# A FPGA-based compute node for the PANDA data acquisition and trigger system.

**Contribution to the proceedings of the XLV International Winter Meeting on Nuclear Physics. Bormio January 14th–21st, 2007 2nd April 2007.**

**T. Pérez, C. Gilardi, M. Liu and S. Yang**
for the PANDA, BESIII and HADES collaborations.
II Physikalisches Institut, Universität Gießen,
35392 Gießen, Germany

E-mail: tiago.perez@physik.uni-giessen.de

**Abstract.** The new proton antiproton annihilation experiment at Darmstadt, PANDA, foresees to run with a frequency of up to $10^7$ interactions per second and will deliver an expected data rate of several tens of gigabytes per second. In this scenario, real time processing, for intelligent data compression and event selection, is mandatory. A general purpose compute node based on field programmable gate arrays (FPGA) is being designed at the University of Gießen. Each compute node will be equipped with five Virtex-4 FX FPGAs, multiple gigabit Ethernet ports, four optical links and 13 RocketIO$^{TM}$ ports connected to the ATCA full mesh backplane. Each node will be realized as a full size ATCA board. The full mesh backplane of the ATCA shelf and the multiple gigabit ports provide the bandwidth necessary to transport the large amount of data delivered by the detectors.

## 1. Introduction

The proton antiproton annihilation at Dardmstadt exeriment, from now on refered as $\bar{\text{P}}$ANDA, is a detector in development that will be installed in the High Energy Storage Ring (HESR) of the future accelerator facility FAIR at Darmstadt[1, 2, 3]. The experiment is being design to exploit the physics of the high intensity phase space cooled anti-proton beams delivered by

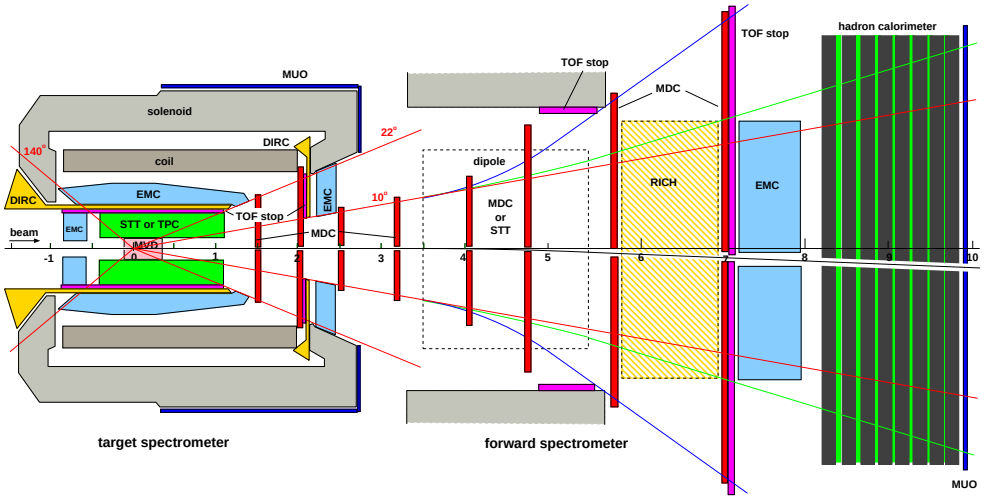


**Figure 1.** Cross section (top view) of the $\bar{\text{P}}$ANDA detector. The total lenght of the detector will be about 12m [3].

the HESR. The HESR will provide high intensity anti-proton beams with momentum resolution of up to $\delta$p/p $= 10^{-4}$ for the stochastic cooling mode and up to $\delta$p/p $= 10^{-5}$ for the electron cooling mode. For both cases the momentum of the anti-proton ranges from 1 to 15 GeV/$c$ . $\bar{\text{P}}$ANDA will be able to perform precision measurements of the anti-proton annihilation in the charm mass region.

$\bar{\text{P}}$ANDA is a fixed target experiment and therefore all particles produce are strongly forward boosted. The experiment, as it can be seen in the figure 1, is strongly asymmetric and has two main detector ensembles named the central target spectrometer and the forward spectrometer.

