Abstract

At the time of this paper, all hardware elements of the CMS Data Acquisition System have been installed and commissioned both in the underground and surface areas. This paper describes in detail the infrastructures and the different steps that were necessary from the very beginning when the underground control rooms and surface building were building sites to a working system collecting data fragment from ~650 sources and sending them to surface for assembly and analysis.

I. INTRODUCTION

The Compact Muon Solenoid (CMS) experiment [1] at CERN’s Large Hadron Collider (LHC) will search for new physics at the TeV scale such as the Higgs mechanism or Super-Symmetry. At its design luminosity of 10^34 cm^{-2}s^{-1} the LHC will provide proton-proton collisions at a center-of-mass energy of 14 TeV with a bunch crossing frequency of 40 MHz. Each bunch crossing will give rise to about 20 inelastic collisions in which new particles may be created. Decay products of these particles are recorded by the sub-detector systems of CMS comprising approximately 55.10^6 readout channels. After zero-suppression the total event size per bunch crossing is expected to be on average 1 MB. A highly selective online-selection process accepts of the order of 10^2 events per second to be stored for offline analysis. In CMS, this process consists of only two levels. The Level-1 Trigger [2], a dedicated system of custom-built pipelined electronics, first reconstructs trigger objects (muons, electrons/photons, jets ...) from coarsely segmented data of the muon and calorimeter sub-detectors. Based on concurrent trigger algorithms which include cuts on transverse momentum, energy and event topology, it accepts interesting events at an average rate of 100 kHz.

All further steps of on-line event processing including the read-out, data transport to the surface and event-building at an aggregate data rate of 1 Terabit/s, high level trigger processing and data storage are handled by the CMS Data Acquisition (DAQ) System [3]. The 55.10^6 readout channels are grouped into approximately 650 data sources by the Front-End Driver (FED) electronics. Full event data are buffered during the 3 µs latency of the Level-1 Trigger and pushed into the DAQ System upon a Level-1 accept. To optimize utilization of available bandwidth, some FEDs with smaller fragment size are merged by the DAQ System resulting in a total number of around 500 balanced data sources of approximately 2 kB, from which events have to be built. The event building process is implemented with a two-stage event building architecture [3]. The first stage is relying on optical Myrinet technology whereas the second stage uses standard copper Gigabit Ethernet technology. This innovative event builder architecture features an optimal usage of the available bandwidth for both technologies and allows a staged deployment of the DAQ system to adapt the analysis power to the luminosity delivered by the LHC. The fully assembled events are passed to the filter farm which executes the high-level trigger decision based on reconstruction algorithms similar to the full off-line reconstruction.

The CMS DAQ architecture includes back-pressure all the way from the filter farm through the event builder to the Front-End Drivers. Through this mechanism, FEDs are prevented from sending data to the DAQ System in case of down-stream congestion. The amount of data received by the FEDs on the other hand is determined by the trigger rate and by the detector occupancy. In order to prevent buffer overflow and data corruption in the FEDs or front-end electronics, a Trigger Throttling System provides fast feedback to the trigger and throttles the rate or disables the trigger when buffers are close to being full.

The present paper focuses on the infrastructures required by all the DAQ devices and their installation both in underground and surface areas.
II. CMS DAQ IN UNDERGROUND AREA

A. DAQ elements

In the CMS cavern, a dual floor counting room identified as USC55-S1 (lower floor) and USC55-S2 (upper floor), has been built to house amongst other services, the sub-detector readout electronics.

Numerous DAQ elements are installed in the counting room. Their tasks are the following:
- Front End data collection and transmission to the surface on-line computing farm
- Front End status collection and elaboration of a smart back pressure signal given to the trigger logic preventing the overflow of the Front End electronics.

