Performance of the CMS Event Builder

Jean-Marc Andre[5], Ulf Behrens[1], James Branson[4], Philipp Brummer[2], Olivier Chaze[2], Sergio Cittolin[4], Cristian Contescu[5], Benjamin G. Craigs[2], Georgiana-Lavinia Darlea[6], Christian Deldicque[2], Zeynep Demiragli[6], Marc Dobson[2], Nicolas Doualot[5], Samim Erhan[3], Jonathan Richard Fulcher[2], Dominique Gigi[2], Maciej Gładki[2], Frank Glege[2], Guillelmo Gomez-Ceballos[6], Jeroen Hegeman[2], Andre Holzner[4], Mindaugas Janulis[2a], Raúl Jimenez-Estupiñán[2], Lorenzo Masetti[2], Frans Meijers[2], Emilio Meschi[2], **Remigius K. Mommsen**[5], Srecko Morovic[2], Vivian O'Dell[5], Luciano Orsini[2], Christoph Paus[6], Petia Petrova[2], Marco Pieri[4], Attila Racz[2], Thomas Reis[2], Hannes Sakulin[2], Christoph Schwick[2], Dainius Simelevicius[2a], Petr Zejdl[5b]

[1]DESY, Hamburg, Germany [2]CERN, Geneva, Switzerland [3]UCLA, Los Angeles, California, USA [4]UCSD, San Diego, California, USA [5]FNAL, Chicago, Illinois, USA [6]MIT, Cambridge, Massachusetts, USA
[a]also at Vilnius University, Vilnius, Lithuania [b]also at CERN, Geneva, Switzerland

presented at CHEP 2016 in San Francisco, California, USA

The CMS Event Builder

The CMS event builder collects event fragments from approximately 700 detector front-end readout drivers (FEDs) and assembles them into complete events at a maximum level-1 trigger rate of 100 kHz. A software-based high-level trigger (HLT) selects O(1%) of these events.

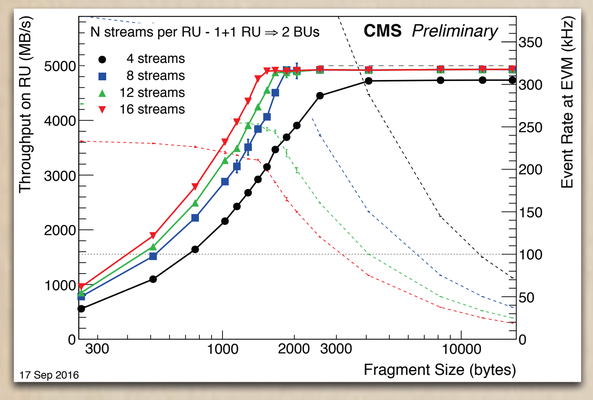
The FEDs deliver for each trigger an event fragment of 0.1-8 kB, depending on the subsystem and/or the instantaneous luminosity, yielding an event size of up to 2 MB. The event building is done in two steps: First, the data concentrator aggregates 1-18 FEDs into a super-fragment, and second, the event builder assembles the super-fragments into complete events.

FEROL

Front-end readout drivers (FEDs) are connected over point-to-point links to custom front-end readout optical link (FEROL) boards. One or two legacy FEDs are connected via copper cables (200/400 MB/s), while the new µTCA-based FEDs uses 10 Gbit/s optical fibers. In both cases a custom protocol is used. The FEROL translates this protocol into an in-FPGA, one-directional 10 Gbit/s TCP/IP connection which is routed to commercial network equipment.

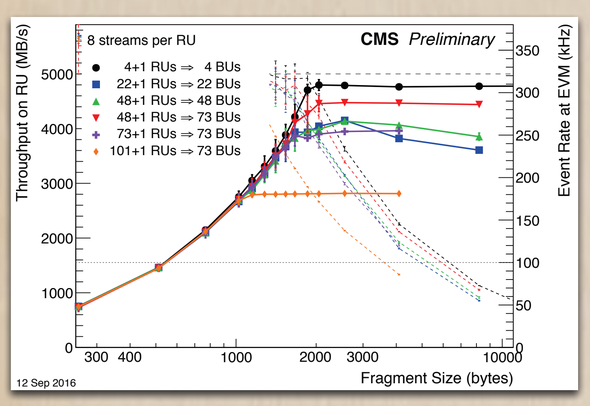
Data Concentrator

The FEROL sends the data from each FED as a TCP/IP stream over optical 10 GbE to a pre-defined readout unit (RU) computer. A series of switches is used to concentrate the data from 1-18 streams into 40 GbE and to transport the data to the surface. The RU splits the streams into FED fragments, checks the consistency and assembles the data belonging to the same event into a super-fragment.