## 2. The Data Acquisition System

The HESR is designed to work in two modes, a high resolution mode with the beam phase space cooled by electrons ($\delta$p/p$\sim$ $10^{-5}$ and $L \sim 2{\cdot}10^{31} cm^{-2} s^{-1}$), and a high luminosity mode ($\delta$p/p$\sim$ $10^{-4}$ and $L \sim 2{\cdot}10^{32} cm^{-2} s^{-1}$) that will use only stochastic cooling. Simulations of common channels of p$\bar{\text{p}}$ collisions at 15 GeV/$c$ and 10 MHz interaction rate deliver a raw data production of around 40 GByte/s (table 1 [3]). But, in case of interaction rates of 20 MHz and assuming that the event size could reach sizes of up to 8 kB, the DAQ system should be prepared to handle up to 200 GByte/s.

An important consequence of the high interaction rate is that the time between two interactions is smaller as the the time needed by some of the

| 15 GeV/$c$ | Multiplicity | #DataBits/hit | #HeaderBits/hit | B/event | MB/s |
|---|---|---|---|---|---|
| FSMDC | 45 | 20 | 16 | 203 | 2025 |
| FSEMC | 102 | 24 | 14 | 485 | 4845 |
| FSHC | 9 | 10 | 14 | 27 | 270 |
| MVD | 16 | 16 | 20 | 72 | 720 |
| EMC | 90 | 24 | 14 | 428 | 4275 |
| STT | 69 | 16 | 20 | 311 | 3105 |
| | | | | | 15240 |

| 15 GeV/$c$ | Multiplicity | #DataBits/hit | #HeaderBits/hit | B/event | MB/s |
|---|---|---|---|---|---|
| FSMDC | 45 | 20 | 16 | 203 | 2025 |
| FSEMC | 102 | 24 | 14 | 485 | 4845 |
| FSHC | 9 | 10 | 14 | 27 | 270 |
| MVD | 16 | 16 | 20 | 72 | 720 |
| EMC | 90 | 24 | 14 | 428 | 4275 |
| TPC | 850 | 12 | 16 | 2975 | 29750 |
| | | | | | 41885 |

**Table 1.** Expected $\bar{\text{P}}$ANDA data rates from GEANT4 simulations. Events representing the dominant annihilation channels at 15 GeV/$c$ are considered for the two different tracking options (top: straw tube tracker, bottom: TPC). The primary interaction rate is $10^7/s$ [3].

slower detectors to discharge. This will be a problem especially for gas detectors like drift chambers or the time projection chamber. Taking some standard figures, drift time of around 1$\mu$s and 107 interactions per second, we can yield an estimation of 10 to 100 events piled up in the gas detectors at any single moment. To recover from this pile up, track from the gas detectors have to be matched with the information coming from faster detectors.

Due to the wide physics program of the $\bar{\text{P}}$ANDA proposal[3], the application of a *hardware level 0* trigger not viable. Usual trigger parameters like multiplicity are here not suitable to trigger on.

It has been proposed for the Panda data acquisition (DAQ) system to use continuously sampling read-out electronics which will take data independently, and without any external *start of read-out* signal This would be achieved by using detector flash analog-to-digital converters (ADC) in the front end electronic modules.

