Innovative developments for real-time data processing in particle physics within IN2P3

June 1, 2022

Dorothea vom Bruch (CPPM)

on behalf of the ATLAS-IN2P3, CMS France and LHCb France communities
and the OWEN and THINK projects

Abstract

This document describes the innovative developments for real-time data processing within IN3P3 required to address some of today's scientific challenges with the ATLAS, CMS and LHCb experiments as well as the OWEN and THINK projects. After a short summary of the scientific and technological challenges in section 1, the developments within every experiment and project are outlined in sections 2 to 6. The document closes with a global summary in section 7.

1 Scientific and technological challenges

The Standard Model (SM) of particle physics describes most fundamental phenomena extraordinarily well, yet many unsolved questions remain. For example, the origin of the matter-antimatter asymmetry in the universe, the nature of dark matter and the origin of neutrino masses are not yet understood. New Physics (NP), able to resolve these questions, has not been found in direct searches yet. So it must be either extremely rare or manifest itself at higher energies, than those probed until now. Searches for NP proceed along two paths: direct searches and indirect searches. It is widely recognized in the particle physics community that both of these complementary approaches must be utilized in the search for NP. The first focuses on directly detecting new particles produced in high-energy collisions. The latter uses high-precision measurements to investigate effects of NP on SM processes induced by virtual particles. Thus, mass scales far beyond the reach of today's colliders can be probed.

The ATLAS and CMS experiments are mainly designed to study the properties of the Higgs bosons and to search for new phenomena at high p_T, while LHCb was designed for high-precision flavour physics measurements. The OWEN project targets experiments searching for double beta decay without the emission of neutrinos, which is the most sensitive way to assess the Majorana nature of neutrinos and therefore the origin of their mass.

For all of these experiments, the current precision is not sufficient for a discovery due to the limited size of data sets. Therefore upgrades and new more sensitive experiments are planned in the near future to observe particle decays at higher rates. But at these rates, efficient selection of signals requires exceptional computing demands. In particular since the increased luminosity at particle colliders leads to saturation phenomena in hardware level triggers. One way of handling this problem is to transfer all data to a server farm where each collision is assembled from the sub-detector data and analyzed in real-time by heterogeneous architectures available in the servers. This comes with two challenges: 1) Connecting the sub-detectors to the server farm. 2) Using the computing architecture best suited for the reconstruction of particle collisions and designing algorithms for this architecture within a heterogeneous software framework.

The LHCb experiment chose this approach to solve the computing challenge it faced for Run 3, starting in 2022. 40 Tbit/s (30 MHz of p-p collision rate) are read out from the detector and received by a server farm equipped with GPU cards, which perform the first reconstruction and selection stage to reduce the data rate by a factor of roughly 30. A second reconstruction and selection stage is performed on a separate CPU farm.

If the entire data rate cannot be readout from the detector, as is the case for CMS and ATLAS, efficient local reconstruction and selection in the hardware-level trigger and close to or on the front-ends is particularly important. Consequently, developments also concentrate on machine learning applications processed on FPGAs and on ASICs developed for data processing.

With the beginning of the high luminosity era (HL-LHC), CMS and ATLAS will face similar software computing challenges in Run 4 (to start around 2029) as LHCb did for Run 3. They will collect pp collision data with a pileup of 200 and increase the hardware trigger output rate from around 100 kHz in Run 3 to 1 MHz, such that around 48 Tbit/s of data have to be processed in software in the HLT stage. LHCb plans a second upgrade (Upgrade II) for Run 5 (to start around 2035), where the data rate will be increased by a factor 5, such that 200 Tbit/s will be analyzed in software.

2 ATLAS

The ATLAS Liquid Argon (LAr) calorimeter measures the energy of particles produced by LHC collisions. This calorimeter has also trigger capabilities to identify interesting events for further processing. The signals from the LAr calorimeter are processed through a chain of electronic boards in order to extract the energy deposited in the calorimeter. An excellent resolution on the deposited energy and an accurate detection of the deposited time, in the blurred environment created by the pileup of many in-time and out-of-time energy deposits, is crucial for the operation of the calorimeters and for the full ATLAS detector to enhance its discovery potential.

2.1 Timeline

The LAr calorimeter trigger readout system was upgraded during the first upgrade phase of ATLAS (2019-2021). The granularity of the new trigger system is ten times larger than the one used during run 2 of the LHC. The improved granularity mainly increases the separation between electromagnetic and hadronic showers leading to a better trigger efficiency for electrons and photons. The new trigger system allows to digitize with on-detector electronic boards the electronic pulses from the calorimeter and send the digitized samples to the off-detector boards where FPGAs are used to compute the deposited energies using optimal filtering algorithms [9]. LAPP designed and produced the main off-detector board for the phase I upgrade. In addition CPPM and LAPP coordinated the ATLAS group responsible of designing the firmware for this board.

