TriDAS Phase 3   (T. Chiarusi – 26/09/2013, v0r1)

The source of the data stream entering the TriDAS is from the FCM Servers that transfer the data coming from the detector to on-shore. The data stream coming to off-shore is distributed among the FCM Servers according to a fixed configuration telling which onshore machines must handle which part of the detector throughput. We refer to such a configuration as *topological*.

The Tower- FCM Server assignment is sketched below in Figure 1.

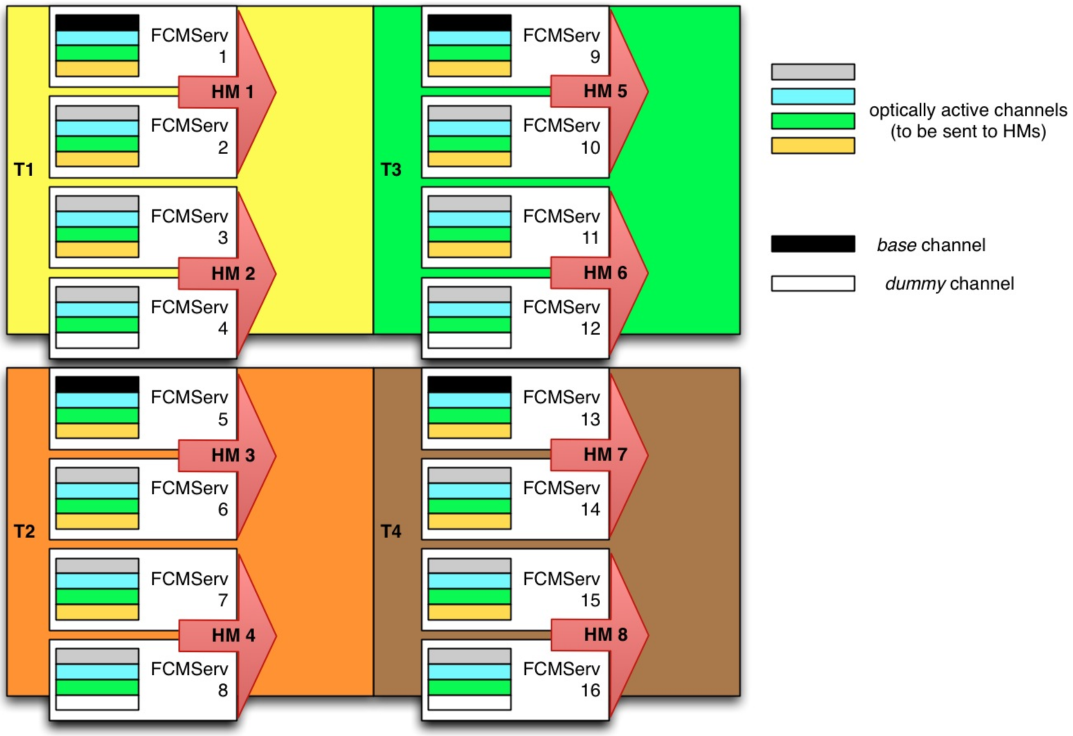


Figure 1. The topological configuration of FCM Server for Sector 1 of the KM3-ITA 8 towers detector.

Please note that the 8 towers of the KM3-ITA detector are logically divided into 2 sector of 4 Towers each. The above picture shows the case of Sector 1, involving Towers from 1 to 4. The big colored squares (yellow, green, orange and brown) represent one Tower each.
Inside each one of the squares one finds those FCM Servers which are supposed to be assigned to the related Tower.

It is apparent from the picture, that any FCM Server box comes with 4 colored rectangles, each one representing the 4 available input channels of the FCM Server ASIC. Each input channel is supposed to receive the data of one floor of the Tower. The KM3-ITA Tower is composed of 15 active floors and, being assigned 4 FCM Server to each Tower for a total of 16 input channels, it is clear that there is one channel per Tower which is not used (in the picture it is sketched as a white filled rectangle). Such channel is referred to as *dummy channel*. Moreover, among the 15 floors of one Tower, the one located at its base (i.e. *floor 0* or *base floor*) doesn't produce any optical data and it is discarded by the TriDAS. The related channel is called *base channel*, and it is sketched as a black rectangle in the shown picture.

