Prototyping a Multi-10-Gigabit Ethernet Data Acquisition for a Cherenkov Telescope Array

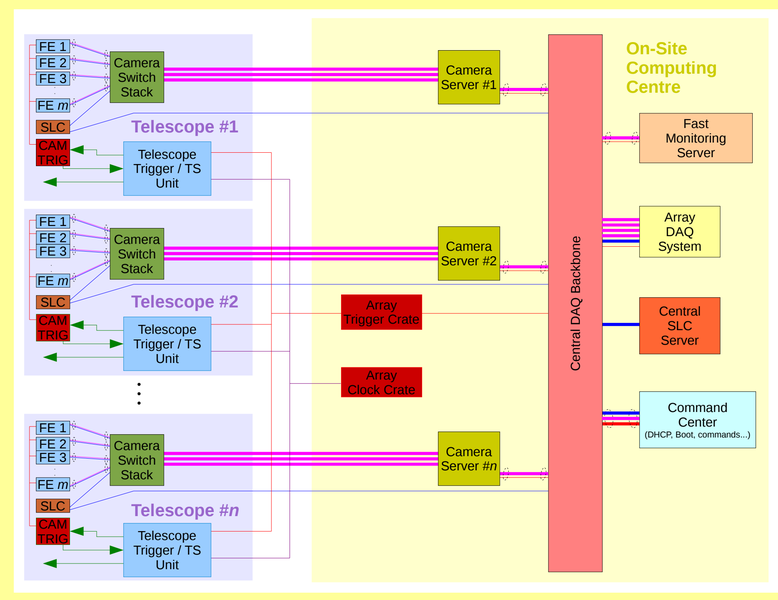
Dirk Hoffmann, Julien Houles, Centre de Physique des Particules de Marseille, Aix-Marseille Université and CNRS/IN2P3, Marseilles, France
for the NectarCAM collaboration within the CTA Consortium

Cherenkov Telescope Array

The Cherenkov Telescope Array CTA is the next generation High Energy Gamma Ray instrument. One observatory is planned in each of the Northern and Southern hemisphere. Compared to previous generation Cherenkov Telescopes, the size of the array and the telescopes will bring an increase in sensitivity and energy range. Each site will consist of about one hundred telescopes of three or four different types, ranging from 4 meters to 24 meters in diameter and using different technologies (single/dual mirror), but with a pixeled, fast photodetection camera (photo-multipliers). Hence the performance and in particular the reliability and performance of each of the telescopes will become
An important ingredient for the success of the project. We made several assumptions for our first studies, which are plausible and focused on support of the LST/MST camera developments in CTA:
The imaging camera of each telescope is composed of 1855 pixels grouped in 265 clusters of seven photo-multipliers (PM) and delivers an average data rate of 20 Gbps for a readout window of 60ns at 10kHz to the downstream DAQ system.
One aim of the project presented here is to deliver a universal DAQ, configurable and usable for similar types of detectors.

Data Acquisition Infrastructure

1866 front-end modules deliver data through 1Gbps Ethernet links, which are connected to a common Camera Server by means of six (off-the-shelf) Ethernet switches. Each of the switches contains 48 1Gbps ports and one or several 10Gbps outputs.
Due to cost considerations, the front-end modules contain only minimal first-level buffering (8 events) and no possibility to re-transmit packets in case of network congestion. Events which cannot reach the Central DAQ system are considered lost (in “dead time”).
The picture sketches the network connections from the front-end modules (on the left) through the Camera network, Camera ethernet switches and the Camera Server to the Central DAQ system. This poster describes Data Acquisition from front-end downstream to the Camera Server. Each Camera Server contains an instance of the Camera Event-Builder application.
Trigger and Slow Control (SLC) system partly share the infrastructure with physics data.