For the phase II upgrade, the full readout electronics of the LAr calorimeter will be exchanged. This allows to cope with the high pileup expected at the HL-LHC and the new level 1 trigger rate which is increased by a factor of 10. The new electronic chain for the phase II upgrade is described in [10]. The off-detector board (LASP) responsible of the data processing and the computation of the energy is being designed by CPPM. The LASP board will comprise two high-end INTEL FPGAs allowing the usage of machine learning techniques to compute the energy.

2.2 Ongoing developments, major roles of IN2P3 people

The CPPM group already produced a demonstrator LASP board containing two Stratix 10 FPGAs. This board was tested and is working successfully. The CPPM team is now working on producing the first prototype of the LASP board which is expected to use the next generation FPGAs from INTEL (AGILEX Family). This prototype is expected to be ready next year. Meanwhile INTEL development kits are being used to develop and test the firmware.

The CPPM group in collaboration with the University of Dresden initiated a project to use Neural Networks (NNs) to compute the energy on the LASP boards. The optimal filter algorithms were perfectly adapted for the ideal situation where there is very limited pileup and no overlap of the pulses in the detector. However with the increased luminosity and thus pileup, the performance of the filter algorithms decrease significantly while no further extension nor tuning of these algorithms could recover the lost performance. Reference [10] shows the increased noise (factor of 2 to 5) and the reduced energy resolution (dominated by the pileup at low energy) for the pileup conditions at the HL-HLC.

The CPPM group focuses on developing Recursive Neural Networks (RNNs) for the computation of the energy reconstruction. RNNs are very well adapted for time series and allow to correct the reconstructed energy at a particular time using the information about past events to probe pileup effects. Several networks based on the Vanilla-RNN and LSTM architectures are developed and shown to outperform the optimal filtering algorithm [11].

The main challenge of this project is to implement the developed NNs on FPGAs. In the current design, one FPGA of the LASP should compute the energy of 384 or 512 LAr readout channels within O(100 ns). The application of NNs on FPGAs is constrained by the limited digital signal processing (DSP) resources, logic and memory available in the FPGA devices and the maximum frequency at which the firmware can run. This, in turn, limits the number and type of mathematical operations that can be used by the NNs. In addition, software tools for converting trained NNs into FPGA firmware are needed. The CPPM group successfully managed to produce very small RNNs capable of computing the energy with improved resolution compared to the optimal filtering algorithms. In addition, through a long process of firmware optimisation, these RNNs could fit within the resource usage and latency requirements of the LASP board. This was an important milestone that shows for the first time that NNs can be used to process in real-time raw detector data within ATLAS. The RNN firmware is developed in HLS which is also compared to VHDL and software implementations. In order to allow other groups to profit from the developed and optimized High-Level Synthesis (HLS) code, the CPPM group added the support for the RNN implementation on INTEL FPGAs into the HLS4ML toolkit [12]. The CPPM LAr group also participates to the THINK project (described in section 6); the LAr energy computation and the HLS developments are used as a use-case for the investigations which are carried on by THINK.

2.3 Future plans

The CPPM and Dresden teams are focusing on further optimisation of the NNs mainly to reduce the FPGA resource usage and latency to a minimum. The first viable networks are available and are being tested on the hardware using INTEL development kits with the Stratix 10 FPGAs. The testing will be also done on AGILEX development kits once these are available. Several networks are implemented in one FPGA with time multiplexing to compute simultaneously the energy of hundreds of channels. Timing violations in the firmware are being scrutinized and the network outputs are being validated with injected data to ensure that the results on the hardware match the simulation.

In parallel further optimisation to the network architecture and training are being followed to ensure that the produced networks are robust against the changing HL-LHC conditions and differences between various cells in the calorimeter. This work should result in a training and calibration procedure for the 180k channels of the LAr calorimeter that should be implemented at the beginning of HL-LHC run. The networks are also being implemented in the ATLAS official simulation software (ATHENA) and we plan to produce a full simulation of physics processes with the energy computed with the implemented networks. This will allow to quantify the effect of the improvements in energy resolution on the physics performance.

