Performance of the NOvA Data Acquisition and Trigger Systems for the full 14 kT Far Detector

2015 J. Phys.: Conf. Ser. 664 082041

(http://iopscience.iop.org/1742-6596/664/8/082041)

Performance of the NOνA Data Acquisition and Trigger Systems for the full 14 kT Far Detector

A. Norman[1], G.S. Davies[2], P.F. Ding[1], E.C. Dukes[3], H. Duyan[4], M.J. Frank[3], R.C. Group[3], A. Habig[5], W. Henderson[3], E. Niner[2], R. Mina[3], A. Moren[5], L. Mualem[6], Y. Oksuzian[3], B. Rebel[1], P. Shanahan[1], A. Sheshukov[7], M. Tamsett[8], K. Tomsen[5], L. Vinton[8], Z. Wang[3], B. Zamorano[8], J. Zirnstien[7]

[1]Fermi National Accelerator Laboratory, Bativa IL, USA
[2]Indiana University, Bloomington IN, USA
[3]University of Virginia, Charlottesville VA, USA
[4]University of South Carolina, Columbia SC, USA
[5]University of Minnesota, Duluth MN, USA
[6]California Institute of Technology, Pasadena CA, USA
[7]University of Minnesota, Minneapolis MN, USA
[8]University of Sussex, Brighton, UK

E-mail: `anorman@fnal.gov`

Abstract. The NOvA experiment uses a continuous, free-running, dead-timeless data acquisition system to collect data from the 14 kT far detector. The DAQ system readouts the more than 344,000 detector channels and assembles the information into an raw unfiltered high bandwidth data stream. The NOvA trigger systems operate in parallel to the readout and asynchronously to the primary DAQ readout/event building chain. The data driven triggering systems for NOvA are unique in that they examine long contiguous time windows of the high resolution readout data and enable the detector to be sensitive to a wide range of physics interactions from those with fast, nanosecond scale signals up to processes with long delayed coincidences between hits which occur at the tens of milliseconds time scale. The trigger system is able to achieve a true 100% live time for the detector, making it sensitive to both beam spill related and off-spill physics.

1. Overview

The NOνA experiment is a long baseline neutrino oscillation experiment, designed to measure the oscillation probabilities for $\nu_\mu \to \nu_e$ and $\nu_\mu \to \nu_\mu$ in both neutrino and anti-neutrino beam configurations. The NOνA detectors were designed as "fully active" range stack/calorimeters using an orthogonal arrangement of X/Y measuring planes of detectors cells[1]. The 14 kt far detector is comprised of over 344,000 active readout cells. The light signals collected from these cells are continuously digitized and readout. The data pathway from these front end digitization boards and into a large event building and buffering compute farm, form the core of the detector's data acquisition chain.

The NOνA experiment presents a unique challenge for both the development and operation of the data acquisition system and for the development and operation of the trigger systems which are used to readout and examine the detector data streams. In particular, the NOνA experiment

has developed a fully asynchronous trigger system which is able to both accept external trigger inputs coming from the Fermilab accelerator complex and simultaneously examine the full unfiltered detector data stream to reconstruct the data, identify specific hit topologies and issue triggers corresponding to these topologies. The NOνA trigger has run at full scale on both the NOνA far detector and the near detector in a physics production mode for more than twelve months. During this period over 1.7 billion physics triggers have been issued corresponding to 91 TB[1] of recorded physics data. This data set correspondings to the examination of more than 24 million seconds of live readout[2]. These data have been used to provide the basis for physics analysis as well as detector characterization, calibration, performance and background studies.

2. DAQ & Trigger Design

The NOνA data acquisition and trigger system is at one level very simple, because there is a single detector readout technology that is used across the experiment. However, that detector readout element is repeated over 344,000 times on the far detector and readout through a hierarchical arrangement of custom electronics, networking equipment and high throughput computing systems. Each readout element in the system is precisely synchronized to every other readout component in the system and to their counter parts at the NOνA near detector 810 km away from the far detector site[2]. These synchronized detector elements are placed in a continuous, free running readout mode where they produce waveform readouts of their corresponding detector cells. Compounding this complexity, the entire system operates on the Earth's surface, where the detector is exposed to over 180 kHz of cosmic ray induce activity.

