Ultrafast linear array detector for real-time imaging


a Institute for Data Processing and Electronics (IPE), Karlsruhe Institute of Technology, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany;

b Institute for Beam Physics and Technology (IBPT), Karlsruhe Institute of Technology, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany;

c Laboratory for Applications of Synchrotron Radiation (LAS), Karlsruhe Institute of Technology, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany;

d Univ. des Sciences et Technologies de Lille, rue Paul Dues 42, 59000, France;

e Now at SLAC National Accelerator Laboratory, Menlo Park, CA, USA;

ABSTRACT

KALYPSO is a novel detector operating at line rates above 10 Mpfs. It consists of a detector board connected to FPGA based readout card for real time data processing. The detector board holds a Si or InGaAs linear array sensor, with spectral sensitivity ranging from 400 nm to 2600 nm, which is connected to a custom made front-end ASIC. A FPGA readout framework performs the real time data processing. In this contribution, we present the detector system, the readout electronics and the heterogeneous infrastructure for machine learning processing. The detector is currently in use at several synchrotron facilities for beam diagnostics as well as for single-pulse laser characterizations. Thanks to the shot-to-shot capability over long time scale, new attractive applications are open up for imaging in biological and medical research.

Keywords: line scan detector, micro-strip detector, ultra-fast imaging, high data throughput readout, machine learning for photon science, artificial intelligent

1. INTRODUCTION

Photon science research depends radically on the developments of sensors and readout technologies for fast imaging. These technologies enable a wide range of applications in e.g. beam diagnostics, tomography, and spectroscopy. The repetition rate of commercially available linear array detectors is a limiting factor for emerging photon science applications. To overcome these limitations, KALYPSO (Karlsruhe Linear arrY detector for MHz rePetition rate SpectrOscopy), an ultra-fast and a wide-field of view linear array detector operating above 10 Mpfs, has been developed. A silicon micro-strip sensor is connected to custom, cutting-edge front-end ASICs in order to achieve an unprecedented frame rate in continuous readout mode. The system is designed for spectral range of near-infrared, visible and near-ultraviolet. The detector is connected to a novel readout system based on the ZYNQ Ultrascale+ MPSoC for high-performance real-time data processing based on artificial intelligence (AI) techniques. The system, operating at megahertz frame rates, is a unique platform combining fast detector electronics and artificial intelligence for photon science research. Novel applications of this system include classification of biological cells and tissues in microscopic images performed with unprecedented precision. This paper is organized as follows: the first section presents the KALYPSO detector and its sensor as well as the front-end ASICs, the second section describes the high-performance heterogeneous FPGA - GPU (Graphics Processing Unit) data acquisition system. The third section discusses the novel Machine Learning (ML) architecture embedded within the heterogeneous FPGA-GPU framework.

*michele.caselle@kit.edu; phone 49 721 608 25903;
2. KALYPSO

KALYPSO has originally been developed for the upgrade of the Electro-Optical Spectral Decoding experiments at the KIT research synchrotron light source KARA [1]. Later KALYPSO was also used at the European XFEL [2] and FLASH [3]. Sensor and front-end electronics are mounted on the detector mezzanine card shown in Figure 1. Two versions of semiconductor linear arrays can be employed, depending on the specific application. The first one is a Si microstrip sensor developed at KIT; the second one is an InGaAs linear array from Xenics [4]. The original version of KALYPSO was based on a slightly modified design of the GOTTHARD front-end ASIC [5] and achieves a maximum frame-rate of 2.7 MHz with 256 pixels. In the new version, several improvements have been included: up to eight GOTTHARD chips are connected to a wider linear array sensor with up to 1024 pixels at pixel pitch of 25 μm. Two multichannel, low-power, high-speed ADCs (Analog-to-Digital Converter) convert the analog samples, with a sampling rate up to 100 MS/s and 14-bit resolution, to digital data which are then processed by the FPGA. The detector board is connected to an FPGA readout card, based on the recent ZYNQ Ultrascale+ technology, by a FMC (FPGA Mezzanine Card) connector. The FPGA readout card is integrated in a custom high-performance and heterogeneous DAQ system consisting of FPGAs and GPUs connected via PCI-Express. When connected to the DAQ system KALYPSO sustains continuous data taking at the maximum repetition rate.

![Image of KALYPSO front-end card](image)

Figure 1. Picture of the KALYPSO front-end card (left), detail of the Si sensor, the two GOTTHARD ASICs and the high-density interconnection technology.

