PROCEEDINGS OF SCIENCE

Stress testing Ethernet Switches for NectarCAM in the Cherenkov Telescope Array with a synchronous UDP frame generator

D. Hoffmann[a], **J. Houles**[a], **F. Louis**[b], **Y. Moudden**[b], **P. Sizun**[*b] **for the NectarCAM and CTA consortia**[†]

[a] *Centre de Physique des Particules de Marseille, Aix Marseille Université, CNRS/IN2P3, CPPM UMR 7346, F-13288, Marseille, France;*

[b] *DSM / IRFU / SEDI, CEA / Saclay, F-91191 Gif-sur-Yvette, France*

W. Gu, B. Raydo

Fast Electronics Group, Jefferson Lab, Newport News, VA 23606, USA

E-mail of corresponding author: Hoffmann@CPPM.In2p3.Fr

The Cherenkov Telescope Array (CTA) will be the next generation ground-based gamma-ray instrument. It will be made up of approximately 100 telescopes of at least three different sizes, from 4 to 23 meters in diameter. The NectarCAM is a Cherenkov camera proposed for the Mid-Size Telescopes of CTA. Its characteristics make it one of the most challenging camera projects for a high speed data acquisition (DAQ) system in CTA has due to its average output rate of up to 40-Gbps on 265 Ethernet 1000baseT links, bundled to 4×10Gbps on four optical links and reduced to 10 Gbps after event-building.

This paper presents results on characterisation and validation procedures carried out on several Ethernet switches, which have been considered as hardware for the camera-internal data traffic of NectarCAM. Two complementary types of data generators, one highly synchronous with up to 64 1-Gbps channels based on an FPGA core, the other with up to 320 1-Gbps channels working on 64 Scientific Linux boards, have been built and used to stimulate the DAQ system with six Ethernet switches and a standard Linux PC for IP packet reception.

The 34th International Cosmic Ray Conference,
30 July- 6 August, 2015
The Hague, The Netherlands

*Speaker.

†Full CTA consortium author list at https://cta-observatory.org

PoS(ICRC2015)697

1. Introduction

The NectarCAM [4] data acquisition system is designed to use commercial switches to funnel 265 Gb Ethernet links from the front end boards into four 10-Gb links to the camera server. Following the event of a camera trigger, each of the front end boards sends out to the camera server the data they captured, framed in one UDP packets. These packets are called event fragments and the assembly of these fragments in the camera server is called event-building.

The schedule between trigger input and data output is deterministic: The UDP frames are built in a data pipeline implemented in programmable logic in the front end FPGA. In the current design, there is no clock distribution to the camera front end boards: each front end board has its own clock source. Given that the trigger pulse is distributed to all camera front end boards simultaneously, the corresponding data frames are sent out to the back end switches nearly simultaneously with a spread of several nano-seconds due to differences in cable lengths.

This design imposes severe constraints on the camera-internal network structure, which are described. We present our conclusions from stress tests and consequences for the selection of network switches for the NectarCAM infrastructure.

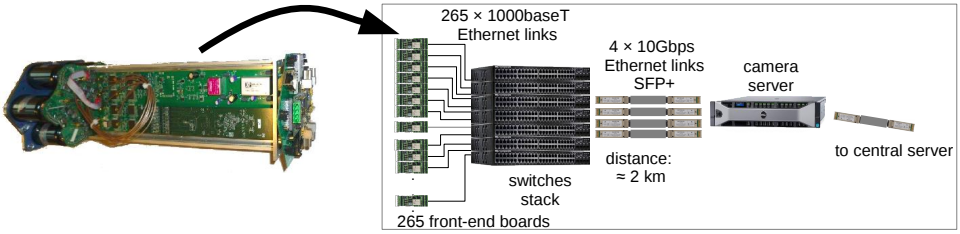

Figure 1: Illustration of the physics data flow from a NectarCAM front-end board to the camera server and further downstream to central CTA DAQ: The 265 NectarCAM modules are connected by groups of 44 or 45 to six Ethernet switches, which connect to a single camera server by 10-Gbps Ethernet links. The number of input ports is limited to 48 for most standard switches, and the number of 10-Gbps downstream links to the camera server may vary from 1 to 6 SFP+ links, depending on the final data bandwidth needed.

2. NectarCAM internal network

The absence of a complex buffer handling in the front-end boards in order to minimise production cost excludes any kind of re-transmit of data packets in case of loss between front-end boards and camera server. Losses due to collisions are excluded with point-to-point connections for the UDP emitters (front-end) and the receiver process in the camera server. However a congestion may occur due to the fact that typical switches concentrate 44 or more inputs into a 10-Gbps line (Figures 1, 2).