The combination of all these factors produces a continuous, unfiltered data stream that is aggregated through multiple levels of event building and then routed to the RAM of a large computing cluster to be temporarily stored until a final trigger decision can be made. This "buffer and hold" design, is able to efficiently store the continuous readout data from the entire detector at an average ingest rate of 6.7-8.1 Gb/s and a peak rate of 33.6 Gb/s. The maximum duration for which this data is held is determined by the total memory capacity of the buffer nodes and the number of nodes that are devoted to the readout during actual running. Three main buffering configurations were used by the DAQ during the first year of operations. The configurations dedicated a total of 45, 225 or 470 GB of memory to live data buffering across the system. This allowed for between 220-1.2K detector frames[3] to be stored on each buffer node as shown in Fig. 1. When aggregated system wide this allowed 44 s, 3.7 min or 7.7 min respectively of "hold" time for the detector data, as shown in Fig. 2. The data is held in a circular buffer fashion, where the oldest data is continually allowed to expire where upon it is overwritten with the most recent data from the detector.

The NOνA trigger systems were designed to take advantage of the long time to live of data in this "buffer and hold" hold design, by operating completely asynchronously to the primary DAQ readout path. In this model, when the trigger systems arrive at a decision for a particular time window of the NOνA readout data, the trigger system instructs the DAQ to search the buffered data for the time interval of interest. Once the data is located it is copied out of the active buffer and sent to the final stage of event building where it is combined with the data from other buffers and recorded to disk.

3. Trigger Subsystems

The NOνA trigger was designed as an integral part of the data acquisition system, but split fundamentally into three separate components subsystems to handle external trigger sources

[1] Excludes minimum and zero bias data collection use for background and calibration. Total triggers collected exceeds 5.7 billion.

[2] Average data acquisition uptime >0.85 at trigger livetime >0.90 for twelve months of running

[3] NOνA readout data is discretized internal to the DAQ system in 5 ms detector frames

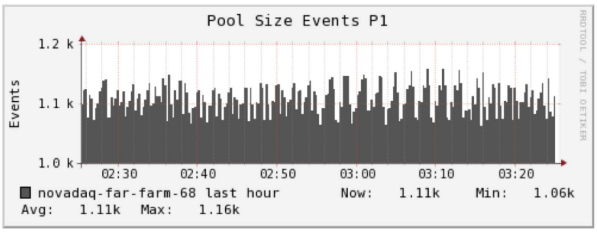

Figure 1. Time profile of the number of 5 ms event frames being held in the buffer memory of a single buffer system under nominal running.

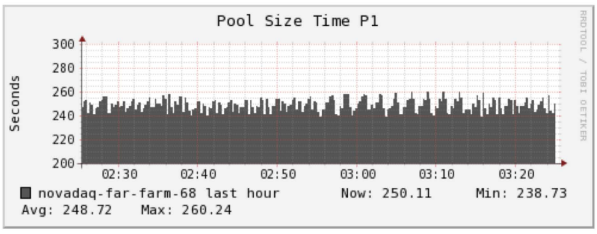

Figure 2. Time profile of the "look back" depth or maximum latency to trigger decision, the buffer farm is holding across the full system.

such as signals from the NuMI accelerator complex, high level data driven triggering, and internal triggering, trigger aggregation, rate balancing and pre-scaling. These subsystems each communicate with each other as well as with key components of the primary event building and data logging chain to produce the final logging of data. However, because the trigger systems were design to operate in a fully asynchronous mode, each systems maintains an isolation layer or interface that prevents faults in the system from feeding back to other parts of the DAQ.

3.1. Beam and External Triggering

The external triggering component of the trigger system is primarily used to propagate information from the NuMI accelerator complex to the different detector locations. The subsystem uses a "spill" or "event" server which runs at a specific location or on specific hardware. The server is designed to decode input information and generate a time window with a start time t_0 and a duration Δt that should be recorded by the NOνA detectors. In order to correlate the time window with the actual data at the detector sites, the times t_0 are required to be derived from GPS based clock source. The GPS derived time is then converted into the 64 MHz NOνA time standard which can be directly used to extract data at the different detector locations.