2.1 Sensor technologies

Three microstrip sensor typologies can be currently be mounted: InGaAs for infrared wavelength with efficiency up to 2.6 μm, three fine-pitch silicon sensors with special Anti-Reflective Coatings (ARC) optimized for spectrum range from near-ultraviolet (300 nm) to near-infrared (up to 1.1 μm), and novel Low-Gain Avalanche Detectors (LGAD) sensor for timing and very high-repetition rate applications. The InGaAs linear arrays are available with different number of sensitive channels with pitch of 25 or 50 μm which makes KALYPSO a very flexible detection system easily adaptable to many experimental conditions. To improve spatial resolution and efficiency at different wavelengths, a custom silicon sensor optimized for imaging applications has been designed and fabricated. The new sensor features an optimized geometry, with a pixel pitches of 25 μm and 45 μm, a variable number of pixels (from 512 to 2048), and three different ARC processes, optimized respectively for visible light, near-ultraviolet (300 nm to 400 nm) and near-infrared (up to 1.1 μm). The sensor wafer has been produced by Fondazione Bruno Kessler (FBK) in a special silicon bulk wafer using a photolithography process dedicated to low-leakage current and low-noise. A wafer produced by FBK and a detail of the sensor layout is shown in Figure 2.

An important limitation in high-speed imaging applications is the response time of the semiconductor detector. With traditional silicon sensors, the time required to collect the charge generated by incoming radiation is typically around a few tens of ns and therefore limiting the maximum frame rate. When exposed to incoming radiation with higher repetition rate, the sensor would quickly saturate. While it is possible to install a gating-intensifier stage in front of the detector, as used in intensified CCD (ICCD) cameras, these can degrade the resolution and the uniformity of the detector. In high-speed imaging detectors, APDs (Avalanche Photo Diodes) are widely employed. These are typically operated in linear mode, meaning that the intensity of the generated signal is proportional to the intensity of the incoming radiation. However, several technological challenges hinder the possibility to realize APD linear arrays with a large number of pixels.
Therefore, this makes them unsuitable for many photon science applications including beam diagnostics since they demand high spatial resolution and high efficiency. In order to overcome the rate limitation of conventional silicon detectors while maintaining high spatial resolution, novel fine pixel segmentation, high gain uniformity and high fill-factor Low-Gain Avalanche Detectors (LGAD) are considered for photon science applications. LGADs have been recently proposed in the literature for 4D tracking applications in high energy physics detectors [6]. With respect to conventional APDs, LGADs use a moderate internal gain, and thus enable the fabrication of finely-segmented microstrip and pixel detectors, and achieve charge collection times down to 3 ns [7]. This would ensure that the sensor is not saturated by the incoming radiation, even if exposed to pulses with repetition rates of a few hundreds of MHz. A novel silicon linear array based on the LGAD technology is currently under development for the KALYPSO system to be used in photon science application. The bulk thickness of only 50 µm will allow a fast charge collection time suitable for high-rate photon science applications. TCAD simulations show a total charge collection time less of 2 ns. The first LGAD demonstrator will possess a 50 µm channel pitch array where the multiplication layer will be insulated by Shallow Trench Insulation (STI) structures. The production of the demonstrator for the KALYPSO detector is scheduled for the mid of 2019.

2.2 Front-end readout electronics

An overview of the front-end electronics of the KALYPSO system is shown in Figure 1. The different channels of the readout ASIC are connected to a microstrip sensors, which can be tailored according to the experimental needs. To be compatible with different types of silicon or InGaAs microstrip sensors, the chip must be able to process signals of both polarities with a detector capacitance up to 6 pF. To further increase the frame rate, a dedicated readout application specific integrated circuit (ASIC) has been developed in a 110 nm CMOS technology from UMC. The ASIC features 128 readout chains with a 50 µm pitch to operate with a continuous frame rate of up to 12 MHz at full occupancy. Several chips will be connected to a microstrip sensor with up to 2048 channels. Each readout channel consists of a CSA (charge-sensitive amplifier), a noise shaping stage and a channel buffer. Multichannel, low-power, high-speed ADCs (analog-to-digital converter) digitize the analog samples that are afterwards processed by the FPGA. The FPGA readout card also synchronizes the operations between different ASICs and with the external timing system. A readout channel is shown in figure 5. The CSA is implemented as a differential folded-cascode amplifier with a variable feedback capacitance and a synchronous reset.
The folded architecture has been selected because of its high gain-bandwidth product, stability and output swing.