As a matter of fact the simultaneous arrival of 44 event fragments on 1-Gbps lines saturates immediately the 10-Gbps outputs, and internal switch buffers are immediately solicited. The underlying physics process of the data triggers (Cherenkov light from secondary particles of VHE gamma rays in the atmosphere) being totally independent, bursts of events with high instantaneous trigger rates may occur according the Poisson statistics. The maximum trigger rate is only limited

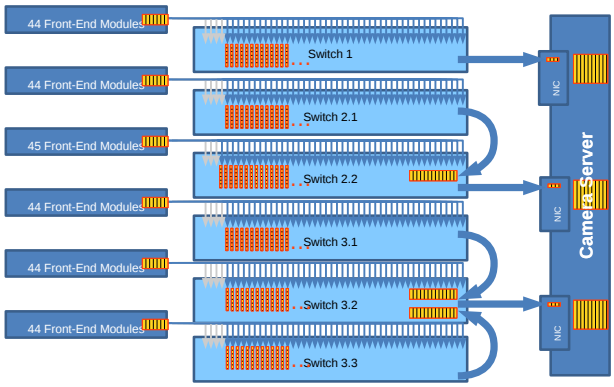

Figure 2: Possible stacking schemes for Ethernet switches within NectarCAM. The picture (unrealistically, for the sake of illustration) combines three different examples for a 1:1, 2:1 and 3:1 concentration of 44 (or 45) 1-Gbps inputs per switch into 1, 2 or 3 10-Gbps down-links towards the camera server. Existing buffers in the front-end boards, in the Ethernet switches (input queues) and in the camera server (network interface hardware buffers and system memory) are drawn in yellow. Thin arrows stand for 1-Gbps links, thick arrows for 10 Gbps. Data flow from the left (front-end modules or data generators) to the right (camera servers), crossing the switch stack in the center of the picture.

by the assumed dead time of front-end modules (2μs, where no new trigger would lead to an emission of new data). Additional high-rate bursts due to astronomical phenomena like a star in the field of vision or a shooting star or airplane are not subject to the constraint of zero packet loss, as they are not supposed to be useful for any genuine measurement with Cherenkov telescopes. But switches must recover reliably within an acceptable timescale from any overflow situation. Hence the buffer structure and dynamical behaviour of the switches used for the camera network is essential to ensure correct function, and a thorough validation is needed. In the absence of a full camera equipped with front-end modules, two devices have been built:

- a full-scale Ethernet packet generator based on Linux single-board PCs serving up to 320 1-Gbps channels on 1000baseT connections with a synchronicity of $\mathcal{O}$(100ns) between channels,
- a 64-channel UDP packet generator based on Virtex-5 FPGAs [1], allowing synchronicity of better than 1 ns between all ports, which is described in this conference contribution.

These simulators or "DAQ stimulators" replace the front-end modules, which do not exist in sufficient number for tests up to now. They also allow for independent, variable and absolutely reproducible tests with a given set of parameters or data samples. An operating mode where real NectarCAM front-end boards are combined with simulating data generators is also possible.

We assume that switches behave identically, and in particular that we can validate a given architecture by decomposing it into identical blocks of two or three switches. Furthermore the behaviour of the connection between switches and camera server is supposed to be independent from the number of connected switches, as long as the total data rate can be simulated and validated

otherwise. Therefore we limit tests with the synchronous FPGA-generator to the validation of a single switch up to an output rate of 10 Gbps.

3. UDP packet generator and reception

Most of the hardware used for this test-bench was designed at Jefferson Labs. A VME/VXS crate including a controller board, 2 SSP (Sub-System Processor) boards [6], 1 SD (Signal Distribution) and 1 TI (Trigger Interface) [5] board were used at IRFU labs in Saclay as part of the data acquisition system of the Micromegas Vertex Tracker for the CLAS12 experiment at JLAB. For the needs of the tests described here, the TI board is used to distribute a single clock source to the 2 SSP boards via the SD board. The TI board is also used to distribute an external trigger signal and from there through the SD board to each of the SSP boards over the VXS backplane. More details on these designs can be found in [2, 6, 5, 3].