The hardware elements installed to perform these two tasks are the following:
- 500 FRL cards receiving the data of one or two sender cards (the sender cards are plugged onto the sub-detector FEDs)
- 650 S-link cables connecting the senders to the FRLs (total length 6 km)
- 56 FMM cards receiving and merging the status from each FED
- 750 RJ-45 cables connecting the FEDs with the FMMs (total length 11 km)
- 500 Myrinet Network Interface Cards (NICs) plugged on the FRLs
- 6 Myrinet switches each with up to 256 input ports
- 1000 optical patch cords connecting the NICs with the switches (total length 38 km)
- 50 Compact PCI crates sub-divided into 60 logical crates (some crates contain dual-backplanes)
- 60 crate controller PCs with their control cables (total length 1.6 km)

- 30 optical cables linking the underground area with the surface area. Each cable is 200m long and is composed of 216 individual fibers grouped in 18 MPO ribbons.

At the exception of the Myrinet switches, all this equipment is installed in 20 racks (18 racks for FRLs and FMMs and 2 racks for PCs) and dedicated cable trays running in the false floors of both levels of USC55. The location of these racks (see figure 2) has been determined to keep the s-link cable length below 11 meters and the CPCI control cable length below 13 meters.

The racks for FRLs/FMMs and the racks housing the PCs/switches are totally different.

The first rack type for FRLs/FMMs are the well known LEP type rack designed for housing electronic crates requiring a vertical air-flow. Depending on the total heat load, several air-water heat exchangers (max. 3) are inserted between crates to maintain a stable air temperature from the bottom crate to top crate. The air-flow is created by tangential turbines located at the very top of the rack moving the air through the crates (from bottom to top) and then the air goes back to the bottom of the rack through channels created in the side walls. At the very bottom of the rack, an air deflector guides the air through the crates again. Usually, a heat
exchanger is located just before the turbines: hence, there is no heat loss in the side walls and cold air is always available for the bottom crate. Temperature sensor and smoke detection devices are located in the top turbine.

The second rack type is a custom design adapted to devices requiring a horizontal front to back air-flow: rack mounted PCs are a good example of such devices. The rack itself is manufactured by one of the well known IT rack providers, Rittal©. The rear door receives a cooling system featuring a heat absorption capacity of 10 kW. The cooling system is composed of an air/water heat exchanger and a set of 3 fans creating the air flow inside the rack from front to back. The room ambient air is sucked-in by the servers (they have their own fans to create internal air circulation). After cooling the internal components, the hot air exhausts at the rear of the servers and goes through the door cooling system and returns to the ambient cold again. Hence, there is never warm air released to the ambient. The air flow created by the door fans is greater than the airflow created by the servers: this avoids warm air from staying between the door and the server’s rear side. The cooling system and its mechanical integration have been studied and manufactured by CIAT ©. This rack type is also used in the surface building to house the computer farm. The advantages and drawbacks of such a cooling method are discussed in the paragraph dedicated to the CMS DAQ surface building.

![Figure 3: Racks for electronic crates and computers](image)

**B. Installation sequence**

The CMS technical office designed the counting rooms with 2 floors, S1 and S2 having respectively the capacity of 100 and 160 racks of 60 x 90 cm² footprint. Each counting room has a false floor of 2 meters for the inter-rack cabling, the cables from detector to racks and cooling water distribution. The electric power is distributed by power bars located on top of the racks (Canalis system).

Starting from this basic scheme, the requirements of every sub-system were collected in order to produce a first rack assignment. Numerous constraints had to be accounted such as the number of racks needed per sub-system, localization of the sub-system rack inside the room (for example to minimize the trigger latency, trigger racks had to be placed at the closest from the detector), inter-rack communications (DAQ s-link cables can be 11 m length max), etc. Reaching the final rack assignment was an iterative process lasting several years which required innovative solutions, compromises and patience to satisfy all the constraints.