Despite the higher intrinsic noise of the differential architecture with respect to single-ended amplifiers, the differential architecture ensures a better shielding against external noise. This is a critical aspect for this development, because the readout ASIC will be closely integrated with noisy components such as the FPGA, PCI-Express data links and the accelerator timing distribution system. A test pulse capacitance $C_{\text{inj}}$ has been implemented on-chip to inject a known charge at the input of the CSA to evaluate the performance of the readout chain. The shaper performs correlated-double-sampling (CDS), reducing the low-frequency and $kT/C$ noise introduced by the synchronous reset mechanism [8]. In order to achieve high framerates, a sample-and-hold channel buffer is necessary to allow integrate-while-read operation. The timing strategy for a frame-rate of 10 MHz is shown in Figure 6.

**Figure 4.** Block diagram of the readout ASIC and the KALYPSO front-end electronics.

**Figure 5.** Simplified schematic of one readout channel.

**Figure 6.** Simulated response of the CSA (left), the CDS (middle) and the channel buffer (right) to a signal of 5 fC and a detector capacitance of 1.3 pF. The different readout phases and their duration are shown for a 10 MHz frame-rate operation. Different traces obtained through Monte Carlo transient noise simulations are superimposed on the plot, to highlight the "reset noise", i.e. the noise introduced after the RST is released.
To optimize the bandwidth while ensuring proper performance a two-stage architecture with cascode compensation has been adopted [9]. To achieve high speed and low distortion with a 100 Ohm load, the output driver has been designed with a two-stage class-AB OpAmp with cascode Miller compensation and Monticelli biasing scheme [10]. This last stage drives large capacitive off-chip loads, with a settling time of less than 4 ns and a non-linearity error of around 0.5% over a dynamic range of ± 900 mV. The layout of high-speed and low-noise readout circuit is the most critical step of the full-custom ASICs. It requires experience and many interactions steps of post-layout simulations and comparison with the expected signals. This is indeed a critical step, as the performance of a circuit can be severely affected by a sub-optimal placement and/or routing. Several design rules have been adopted to achieve superior signal-to-noise performance. Interdigitated transistor with common centroid and dummy-edge transistors are employed in analog design to improve the performance and compensate the gradient effect typical of the sub-micron CMOS technology. Furthermore, because of the increasing demand of using KALYPSO in soft X-ray applications, the digital circuits have been designed according to radiation hardness layout techniques as shown in figure 7.

Figure 7. Layout of two readout chains (left). Layout of digital circuits by Enclosed Layout Transistor (ELT) and P+ guards radiation hardness techniques (right).

3. HETEROGENEOUS HIGH DATA THROUGPUT FPGA-GPU ARCHITECTURE

Modern photon science experiments generate very large data volumes. To cope with this so called “big data” challenge, a high-performance heterogeneous FPGA-GPU-based computing infrastructure has been developed [11]. This infrastructure has been proposed for future generations of real-time data systems in photon sciences and high energy physics. General purpose computing on graphics processing units (GPGPU) is an efficient technology that allows programmers to leverage the massively parallel computing pipelines on modern graphics cards. To meet real-time constraint, the data is directly transferred from the FPGA into the GPU memories bypassing CPU and CPU memory by high-performance ad-hoc direct memory access (DMA), as outlined in figure 8. Using DMA it is possible to move data from FPGA to GPU memory without any intermediate buffering, off-loading the CPU and avoiding operating system jitter effects [11]. Traditionally, the data is first transferred from the FPGA to the main system memory and then sent to the GPUs for final data processing (dashed line in figure 8). The main memory is involved in a certain number of read/write operations, depending on the specific hardware or software implementation. Using direct FPGA-GPU communication (solid line in figure 8), the DMA engine has direct access to the GPU memory, therefore the total latency is drastically reduced down to only a few microseconds.

Figure 8. Heterogeneous FPGA-GPU data acquisition system based on direct FPGA↔GPUs data access.
The FPGA→GPU-based processing framework enables user-friendly and on-line monitoring of the data produced by the detector systems [12]. As demonstrated in [13], our architecture allows scientists to develop high-performance processing algorithms without specific knowledge of the underlying hardware architecture. Moreover, the low-latency performance of the DAQ system meets the requirements of fast feedback systems. With specific optimization of the data processing software components, latency as low as a few microseconds has been achieved [14]. The data throughput performance of the heterogeneous system is shown in figure 9 for two different payload block sizes.