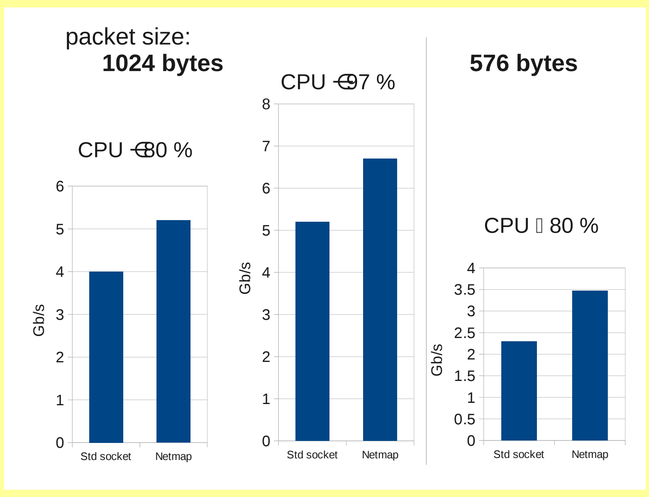


From boundary conditions to design

Event Builder

- One (logical) Event-Builder per camera, several Event-Builders can be hosted in one (physical) Camera Server.
- Software architecture and performance have been published in CHEP2012 [1].
- Even if several logical Event-Builders are hosted in the same hardware, extensions are possible, depending on analysis needs, for local computing, data treatment as well as an interface to delegate parallelisable operations to GPUs.

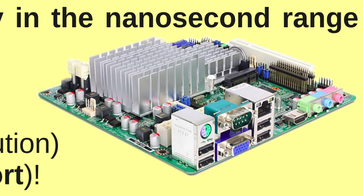
Tests of the Event-Builder application have shown that hardware and software (with two different program designs) are capable of receiving and reconstructing camera event fragments from 300 Ethernet nodes, which are assembled to full camera events on the fly, not exhausting more than 300% CPU equivalents (cores) out of the 12 installed.



- Theoretical bandwidth of 20Gbps (two SFP+ links) cannot be reached
- Standard Linux Ethernet drivers generate overhead with memory allocation in user and kernel space
- Also *number* of packets per time unit is limited.
- Small packets and even standard MTU values of 1500 bytes barely exploit **less than 40% of the physical layer bandwidth**.
- Jumbo Frames (9 kB) as well as specially developed network drivers like **netmap** [2] can help circumvent this limitation, if the front-end data can be bundled appropriately.

Camera Simulator

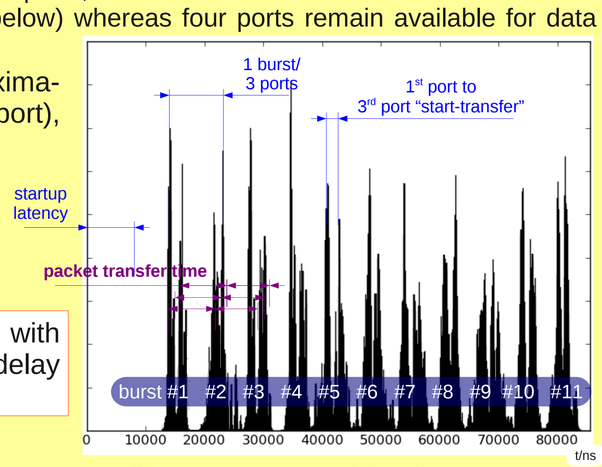
- Validation of concept and performance of the Event-Builder and infrastructure
- Delivery of **2000 data packets** of 1kB with **synchronicity in the nanosecond range** implies:
  - Synchronicity among dozens of boards
  - Synchronicity among ports on each board
- Random event trigger (independent events, Poisson distribution)
- Minimal cost, as it is a validator (approximately **50 EUR / port**)!



Test and Validation

One **Single-Board Computer (SBC) replaces up to five front-end boards**. After evaluation of several alternatives the final choice has been made for a JetWay NF9I SBC with Atom D2550 dual-core processor. It contains five Ethernet ports, out of which one can be reserved for the synchronisation of all boards via PTP (see below) whereas four ports remain available for data generation.

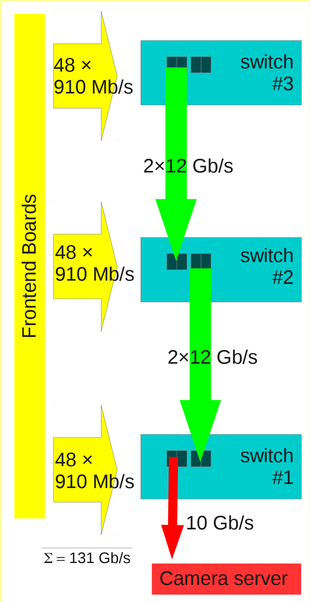
- Synchronicity condition fulfilled approximatively (at the order of 1µs per used port), due to use of commun busses
- Latency 14µs for first outgoing packet
- Time between bursts 3.5µs
- Time between packets in one burst: (1.2±0.5)µs on average, 2.3µs max.
- Power consumption 15W
- Need elaborate timing algorithm with predefined set of pseudo-random delay values to synchronise *O*(60) SBC.