Then, once the rack assignment was more or less defined and stable, all DAQ cables/fibers were individually listed and declared in the CMS cabling database, along with their start and end points. Concurrently, dedicated cable trays for DAQ cables were included in the global 3D model of CMS. These cables trays are running immediately under the false floor tiles in order to minimize the cable length. The 3D model computes the length of every cable and the cross-section of cable within the cable tray in order to know if the tray is correctly dimensioned. However, even with such tools, a significant contingency was added everywhere to accommodate late changes or requests from the users.

Once the under floor cable trays and the LEP type racks were put in place the racks were equipped with the cooling system (distribution manifolds, heat exchangers, tangential turbines, side panels and air deflector) and the vertical cable trays to guide the cables/fibers from the crates into the false floor. At this point, the inter-rack cabling could start. However, before being able to start the cabling and install our electronics, defects in the manifolds were discovered which gave water leaks under the nominal pressure of 5 bars. The manifolds had to be dismounted, fixed by the manufacturer and remounted in the racks. As heavy work was still taking place in the control rooms (soldering, grinding and all sorts of mechanical tasks) mock-up crates and fake cards instead of our real hardware have been used to properly arrange the cables. Given the high density of cables, they have been packed in the cable tray individually. All the s-link and FMM cables have been pulled to the destination racks and connected at the FRL/FMM side, the other end of the cables waiting for the sub-detector readout electronics to be available. The installation of the real crates and their electronic cards was done later. Then, these elements have been switched on and retested in-situ. Every single cable was also tested with a portable custom pattern generator plugged at the FED side and a crate controller computer on the other side.

After the copper cable installation and test came the installation of the optical patch cords. Before installation, patch cords have been assembled in bundles to give them a protection and to ease the installation in the cables trays. In order to avoid the damaging the fragile fibers, cable trays for copper and fibers are separated. The length of the bundle depends on the distance between the crate and the switches (20m, 30m, 35m and 40m) and the number of fibers in a bundle depends on the number of FRLs in a crate. Two duplex fibers are added to each bundle as spares in case of fiber damage leading to 8 to 18 fibers per bundle. The installation of these bundles was done with the help of a larger number of people from our group and concentrated in a few intensive days. Once all optical patch cords were in place and connected at both ends, data delivery from the FED end of the s-link cable up to the output of the underground part of the Myrinet switches was checked by using the pattern generators as before. At this stage, the underground DAQ part is ready to be connected to sub-detector electronics and start the detector commissioning at slow trigger rate (a few tenth of Hz). The limitation on the rate is given by the small number of PCs (maximum 18) reading out the Myrinet switches. These PCs are located next to the switch racks in S1 and known as “MiniDAQ”.

The third rack type is also used for the basis of the surface building to house the computing and DAQ electronics. The rack type is also used in the surface building to house the computer farm. The advantages and drawbacks of such a cooling method are discussed in the paragraph dedicated to the CMS DAQ surface building.
The last elements installed were the optical cables linking the underground counting rooms with the DAQ surface building where the rest of the event builders and the filter farm are located. This installation has been sub-contracted to a specialized company.

The installation schedule was the following:
- rack welding in the counting room: up to November’05
- CPCI crate controller PC install: December’05
- cable tray installation and rack equipment: January’06
- manifolds repair: up to July’06
- copper cabling and electronic installation: Q3-Q4 2006
- optical patch cords: 7 days between November’06 and January’07
- optical cables installation, underground to surface: March and April 2007

During the systematic tests carried out after installation, very few broken devices were discovered: 2 FMM cables, one optical patch cord, 3 FRLs and 2 Compact PCI backplanes. All these items have been changed or fixed.

Today, the sub-detector readout electronic commissioning program is well on the way using all DAQ elements previously installed and the MiniDAQ. A temporary underground control room has been setup and sub-detector teams are going through the integration/commissioning steps systematically.