2.4 Resources

This project (called AIDAQ) started in 2019 with an AMIDEX funding at CPPM in collaboration with the university of Dresden. This funding allowed to hire a postdoc for 20 months followed by an engineer for 10 months. Since then, additional funding from AMU and CNRS allowed us to hire 3 PhD students to work on this project. Two of these students are electronic engineers that performed the firmware implementation. Last, an ANR JCJC funding

Physicists	G. Aad, E. Monnier
Postdoctoral researchers	T. Calvet (2020-2021), N. Sur (2021-2024)
Electronic engineers	R. Faure (2022)
Doctoral students	N. Chiedde (2020-2023), E. Fortin (2020-2022), L. Laatu (2020-2023)

Tab. 1: People involved in the AIDAQ project at CPPM. The covered period is shown for non-permanents.

was obtained for a 3 years postdoc that started end of 2021. All the hired postdocs, students and engineers are at CPPM and are listed in table 1. Since 2019, seven students in physics, informatics and electronics at different levels in their studies performed their internship on the AIDAQ project. The project is fully relying on the above-mentioned dedicated funding and does not use any significant resources from the permanent engineers at CPPM. Two permanent physicists at CPPM are involved in this project part time. An ERC project including this development was submitted in 2021 but was not successful.

2.5 Visibility, strengths, weaknesses

The CPPM and Dresden teams are leading this effort within the ATLAS LAr collaboration. The CPPM team is very visible with regular presentations within the LAr collaboration meetings. A first paper [11] showing the feasibility of the project was published during the second year of the project. In the last year, 3 presentations (2 CPPM) and 7 posters (4 CPPM) were shown on this subject at international conferences (more already scheduled for the remaining of this year). Also the optimisations that we added to reduce the resource usage were very well received by the HLS4ML group. This allowed the CPPM group to join the HLS4ML developers and gain additional visibility in this community.

The relatively fast development of NNs that are viable for the LAr energy computation allowed the AIDAQ project to be taken seriously within the LAr collaboration after a doubtful period. This, added to the fact that the CPPM is designing and producing the electronic boards where the NNs will run, represents one of the main strengths of the project.

The most important weakness of this project is the lack of permanent resources especially electronic engineers working on the firmware. Developing such a firmware requires specific knowledge that can be acquired only by experience. Unfortunately this knowledge will be partially or totally lost with the departure of the students if no additional resources are found well in advance to allow the transfer of this valuable experience of implementing NNs on FPGAs.

3 CMS

3.1 Timeline

French CMS groups have been involved in the development of the trigger for many years. On one hand IPHC Strasbourg has been involved in the implementation of b-tagging algorithms for the HLT, and in the development of trigger paths for analysis purposes for Run 2. On the other hand, the LLR group has been a leading protagonist in the development of the CMS L1 trigger system. It has major contributions within the existing system, on the ECAL trigger primitives generation and on the calorimeter trigger. It is also one of the main drivers of the developments of the future L1 trigger system for the HL-LHC era (starting with Run 4), and more precisely on the reconstruction of the HGCAL trigger primitives.

The detector upgrades for the HL-LHC will be installed during the long shutdown 3 (LS3) between 2026 and 2028. Run 4 is then expected to start in 2029 and will last for a period of 4 years. In this context the production of the components of the future L1 trigger system should be completed at the end of 2025, and their integration should then be finalized in 2028.

3.2 Ongoing development, major roles of IN2P3 people

The LLR contributions to these different trigger systems are related to algorithm developments. These algorithms are real-time reconstruction algorithms running on FPGAs with latency constraints of a few micro-seconds. The inputs of these algorithms are typically calorimeter cells and the output are higher-level reconstructed objects such as clusters of energy or electrons, photons, hadronic taus, etc. In this context the LLR CMS group has been interested very early in integrating machine learning algorithms within L1 trigger processing chains. BDTs have for instance been used within the reconstruction of electrons, photons and hadronic taus already in 2013 during the development of the Phase 1 calorimeter trigger. These BDTs, encoded into lookup tables in the FPGAs, are running in the current trigger system. Although this was a major achievement at that time, such methods to synthesize machine learning models into an FPGA is limited by the available number of Block RAM in the chip. And only models based on a very low number of inputs can be implemented in this way.