It has been said earlier that a very large amount of data will be produced by

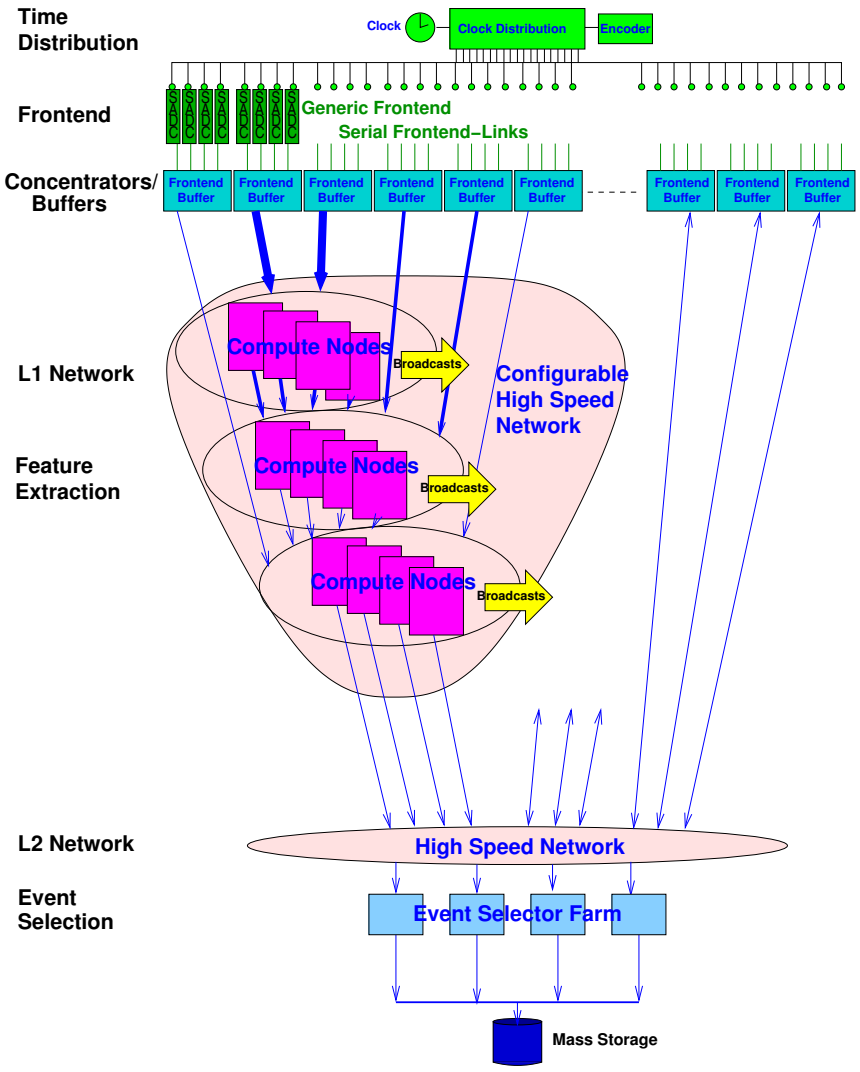


**Figure 2.** Schematic overview of the $\bar{\text{P}}$ANDA DAQ. Data flow from top (detector data concentrators) to bottom (mass storage) [3].

$\bar{\text{P}}$ANDA. Those figures shown in the table 1 are too large to be sent directly to permanent storage. Two main methods are foreseen to reduce the amount of data: Feature extraction at the front end level and event selection on dedicated compute nodes.

*2.1. Feature Extraction*

Read-out modules will perform a pre-processing of the data to locally extract interesting parameters. This extracted features can be used to reduce the *raw*

data (for example: the samples of a detector signal with a Gaussian shape can be substitute by the width, height and position of the maximum) and to help in the processing of the event selection later on.

To keep the synchronization, a high precision clock will be distributed to every detector read-out module. This will be done in a similar way as it is done by the COMPASS experiment [4, 5]. Based on the COMPASS experience we expect a clock jitter at detector read-out level in the order of only 25 ps [6]. Every signal will be tagged with a *timestamp* derived from the distributed clock.

*2.2. Intelligent Real Time Event Recognition*

It has been already mentioned that due to the wide physics program of $\bar{\text{P}}$ANDA, a *hardwired* trigger is not possible. For the $\bar{\text{P}}$ANDA data acquisition system, several successive stages of intelligent triggers are foreseen. These triggers will be calculated by high intensity computing networks that will be reconfigured/reprogrammed for every run depending upon the different physic topics addressed.

In the current plan, FPGA-based compute nodes will be used for the first stages of the selection networks (levels 1 and, maybe, 2). For this stages, low latency and deterministic time are strong requirements that can be more easily fulfilled with programmable logic devices than with normal CPU's. Typical task that need to be performed here include image processing for pattern recognition. As an example we can cite the online ring recognition of the HADES 2nd level trigger [7, 8]. The image processing unit of the HADES ring image Cherenkov detector was realized a 6U VME equipped with 12 Xilinx XC4000 FPGAs. This unit can analyze, in parallel, all 96 columns of one sector of the detector's pad plane searching for possible ring centres, and produce a trigger decision in less than $10\mu$s .

