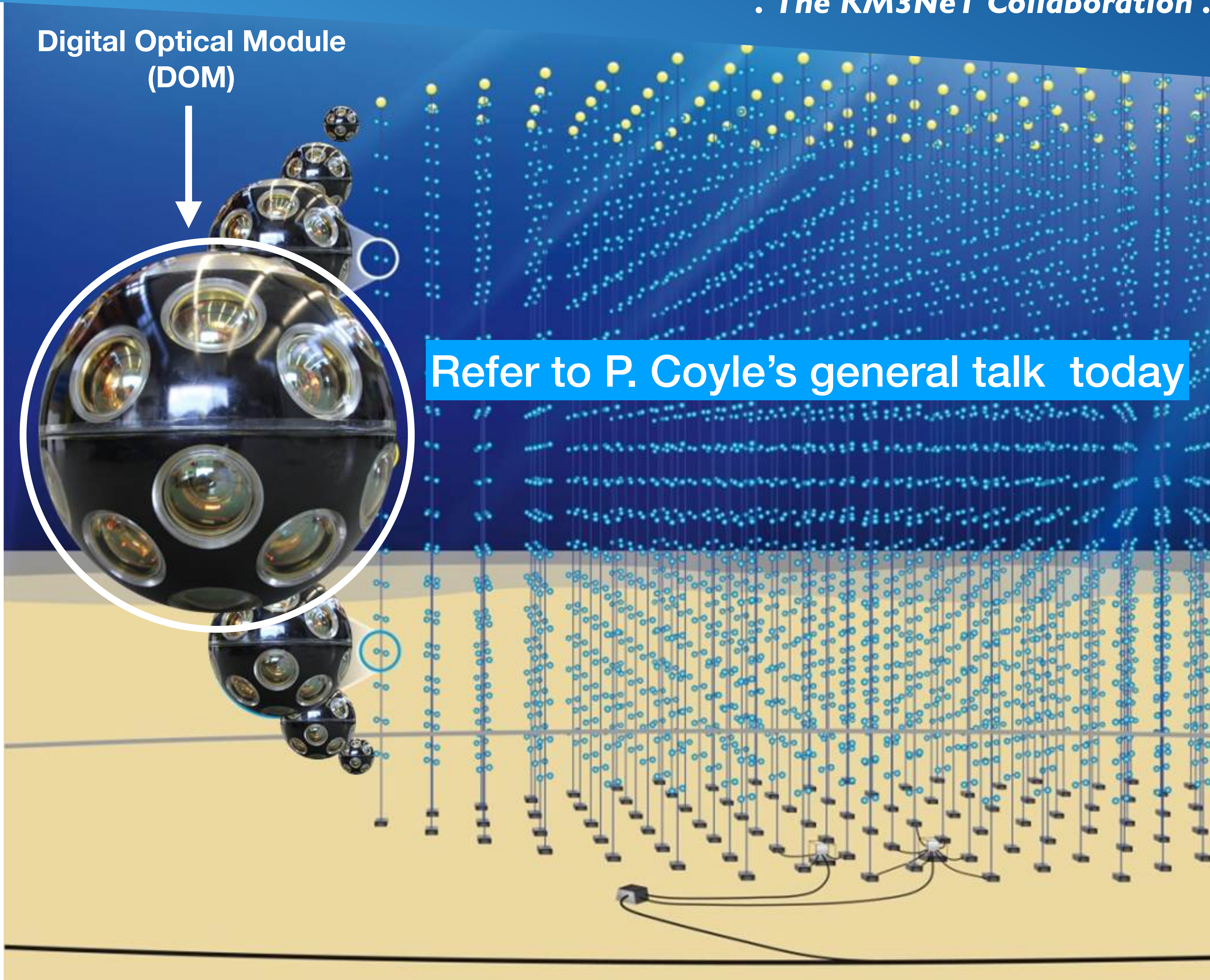
- 2 submarine detectors :
  - **ARCA** (Portopalo) - 3500m u.s.l.
  - **ORCA**(Toulon) - 2500m u.s.l.