The external triggering systems do not add any delays to their t_0's or in any other way modify the times they report. Instead offsets and delays are added to the base trigger times by the systems receiving the signals. This allows for a spill server to operate without any a priori knowledge of the endpoint systems that will use its information and a the same time allows a single spill server broadcast its information to multiple recipients.

NOνA has created dedicated spill servers which run on the powerpc platform[4] used by its timing system[3] under an embedded linux operating system. These devices are synchronized to GPS and are able to decode and time stamp the Fermilab accelerator clock signals. These spill servers determine the beam spill windows used by the oscillation analyses, as well as providing other event markers that allow for calibration and alignment of the DAQ systems.

Separately, NOνA has created an event server, based on the same modular architecture as the spill server, which runs on an x86 linux platform at Brookhaven National lab. This event server runs as part of the Super Nova Early Warning System (SNEWS)[4] and takes as input messages from neutrino experiments around the world that are attempting to detect the neutrino flash from a supernova[5]. In the event of a supernova signal from one of the participating experiments, it generates a $t_0 + \Delta t$ time window in the NOνA time base similar to other NOνA triggers. This event server also generates regular "test" and heart beat triggers that are used to validate the communications chains used by the system.

In addition to the spill server, the external triggering system utilizes a series of receivers and

[4] Custom system with embedded Freescale 8347E microprocessor

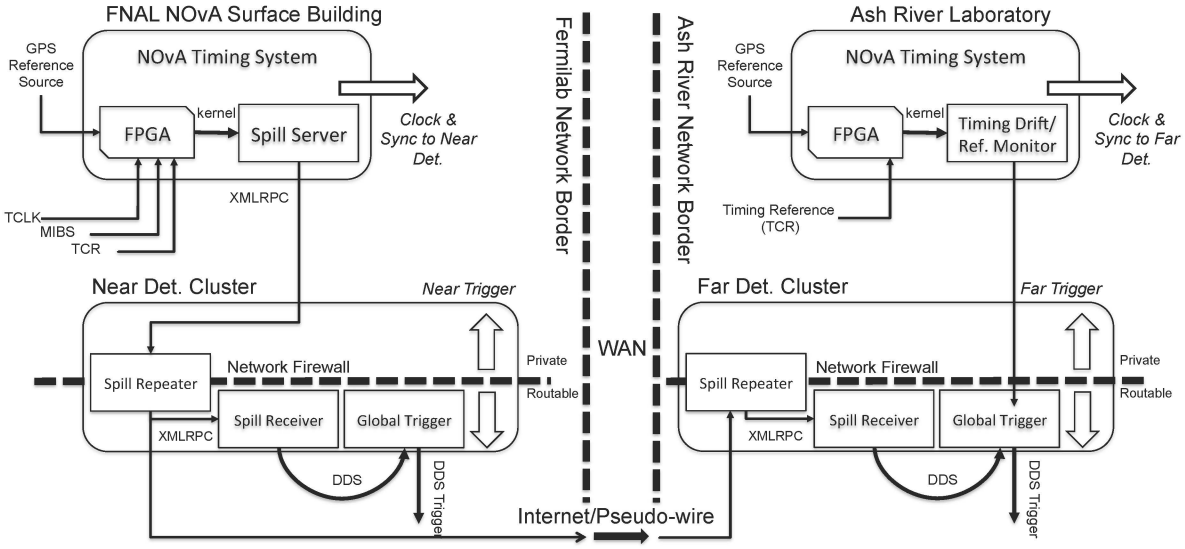

Figure 3. External triggering subsystem used by the NOνA experiment. Spill servers decode the Fermilab accelerator signals and then broadcast information to a series of repeater and receiver stations using the XMLRPC transport and Data Distribution Service (DDS) protocols. This system is used to cross network boundaries within the DAQ systems and to send information between sites using the internet.

repeaters. The receiver/repeater design allows for triggers generated within private LANs to traverse gateways and firewalls and allow for communication between the trigger source and the different NOνA detector sites to be carried over public WANs, as shown in Fig. 3.

3.2. Data Driven Triggering

The data driven triggering (DDT) system that NOνA has developed exploits the deep data buffering that the DAQ systems uses, to examine the full unfiltered data stream that is generated by the detector. The DDT system executes a full featured reconstruction and analysis suite based on the art/artdaq[6, 7] frameworks. The results of the analysis modules are then issues as trigger decisions to the DAQ and recorded through the normal event building and logging chain.