For the higher stages (level 3 and, maybe, level 2) of the trigger a network based on standard computers is foreseen. Here, the ease to program complex algorithms is the deciding feature as latency can be overwhelmed by large memory buffers.

Figure 2 shows a schematic overview with the main features of the panda DAQ system. The pipeline structure of the data acquisition system is apparent in the figure. The data flows from top to bottom: front-end modules, concentrators, compute node networks and mass storage. In the figure only two trigger networks are shown, though this issue is still under discussion at the present time.

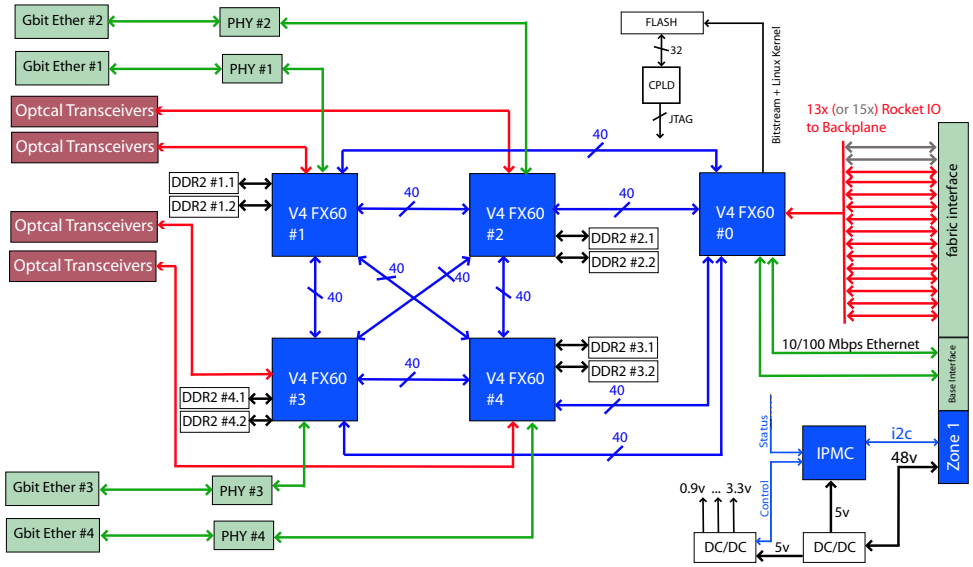


**Figure 3.** Main logic blocks of the compute node. The Ethernet ports are on the left depicted with black letters and the RocketIO$^{TM}$ links are shown on the right connected to the ATCA backplane. On the bottom right corner the IPMC is schematically depicted.

## 3. The Compute Node

To fulfil the requirements on bandwidth and compute power of the $\bar{\text{P}}$ANDA data acquisition system, a compute node base on FPGAs is being design. The compute node will be equipped with five field programmable gate array (FPGA) from Xilinx [9] which can provide large compute resources, enough to execute the trigger algorithms. In addition each compute node will have four 1 Gbit Ethernet ports, four optical fiber ports plus 13 serial RocketIO$^{TM}$ [9] connections to the backplane (fig. 3). Altogether, the compute node will provide an aggregate bandwidth of around 26 gigabits per second. From the point of view of the data transport possibilities, the $\bar{\text{P}}$ANDA data acquisition system could be realized already with only 60 compute nodes.