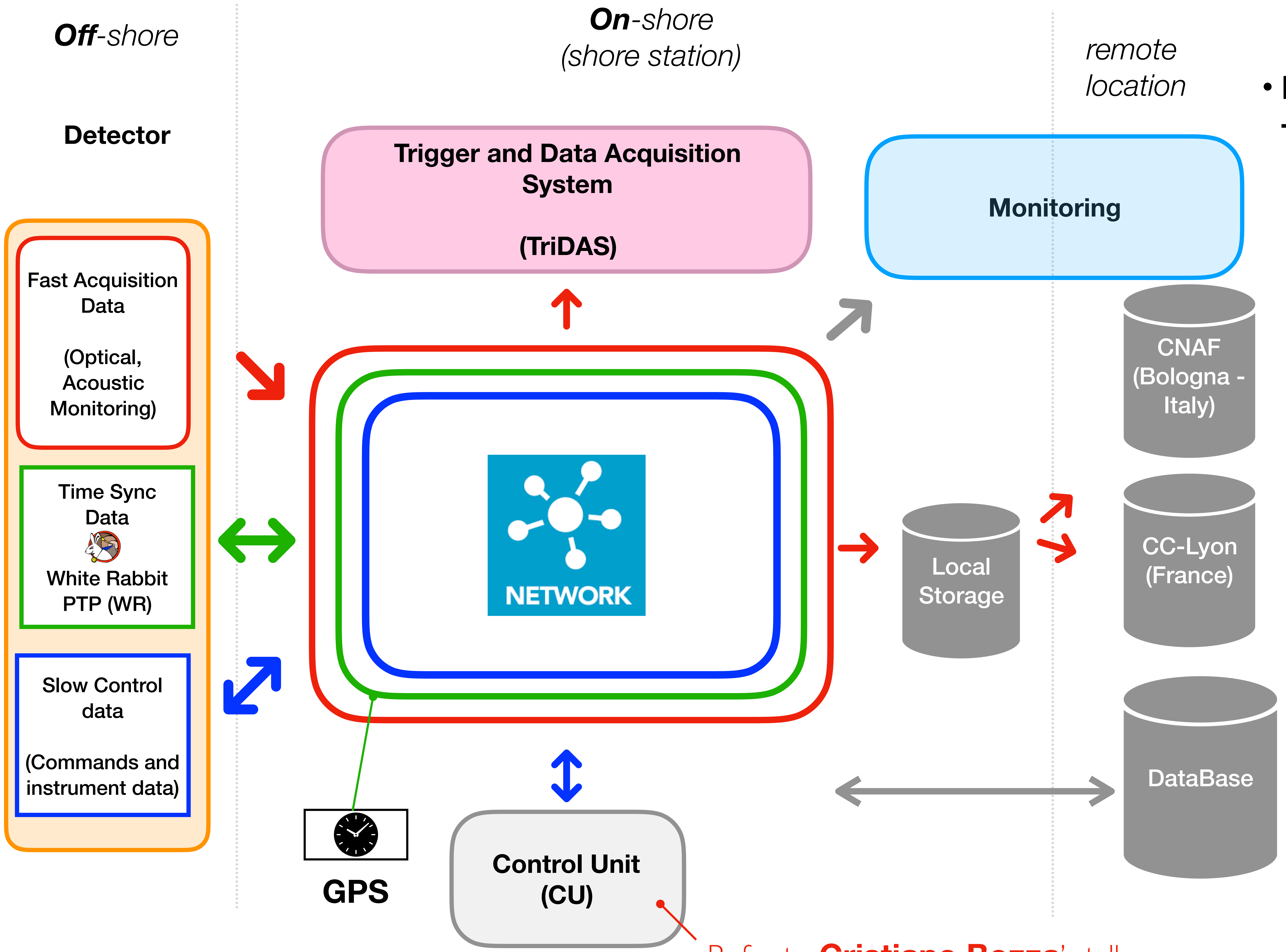
- Building Blocks (2 ARCA, 1 ORCA)
  - 115 Detection Unit (DU)
    - 18 DOM + 1 Base Module/DU
      - 31 x 3"PMT/DOM

$\Longrightarrow$ **2185 nodes / BB**

- All data to shore DAQ (see Ronald Bruijn's talk)