Event Builder

The super-fragments are stored on the readout unit (RU) computers. Upon an event request, the super-fragment is send over the event-builder switch to the builder unit (BU) machines. The event-builder switch uses Infiniband FDR at 56 Gb/s. It employs a Clos-network structure with 12 leaf and 6 spine switches. The BU assembles the super-fragments into complete events and checks their consistency. Finally, the BU writes the event to files residing on a local RAM disk. Each BU has 12 or 16 filter-unit machines assigned to run the high-level trigger code. The event selection is using standard CMSSW processes. They read the events from files on the RAM disk mounted via NFSv4.



Event Building Protocol

(1) Any builder unit (BU) which has resources to build events sends a request for a predefined number of events to the event manager (EVM).

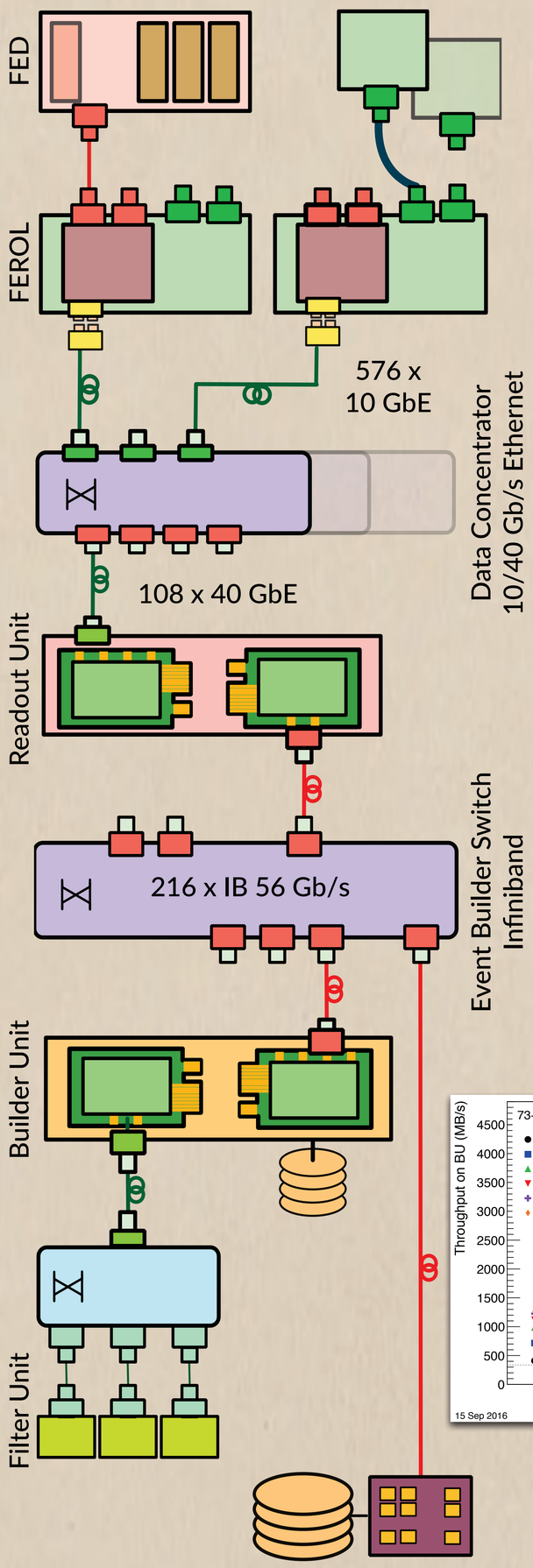
(2) The EVM and readout units (RUs) receive asynchronously the event fragment from the assigned FEROLs. The EVM is connected only to the FEROL sending the level-1 trigger fragment.

(3) Once the EVM received enough event fragments from the trigger to satisfy the number of requested events, or if a timeout is reached, the EVM sends a message to all RUs. This message contains the level-1 trigger numbers of the events assigned to the BU which requested the events.