The DAQ and DDT systems interact through an interface that is made on the compute nodes that serve as the data buffers for the raw data stream. The second stage DAQ event builder applications, which run on the buffer farm computing, first aggregate and assemble all of the data from the different regions of the detector into a single 5 ms long time frame representing the complete readout of the detector. The event builder then pushes a copy of this data both to its own internal circular buffer cache and to a System V style shared memory segment. The shared memory segment has implemented on top of it a "single writer/multiple reader" queue. The queue is presented to the writer as a circular buffer, while the queue is presented to the readers as a fixed length FIFO with a destructive read. This architecture, shown in Fig. 4, allows for a large number of reading processes to concurrently access the event queue without there being a duplication in the data that each reader receives. The design also prevents read/write contention and allows the writing process to keep statistics on the number of data frames that have been examined or that have been overwritten/dropped without being examined. The shared memory design insulates the DAQ from the DDT and prevents run time anomalies in the analysis of the data from providing back pressure on the readout.

The trigger filter applications are all independent of each other and operate in parallel on the

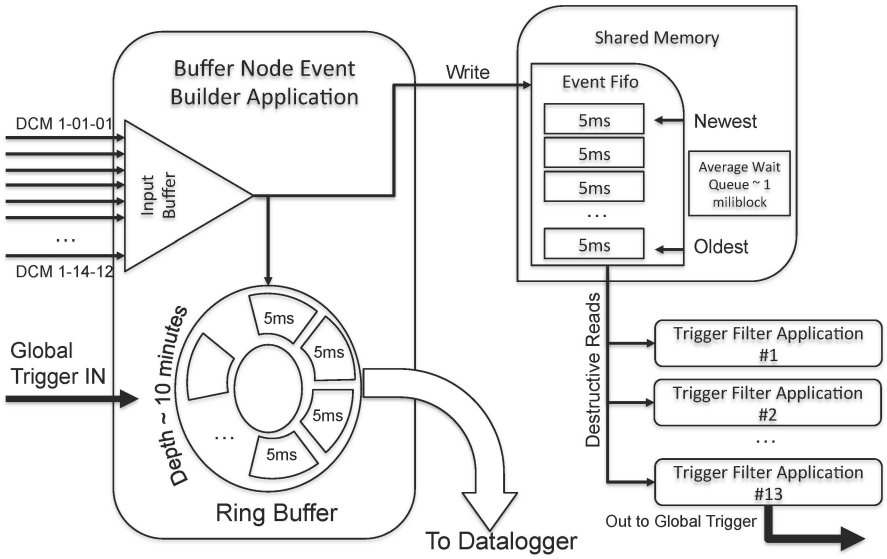

Figure 4. The DAQ event builder application shares detector with the DDT system through a shared memory interface with a single writer, multiple reader FIFO interface.

host buffer computer. Each filter application executes the same suite of reconstruction/analysis code. The DDT is designed to take full advantage of modularity and as a result each physics trigger is defined as a series of "paths" through a set of algorithm modules. This design allows for complex triggers to be built from smaller units of work. The algorithms modules can be shared between different trigger paths and at run time the different active trigger paths are cascaded in such a way that each module is run at most once. This cascade is shown in Fig. 5 for a subset of the physics triggers run by NOνA. This prevents common but expensive operations like hit sorting, pattern recognition and tracking finding from being duplicated and inadvertently run multiple times. This also allows for multiple branches or entire trigger paths to be aborted early in their execution if earlier reconstruction or tracking fail to produce candidates.

One of the design choices that was made in the NOνA DAQ and Trigger, was that trigger decision would be independent of each other decision that is being run or could be run on a block of data. This results in each trigger that is issued having one and only one trigger bit set. Blocks of data that satisfy more than one trigger have multiple trigger messages generated for them. This means that there is no need to hold off decisions until all trigger paths have completed, and allows for out of order execution of modules to minimizes the total time to decision for each trigger path, as shown in Fig. 6 for the primary physics triggers run during the first twelve months of data taking.