KM3NeT 2.0 Letter of Intent:
(arXiv:1601.07459)
J.Phys. G43 (2016) 084001





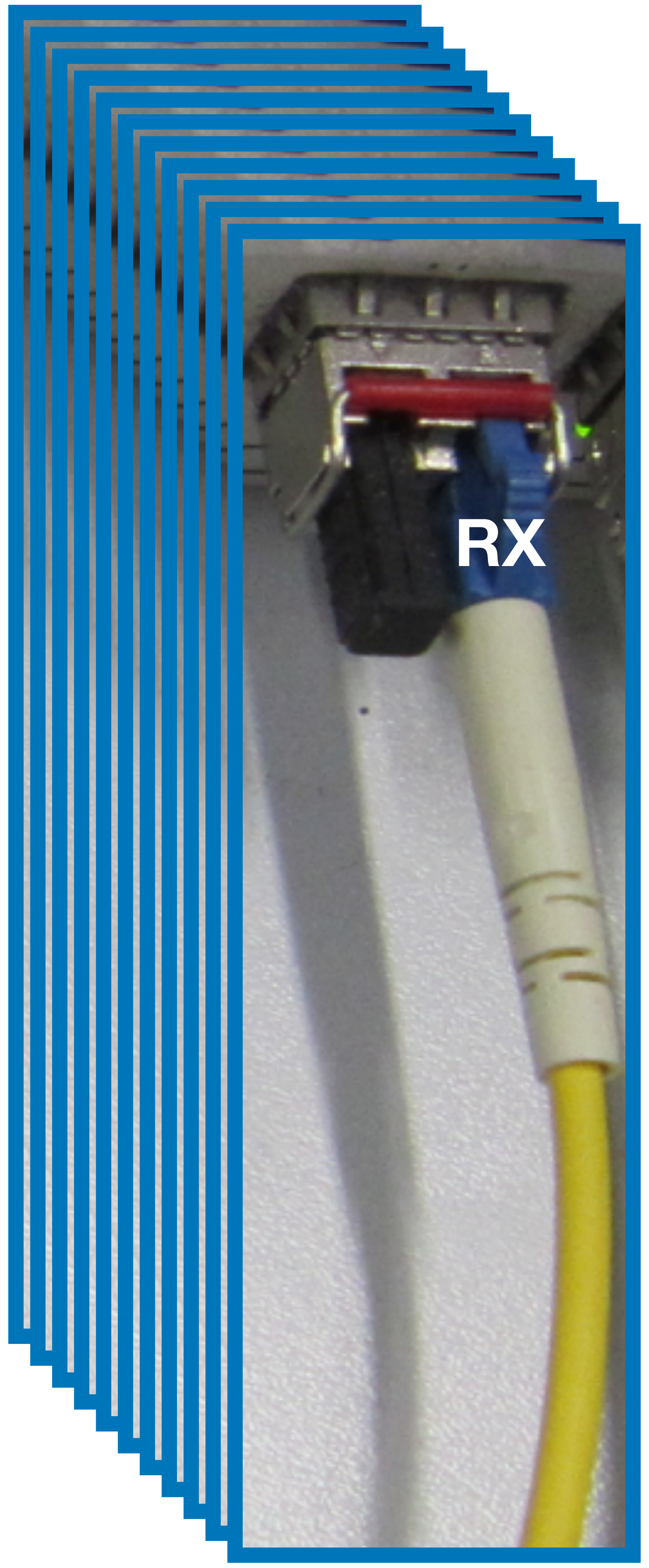
Refer to **Cristiano Bozza**'s talk

- DOMs (and DU bases) bitstream is processed by **TriDAS**
  - *DataQueues* reassembles ethernet frames sent by DOMs
  - *DataFilter* applies online trigger and selects ''good'' events
- **CU** is process orchestrating on/off-shore resources through SlowControl commands (**SC**)
- **Type of network streams:**
  - O- , A-, M-Data: optical, acoustic and monitoring data, respectively (no optical data for Base-modules)
  - SC-CMD, SC-FBK: the slow control commands and feedbacks exchanged between the CU and the detector;
  - WR-PTP: for the time synchronisation

Network Asymmetry

1 sender

Multiple receivers on separated switches



Hybrid Infrastructure

White Rabbit Switches
(for time synchronisation)
1 GbE
Master / Slave (DOMs/Bases)

Higl level/performing Layer 2 Switches
S6000, S4048-ON, S3124F, N3024F
40/10/1 GbE

Beware : in this case *broadcast* is **not** an “ethernet broadcast” (ff:ff:ff:ff:ff:ff)…