A logic block diagram of the main components of the compute node is shown in the figure 3. The main components are clearly visible in the figure. The boxes in the centre with the white V4 FX60 depict the five FPGAs, the Ethernet ports are drawn on the right and the serial links are the red lines to the backplane. In addition the IPMI Controller system and the dual-star

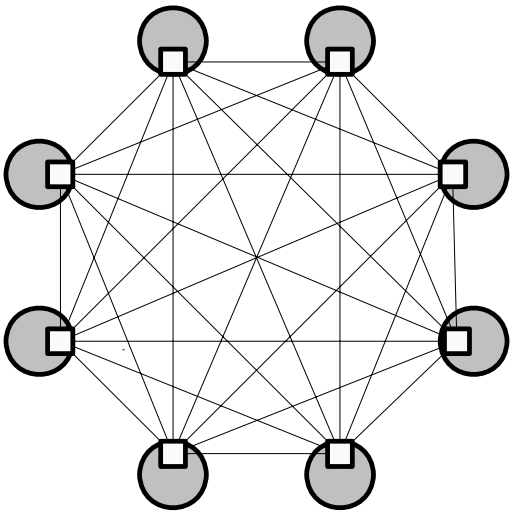

**Figure 4.** Diagram of a full mesh interconnection between 8 nodes.

base interface are also shown on the bottom right corner of the figure.

*3.1. The Shelf System*

The nodes will be built to comply with the specifications of the Advanced Telecommunication Computer Architecture (ATCA)[10]. This is a new shelf system (the first draft of the specification was issued by the PICMG group late in 2001 [10]) which provides important advantages over other widely spread crate systems used until now in nuclear and particle physics experiments like NIM or VME .The main assets of ATCA are the big power per slot available, up to 200W, and the high speed backplanes. The latter use differential copper lines that can provide a per channel bandwidth of up to 10 gigabits per second. The ATCA specifications define three topologies for backplanes: full-mesh, star and dual-star. All of them are suitable for Ethernet, InfiniBand and Serial Rapid IO protocols among others.

The Panda compute node will make use of a full mesh backplane. As it is shown in the figure 4, the full mesh topology provides point to point interconnections between any two modules in one shelf. With this kind of backplane, several configurations of the Panda DAQ are possible. These range from a pure *vertical* topology with data flowing in a pipelined manner from the front end electronics to mass storage, to a more complex system where data can flow not only *vertically* but also *horizontally* in every layer. The interconnection topology for $\bar{\text{P}}$ANDA is not yet settled, though most likely the final interconnection scheme will be an intermediate

solution, i.e. basically pipelined but with some restricted horizontal paths. Communication in-layer are necessary to solve certain problems like event event building. These horizontal communication provide more flexibility in the data path but introduce an extra grade of complexity, what implies more resources need to be used for network control.
The full mesh backplane gives us the possibility to adapt the data path by reconfiguring only the FPGAs with no need to build new hardware.

*3.2. The Selection of the FPGA*
The FPGA chosen for the compute nodes is the Xilinx Virtex-4 FX60 [9]. This selection was based on the number of features these devices offer on top of the traditional FPGA fabric. These additional features include:

(i) 2 embedded PowerPC *hardcores*.
These is a standard processor architecture supported by the Linux kernel [11] and therefore an almost standard GNU/Linux operating system can run on it. Although not powerful enough to perform any serious calculation, the embedded PowerPC's provide an easy way to have networks access to the FPGA and perform slow control routines, like remote reconfiguration of the devices or status monitoring.

(ii) 128 digital signal processor (DSP) slices.
Integer complex operations like multiplication-accumulation can be perform in Virtex-4 devices as fast as any modern digital signal processor.

(iii) 4 gigabit Ethernet media access controllers.
On-chip tri-mode Ethernet allows to produce more compact and simple designs. With embedded media access controllers, the only pieces missing for a successful Ethernet connection are magnetics and physical layer chips.

(iv) 16 *RocketIO* $^{TM}$ serial transceivers.
The RocketIO is a proprietary technology of Xilinx Inc. It provide serial communication differential channels at speeds from 22 Mbps to 6.5 Gbps in full duplex operation.

Thanks to all the embedded extra features the design of the compute node boards is not so complex as it may look like. Almost all what we need is already in the FPGA, only very few extra modules (RAM, physical layer interfaces and optical transceivers) have to be added on top.