Therefore, more recently some preliminary work has been done in order to study the implementation of actual machine learning models such as BDTs and neural networks in FPGAs. This is part of a global effort, and this is a topic that has gained interest worldwide thanks to modern FPGAs with larger resources and the growing maturity of C++-based HLS tools. The focus of the work at LLR has been on the multi-objective optimization of both the model performance and their resource utilization. Typically two aspects of a model can be optimized in terms of FPGA resources: 1) the number of operations performed in the model, or its size and 2) the precision of these operations (the number of bits on which the operations are done). These two aspects have been studied at LLR, for both BDTs and neural networks. Optimization techniques such as Bayesian optimization and genetic algorithms have been used to optimize jointly the model performance and their size and precision. In the case of neural networks, the QKeras quantization library has been used to perform quantization-aware training. FPGA implementations have been performed thanks to the HLS4ML and Conifer tools for neural networks and BDTs, respectively. These techniques have been applied to a variety of use cases, using classification and regression models based on high-level features (BDTs and fully connected networks) or low-level features (convolutional neural networks). Given the irregular geometry of the HGCAL, there is also a research activity at LLR on graph neural networks. It has mostly been applied in the context of the offline reconstruction so far, but it is planned to extend these studies for real-time reconstruction at the L1 trigger level.

3.3 Future plans

So far, BDTs and neural networks have been mainly developed and studied as isolated components. One major step forward will consist in integrating these components into the real hardware system and connecting them with the rest of the firmware. This will be a challenging task given the extremely high data throughput processed by the system. In parallel, the implementation at the L1 trigger of more sophisticated models, such as graph neural networks, will also be studied. The integration of such models is expected to be even more challenging with the current state-of-the-art technologies. It is nevertheless important to start these studies already now in order to pave the way towards the next trigger generations.

3.4 Visibility, strengths, weaknesses

The LLR group has highly visible contributions and major roles within the L1 trigger system of CMS, of which A. Zabi (LLR) is the project manager, and in the HGCAL project, where J.-B. Sauvan (LLR) is coordinator of the algorithm developments for the HGCAL trigger primitives generation (TPG). Its major strengths are its longstanding involvement in these projects and its strong bounds with other contributing institutes. The developments of machine learning algorithms for the CMS L1 systems is still ramping up as the group has important commitments in the HGCAL TPG project and its efforts are driven by the simultaneous concerns of building a system that matches the hardware constraints and will be operational on day 1 and boosting its performance thanks to machine learning. Nevertheless, the group is spending a significant effort in strengthening its level of resources on the latter.

3.5 Resources

The studies described above are made possible thanks to the availability at LLR of several GPU and FPGA platforms. GPUs are being used for the training of neural networks, while FPGAs are used for the implementation and test of the models. These platforms have been funded from various sources (P2IO Labex, ANR and TGI HL-LHC).

In addition, the CMS group at LLR has spent a significant effort in strengthening its level of human resources on these subjects in the past years. Strong collaborations with other institutes are being built, in particular with the University of Split (Croatia) and the Imperial

LLR	Physicists	F. Beaudette, J.-B. Sauvan, A. Zabi
	Research engineers	E. Becheva, F. Magniette
	Doctoral students	A. Hakimi, J. Motta

Tab. 2: People involved in real-time analysis for CMS at LLR.

College (London, UK). Effective and successful collaborations have already been built with these institutes in the past on L1 trigger related projects. The level of resources has also been strengthened through external funding. A first ANR project (HiGranTS, PI: J.-B. Sauvan) started at the end of 2018 and includes one work package on the implementation of machine learning models on FPGAs. It has been funding one PhD thesis and two years of postdoc. A second ANR project (OGCID, PI: F. Magniette), which started at the end of 2021, is focusing on detector reconstruction based on graph neural networks with a part devoted to online reconstruction. It will fund one PhD thesis. Finally a joint international project between Imperial College and LLR is starting now and will fund one PhD student at LLR. A snapshot of the personnel involved in real-time analysis for CMS at LLR is shown in table 2:

4 LHCb

4.1 Timeline

During Run 2, LHCb employed first a hardware-level trigger to reduce the event rate from 40 MHz to 1 MHz, followed by a two-stage high level software trigger: HLT1 and HLT2. The high level trigger ran on a server farm where alignment and calibration was processed in near real-time between HLT1 and HLT2 stages. For Run 3 starting in 2022, LHCb underwent a major upgrade (Upgrade I). Since the luminosity is increased by a factor 5, a hardware level trigger is no longer efficient due to signal saturation. Therefore, the full detector is read out at 40 Tbit/s and reconstruction occurs at the 30 MHz inelastic collision rate in LHCb. For Upgrade II [1], another increase in luminosity of a factor 10 will result in $\sim$ 200 Tbit/s data rate to be analyzed.