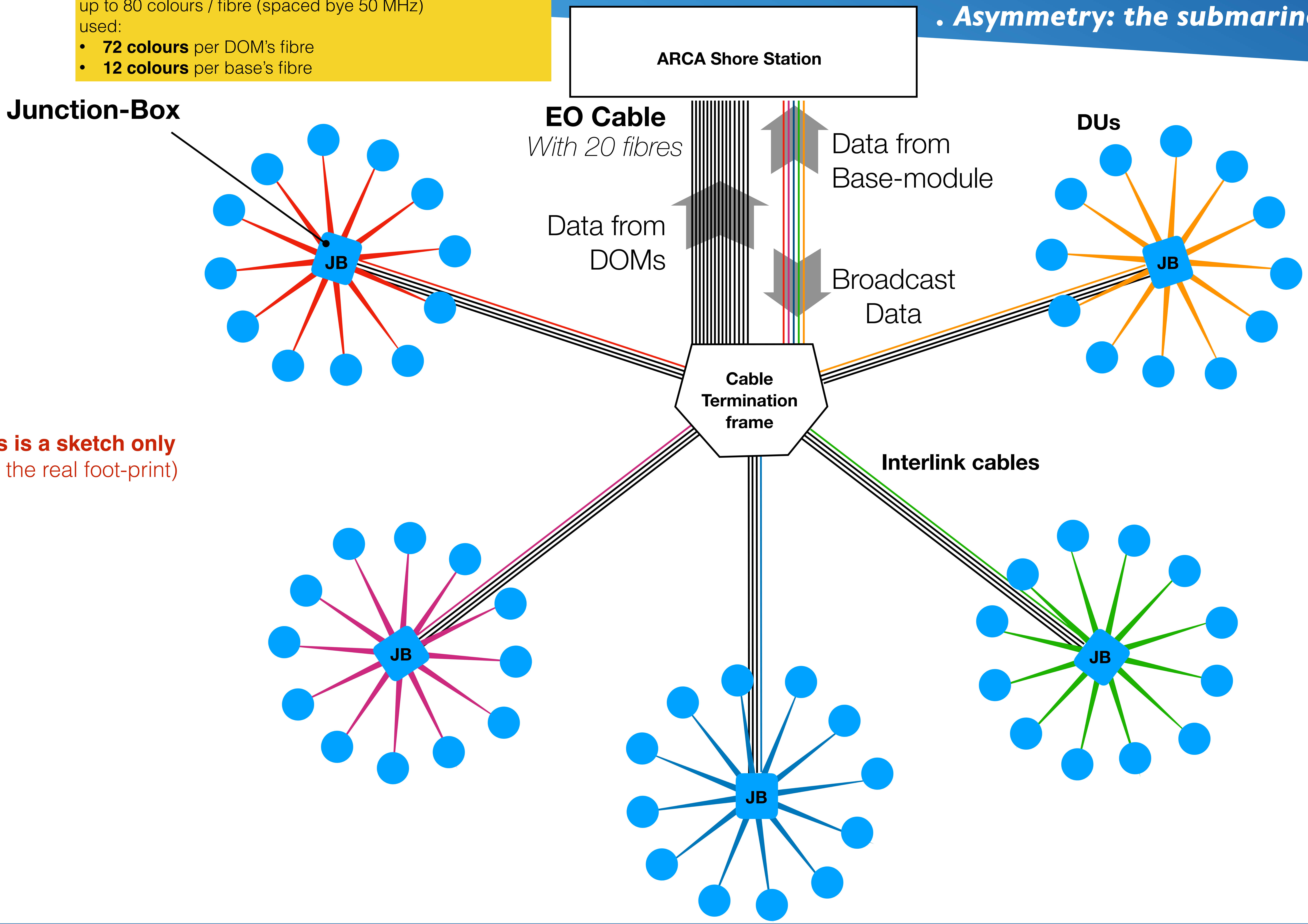
Broadcast channel: one **single stream** composed of
- *WR-PTP* packets
- *Unicast slow-control UDP packets* (i.e. addressed to a specific DOM)

It is produced in the shore station and sent off-shore

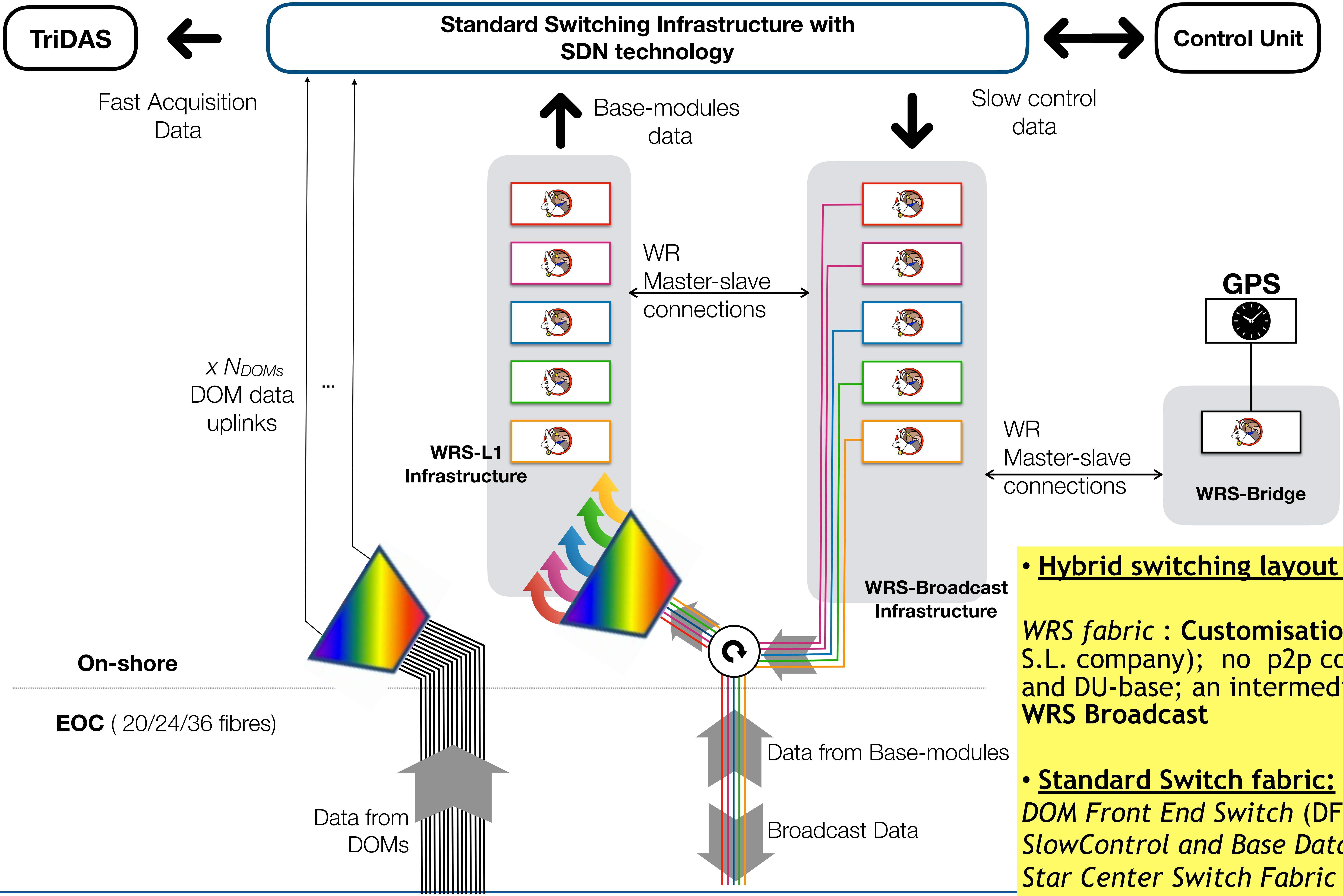
The original stream is optically split into different “SC” lines, serving the modular detector => Every end-point (DOMs & Base-modules) receives all the packets! Dedicated FPGA filters the packet.
Base-modules are kept in the White-Rabbit loop, on the same fibre;
DOM SC replies + Fast Acquisition DATA are routed back to shore along different fibres

