CMS data to surface transportation architecture

E. Cano[a], S. Cittolin[a], A. Csilling[a], S. Erhan[b], D. Gigi[a], F. Glege[a], J. Gutleber[a], C. Jacobs[a], M. Kozlovszky[a], H. Larsen[a], I. Magrans[a], F. Meijers[a], E. Meschi[a], S. Murray[a], A. Oh[a], L. Orsini[a], L. Pollet[a], A. Racz[a], D. Samyn[a], P. Scharff-Hansen[a], P. Sphicas[a,c], C. Schwick[a], T. Strodl[a]

[a]CERN, Div. EP, Meyrin CH-1211 Geneva 23 Switzerland
[b]University of California, Los Angeles, USA
[c]University of Athens, Athens, Greece

Abstract

The front-end electronics of the CMS experiment will be read out in parallel into approximately 650 modules which will be located in the underground counting room. The data read out will then be transported over a distance of ~200 m to the surface counting room where they will be received into deep buffers, the "Readout Units". The latter also provide the first step in the CMS event building process, by combining the data from multiple detector data sources into larger-size (~16 kB) data fragments. The second and final event-building step merges 64 such super-fragments into a full event. The first stage of the Event Builder, referred to as the Data to Surface (D2S) system is structured in a way to allow for a modular and scalable DAQ system whose performance can grow with the increasing luminosity of the LHC

I. INTRODUCTION

The CMS DAQ system has been designed to collect event fragments from approximately 650 data sources at a first level trigger rate of 100 kHz.

The data sources (or Front End Drivers, FED) are generating for each triggered event an event fragment whose average size is between 300 bytes and 2.3 kB. Fragments are sent through a short distance link (max 15 m.) to the Front End Readout Link card (FRL) that can merge up to two data sources. There is a maximum of 512 FRLs. From there, the data are sent over a 200 m optical link to the surface where the first stage of event building (FED builder) takes place (see Fig. 1).

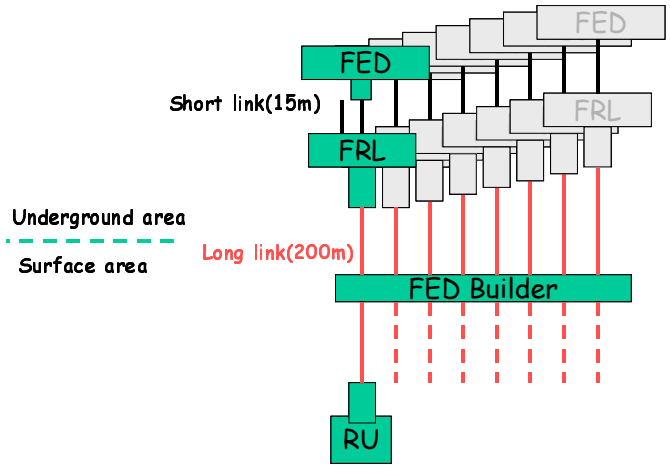

Fig. 1 D2S transportation devices

At this stage, 8 fragments are merged by a 8 x 8 switch to form a super-fragment. There is a total of 64 FED builders. Super-fragments are then stored into the readout units (RU) waiting for the second stage of event building (RU builder). The RU builder is implemented with multiple 64 x 64 switches: a single switch is also called "slice". The number of slices can be 1 to 8 and depends on the number of RUs attached to each FED builders: hence, the DAQ capacity can be adapted.

Super-fragments are assembled into entire events by the RU builder and transferred to the Filter farm where higher-level trigger decisions are made by software. Accepted events are permanently stored for further physics analysis.

With an event size of 1 MB, the system has to sustain a data throughput of 100 GB/s. This is achieved with the parallel architecture shown in Fig. 2.

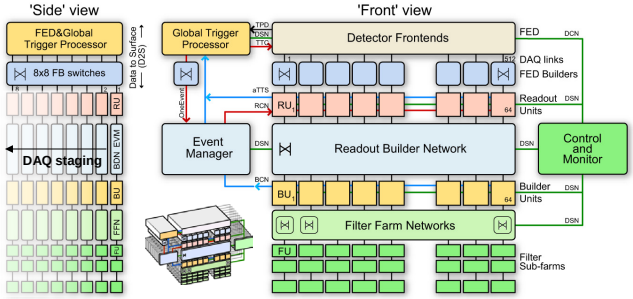

Fig. 2 CMS DAQ block diagram

II. D2S REQUIREMENTS

The D2S is in charge of collecting the data from the front-end data sources (see Tab. 1.) and to transport them to the surface where the multi-stage event building process takes place.

