REadout a DIeci Giga bit al secondo – REDI-GO

Impiego di reti ethernet a 10Gb/s per il readout negli esperimenti di fisica nucleare e delle alte energie.

Proposta di esperimento di Gr. V

M. Bellato, D. Bortolato, R. Isocrate, F. Montecassiano

INFN – Sezione di Padova

M. Gulmini,G. Maron, N. Toniolo, A. Triossi

INFN – Lab. Nazionali di Legnaro

D. Galli, I. D'antone, I. Lax, U. Marconi, V. Vagnoni

INFN – Sezione di Bologna

Fast data transmission technologies (in particular optical) are being deployed on an unprecedented scale in the LHC era experiments. Most of the resultant optical links are used to connect the detector front-ends for both data readout and detector control, while a smaller number are used for data transmission between cards, crates and racks in the counting rooms.

Fore coming experiments require a new generation of data acquisition systems that are cost-effective and maximize the use of commercially available components. Using standard networking protocols and hardware will ensure compatibility between different components of the detectors, whilst allowing for seamless incremental upgrades to individual systems.

Some new experiments (e.g. ILC) will take data without the use of a hardware trigger. Consequently, to minimize dead time, all experimental data must be read from the detector in the 200 ms gaps between beam bunch trains. This semi-continuous data stream must be routed to DAQ computers, currently assumed to be PCs, processed and then sent to offline storage.

Figure 1 shows a typical readout and DAQ system envisaged for modern detectors. The front-end electronics are connected to concentrators over custom, high-speed links. These concentrators then feed the data over a network to compute nodes where the data are processed and stored. These concentrators may do some re-formatting or zero suppression. Whilst the use of some application-specific devices is inevitable, it is highly desirable to use commercial networking devices and protocols as much as possible. One possible option would be the use of 1 and 10 Gigabit Ethernet.

In particular, since much of the front-end electronics is actually and likely will be controlled by Field Programmable Gate Arrays (FPGA), it is vital to assess the suitability of FPGAs for directly driving network traffic and to optimize the performance of the hardware and protocols used to send the data over the network.