As it will be discussed in the next section, the first stage of the TriDAS is the *aggregation level*, where the data are gathered into subsequent *time slices*. Many aggregation nodes, called *Hit Managers*, are required to handle the incoming throughput. Any Hit Manager is interfaced to a certain number of FCM Servers. We anticipate here that one conservative solution is to assign 2 FCM Server to one Hit Manager, so that the entire throughput from a Tower is handled by 2 Hit Managers. In order to equalize the FCM Server data stream to the Hit Managers, the FCM Server with the dummy channel doesn't send its data to the same Hit Manager which is receiving data from the FCM Server with the base channel.

TriDAS dimensioning and required network performances

The TriDAS design follows what has been implemented for NEMO Phase2.
We can sketch it as follows:

1) **HitManager (HM)**. Each HM is interfaced to a group of FCMServers, perform the TimeSlice aggregation (over a time window of 100-200 ms) and send them to the TriggerCPUs.

Being conservative, it is feasible for 1 HM to serve ½ of a KM3-ITA tower, i.e. to receive data from 2 FCMServer, with an input throughput of about 2 Gbps.
In principle 16 HMs could be enough, but we suggest to extend the number to **20 HMs**. In this case, some occasional but possible HM faults can be promptly solved, by redirecting the related FCMServer datastream to a different machine.

HM input throughput (Gbps)	**HM output throughput (Gbps)**
2(4) FMCserver/HM x 1 Gbps = 2(4)	2-4

2) **TriggerCPU** (TCPU). Each TCPU is aggregating data belonging to the same TimeSlice, so it must be interfaced with all the HMs.

Number of TCPUS
Simple conservative considerations can motivate the proposed number of **50 TCPUs**.

The Nemo Phase2 TriDAS works stable with 2 TCPUs only. Note that this result is strongly determined by the NEMO Phase2 detector and the measured hit rate on top of the PMTs. The NEMO Phase2 Tower is made of 32 PMT, hitted at a rate of ~ 50 kHz. The KM3-Ita Tower is instead made of 84 PMTs (enhancement of a factor ~ 2.63). So, it is clear from this trivial “computation on the back of the envelope ” that for a single new KM3-Ita tower, more TCPUs per tower are needed. In this picture 5 TCPUs are required. Now, supposing to have 8 **independent** acquisitions, one per each one of the KM3-Ita Tower, we should get a total number of TCPU of ~40 TCPUs. It is clear that combining hits from the whole detector, the number of operations to get ordered lists and what is needed to perform the triggers increase not linearly.

To be conservative we propose to set the TCPUs number to 50 TCPUs.

Further investigations are under development with MonteCarlo simulations, and will be reported within few weeks.

TCPU input/output throughput

- *Input throughput*. In principle more than 1 HM are supposed to simultaneously talk to the same TCPU. So the TCPU input throughput can be more than 2 Gbps up to 10 Gbps. In fact, the choice to have point to point connections with a 10 Gbps bandwith allow to optimize such a data transfer from the HMs.

- *Output throughput*. Let’s do very conservative assumptions:

  i) continuous single hit rate : 150 kHz / PMT
  ii) muon trigger rate: 800 Hz
  iii) hit size: 200 bit
  iv) event window duration: 6 μs

We get total datastream from the whole DAQ system which is <12 MBps.

So in the assumption of 50 TCPUs, we compute a single TCPU output throughput < 2 Mbps.

TCPU input throughput (Gbps)	**TCPU output throughput (Mbps)**
≤ 10	< 2

3) **EventManager (EM).** The EM input throughput is the summation of the output throughput from each TCPU. According to a conservative approach, this is expected not to exceed the 12 MBps. Apart from some overhead this is a reasonable expectation for the output throughput.

EM input throughput (MBps)	**EM output throughput (MBps)**
≤ 12	< 12

4) **TriDAS System Controller (TSC).** The TSC is supposed to exchange information with all the on-shore servers. We can divide the information types into two groups.

Slow Control (required bandwith at 1Gbps)
The TSC must talk to

- the DataManager, since the TriDAS should be driven by the Global Control Unit, so via the DataManager;
- the FCMServers , to dynamically configure the data sending to the available HMs;
- the HMs and TCPUs for the same reasons;
- the HMs, TCPUs and EM to configure the machines

Monitoring or data dumping (required 10 Gbps bandwith)
In this case the TSC could be used to dump data from the HMs (e.g. in case of occasional rawdata direct dump for some off-line quick data analysis). The TSC could also be the gate for the TriDAS monitoring.