Reducing the diversity in the electronic devices is necessary in order to facilitate the system integration (especially during the initial test phase) and the maintenance operations. Therefore, the decision to use a common interface for all sub-detectors was made at a very early stage of the DAQ design.

The required data throughput per data source is 200 MB/s (2 kB @ 100 kHz trigger rate) over a distance of 200 m. The data transportation hardware must be able to absorb stochastic fluctuations on the event size and provide enough contingency to cope with large uncertainties on the LHC luminosity and the detector occupancy/noise.

In order to have a good working efficiency, the event builder must receive a balanced traffic through its input ports. As shown in Tab. 1., some detectors have a large spread of data sizes at the output of their data sources. Therefore, the data transportation hardware must be able to average the traffic

over several FEDs by appropriately grouping FEDs with low and high data volumes per event.

Tab. 1. Front-end readout requirements

Detector	Number of data sources	Number of DAQ links	Event size per DAQ link KB	Fluctuations RMS	Nominal event size KB (pp run)
Pixel	40	36 (merg)	2.0 - 2.4	30%	72
Tracker	440	272 (merg)	0.4 - 1.5	0.27 KB	300
Preshower	47	47	2.3	?	110
ECAL	52	52	2	1 KB	~100
HCAL	24	24	1.7	?	48
Muons CSC	9	9	2	Huge...	16*
Muons RPC	5	5	.3	?	1.5
Muons DT	5	5	1.6	?	8
Glob. trig	4	4	2	none	8
DT track f.	4	4	2	none	8
Total	630	458			671.5

* most of the time, the detector is empty, one muon track generates ~5 kB of data

The full capacity of the DAQ will be needed only at full LHC luminosity. A capacity of 25% of the nominal one is planned to be available at startup, doubling after 6 months of data-taking to reach 100% after one year of operation. Therefore, the data transportation architecture must allow a progressive deployment of the DAQ.

III. D2S IMPLEMENTATION

To fulfill all the above requirements, the D2S transportation devices are the following (see Fig. 3):

- a common DAQ interface housed by the data source (see Fig. 4)
- an S-LINK64 [1] data link made out of a source card, a copper cable (max. length 15 m.) and an S-LINK64 receiver/merger card
- a Front-end Readout Link card (FRL) (described below)
- a PCI Myrinet [4] network interface card
- a 200 m duplex optical fiber
- a Front-end builder switch (Myrinet technology)

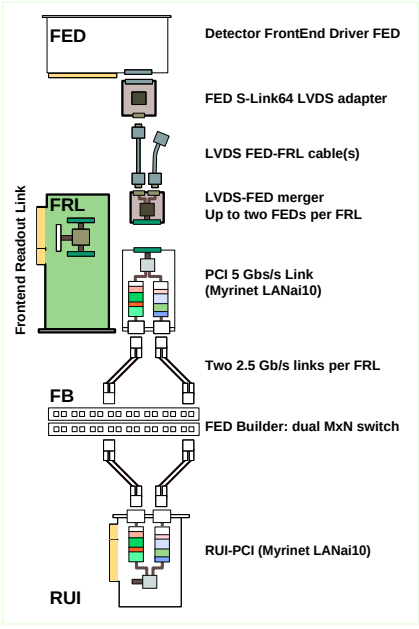

Fig. 3 D2S physical implementation

A. The common FE/DAQ interface

The interface specifies a common hardware platform and a data format which is essentially a header and a trailer with information relevant to the event building process [2].

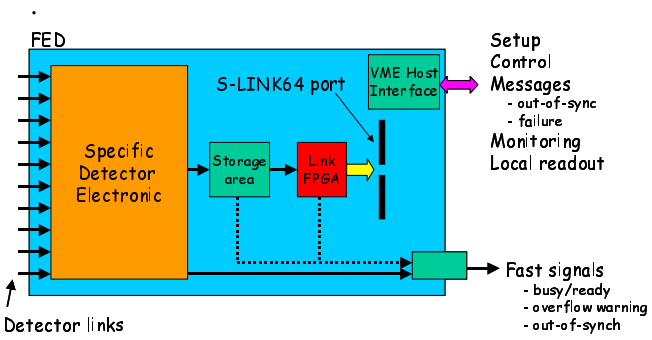

Fig. 4 FED block diagram

The payload of the event fragments is only inspected in the filter farm, i.e. it is not considered during event building. As a result of this common interface, there is no sub-detector specific hardware after the FED in the CMS read-out chain.