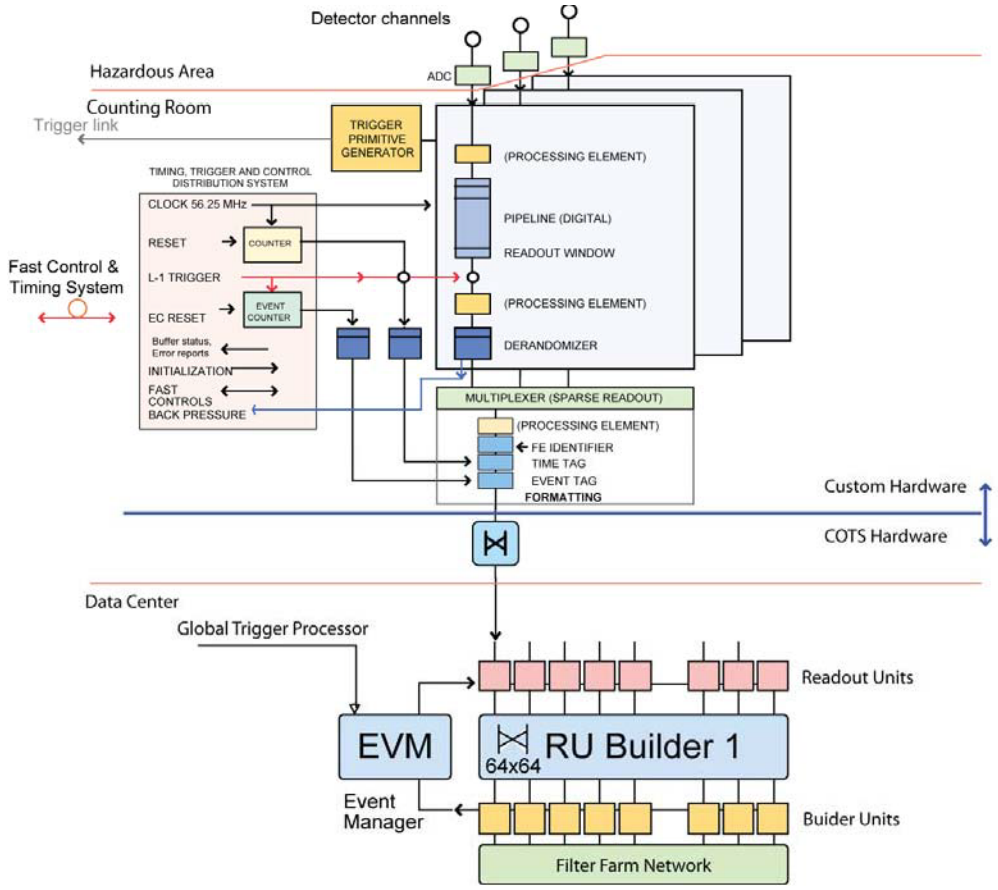

Figure 1 – Block diagram of a typical modern frontend readout and data acquisition system for physics experiments.

As indicated in Figure 1, without some form of traffic shaping between the concentrators and the destination PCs, there would be the classic bottleneck problem on the egress of the Ethernet switch, with data queuing for transmission to the processing node and the possibility of packet loss [1].
The growing convergence of storage protocols (iScsi, FibreChannel over Ethernet,iWarp, RDMA), around 10 Gigabit Ethernet standard makes it attractive for deployment in new data acquisition systems for a number of reasons.

IEEE has been developing a few standards they collectively refer to as “Data Center Bridging” (DCB) and that are also sometimes referred to as “Converged Enhanced Ethernet” (CEE). This refers to high speed Ethernet (currently 10 Gb/sec, with a clear path to 40 Gb/sec and 100 Gb/sec), plus new features. The main new features are:

- ✓ Priority-Based Flow Control (802.1Qbb), sometimes called “per-priority pause” [2]
- ✓ Enhanced Transmission Selection (802.1Qaz) [3]
- ✓ Congestion Notification (802.1Qau) [4]

The first two features let an Ethernet link be split into multiple “virtual links” that operate pretty

independently — bandwidth can be reserved for a given virtual link so that it can't be starved, and by having per-virtual-link flow control, one can make sure certain traffic classes don't overrun their buffers and avoid dropping packets. Then congestion notification means that one can tell senders to slow down to avoid congestion spreading caused by that flow control, thus achieving (at least part of) the traffic shaping that is required in event builders for physics experiments.
But event building would benefit even more from RDMA [5]. RDMA stands for Remote Direct Memory Access, but the term "RDMA" is usually used to refer to networking technologies that have a software interface with three features:

1. **Remote direct memory access (Remote DMA)**
   Remote DMA is what the name implies: Direct Memory Access on a remote system. An adapter on system 1 can send a message to an adapter on system 2 that causes the adapter on system 2 to DMA data to or from system 2's memory. The messages come in two main types:

   RDMA Write: includes an address and data to put at that address, and causes the adapter that receives it to put the supplied data at the specified address

   RDMA Read: includes an address and a length, and causes the adapter that receives it to generate a reply that sends back the data at the address requested.

   These messages are "one-sided" in the sense that they will be processed by the adapter that receives them without involving the CPU on the system that receives the messages. This is particularly important for 10 Gb/s applications where the involvement of the CPU would be massive. RDMA gives fine-grained control over what remote systems are allowed to do: protection domains and memory keys allow the control of connection by connection and byte-by-byte with separate read and write permissions.

2. **Asynchronous work queues**
   Software talks to RDMA adapters via an asynchronous interface (which is called a "verbs" interface for historical reasons).When using an RDMA adapter, one creates objects called queue pairs (or QPs): a send queue and a receive queue, and completion queues (or CQs). Operations are posted to one of the work queues, the RDMA engine executes it asynchronously, and when it's done, the adapter adds work completion information onto the end of the CQ. Operating asynchronously is important because it allows the overlap of computation and communication, a mandatory feature in event building.