(4) At the same time, the EVM packs the event fragments corresponding to the requested events into an Intelligent Input/Output (I2O) message and sends it to the BU.

(5) When the RU receives the event assignment from the EVM, it combines all FEROL fragments into a super-fragment. Each RU knows which FEROLs participate in the readout and waits until it has received all fragments. The super-fragments for all events assigned to the given BU are packed into one or several I2O messages with a fixed size of 128 kB. They are then send to the BU.

(6) Once the BU has received the super-fragments from all RUs, it builds the event. Each super-fragment message contains the identifiers of all RUs participating in the event building. Therefore, the BU knows on an event-by-event basis if it got the super-fragments from all RUs. The BU checks the consistency of the event by verifying the correct structure of the FED data, the trigger number embedded in each FED fragment, and the CRC of the FED payload. Finally, the event is written to a file residing on the RAM disk.

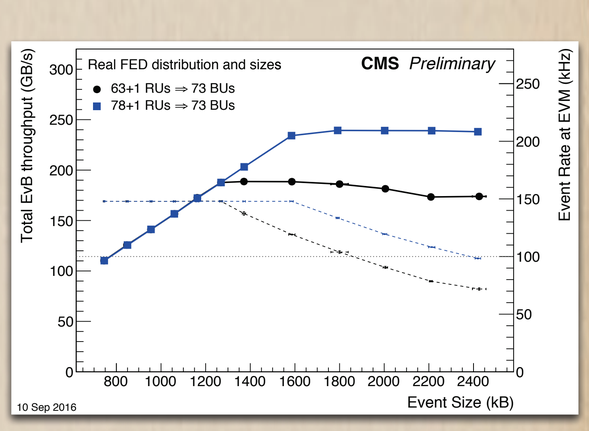


DAQ2

The CMS data-acquisition system from run 1 (DAQ 1) was upgraded during the first long-shutdown of the LHC (2013/14) because of the aging of the existing hardware (both PCs and network equipment older than 5 years), and because sub-detectors with upgraded off-detector electronics had needs that exceed the original specification of the DAQ system. The run-2 DAQ system is an order of magnitude smaller than DAQ 1.

System Performance

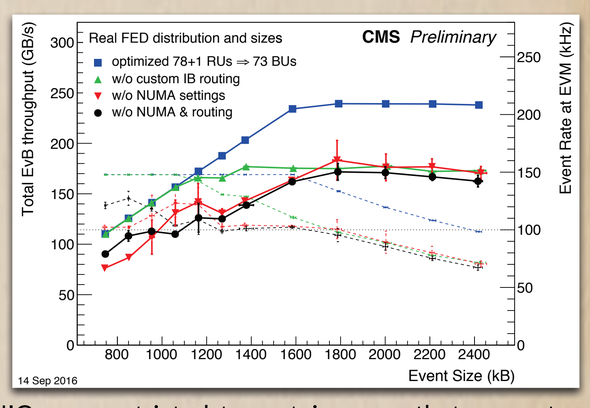
The full system performance is measured by generating event fragments on the FEROLs. The fragment sizes and their log-normal distribution has been measured as function of instantaneous luminosity during physics data taking. These parameters are used to extrapolate the system behavior to larger event sizes. Events are assembled by the BU, but not written to disk. The bandwidth to the HLT would limit the total throughput to ~175 GB/s.



Performance Tuning

The event-building system has to be optimized in order to exploit the full capability of modern computing and network hardware. The applications use multiple threads to assemble and handle events in parallel. In addition, the number of memory copies needs to be kept to a minimum. In order to cope with the non-uniform memory architecture (NUMA), we had to bind each thread and memory structure to specific CPU cores.

Moreover, the interrupts from the NICs are restricted to certain cores that are not used for any data handling. Finally, we worked out a custom routing scheme for the Infiniband to account for the uni-directional event-building traffic.



Builder Unit

The BU builds and writes event using multiple threads working in parallel. Each thread writes its own file to avoid any locking. The event fragments are kept and checked in the RDMA buffers of the Infiniband NIC and then assembled directly into files on the RAM disk without leaving the kernel space. The useable throughput is limited by the speed with which the filter units can read the data from the RAM disk. The BU throttles the event requests when the HLT cannot keep up with processing the files.