The DAQ port is an S-LINK64 interface. The S-LINK64 protocol is based on S-LINK which has been extended to match CMS needs of higher bandwidth (64 bits data width, maximum input frequency 100 MHz). Both S-LINK and S-LINK64 only specify a set of connectors for sending and receiving data, along with mechanical constraints for the sender and the receiver cards. The physical link between them is not specified beyond these constraints. In addition, both interfaces specify the protocols for writing into the sender card and reading from the receiver card. Both protocols resemble standard protocols that read from and write to a synchronous FIFO. The design and the implementation of the S-LINK64 port on the FED is under the responsibility of the sub-detector.

B. The short distance link

The current implementation of the short distance link is fully compliant with S-LINK64 on the sender side. The receiver side (housed by the FRL) is designed to receive data from one or two sender cards hence performing the merge of two data sources. Detailed information on hardware implementation and performance can be found in [3].

C. The Front end Readout Link card

A significant number of FEDs are expected to provide event fragments with less than 1 kB average size (see Tab. 1.) In order to reduce the number of inputs to the FED builders and to better exploit the available bandwidth in the event builder, the FRL is able to merge data from two FEDs into a single event fragment.
The current design of the FRL is a Compact-PCI card (see Fig. 5) with three interfaces:

- Input interface which handles one or two S-LINK64 cables.
- Output interface to the long distance link, currently implemented as a PCI connector since a PCI-Myrinet Network Interface Card (NIC) has been chosen. This NIC, along with the 200 m optical fibre, implements the data link to the counting room on the surface.

- Configuration and control interface which is implemented as a second PCI interface connected to the compact-PCI backplane.

The FRL is designed with a form factor compliant to Compact-PCI: the depth of the board is extended to host a standard PCI card on a PCI edge-connector: when assembled, the FRL and the PCI NIC are co-planar. Hence, the ensemble can be plugged in standard Compact-PCI backplanes mounted in standard Europe mechanics.

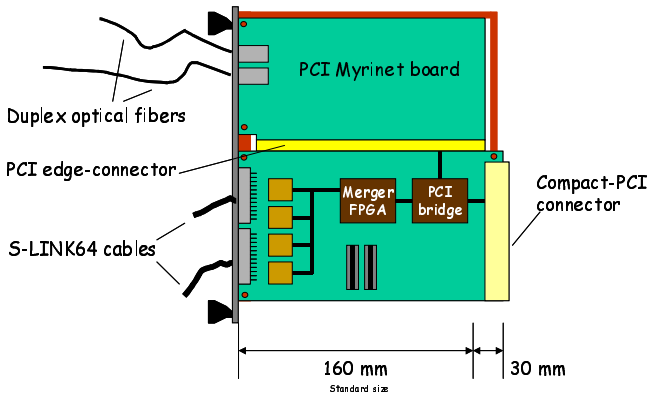

Fig. 5 FRL physical implementation

D. *Long distance link*

The long distance link is implemented with commercial devices. The technology chosen for the baseline design is Myrinet [4]. The PCI board features two "rails" running each at 2.5 Gb/s. Although a data source provides an average of 200 MB/s, the total bandwidth of the long distance link is more than twice the required throughput. This is needed for compensating the FED builder utilisation efficiency that is about 50% under the traffic conditions generated by the FE data sources: variable and unbalanced event size and random traffic.

E. *The front-end builder*

As each FRL hosts a dual rail NIC, the front-end builder is implemented with a pair of 8 x 8 crossbar switch. Measurements done with different traffic conditions show that the usage efficiency of the switch is 95% (see Fig. 6) when the events have constant size and drops to 55% when the events have a size variation equal to their average size (log-normal distribution).