The average time to decision $< t_{dec} >$ that the DDT system needs to maintain in order to keep up with the input data rate from the detector is determine by the frequency, f_{DAQ} with which new data arrives at each buffer node, the number of buffer nodes, N_{buff}, in the system and the number of independent analysis applications, M_{filt} that are running in parallel,

$$< t_{dec} >= N_{buff} M_{filt} / f_{DAQ} \tag{1}$$

The number of DDT filter applications that could be run concurrently was set to 13 based on optimization of the number of cpu cores that were available on each of the buffer nodes and the number of cores that were strictly reserved for both the DAQ event builder and for the general system processes. The NOνA DAQ system was set to nominally round-robins data to a different buffer nodes every 5 ms during physics data taking. The number of buffer nodes that were used

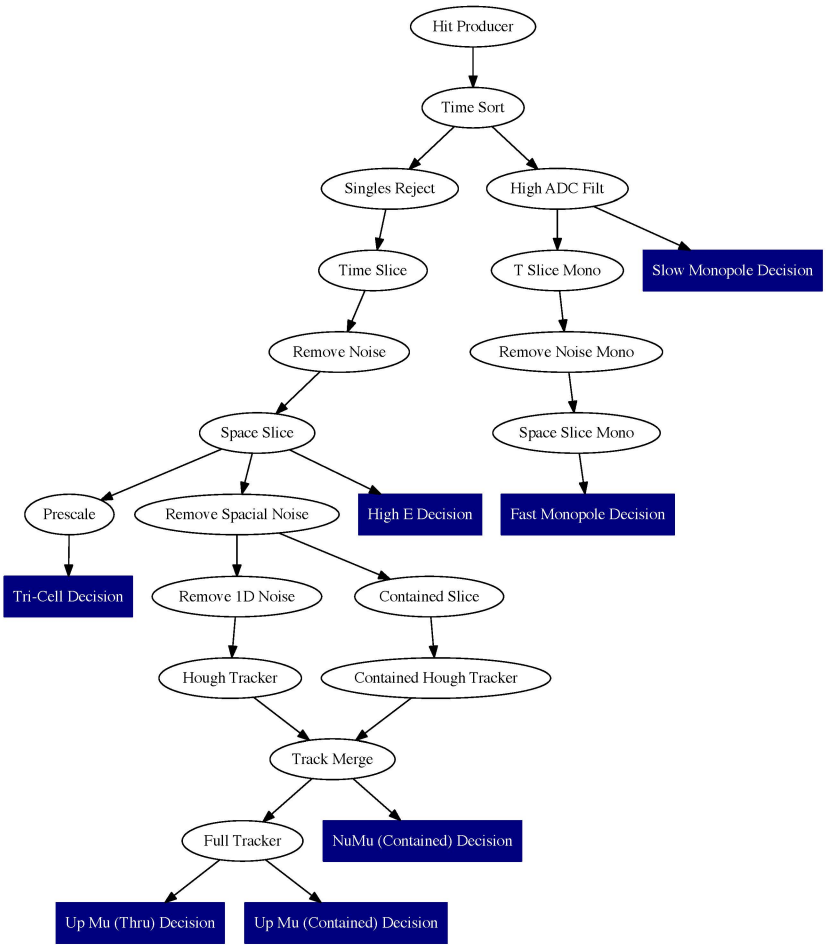

Figure 5. Cascade of trigger/reconstruction modules used for for upward going muon, magnetic monopoles, calibration and high energy deposition triggers during the 2014/2015 run period. Trigger decision points shown in blue. The modular design allows for different paths and branches within paths to be aborted after any module in the cascade.

during the first twelve months of operations were limit by the scaling behavior of the DAQ computing cluster and by the cooling that was available in the far detector's computing center. The experiment ran primarily with 45 buffer nodes during 2014 and expanded to 94 buffer nodes during the spring of 2015. Under these conditions the target average time to decision for the DDT system was 2.9 s and 6.1 s respectively. The number of different physics triggers and operational parameters of the triggers were tuned to account for the time budget. The plot in Fig. 7 shows the resulting live times that the DDT system ran under during different periods. These live times represent the total fraction of continuous readout for the full detector that was fully examined and analyzed by all the active DDT triggers.