*3.3. Mode of Operation*

For the $\bar{\text{P}}$ANDA system we will use 19 inches ATCA shelves with 14 slots. On top of the full mesh fabric interface which we will use for high density data transport. These shelves also have an additional dual-start *base interface* [10] with the centers of the dual-star located in the logical slots 1 and 2 (fig. 5). In our current design, the full mesh *fabric interface* will carry the detector and trigger data. Each module will have a serial RocketIO$^{\text{TM}}$connection (1.25 Gps) with any other module in the same shelf. For inter-shelf communications there are two possibilities foreseen. These will be either gigabit

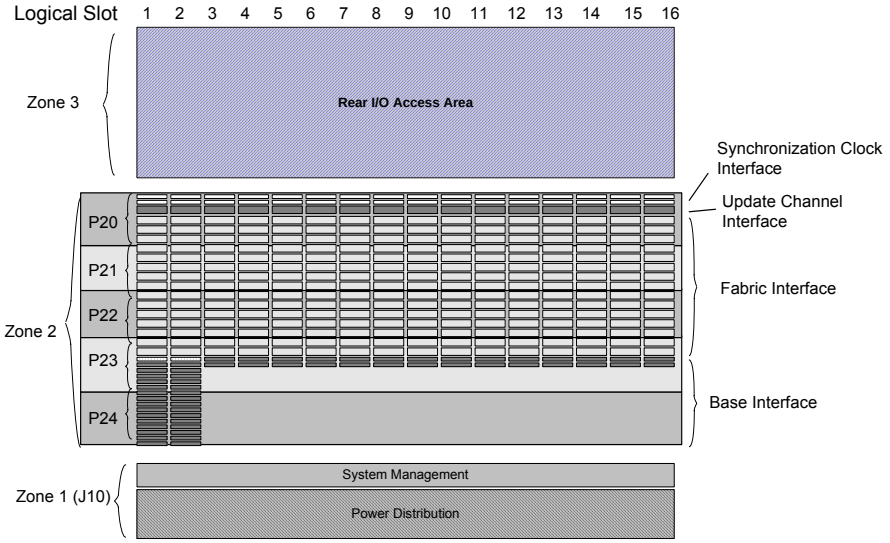


**Figure 5.** Logical structure of the ATCA backplane [10].

Ethernet or the optical fibers. For the latter option, the optical transceivers would be directly routed to the RocketIO$^{\text{TM}}$ports in the FPGAs. This is a method that has been already successfully implemented in the data acquisition system of the BESIII experiment [12, 13, 14].

As it is apparent in the figure 3, all backplane lines are connected to a single FPGA (tagged as `#0`). This FPGA will act as a network interface of the node. It will take care of the position of the module in the shelf and will present data to process to the other FPGAs. This system allows to have an identical designs for 4 FPGAs on each module. A design which will be completely oriented to data processing. For each compute node only the FPGA #0 needs a custom firmware to account for the interconnections topology.

*3.4. Remote Configuration*

Data acquisition equipment of large detectors need to have to be operated completely remotely. This includes the ability of perform power cycles and firmware updates. For PANDA compute node the embedded PowerPC in the FPGA #0 will have the duty to perform the firmware upgrades. The FPGA #0 will have an Ethernet connection via the base interface to a central module which will be most likely a commercial computer. Via this channel updates of the firmware and of the Linux kernel will be distributed. Once new data is received, they will be written to the flash device (fig. 6) containing the configuration data. After the next power cycle the FPGAs will be loaded with a new firmware [15].

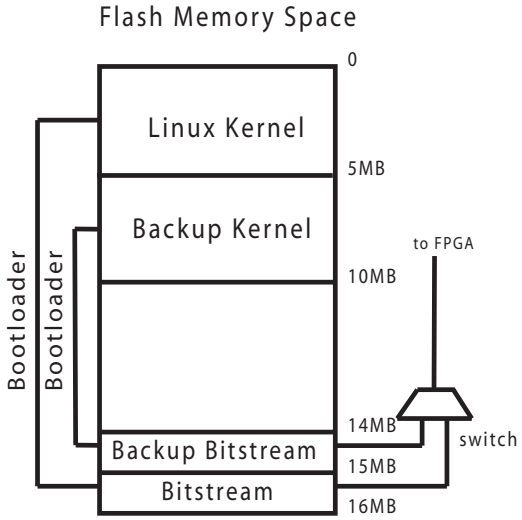


**Figure 6.** Structure of the non volatile flash memory. Firmware and Linux kernel are store here [15].

## 4. Current Status

Although the compute node is still in its developing phase, most of the subsystems have been already successfully tested. Connections of RocketIO$^{TM}$ports via optical fibers has been implemented for the data acquisition system of the BESIII experiment [14].