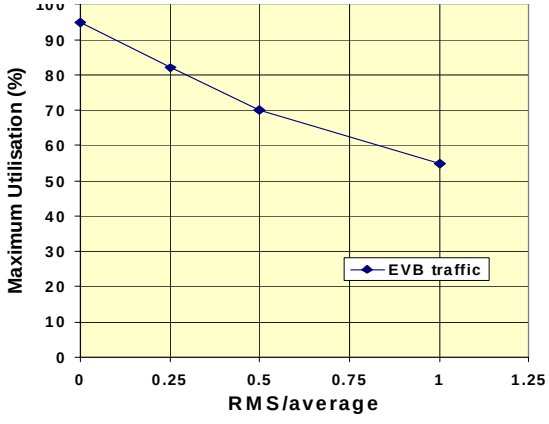

Fig. 6 Switch efficiency for different event size variations

By extrapolation, the available bandwidth for each FRL is hence 275 MB/s, well above the required 200 MB/s.

The switch behavior for unbalanced data sources has also been measured: the results are shown in Fig. 7. There is no significant efficiency loss when data sources are unbalanced: the efficiency is 45 % when four sources provide events with an average size of 1 kB along with four sources providing events with an average of 3 kB (unbalance ration of 3). By extrapolating on this later case, the available bandwidth for each FRL is 225 MB/s, still above the required 200 MB/s.

With these measurements in realistic conditions, we see that the D2S provides the required bandwidth for the Front-end data sources as well as traffic balancing. This contributes to good working conditions for the next stage of the event building process.

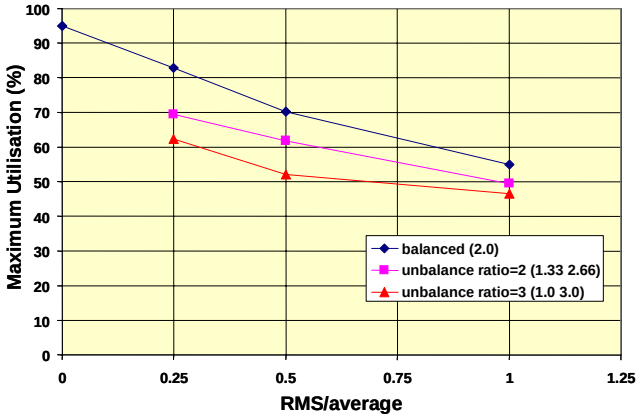

Fig. 7 Switch efficiency for unbalanced data sources

IV. Deployment Scenario

Each FED builder has 8 output ports. The system is scalable with the number of used ports. A DAQ reading out all data sources but at 1/8th of maximum trigger rate is obtained (1 MB @ 12.5 kHz) by connecting one RU to each FED builder and connecting these RUs to a RU builder slice.

By connecting a second RU to each FED builder and a second RU builder slice (odd events assigned to output 1 and even events to output 2 for example), a DAQ with 1/4th capacity is obtained (1 MB @ 25 kHz).
This principle works up to the full DAQ capacity (1 MB @ 100 kHz) by connecting 8 RUs to each FED builder and using the event number as base for output assignment.

The D2S architecture allows a progressive deployment of the rest of the DAQ devices. It is also possible to use different technologies when new slices must be installed and hence profit from the future technological evolutions.

V. Summary

After reviewing the requirements of the readout of the CMS Data Acquisition system as well as the main characteristics of the data producers, the architecture of the Data to Surface (D2S) system is presented. The average amount of data produced varies among the data sources whereas the operation of the event builder with high efficiency requires that all inputs carry the same amount of data. The D2S is designed to solve this problem by providing a first stage in the event building process. The D2S concentrates several data sources into an

output channel and multiplexes the event data to different streams in the second stage of the event building process. The D2S output channels therefore provide more evenly distributed data sizes to the second stage of the event builder. Moreover, the multiplexing allows for a scalable design for the second stage of the event builder, resulting in a system that can be procured and installed in phase with the requirements arising from the performance of the accelerator and the experiment itself.

VI. REFERENCES

[1] The S-LINK64 bit extension specification: S-LINK64
A. Racz, R. McLaren, E. van der Bij
http://hsi.web.cern.ch/HSI/s-link;
The S-LINK Interface Specification
O. Boyle, R McLaren, E. van der Bij
http://hsi.web.cern.ch/HSI/s-link

[2] Read-out Unit Working Group WEB pages
http://cmsdoc.cern.ch/cms/TRIDAS/horizontal
then click on “RUWG” in the main banner

[3] FED-Kit design for CMS DAQ system
E. Cano et al. in these proceedings

[4] http://www.myri.com