3.3. Global Trigger

The NOνA global trigger subsystem is responsible for the aggregation of input coming both from the external trigger systems and from the DDT trigger system, as well as internally generated triggers for minimum and zero bias readout. The global trigger is responsible for issuing the

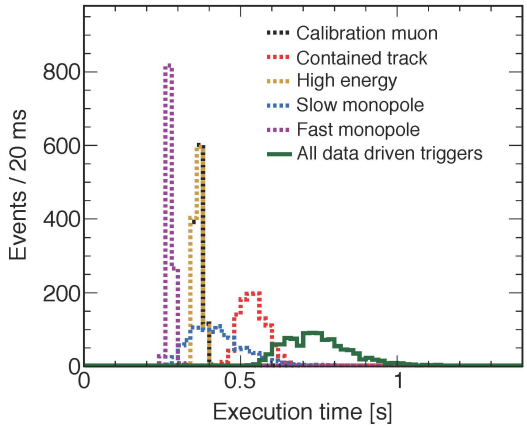

Figure 6. NOνA trigger execution time for leading trigger processes. Total average time to decision < 1 s.

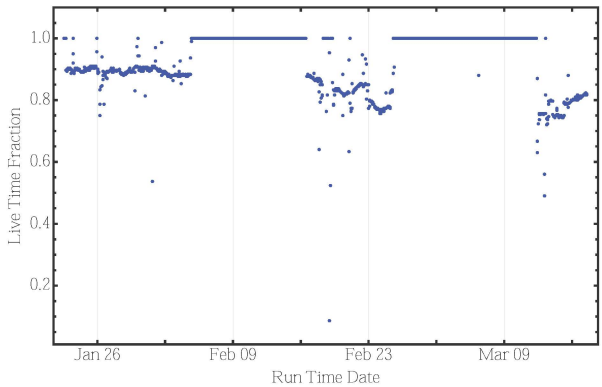

Figure 7. NOνA trigger live time corresponding to the faction of absolute wall time that the trigger was able to process to decision across all paths.

message that are directed to the data buffers and to the data loggers which cause the information being held in the buffers to be searched and the data corresponding to the time window of interest to be copied out to the final stage of event building and logging.

The aggregation of trigger across the internal, external and DDT systems is performed through a series of trigger message reception pools which listen for incoming trigger information from the beam spill servers, SNEWS network and DDT systems. These pools use the Data Distribution Service (DDS) messaging protocol which provides a publish/subscribe model for sending messages between applications. When a trigger messages is received, the corresponding data is unpacked and examined to construct an appropriate trigger object for the DAQ system to use for data logging. At the same time the trigger is classified and, if requested, relevant per trigger prescaling and dynamic throttling is performed. The trigger objects are then pushed into a series of prioritized queues corresponding to their general classification and assigned priority. These queues act as standard fixed length double ended FIFOs with atomic push and pop operations. This design allows for high levels of concurrent read and write access to the FIFOs without incurring latency penalties related to locking operations.

The trigger queues are serviced by an algorithm that sequentially drains the queues based on their priority. In this algorithm the queues are checked in priority order until a queue with non-zero occupancy is found. A fixed number of triggers in that queue are then drained into the pipeline for broadcast to the rest of the DAQ. The algorithm then moves to the next non-empty queue in the chain and repeats the process, returning to the highest priority queue only after all other queues have been serviced. This design ensures that high priority triggers are issued within a maximum latency window, while simultaneously preventing lower priority triggers from being "squeezed out" by higher priority triggers with high real time rates.

Each trigger pushed into the pipeline for broadcast can be individually prescaled and throttled so that the rates for groups of related triggers can be controlled. The broadcast pipeline shapes the final trigger traffic by imposing an intra-trigger transmission delay along with global prescaling and dynamic throttling of the trigger stream. This traffic shaping is done to prevent collisions and saturation at the network level with transmission of data between buffer nodes and datalogger and with overrunning the message queues of the buffer nodes and datalogger applications. This allows the full bandwidth of the system to be used even during bandwidth