Tests on the performance and usability the embedded PowerPC to control the gigabit Ethernet ports and the firmware update routines have been carried out using a Xilinx ML403 development board [16]. The ML403 is equipped with a Virtex-4 FX12 which lacks of RocketIO$^{TM}$capabilities, therefore only the gigabit Ethernet and the firmware upgrades has been tested.

*4.1. Performance of the Gigabit Ethernet*

Due to the very demanding requirements of the $\bar{\text{P}}$ANDA data acquisition system concerning bandwidth, thorough performance test of the gigabit Ethernet ports were carried out. Tests were done using the gigabit Ethernet Xilinx IP core version 3.00 and the embedded Linux kernel version 2.6.10. The final figures show a maximum throughput of 300 Mbps with TCP and up to 375 Mbps when the protocol used was UDP [15].

## 5. Summary

The computational capabilities of FPGAs are not a new thing. The have been pushing through in the last years. Nowadays is normal that almost every supercomputer has a FPGA based co-processor (CRAY XD1 for example). The advantages of FPGA technology to build high intensive integer computational units have no doubt. The drawbacks arise from the lack of floating point support and, and this is probably the biggest issue, the lack of a simple-to-learn-yet-powerful high level language like C++, to program FPGAs. Although there are several projects to bring C-like languages to FPGA (Handel-C and SystemC among others), none of them has yet been able to get close to the performance, reliability and portability of VHDL or Verilog.

Once the compute node is ready, the challenge for the $\bar{\text{P}}$ANDA collaboration will be to find the appropriate methodology or language so that the members of the collaboration, non experts in logical design, could also develop software for the compute nodes. This way the scientific goals of the Panda program could be addressed much faster.

## Acknowledgements

The authors sincerely acknowledge the BMBF, the EU and the DFG for their support.

## References

[1] FAIR - Facility for Antiproton and Ion Research:
http://www.gsi.de/fair/index_e.html
[2] Panda homepage:
http://www-panda.gsi.de/auto/_home.htm
[3] Abazov *et al.* (2005), Technical Progress Report for PANDA.
http://www-panda.gsi.de/archive/public/panda_tpr.pdf
[4] P. Abbon *et al.* CERN-PH-EP/2007-001 (2007).
[5] L. Schmitt *et al.* IEEE RealTime 2003, Montreal, (2003).
[6] Igor Konorov (2006), *Private Communication.*
Panda DAQ workshop, Munich November 2006.
[7] J. Lehenrt *et al.*, NIM **A433** (1999)268.

[8] M. Traxler *et al.* IEEE Trans. in Nucl. Sci. **47** (2000) 376.
[9] Xilinx Inc. San José, CA (USA)
http://www.xilinx.com
Virtex-4 Multi-Platform FPGA:
http://www.xilinx.com/products/silicon_solutions/fpgas/virtex/virtex4/index.htm
[10] PICMG (PCI Industrial Computer Manufacturers Group).
AdvancedTCA specifications:
http://www.picmg.org/v2internal/newinitiative.htm
[11] The Linux Kernel Archives.
http://www.kernel.org
[12] Beijing Spectrometer III (BESIII) located at the IHEP, Beijing.
Web: http://bes3.ihep.ac.cn/bes3/index.htm
[13] IHEP:Insitute of High Energy Physics, Beijing (CN)
http://www.ihep.ac.cn/english/index.htm
[14] Z. Liu, The BESIII data acquisition system, BESSIII coll. meeting, Beijing, January 2007.
[15] Ming Liu, University of Gießen (2007). *Private Communication.*
T. Pérez, M.Liu *et al.*. DPG Spring Meeting, HK40.30, Gießen March 2007.
[16] Xilinx ML403 development Board.
http://www.xilinx.com/products/boards/ml403/reference_designs.htm