up to 80 colours / fibre (spaced bye 50 MHz) used:

- **72 colours** per DOM's fibre
- **12 colours** per base's fibre



This is a sketch only
(not the real foot-print)



- <u>**Hybrid switching layout :**</u>

WRS fabric : **Customisation** of WR protocol (Seven Solutions S.L. company); no p2p connection from WR switch to DOM and DU-base; an intermediate layer (**WRS-Layer1**), fed by **WRS Broadcast**

- <u>**Standard Switch fabric:**</u>

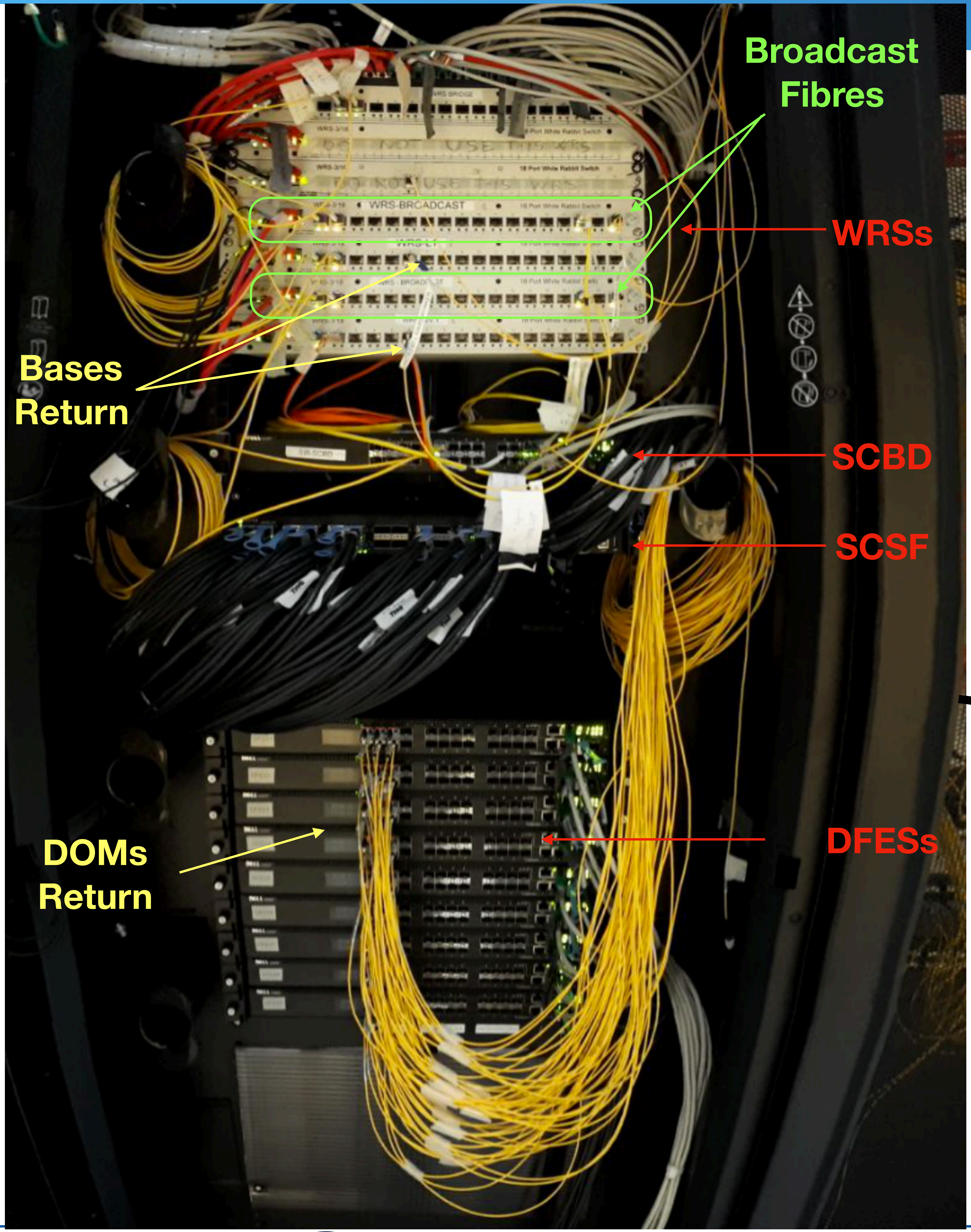
DOM Front End Switch (DFES)
SlowControl and Base Data Switch (SDBC)
Star Center Switch Fabric (SCSF)