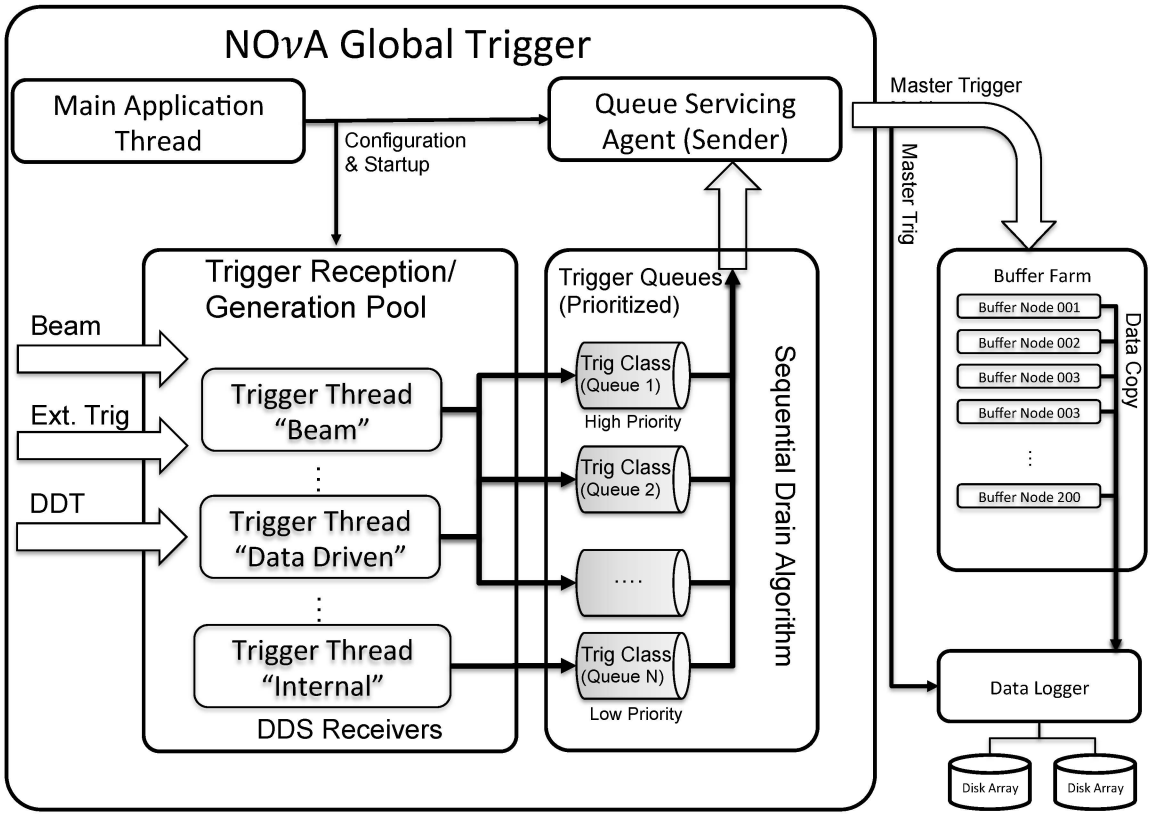

Figure 8. The NOνA global trigger uses a series of trigger reception pools and prioritized queues to distribute triggering instructions to the DAQ buffer and logging systems.

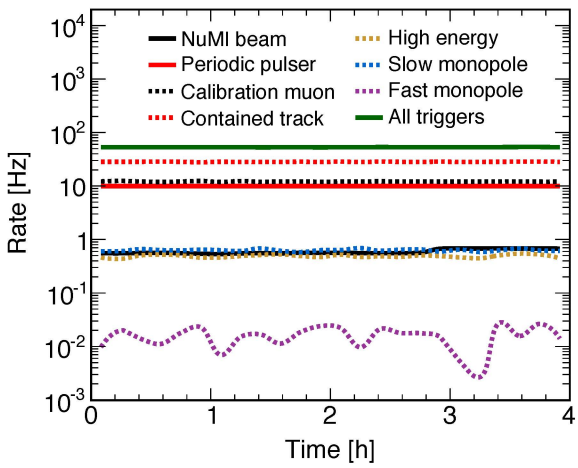

Figure 9. Trigger rate profile and stability for primary trigger streams after shaping by the global trigger system.