TriDAS Networking

BASIC CONFIGURATION

Figure 2. shows the configuration of the TriDAS connections. The Towers are sketched on the left, in the light green vertical boxes.

As already mentioned, the whole detector is logically divided into two sectors, each one with 4 Towers.

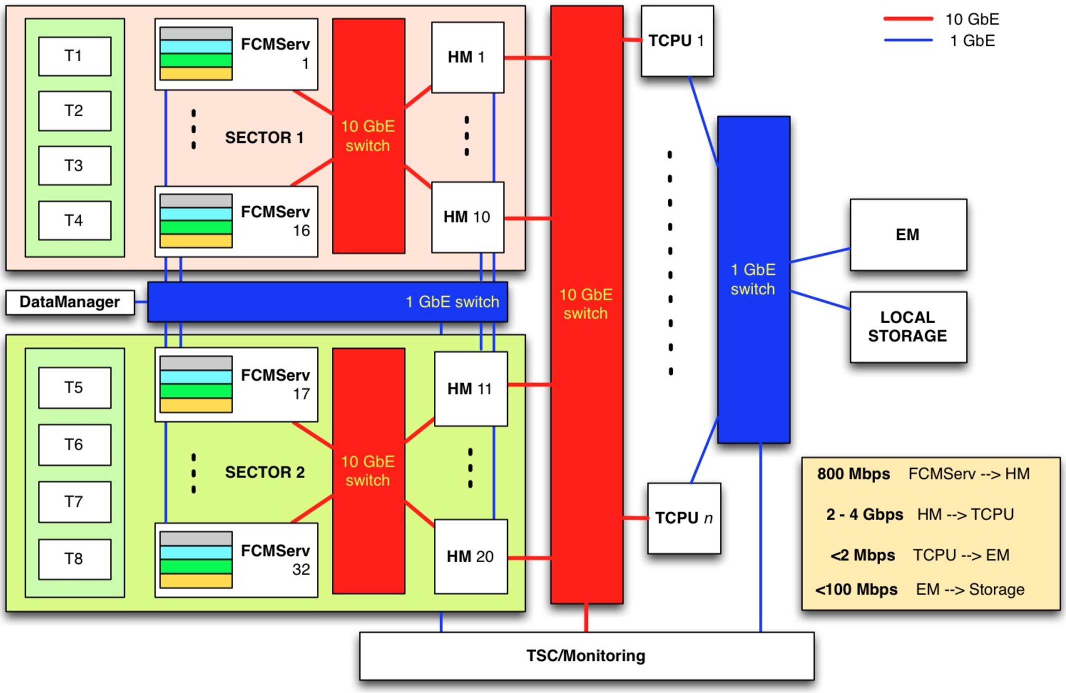


Figure 2. Logical networking design (Basic Configuration) for the on shore data center. The number of the estimated TCPU is for *n* = 50.

The FCMServer boxes present the 4 colored input channels, one per floor.

The onshore network facility is Ethernet based. The case represented in Figure 2 implies the use of connections at 1 GbE (blue lines) and 10 GbE (red lines).

There are 2 classes of 10 GbE switches:

- the *Sector switches* which connect the FCM Servers to the HMs; the figure shows that each Sector switch allows 16 FCM Servers to connect to 10 HMs
- the *Trigger switch*, which interconnect the 20 HMs with the 50 TCPU.

IMPORTANT: in this design it is missing the connections with the nodes for the acoustic DAQ.

10 GbE ports

The global number of high bandwidth ports is expected to be at least **124**:

- 32 ports @ FCMServer level
- 20 x 2 ports @ HM level
- 50 ports @ TCPU level
- 1 port @ TSC level
- 1 port @ DataManager (as an option)

Connection between FCMServer and HM

26 ports @ 10 GbE per each sector (52 ports total).