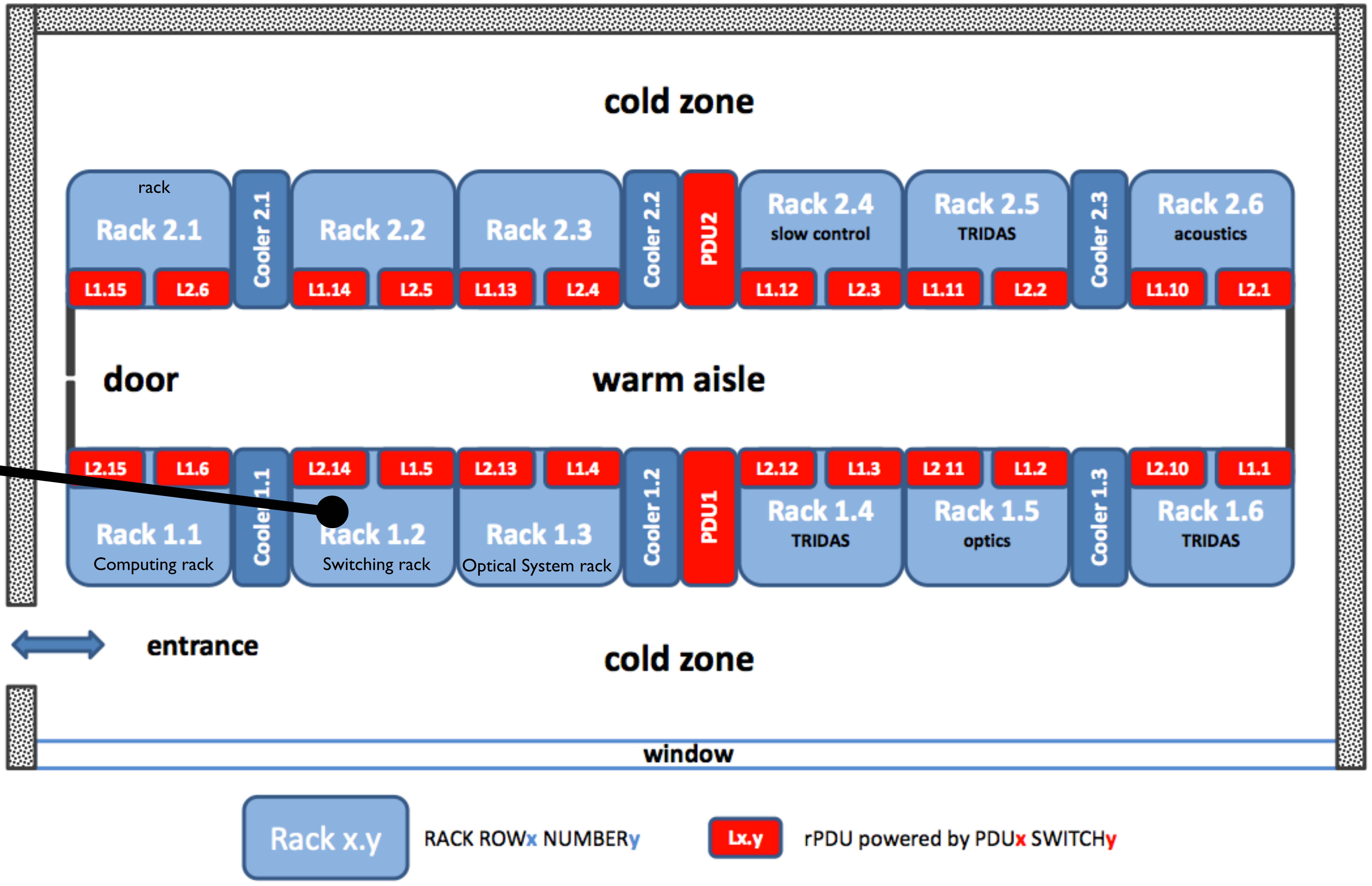
SDN is used to define routes for specific data flows