![Figure 9](image)

**Figure 9.** Dual core DMA architecture based on PCIe generation 3 using two different payload block sizes: the red curve represents 256 Bytes and blue one uses 128 Bytes.

### 4. PCIE REDAOUT CARD

To prepare for the upcoming demand of high data throughput and fast data processing close to data source, a novel readout system based on PCIe generation 4 is currently under development. A sketch of the card is shown in the Figure 10.

![Figure 10](image)

**Figure 10.** New PCIe readout card for next generation of heterogeneous FPGA-GPU architecture. In addition to previous requirements the system is dedicated to machine learning applications.

The main processor is based on the ZYNQ UltraScale+ targeting the xcu11eg Xilinx device; it includes a 64-bit quad-core ARM processor with up to 1.5 GHz and a dual-core ARM with up to 600 MHz for Real-Time Operating System (RTOS). A Mali-400 GPU is also available for simple parallel data processing. The ZYNQ is equipped with a large FPGA with about 600 k configurable logic blocks (CLB) and several tens of megabytes of RAM and UltraRAM logic blocks. The selected FPGA contains more than 2900 DSP slices which are necessary for the synthesis and implementation of machine learning on FPGA. Furthermore, the FPGA support high performance interfaces i.e. four hardware implemented PCIe IP cores, one 150 G Interlaken, 32 GTH transceivers each one operating with up to 16.3
network weights are zero or near zero, they can be eliminated.

Results at 8 bits of precision, and in some cases even binary,
with considerably less precision.

Training flexible arithmetic capabilities
typically done on GPUs.

Computation functions
n (Facebook), TensorFlow (Google), Torch, PyTorch,
libraries that implement the common training and inference operations. Examples of
flexibility, compared
T
unique benefits to the deep learning problem, and a properly designed system can take advantage of this
technologies in a unified deep learning architectur
challenges of deep learning training and inference,
has enable
The combination can provide unique capabilities that result in higher
accuracy.

Deep learning (DL) techniques possible, led to wide usage in many applications. Deep learning and complex machine learning
have quickly become one of the most important and computationally intensive applications for a wide range of fields.
The combination of large data sets, high-performance computational capabilities, and evolving and improving algorithms
has enabled many successful applications which were previously difficult or impossible to realize. To cope the
challenges of deep learning training and inference, a heterogeneous system, which combines GPU and FPGA
technologies in a unified deep learning architecture, is under development. Each of these hardware technologies offers
unique benefits to the deep learning problem, and a properly designed system can take advantage of this combination.
The combination can provide unique capabilities that result in higher performance, better efficiency and greater
flexibility, compared to GPU and FPGA systems designed separately. Deep learning frameworks are sets of software
libraries that implement the common training and inference operations. Examples of these include Caffe/Caffe-2
(Facebook), TensorFlow (Google), Torch, PyTorch, etc. All of these are available as open source software. Deep neural
networks (DNN) training is very computationally intensive and typically makes use of common floating-point
computation functions, such as basic linear algebra subprograms (BLAS), therefore, in the proposed architecture, it is
typically done on GPUs. While the inference implementation, which requires large and low-latency on-chip memory and
flexible arithmetic capabilities, is on FPGAs.

5. MACHINE LEARNING IMPLEMENTATION

Artificial Intelligence (AI) has been around for 50 years, but only recent technologies have made advances in deep
learning (DL) techniques possible, led to wide usage in many applications. Deep learning and complex machine learning
have quickly become one of the most important and computationally intensive applications for a wide range of fields.
The combination of large data sets, high-performance computational capabilities, and evolving and improving algorithms
has enabled many successful applications which were previously difficult or impossible to realize. To cope the
challenges of deep learning training and inference, a heterogeneous system, which combines GPU and FPGA
technologies in a unified deep learning architecture, is under development. Each of these hardware technologies offers
unique benefits to the deep learning problem, and a properly designed system can take advantage of this combination.
The combination can provide unique capabilities that result in higher performance, better efficiency and greater
flexibility, compared to GPU and FPGA systems designed separately. Deep learning frameworks are sets of software
libraries that implement the common training and inference operations. Examples of these include Caffe/Caffe-2
(Facebook), TensorFlow (Google), Torch, PyTorch, etc. All of these are available as open source software. Deep neural
networks (DNN) training is very computationally intensive and typically makes use of common floating-point
computation functions, such as basic linear algebra subprograms (BLAS), therefore, in the proposed architecture, it is
typically done on GPUs. While the inference implementation, which requires large and low-latency on-chip memory and
flexible arithmetic capabilities, is on FPGAs.