III. CMS DAQ IN SURFACE AREA

A. DAQ building

The DAQ building on CMS experimental site contains the general detector control room, the DAQ farm control room, a sub-detector control room, a conference room and the DAQ farm itself. Everything but the farm is located on the ground floor. The farm occupies the whole of the second floor. The first drawings of the building featured a third floor and the farm was foreseen to be on the ground floor, leaving the other floors for control rooms, conferences rooms and some offices and laboratories. However, due to CERN safety rules, escape stairs would have to be present for all floors above the ground floor computer room. Adding external escape stairs was refused by the architect for aesthetic reasons, hence the computer room has been moved to the first floor at the expense of substantially re-enforcing the concrete to hold the weight of thousands of rack-mounted servers.

Around 2001-2002, we had to make a choice about the way to absorb the dissipated power in the computer room. The driving concept was to have a high ceiling for the accommodation of hot-air and huge air ducts to drive the hot air to cooling units and then pushes the cooled air into the false floor where transparent tiles would release it to the front of the server racks. So, according to this concept, the third floor of the building has been removed to create a 6 meter high ceiling. Later in 2003, a first estimation of the number of servers with their projected power consumption has been made: ~140 racks and ~750 kW of dissipated power. The power density in the room was about 2 kW/m2 and depending of the PC type (readout/builder or filter unit), the power per rack was ranging from 4 kW to 10 kW. These are still our working parameters today. When injecting these parameters in the design of the air cooling plant, we quickly came to the conclusion that the scheme was too inefficient (air-flow of 650 000 m³/h to be created) for such a power density and water-cooled racks were the only satisfactory way to go. So there was no need anymore to keep a 6 meter high ceiling but it was too expensive to ask for an engineering change request at this stage of the project and the building stayed like that (See fig 4).

B. Computer racks

Before 2003, cooling rack mounted PCs with water directly in the rack were simply out of the question with bad memories from the water cooled super computers. However, at CERN, there is a long and successful experience of cooling electronic crates with air/water heat exchangers inserted between crates and so bringing water close to a computer was not considered a problem.

But for rack-mounted PCs, a horizontal air flow was required. As there were no satisfactory products on the market at this time, all 4 LHC experiments set up the “LHC PC Rack Cooling Project” [4]. Its mandate was to produce a common specification for a cooling system able to be integrated into various racks and later select the best company through a tendering process to produce and integrate the cooling unit. The result of this work has been summarized in [yy]. CMS purchased about 150 water cooled PC racks from CIAT [xx]. Its main features are:

- 47 U high, 19 inch mounting standard
- 44 U usable space
- 60 cm x 90 cm footprint, 106 cm total depth
- 10 kW thermal capacity
- 2 m³/h water flow, ΔT 4 deg.
- 3 fans, 2450 m³/h air flow, front to back

To maintain a very low cost, we opted for very simple monitoring systems at the rack level: a fan failure signal per fan and a basic thermostat with a threshold at 40 deg. C. At regime, the air temperature at the back of the PCs is around 25 deg. C and the temperature is monitored by the sensors present in the server and read through the IPMI interface of the PC. The thermostat is present to cut the power supply in case of IPMI failure and cooling failure at the same time.
1) Advantages and disadvantages of water cooled racks

Traditional data center cooling (forced cold air through plenum floor) is adequate for power densities of 3-4 kW per rack and 500 W/m² but with today’s server power dissipation of 10 kW per rack and beyond is readily obtained (a rack full of blade servers reaches 20 kW). With such power densities, there are three main difficulties:

- to bring enough cold air in front of the servers, the surface of transparent tile is important leading to a poor space utilization
- the hot air exhausted by the servers is mixed with the cold air at the front and this is increasing with the power density
- air flow through the transparent tiles is hard to balance and changing the configuration in one place of the room can create hot spots somewhere else.

Several workarounds like hot/cold aisle layout (PC rack rows are placed alternatively front to front and back to back) or recently cold aisle containment (the cold aisle is isolated by physical walls from the rest of the room) have been implemented and improve the situation but never solve the above difficulties (see fig. 5).