- Without SDN, CU had to be connected to WRS-Broadcast
  - Loops followed otherwise
- Same for SC-FBK and A-Data
  - As the detector scales up, WRS performance degrades

High Level Standard Switches Implementing *OpenFlow*

OpenFlow





Similar switching infrastructure for the ORCA shore-station

Openflow 1.3
OpenDaylight Nitrogen

SDN works as a Layer 2 Router

#ID	Source	Destination	Action
SCSF-1	any	ff:ff:ff:ff:ff:ff (broadcast)	To raw-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 (DAQ server)	O+A Data to TriDAS
SCSF-4	CU	08:00:30:00:00:00/ff:ff:ff:00:00:00	SC-CMD to SCBD

#ID	Source	Destination	Action
SCBD-1	08:00:30:00:00:00/ff:ff:ff:00:00:00	any	(SC-FBK,A-Data) to SCSF
SCBD-2	Everything from SCSF uplink	08:00:30:00:00:00/ff:ff:ff:00:00:00	SC-CMD to multiple WRS-B/cast
SCBD-3	08:00:30:00:00:00/ff:ff:ff:00:00:00	ff:ff:ff:ff:ff:ff (broadcast)	To SCSF

The **SDN Controller** runs on a virtual machine.

It is implemented with OpenDaylight (release - Nitrogen https://www.opendaylight.org/).

Rules are loaded/erased into/from the Controllore via the *RESTCONF* protocol.
Afterwords and automatically, the Controller pushes them into the switches

Once the rules are pushed on the SDN switches , they remain active until the switches are powered down.

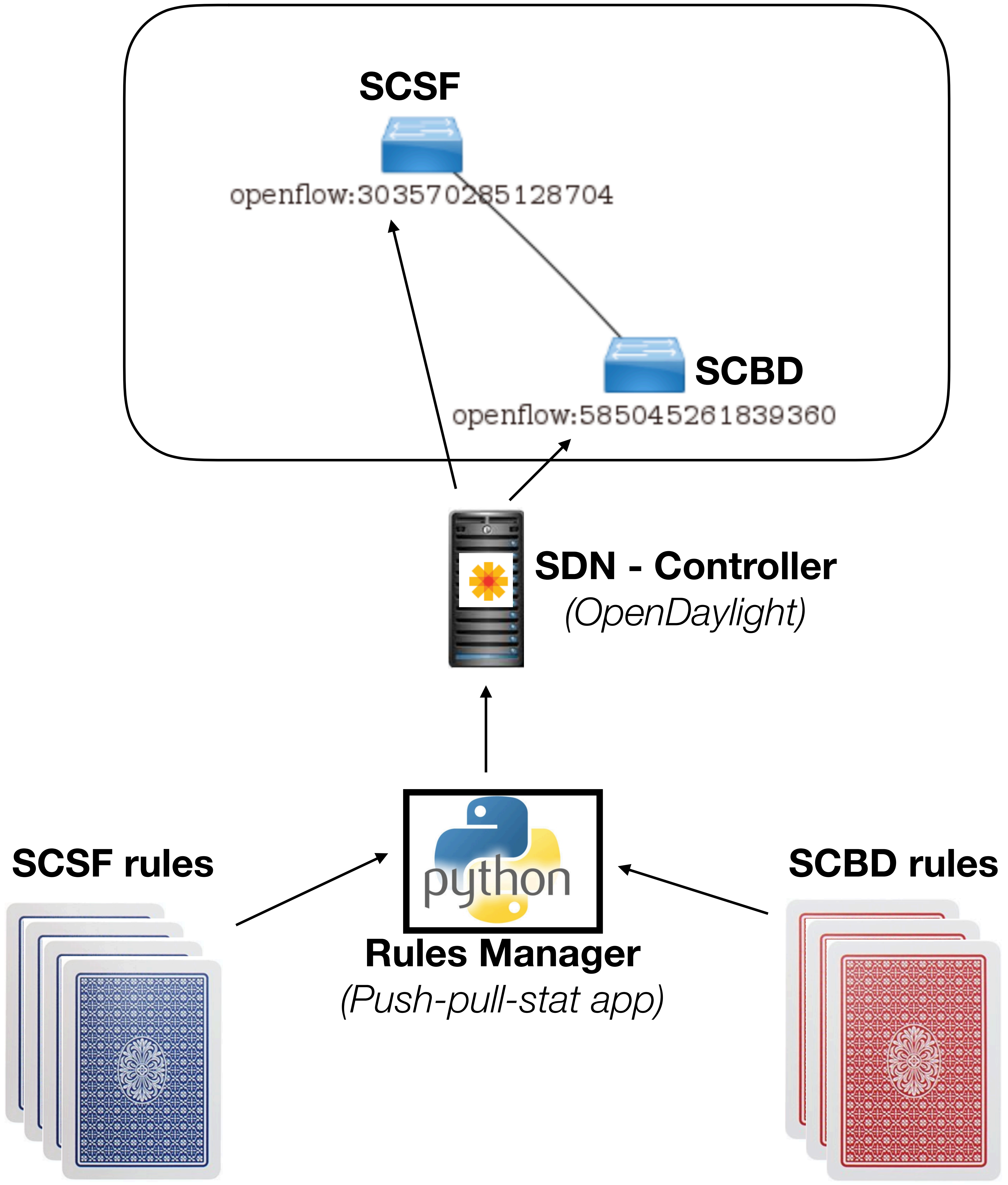
Once the SDN switches are back operational, the Controller automatically pushes the rules in.