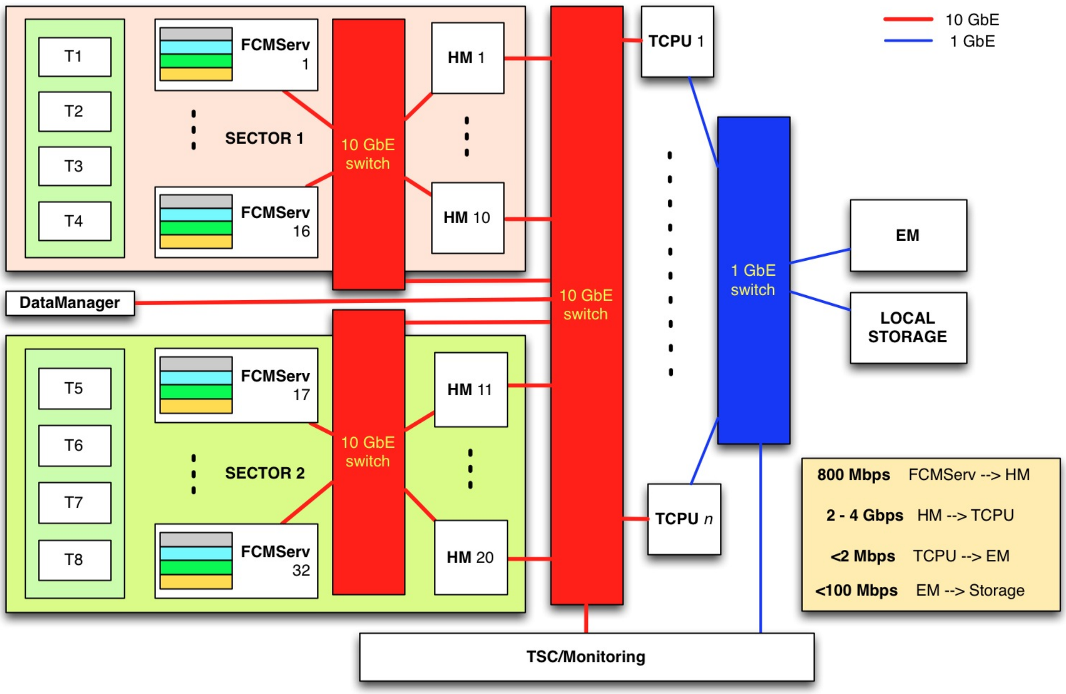
Connection between HM , TCPU and TSC

72 ports @ 10 GbE

VARIATION TO THE BASIC CONFIGURATION

It is worthy to mention that if the DataManger is interfaced to the FCM Servers via 10 GbE connections, instead having dedicated connections at 1 GbE, we remove the 1 GbE switch on the left (horizontal blue rectangle). We have 3 possible options in order to grant all the needed connections.

Variant 1 is sketched in Figure 3, where the Data Manager and the 2 sector switches are connected to the Trigger switch .



800 Mbps	FCMServ --> HM
2 - 4 Gbps	HM --> TCPU
<2 Mbps	TCPU --> EM
<100 Mbps	EM --> Storage

Figure 3. Logical networking design for *variant 1*.

Variant 2. In order to minimize the number of 10 GbE connections one further evolution could be the one sketched in Figure 4, where the HM-TCPU connectivity is kept still separated from the previous stages.