![Diagram](image)

Figure 11. Heterogeneous FPGA GPU machine learning architecture.

Training mainly uses floating-point operations, but for inference operations many network weights can be represented
with considerably less precision compared to the precision used in the training data. Recent studies show very good
results at 8 bits of precision, and in some cases even binary networks can obtain the desired accuracy. In cases where
network weights are zero or near zero, they can be eliminated, or “pruned,” from the graph. Such pruning can
considerably reduce the computation requirements for inference operations. Once a model for deep learning is trained, deploying it for inference in deep learning applications typically has additional requirements. While the computational demand for the inference is less in terms of complexity and training data, there are several key requirements for inference. In photon science application that requires a “fast decision”, inference requirements fall into two categories: low latency and high data throughput. FPGAs take advantage of their flexibility to provide broad performance, latency, and power envelopes. Beside the complementary strengths of CPUs, GPUs, and FPGAs for different types of deep learning operations, there will be a growing benefit of unified deep learning configurations combining all three types of hardware elements into single heterogeneous system. The training and inference are typically thought of as very distinct operations, but there are emerging applications that will combine them. Continuous and reinforcement training can be used when a deep learning system is deployed in an environment where regular retraining and updating of a DNN is required. The proposed architecture that controls a device or analyses datasets in real-time is shown in Figure 11. It is based on continuous learning and adapting in according to the results of their previous actions. Similarly, systems that interact with humans can use feedback from humans to continuously improve their training. The inference is implemented in the FPGA and it is optimized in terms of performance and response latency, whereas the training is executed on GPUs. During the data processing by ML on FPGA, two cases are possible: the result of inference is good, meaning that the feedback signal to the experimental station has been predicted with high grade, or not so well predicted. In this last case, the inference ML in the FPGA will send to the GPU the ML status as well as the prediction of the next ML status together with the Action taken for a given dataset (red arrow), the GPU will evaluate the answer by an independent and massive floating-point ML network and then re-configure the FPGA in order to improve the prediction of the inferred ML network in the FPGA (blue arrow). A unified deep learning platform that simultaneously employs GPUs for training updates and FPGAs for inference enables efficient implementation of such continuous training systems.

6. SUMMARY AND CONCLUSIONS

Modern photon science detectors requires many cutting-edge technologies: custom ASIC and semiconductor design and fabrication, low-noise analog circuits, high-density interconnect technologies, high-throughput readout electronics combined with advanced real-time data processing based on new generation of artificial intelligence. KALYPSO is the result of a close collaboration between engineering and beam line scientists. The collaboration has produced one of a fast line-camera available in the scientific communities. To the best of the author’s knowledge, this is the highest frame rate achieved by high-resolution imaging detectors with continuous data acquisition. The detector has been installed at the EOSD experimental setups at KARA and at the European XFEL. Several technical developments have been started to further improve the performance of the detector like the possibility of replacing the CDS noise shaper with a trapezoidal time-variant shaper, in order to further optimize the noise performance of the GOTTHARD chip [18]. Furthermore, novel LGAD silicon sensor will open new possibility for high resolution and repetition rate applications. The continues acquisition for long observation time by high data throughput readout electronics combined with artificial intelligence framework will open new scenario for the real-time data processing and fast detection of rare event today considered very difficult or almost impossible to observe.

7. ACKNOWLEDGMENTS

The authors would like to thank A. Mozzanica and B. Schmidt for providing us with the GOTTHARD chip, Lorenzo Rota for the design of the ASICs, Fabio Colombo, Pia Steck and Simon Kudella for their help with the production of KALYPSO, and Alexander Dierlamm, Marta Baselga and Daniel Schell for their assistance during the design of the silicon sensors. This project is partially funded by the German Federal Ministry of Education and Research (BMBF) contract number 05K16VKA and by the Helmholtz President’s strategic fund IVF “Plasma Accelerators”.

REFERENCES