Provided that the cooling system is well dimensioned for the installed PCs, a water cooled rack is not sensitive to the distance between rack rows as the exhausted air is cooled down before release to the ambient. The room surface usage does not depend on the rack power but only on its footprint. For the same reason, as there is no hot air released to the ambient, there is no mixture of hot/cold air.

The only difficulty of such system is to provide the adequate water flow to each rack. This requires careful studies of the water distribution and the presence in the false floor of the tubing network.

Of course, the presence of water close to the PCs is a risk that does not exist with the traditional forced cold air system but a high level of production quality and systematic pressure tests maintain the risk to an acceptable level.

The tendering process for the rack purchase (USC and SCX area) has been launched the end of March 2005 and the racks for USC (26 units) have been delivered the beginning of October 2005 and the remaining racks for SCX (108 units) have been delivered in March 2006. Up to now, we had only one water leak but no PCs were in direct contact of the water.

C. Installation sequence

Prior to install the racks, the support framing in the false floor and the water distribution system had to be designed and installed. The racks have been arranged in group of 8 (7 when pillars were present) and placed according to the hot/cold aisle principle: although this arrangement is not needed for cooling efficiency, we found it practical for accessing the front faces of the PCs without having a permanent cold wind blowing on our back. The distance between rows of racks is 1.50 m front to front and 1 m back to back.

The room has a total capacity of 180 racks for a total power dissipation of 800 kW. Today, about the half of the capacity is installed (106 racks) which corresponds to a data processing capacity of 50 kHz trigger rate. The remaining half will be equipped when the luminosity of the LHC ramps up. However, the entire plumbing infrastructure has been installed for the whole the room.

Once the racks were put in place and secured on the framing, the electrical distribution was installed. We opted for a Canalis system feeding each group of 8 of 7 racks. A Canalis power bar brings a maximum of 64 Amps on 4 phases in each rack for a maximum power of 14 kW. Each phase is equipped by a D-type breaker of 16 amps. Then per phase, there is a 10 outlet power distributor including a sequencer on 3 groups of outlets with a 200 ms delay between groups. Hence the phase breaker does not trip because of the inrush current when all servers are switched on at the same time. The power strip form factor is a 1U box: they are placed amongst the servers every 10 units. With this powering scheme, a rack
can house up to 40 servers of 1U form factor within the limit of 10 kW of heat dissipation.

After installing the power supply, the different communication networks have been laid down in cable trays running in the false floor. Mostly, the cable trays are organized in three layers of 40 x 10 cm². The main networks are:

- the optical Myrinet switch-RUBU network. This network distributes the data coming from the underground counting rooms to the 640 RUBU machines
- the service network (copper). This network is used to access the machine for maintenance/monitoring purposes. For example, the IPMI temperature monitoring process runs over this network.
- the data network (copper). This network is used to exchange data packets during the event building process between the RUs and the BUs.

Exhaustive information on the network topology can be found in [5]. Currently, each of the 640 RUBU machines is connected with a couple of duplex Myrinet fibers, 1 RJ45 for the service network and from 2 to 4 RJ45 for the data network. Given this high number of fibers/cables in a rack, the routing has been done very carefully as you can see in picture 7.

Finally the last element was the purchase of the 640 RUBU machines and about 160 identical machines acting as servers for different general services like data storage, data quality monitoring, databases, etc. The tendering process was launched late December’06 and the machines were delivered in 2 batches of 400 units the end of April and beginning of May’07. The unpacking of the servers and removal of packing material by the selected company was included to the purchase contract. Our experience with the underground PCs showed us that the longest task when installing a PC in a rack was to remove the machine from its protective box and discard the packing materials.

D. What next in DAQ surface building?

From July’07, all the RUBU computers and other servers/services are being commissioned. In September’07, the sub-detector commissioning will be done with the hardware available in the DAQ surface building hence giving access to faster triggering rates.