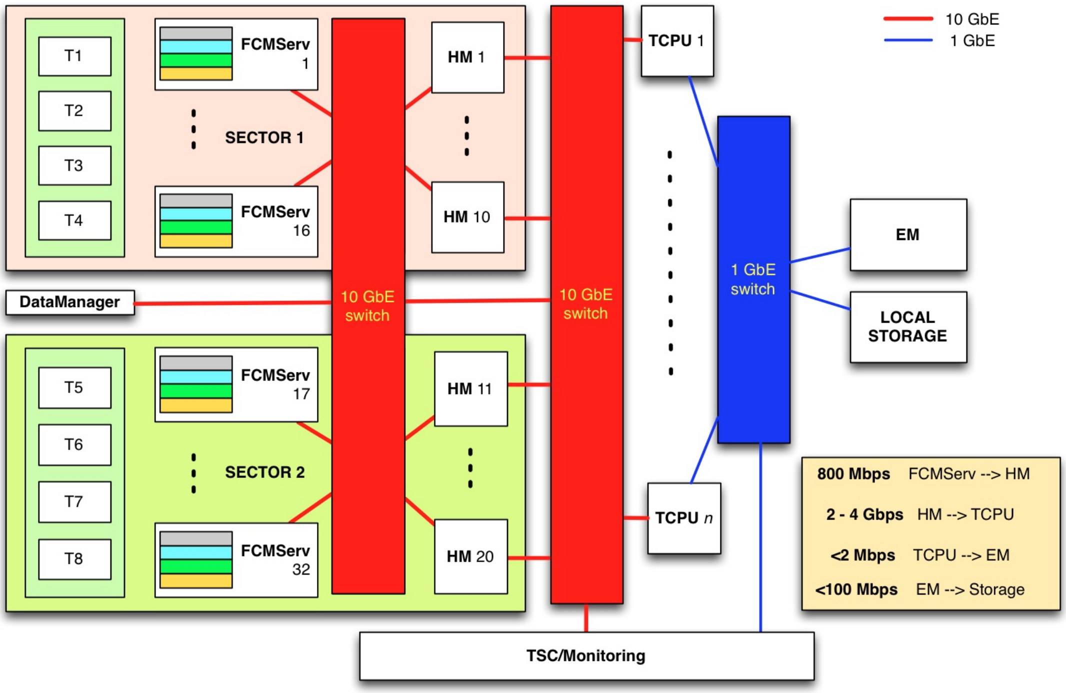


Figure 4. Logical networking design for *variant 2*.

One possible complication in the previous two variants is making the DataManager talk to the TSC, which is supposed to lay in the subnet of HM – TCPU interfaces, different than the FCM Server – HM one. This could be made possible by opportunely defining the FCM Server – HM and HM – TCPU subnets and by setting the DataManager subnet-mask correspondently.

Variant 3.   A star-center network topology may be the easiest way to connect all the on-shore nodes. The Basic Configuration can be obtained by grouping the switch ports into virtual LANs.

The final choice is clearly a matter of affordability.

Costs and power consumption (still rough) estimations

SWITCHES
We've asked to 3 different Vendors (CISCO, DELL and Hewlet Packard) to give a reliable estimation of the costs and power consumption.

CISCO option

Variant	**Sector Switch (x 2)**	**Trigger Switch (x1)**	**Cost (k€)**	**Power Cons. (kW)**
Basic Config.	Nexus 5548 with 32 10GbE Ports	Nexus 7009 with 2 slot F2-Series 48 Port 1/10G (SFP+)	140	5
Variant 3	Nexus 7009 with 3 slot *F2e* non-blocking 10 GbE 48 port		140	5

Pros : completely reliable, redundant, quick aid intervention (within 8h from the call) package for 1 year, expandable for a detector rescaling.
Cons : it is expected to be about twice more expensive with respect to the other solutions.

DELL option

Variant	**Sector Switch (x 2)**	**Trigger Switch (x1)**	**Cost**	**Power Cons. (kW)**
Basic Config.	???	Z9000 (32 ports @ 40 Gbps → 128 ports 10 Gbps con breakout)	??	?? + 1
Variant 3	Z9000 (32 ports @ 40 Gbps → 128 ports 10 Gbps con breakout)		??	1

Pros : it is about the cheapest solution, simple stand alone solution
Cons : not expandable, not redundancies

HP option

Variant	**Sector Switch (x 2)**	**Trigger Switch (x1)**	**Cost (k€)**	**Power Cons. (kW)**
Basic Config.	HP 5900 (48 port 10 GbE not blocking);	HP 10504 modular -slot 32 port 10GbE -slot 48 port 10GbE	50 +70	0.6 + 2.5
Variant 3	HP 10504 modular 128 port (exp. >70)		112	<4

Pros : each optical port costs 337 € (the cheapest available in the market **and present in CONSIP**)
Cons : HP is new in switch production.

SERVERS

The servers are grouped in three categories:

Category	**Brand/model**	**Cost per unity (k€)**	**quantity**	**Total cost (k€)**
FCM Server	Supermicro *1027GR-TRFT+.*	2.5	32	**80**
HM, TSC, EM	??	2	22	**44**
TCPU	??	2.5	50	**125**
Grand total				**249**

We assume an average **150 W / server** as power consumption.

10 GbE INTERFACES

1 board interfaced to the bus + transceiver (i.e. the optical port): 600 € / interface.

Total cost for all the required ports: 74 k€.

10 GbE Cables (5 meters)

Total costs for all the connections: 10 k€

TERASIC DE5

Total costs for all the asics: 10 k€: 40-70 k€

TOTAL ESTIMATIONS

	with Dell Basic Config. (??)	**with HP/CISCO**
Costs (k€)	**~ 400**	**~ 510**
Power Cons. (kW)	**18**	**20**

IMPORTANT NOTES

- **A more accurate estimation of performances and prices of all the servers, with major attention to the TCPUs, is currently ongoing.**

- **this estimation does not include yet the costs for:**
  - **the racks (even though it should be globally limited to 20 k€ for 3 racks)**
  - **the Acoustic nodes (and relative ports on the switches)**