Firmware has been specifically developed for the SSP boards, in order to generate UDP packets similar to the NectarCAM data format. Validity of the generated traffic has been checked by means of Ethernet traffic analysers in hardware (MTS 5800 by JTSU) and for low traffic bandwidth also in software (**WireShark** application). Synchronicity of packets emitted on different output channels of the SSP are measured to be better than 1 ns (between 90 ps for the same board and up to 580 ps for channels of different boards). The transfer time of a 1024-byte (payload) packet is measured to be compatible with the expected transfer time of 8.5μs

Triggers are provided by a programmable pulse generator (Agilent 33500B), which allows continuous, constant frequency triggers or a precise number of pulses, as well as a combination of short constant-frequency bursts, which are repeated at a secondary frequency value.

All packets transiting the switch under test successfully are received by a dual-CPU multi-core PC with (dual) SFP+ interface card (X520 by Intel). In order to optimise reception of small (1 kB) packets over 10-Gbps links, we are using the **netmap** framework [7] and a prototype event-builder application. Crosschecks of the number of received events are carried out by **WireShark** and by reading the packet counters in the switch.

4. Stress tests of Ethernet switches

The set-up was used to examine different types of switches with significantly different characteristics. We carried out simple tests to prove a linear relationship between input rate, packet size and the number of connected ports on one side and the total output rate on the other. More interesting tests concerned the behaviour under steady load, just above or below the output saturation bandwidth (**static** behaviour). Finally we put the data handling inside the switches under test by increasing the input rate to the maximum wire speed, which would obviously exceed the average output speed by several factors and can only be compensated by correctly working buffer mechanism of sufficient size (**dynamic** behaviour).

All analysed switches allow to obtain continuous lossless transmission at 10-Gbps wire speed, as long as the input rate does not exceed 10 Gbps divided by the number of connected input ports, e. g. 27 kHz for 1024-kB packets on 44 ports. Overnight runs have proven reliable function of all elements in this respect, and variation of input parameters (packet size, loaded ports, trigger

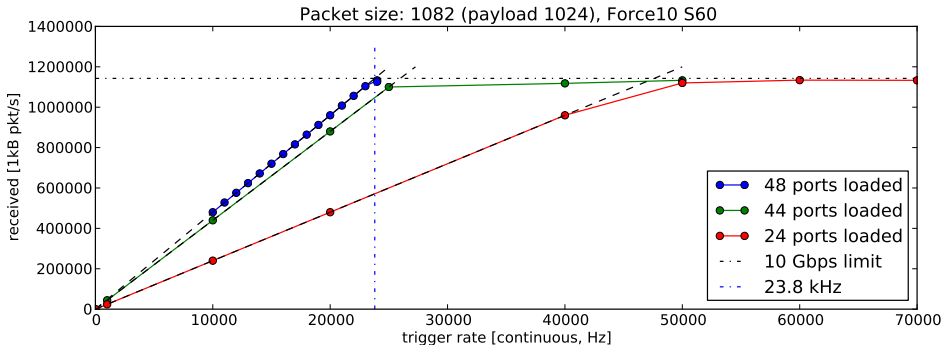

Figure 3: Saturation extrapolated from various frequencies for a varying number of input ports loaded by the generator. The theoretical bandwidth of 10 Gb/s corresponds very well to the measured values.

frequency) agree with the expected saturation behaviour (Figure 3). However a simple comparison between the "deep-buffer" switch DELL S60 and a cheaper model of the same vendor ruled out the latter by reducing the gap between two successively triggered events to the technically possible minimum (8μs, due to 1-Gbps wire speed). This configuration leads to two successively received packets on each of the input ports, and some of these event fragments of the second event are lost, as soon as more than 28 input ports (out of 48) are connected. Obviously this simple situation of "trigger burst" must be handled correctly to ensure data integrity during standard operation.

Starting from these extreme corner-points, critically quick successive events (bursts) and steady lossless operation at the output saturation level, we have explored several intermediate situations in the parameter space of "event separation" (or inverse instantaneous burst rate) and number of received events (proportional to the burst length). Figure 4 shows the exploration map of complete reception (•) and loss (×) of data as a function of event separation (abscissa) and the number of events in the test burst (ordinate), compared to the theoretically expected hyperbolic curve, which separates the area of lossless reception from the critical domain, where at least one packet has been lost in transmission.