3. **Kernel bypass**
   Kernel bypass allows user space processes to do fast-path operations (posting work requests and retrieving work completions) directly with the hardware without involving the kernel. Latency-sensitive applications (like an event builder) greatly benefit from this, thus saving the system call overhead impacts on the memory staging on the front-end side, possibly in a radiation environment, where buffering is difficult and expensive.

The RDMA/TCP specification enables 10 Gigabit Ethernet RDMA implementation at the physical layer and TCP/IP as the transport, combining the performance and latency advantages of RDMA with a low-cost, standards-based solution. The specification is intended to be implemented in hardware, on

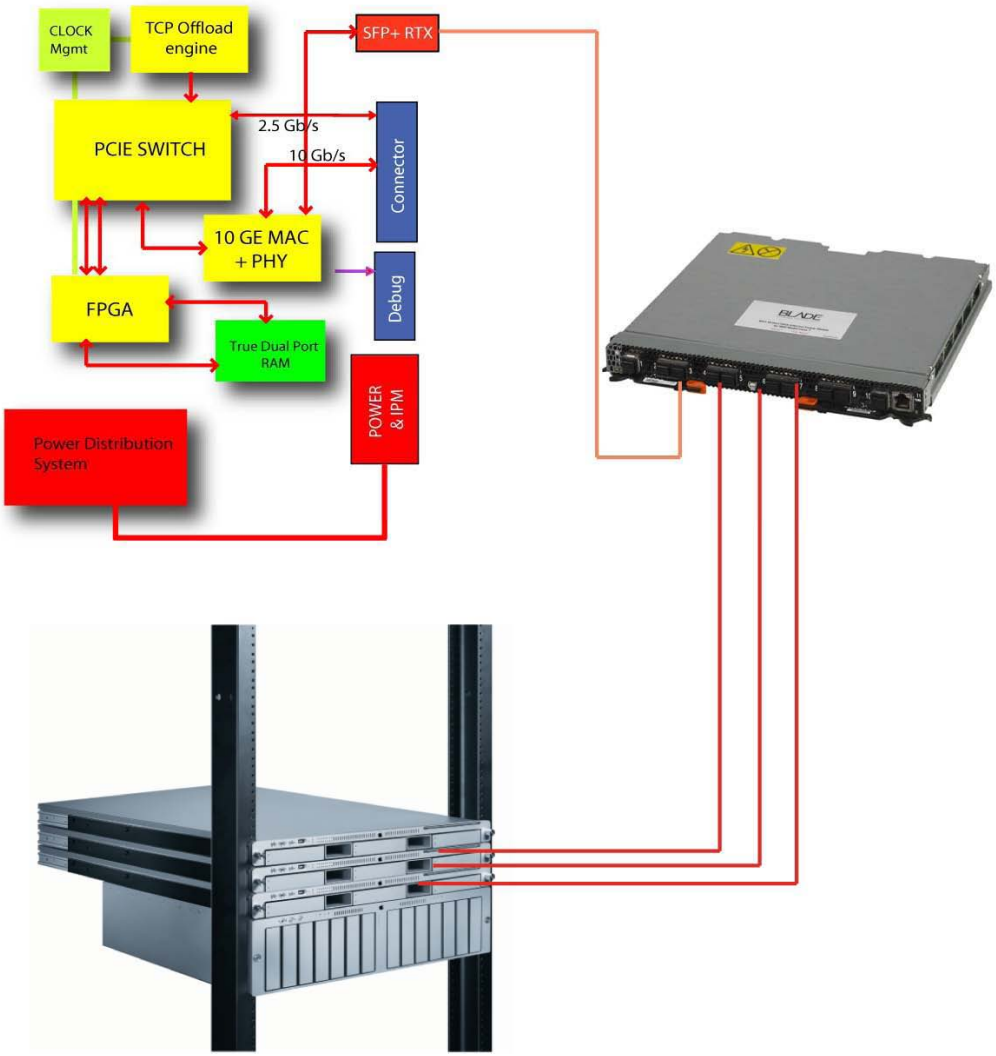

Figure 2 – Block diagram of the hardware testbed for the proposed experiment

top of a so called TCP Offload Engine (TOE), a technology used in network interface cards to offload processing of the entire TCP/IP stack to the network controller.
Figure 2 shows the hardware setup of proposed experiment which would investigate the potential benefits of 10Gigabit Ethernet with RDMA and TCP offload technologies for the readout and event building of physics experiments.