A redundant Controller is highly recommended, to apply a failover strategy.

The Controller installation and configuration is managed via ***ANSIBLE***
(https://www.ansible.com/).

ANSIBLE (Foreman and/or Kickstart) is the way the full DAQ system is deployed and maintained in the shore-stations as well as in various test-stations distributed in different laboratories throughout Europe.

```
[ {  "flow": [   {
          "id": "3",
          "match": {
             "ethernet-match": {
                "ethernet-source":
                   {"address": "08:00:30:00:00:00",
                   "mask": "ff:ff:ff:00:00:00"},
                "ethernet-destination": {
                   "address": "00:26:18:2c:73:91"
                }}},
          "instructions": {
             "instruction": [
                {  "order": "0",
                   "apply-actions": {
                      "action": [
                         {   "output-action": {
                               “output-node-connector": "openflow:
303570285128704:86",
                               "max-length": "60"
                            },
                            "order": "0"  } ]}}]},
          "flow-name": "SCSF_DOMtoDQ",
          "installHw": "true",
          "idle-timeout": "0",
          "cookie": "3",
          "table_id": "1"}]}
```


Network instabilities leading to loss of WR synchronisation
LEGACY layout
SDN Layout
SCBD Traffic (bit per second)
SC-CMD received from the CU
A-Data + SC-FBK received from the Base-Modules
Monitoring of SC-CMD transmitted by the WRS-Broadcast to the off-shore detector
SC-CMD transmitted to the WRS-Broadcast
A-Data + SC-FBK transmitted to the SCSF
Day of month

Conclusions:

Challenging requirements for KM3NeT networking in terms of
- high throughput
- scaling with detector components and computing resources

Broadcast scenario: a not standard, strongly asymmetric, ethernet layout

Hybrid switch-fabric combining White Rabbit switches for time calibration and standard switches.

SDN technique is the answer for it allows:
- deterministic configuration of the network
- no degradation of WRS-fabric performances due to traffic enhancement with detector scaling
- stable Layer 2 routing of various data-flows
- exploiting standard (JSON) scripting languages for flow configurations

KM3NeT is the first actual SDN use-case in High Energy Astrophysics community.

Outline:

New utility-SDN-rules better control the flows (e.g. broadcast-storm dumper, ARP handlers)
Custom management API/clients to optimise the creation of rules and the interaction with the SDN Controller

SPARE SLIDES

Optical Throughput

Case		{Expected (ν_{single}=, 7 kHz)}	{Conservative (ν_{single}=, 15 kHz)}
3" PMT (0.25 p.e. thresh.)	(Mbps)	0.4	0.8
DOM (31 PMT)	(Mbps)	11.0	23.0
String (18 DOM)	(Mbps)	200.0	420.0
Phase 1,It (24 strings)	(Gbps)	4.7	10.0
Phase 1,Fr (7 strings)	(Gbps)	1.4	2.9
1 Block – Phase 2 Fr (115 strings)	(Gbps)	22.0	48.0
2 Blocks – Phase 2 It (230 strings)	(Gbps)	45.0	96.0
Phase Next (690 strings)	(Gbps)	130.0	290.0

Acoustic Throughput

Case	Raw Thp/Sensor (Mb/s)	Raw Thp/DU (Mb/s)	Raw Thp/Detector (Gb/s)	TOA (Mb/s)	Positions (Mb/s)	Storage (TB/y)
Phase 1–*It*	4.6	88.0	2.1	0.20	0.08	1.10
Phase 1–*Fr*	4.6	88.0	0.6	0.06	0.02	0.32
1 Block, Ph2 *Fr*	4.6	88.0	10.0	0.94	0.38	5.20
2 Blocks, Ph2 It	4.6	88.0	20.0	1.90	0.75	10.00
Phase Next	4.6	88.0	61.0	5.70	2.30	31.00