5. Conclusion

In a joint effort across labs, institutes and collaborations we have built a Ethernet packet generator in order to test and validate deep-buffer switches for the internal network infrastructure of the NectarCAM, which is proposed to equip some of the telescopes of the Cherenkov Telescope Array. The generator emits UDP packets on 1000baseT (RJ45) lines at conditions corresponding exactly to the IP standard specifications. It has been used to fully qualify the behaviour of a DELL S60 deep-buffer switch with 10 Gb (1.25 GB) memory. Our tests have shown that this type of switches allows buffering of approximately 40×8000 1024-byte payload UDP packets at maximal input rate, before saturation and loss of single packets. This corresponds to the average value of NectarCAM event triggers within one second and therefore validates the hardware for our purpose. The corresponding memory size of 330 MB is about a factor four below the announced buffer memory and plausibly due to the fact that we did not configure any non-factory-reset features of input port queues like dispatching incoming packets on several independent queues. The dynamic behaviour

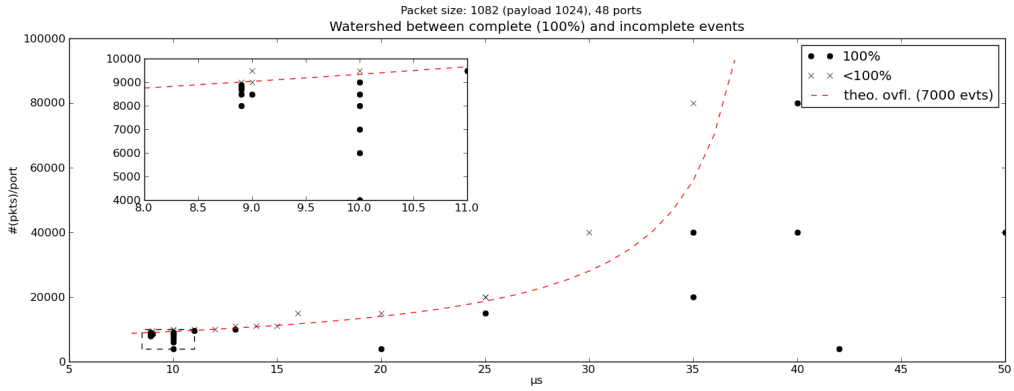

Figure 4: Empirical "watershed" (separating line) of event integrity compared to the theoretical line, at which where buffer overflow should occur. The inlay shows a zoom on the small-gap/low-number region surrounded by the dashed line. Measurement points with packet loss are marked with crosses (×), measurement points without loss with bullets (•). Agreement between the extrapolated (fitted) value of 7000 buffered packets (per input queue) and all measurements is very good.

of saturation and overflow over a large range of input rates is well reproduced by a simple queuing model of available buffers in the switch.

Similar switches are not available from many other vendors, and DELL announces to stop commercialisation of the S60 product next year. When it comes to alternatives, these usually contain 48 front inputs for 10-Gbps links, which increases the price significantly. For our project we may have to continue market research and technology watch, but our test tools are well prepared to qualify replacement candidates within optimal time.

6. Acknowledgements

This work has been carried out thanks to the support of the OCEVU Labex (ANR-11-LABX-0060) and the A*MIDEX project (ANR-11-IDEX-0001-02) funded by the "Investissements d'Avenir" French government program managed by the ANR. We gratefully acknowledge support from the agencies and organisations listed in this page: https://portal.cta-observatory.org/Pages/Funding-Agencies.aspx. We appreciated the help of C. Cuevas, E. Jastrzembski and B. Moffit from the Fast Electronics Group at Jefferson Lab, putting at our disposal the CLAS12 hardware based upon work supported by the U. S. Department of Energy on our publications under U.S. DOE contract No. AC05-06OR23177.

References

[1] http://www.xilinx.com/support/documentation/virtex-5_user_guides.htm.

[2] D. Abbott, C. Cuevas, D. Doughty, E. Jastrzembski, F. Barbosa, B. Raydo, H. Dong, J. Wilson, A. Gupta, M. Taylor, and S. Somov. A 250 mhz level 1 trigger and distribution system for the gluex experiment. In *Real Time Conference, 2009. RT '09. 16th IEEE-NPSS*, pages 548–551, 2009.

[3] D. Attie et al. The readout system for the clas12 micromegas vertex tracker. In *IEEE-NPSS 19th Real Time Conference*, 2014.

[4] J. F. Glicenstein et al. NectarCAM: a camera for the medium size telescopes of the Cherenkov Telescope Array. In *ICRC 2015 Proceedings*, 2015.

[5] W. Gu and B. Moffit. Description and technical information for the VME trigger interface (TI) module — Jefferson Lab — nuclear physics division - data acquisition group. pages 548–551, 2009.

[6] B. Raydo. Sub-system processor manual - jefferson lab - nuclear physics division - fast electronics group. 2014.

[7] L. Rizzo. netmap: a novel framework for fast packet I/O. *Proceedings of the 2012 USENIX Annual Technical Conference*, June 2012.