An FPGA driven 10 Gigabit Ethernet Media Access Controller(MAC) injects data via a 10 Gb/s Ethernet switch into a PC server farm for event building purposes. The FPGA emulates the synchronous interface of a generic experiment readout system.
The proposed activity should envisage the design and construction of a board or adapter with a suitable FPGA and a 10Gb Ethernet MAC and PHY chips to be used as a front-end readout emulator.
The emulator lends itself to test different approaches to data injection; besides the complex programming model of RDMA with a full-featured TCP transport, It makes sense to investigate a custom

developed TCP stack to be embedded in a field programmable gate array. In the context of a front-end readout it should be noticed that a stream oriented reliable protocol is needed for the following reasons:

1. the N:1 configuration (N sources to 1 destination in turn) that is at the basis of the event building process lends itself to traffic congestion at the destination (the event builder farm machine) so a flow control mechanism is mandatory.
2. Reliable delivery of data to the destination is required; hence the underlying protocol should provide retransmission in the case of data loss.
3. The mechanics of changing the event builder destination is usually pursued by deploying a switch between the front-end readout cards and the event builder farm, so congestion control is an issue at the switch level also.

The effectiveness of this custom built TCP stack in an FPGA is to be tested on its own and against a traffic shaping mechanism. The testbed will allow investigations of latency, throughput, buffering schemes and global event building bandwidth, but will allow also the evaluation of novel schemes of traffic shaping based on the new IEEE extensions concerning virtual links and per link congestion signaling driven by hardware.

The proposed R&D activity is geared towards the adoption of the 10Gethernet technology in fore coming experiments. A good candidate is the SuperB project [6], in which the proponents have involvements in the specification of the readout electronics and data acquisition system. The Warp [7] collaboration is willing to test the demonstrator in its present experimental setup at Gran Sasso laboratories of INFN. Finally, an expression of interest to collaborate in this development has been done by the CERN CMS [8] DAQ group in view of the fore coming SLHC upgrade.

The R&D activity should span a period of three years, with the contribution of different INFN sections and laboratories. A mandatory qualification of commercial devices arranged in a 2x2 switched configuration should take place during the 1st year; the subject of investigation should include the use of TOE's in a prototype 2x2 event builder. In parallel the design and development of a custom TCP stack in a Field Programmable Gate Array should take place using a suitable evaluation board as a test-bed for firmware assessment. The 2nd year of activity includes the design and prototyping of an adapter that should serve as a front-end readout emulator with the support for a commercial and custom TOE and its integration in the prototype 2x2 event builder. The 3rd year is devoted to qualification of a 4x4 event builder, possibly with the inclusion in the data acquisition chain of a real experiment setup.

The following table summarizes the activities.

Sezione/Anno	**2010**	**2011**	**2012**
Padova	Sviluppo TOE	Sviluppo TOE + test integrazione	Qualifica EVB 4x4
LNL	Qualifica link + adapter 2x2	Sviluppo EVB	Qualifica EVB 4x4
Bologna	Costruzione adapter	Sviluppo + qualifica EVB 2x2	Qualifica EVB 4x4

References

[1] Richard Hughes-Jones, Peter Clarke, Steven Dallison, *Performance of Gigabit and 10 Gigabit Ethernet NICs with Server Quality Motherboards*, Grid Edition of Future Generation Computer Systems Volume 21, Issue 4, April 2005 p 469-488
[2] http://www.ieee802.org/1/pages/802.1bb.html
[3] http://www.ieee802.org/1/pages/802.1az.html
[4] http://www.ieee802.org/1/pages/802.1au.html
[5] http://www.rdmaconsortium.org
[6] http://www.pi.infn.it/SuperB/
[7] http://warp.lngs.infn.it/
[8] http://cms.web.cern.ch/cms/index.html