The figure shows the timing pattern of the “*start transfer*” flags (of physical IP traffic), measured for 11 bursts of 1kB-packets (broadcast time 8µs over 1Gbps) sent out over three ports each.

Technology Highlights

Queuing theory and buffer handling



- Little availability of buffers in the front-end (cost)
- De-randomising incoming event blocks in network
- Instantaneous input rate (1Gbps per front-end) approximately 13 times max. output rate (10Gbps)
- Relaxation time 1.15ms for one event
- Buffer overflow will lead to (small, acceptable) data loss.

First system analyses confirm that the critical working point can be reached within the required event rates (5 to 9 kHz) for an average IP packet size of 1kB.
Detailed system information about buffer allocation and dynamic behaviour of IP switches is difficult to impossible to obtain from vendors. All assumptions must be confirmed by measurements.

Schema of a half-system, assuming that Data Acquisition will be composed of two identical 10Gbps blocks serving half of the front-end modules.

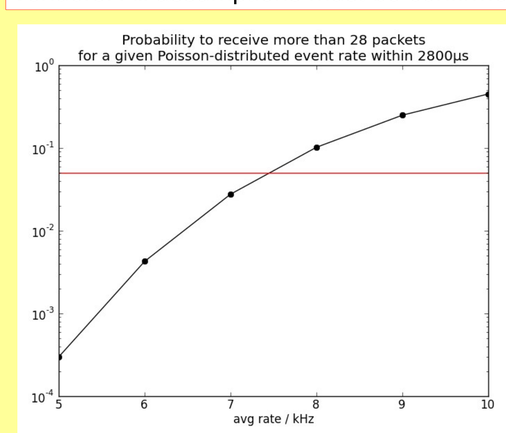
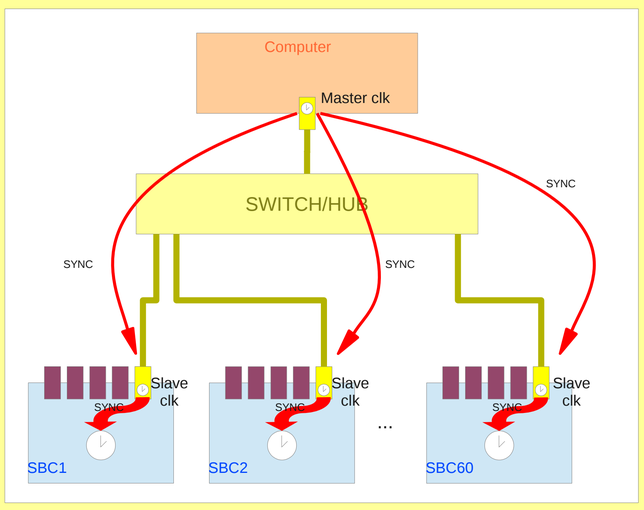


Diagram of example case:
A transitory event burst at 10kHz (100µs per event) will saturate the lowest switch after 28 events (2.8ms). The Poisson probability to reach this case becomes significant for a nominal random rate of 7.5kHz (crossing of the 5% dead time limit).

Precision Time Protocol (PTP)

- External triggers (like parallel port or GPIO) imply too large jitter and interrupt latency compared to the O(1ns) synchronicity required for the validation.
- SBC equipped with Intel 82574L Ethernet controllers complying to Precision Time Protocol with hardware time-stamping
- One port of each board dedicated to time synchronization via PTP
- Same timing scenario containing the packet output times generated ineach SBC
- The master clock synchronises each PTP hardware clock of dedicated Gb port on each board:
  - Dynamic measurement of the packet transmission delay only between Master and SBC1 (assumed identical for the other SBCs)
  - When sufficiently stable, propagation of the measurement to each SBC
  - Inhibition of delay requests on each slave, start receiving SYNC packets from the master
- SBCs synchronise their Real Time clocks on PTP hardware clocks locally
- Poll on the Real Time clock and packet sent as specified in the timing scenario



References:
[1] D. Hoffmann, J. Houles, Journal of Physics: Conference Series 396 (2012) 012024, doi:10.1088/1742-6596/396/1/012024
[2] **`http://info.iet.unipi.it/~luigi/netmap/`**

Dirk.Hoffmann@In2p3.Fr, October 2013