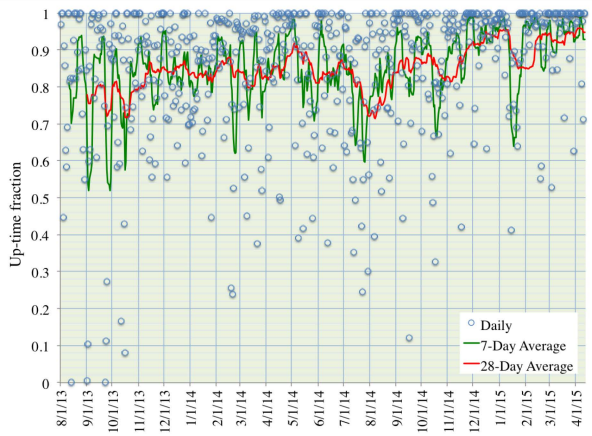

Figure 10. DAQ and trigger up time fractions for the first 22 months of data taking. Fractions include downtime for regular detector and accelerator maintenance as well as downtimes caused by systems outside of the DAQ/Trigger environment.

intensive readout periods without crowding out individual trigger streams, or causing back pressure on the rest of the DAQ systems. In particular the ability to rate limit and shape the DAQ traffic has allowed for the readout of continuous 60 second long time blocks of data, as are required by the SNEWS supernova triggers, while at the same time reading out a high rate of 50 μs time windows corresponding to physics objects in the detector. This allows the system to operate stably with triggers that operate with rates that differ by four orders of magnitude, as shown in Fig. 9 and data volumes that differ by more than six order of magnitude.

4. Summary

The NOνA experiment has successfully created, commissioned and is operating, a continuous readout, free running, dead-timeless DAQ system that is able to scale to readout the more than 344,000 detector elements of the full 14 kt NOνA far detector. In addition to the core DAQ readout systems, the experiment has also created a robust trigger system that is able to leverage the deep readout buffers of the DAQ to perform highly a synchronous triggering and recording of data from the detector even minutes after the initial signals.

This design has also lead to the successful creation of a high level data driven trigger system that is capable of analyzing the full, unfiltered NOνA data stream and performing full pattern recognition and event reconstruction on the data. This data driven approach has allowed for the development of trigger algorithms that can identify prompt physics topologies, like the upward going muon topology[8], which require single hit timing resolutions at the level 10 ns, to extremely long time duration topologies, like the search for slow magnetic monopoles and for supernova neutrino bursts, that can only be seen on millisecond time scales.

The DAQ and trigger have combined to provide a system that has operated at high efficiency and near 100% live time while reading triggers windows with durations 50 μs to 1 minute in length. The systems have combined to collected over 5.7 billion physics, background and calibration events which are currently under analysis.

Acknowledgements

The author acknowledges support for this research was carried out by the Fermilab scientific and technical staff. Fermilab is Operated by Fermi Research Alliance, LLC under Contract No. De-AC02-07CH11359 with the United States Department of Energy

References

[1] Ayres D *et al.* 2007 The NOvA technical design report Tech. rep. FERMILAB-DESIGN-2007-01
[2] Niner E *et al.* 2014 *J.Phys.Conf.Ser* **513** 012028
[3] Norman A *et al.* 2012 *J.Phys.Conf.Ser.* **396** 012034
[4] Antonioli P, Fienberg R T, Fleurot F, Fukuda Y, Fulgione W, Habig A, Heise J, McDonald A B, Mills C, Namba T, Robinson L J, Scholberg K, Schwendener M, Sinnott R W, Stacey B, Suzuki Y, Tafirout R, Vigorito C, Viren B, Virtue C and Zichichi A 2004 *New Journal of Physics* **6** 114 URL `http://stacks.iop.org/1367-2630/6/i=1/a=114`
[5] Habig A and Zirnstein J 2015 Integration of the super nova early warning system with the nova trigger *21st International Conference on Computing in High Energy and Nuclear Physics*
[6] Green C, Kowalkowski J, Paterno M, Fischler M, Garren L and Lu Q 2012 The art framework *Journal of Physics: Conference Series* vol 396 (IOP Publishing) p 022020
[7] Biery K, Green C, Kowalkowski J, Paterno M and Rechenmacher R 2013 *Nuclear Science, IEEE Transactions on* **60** 3764–3771
[8] Mina R *et al.* 2015 Implementation of an upward-going muon trigger for indirect dark matter searches at the noνa far detector *21st International Conference on Computing in High Energy and Nuclear Physics*