However, Filter Units (FUs) are still to be purchased and installed in order to provide the needed processing capability. In accordance with the LHC luminosity and schedule, a first phase with 50 kHz trigger rate is foreseen for next summer. With the available computer technology (at least dual CPU quad core 1U machines), about 1200 1U servers will be purchased and installed for June’08, using the remaining rack space currently available in the surface building. Depending on their power dissipation, the number of machines per rack will vary between 30 and 40.

For the second phase (100 kHz trigger rate), racks with higher power dissipation capability (close to 20 kW) are already under study and will be needed to house the servers. This second phase will take place very likely in the 2009-2010 time frame. The technology and the form factor (pizza boxes, blades or other) are not yet known but one thing is sure, the total power per rack will increase.

E. Some learnings…

During these 20 months of installation, some organizational problems created delays. Here are below some of them.

The supply chain must be carefully watched to be sure that no component outage will stop the work. A lack of basic cable ties can stop the installation of complex optical switches for a few days or more, the time needed to procure new ones. The ideal is to proactively purchase all components/parts, anticipate on the consumption and add a significant contingency on quantities because the delivery times announced by salesmen are often optimistic…
Some tasks are simply omitted in the sequence because of their banalities. For example, in our case, the labeling of the FMM cables has not been included in our task list: this task has required 2 weeks of work for 2 persons. A few omission like this and a nice schedule can shift severely.

In project like ours (installation of many electronic systems), we depend very often on external chained milestones to start a task. For example, “Ready for crate install” implies that the cooling devices are mounted in racks and pressure tested. To be able to perform the pressure test, all the piping work must be finished and commissioned, etc, etc. Delays in such chains are almost unavoidable and our project has to live with them. Our solution to overcome these waiting periods was to have a list of “floating” tasks, unconnected from any active chains and being executable any time. Hence even if one of our tasks was delayed, we could use this delay for another of our task.

IV. SUMMARY

In this paper, we went through the different steps needed to install the CMS DAQ electronics in the underground and surfaces areas. The underground part is performing the reception of event fragments produced by about 650 detector data sources and transmission to the surface DAQ computer farm for full assembly and on-line analysis.

Currently, all underground DAQ elements are used successfully for detector commissioning. The Readout Unit/Builder Unit PCs on surface are fully deployed (640 servers) and being setup to perform their assembly task. For very low trigger rate, they can also perform the on-line analysis tasks. In view of LHC startup next year, the purchase and installation of ~1200 Filter Units is being prepared to reach the 50 kHz trigger rate analysis capacity.

V. ACKNOWLEDGMENTS

When we started to prototype the first hardware elements of CMS DAQ, we were far from thinking that the same small team of designers would be still present for the production and the installation 15 years after. Having the same persons doing and coordinating the installation was of great help when unexpected or unclear situations have been encountered. I want here to thank warmly Dominique Gigi, Lucien Pollet and Sham Sumorok for their commitment in the project and the daily work we have gone through these 20 months.

For the delicate installation of the fiber optic bundles in the underground false floors, many software designers of our group gave their time to pull the fragile bundles through the curvy cable trays: many thanks to Vincent Boyer, Esteban Gutierrez, Elliot Lipeles, Frans Meijers, Steven Murray, Steve Pavlon, Jonatan Piedra, Matteo Sani, Hannes Sakulin and Christoph Schwick. Although the work was intense and difficult (crawling for hours in dusty false floors is never a pleasure), the working atmosphere made these seven days of installation almost a pleasant time.

Special thanks to the CMS technical office, Gerard Faber, Stephane Bally and Patrick Ingenito for always answering the best they could our unusual requests.

Finally, many thanks to Robert Lassausaie, Thierry Bourel (electric distribution), Guillermo Peon, Bernard Pirollet, Richard Ecclestone (cooling and ventilation), Eric Dugue, Regis Boscus (material transport), Alain Meynet (safety systems), Sylvestre Dos Santos (mechanical/electrical support) working on the experimental site and always present and helpful when we were calling for help.

VI. REFERENCES