LHCb-France scientists were involved in the reconstruction and real-time processing of LHCb data since the construction phase of the original LHCb detector in the early 2000s. The French LHCb teams were responsible for the L0 trigger (R. Le Gac) and designed, built and exploited most of the hardware processors: L0 Calo (IJCLab), L0 Muon (CPPM), L0DU (LPC), which performed well during Runs 1 and 2. All the charged particle reconstruction algorithms for the LHCb detector were originally written by O. Callot (IJCLab). V. Gligorov (now LPNHE) wrote most of the original selection algorithms for the first-level software trigger while at CERN in 2010, which remain the basis of LHCb's Run 3 selections. Together with M. Williams (MIT) he also developed the primary inclusive second-level software trigger. It was the first of the main LHC triggers fully based on a machine learning approach, and lasted the whole of Runs 1, remaining the baseline for the Upgrade I. Building on this long tradition of work in reconstruction and real-time processing, French teams have been heavily involved in preparations for Upgrade I (being commissioned now) and plan to continue in view of

Upgrade II [2].

4.2 Ongoing development, major roles of IN2P3 people

For Upgrade I, French teams are involved in two projects: The PCIe40 cards and the Allen project. The former connects sub-detectors with the server farm, while the latter is a natively cross-architecture framework for heterogeneous real-time data processing used for LHCb's first selection stage on GPUs. The PCIe cards receive the data at 40 Tbit/s from the sub-detectors and transfer it to the server farm for event building. Local data processing occurs on a card using only the information provided by the links from the sub-detectors connected to it. The type of processing varies depending on the sub-detector, for some of them clustering of measurements is processed for example. The card was developed at CPPM with R. Le Gac (CPPM) as scientific project leader and J.-P. Cachemiche (CPPM) as technological project leader. The card is generic enough to be re-used similarly by the ALICE, Belle-II and Mu3e experiments.

Within Allen [14, 15], HLT1 is processed entirely on GPUs. This includes decoding of raw data, clustering for some sub-detectors, track reconstruction using all tracking detectors, vertex reconstruction, muon and electron identification. Selections of collision events are based on single track and vertex properties to extract both signal processes and control signals. The latter are used to align and calibrate the detector in quasi real-time. D. vom Bruch (CPPM) and V. Gligorov (LPNHE) led this project together with R. Aaij (Nikef) and D. Campora (Maastricht University). Building on pioneering R&D work performed during D. Campora's thesis, Allen delivered a full implementation of LHCb's first-level software trigger, running fully on GPU cards, in time to be chosen by the collaboration as baseline for Run 3. This work has been accomplished by combining the competences of several IN2P3 laboratories (CPPM, IJCLab, LAPP, LPNHE) in terms of reconstruction software, trigger system development and integration in the DAQ system. The majority of LHCb France's involvement with real-time processing in Run 3 will be to develop and maintain Allen, which is generic enough to be utilized also by other experiments.

Allen has already shown that LHCb will be able to perform tracking to far lower momenta than foreseen in the original upgrade trigger TDR [5], as well as to deploy full electron identification and Bremsstrahlung recovery already at the first trigger stage. With further developments, it would be possible to run the full LHCb tracking, including finding tracks which originate outside the vertex detector, for the first time at the first trigger stage.

A secondary but nevertheless important area of activity is calorimeter reconstruction and calibration for the second trigger stage, led by J.-F. Marchand (LAPP). In the past years the calorimeter software has undergone a complete revision in order to modernize and optimize the reconstruction algorithms of calorimetric objects for multi-threading and parallel processing. The calibration procedure, fully automatised during Run 2, consists of an absolute and a relative calibration. The absolute one is based on an iterative computation of the π^0 mass for each calorimeter cell and is processed online. The relative calibration uses injected LED signals to tackle the PMT ageing during physics fills. The same approach has been adapted for Run 3 with revisited and modernized selections to handle the larger pile-up and to fit the new software environment.

4.3 Future plans

PCIe card
For Upgrade II, the bandwidth and processing power need to increase by a factor 10 compared to the PCIe40 card. In Addition, to keep the event-builder compact and to minimize its cost, the network interface card (NIC) has to be included directly in the successor of the PCIe40 card. Therefore, the board serves directly as source of the event building process. To achieve this, it is foreseen to develop cards in two-phases: For LS3, the PCIe400 card transferring 400 Gbit/s to the PC-server using the PCIe bus and the FPGAs from Intel (Agilex) available in 2022. For LS4, the PCIe800 card transferring 800 Gbit/s of data via Ethernet and using a more powerful FPGA.

The PCIe400 card is currently under development by CPPM, CENBG, IJCLab, LAPP and LPC Caen (IN2P3 R&T project "PCIe400"). The wide involvement shows the interest of various laboratories in this innovative technology. It will ensure the transfer of knowledge in terms of high-bandwidth data acquisition, production, software for processing and control. In addition to the aggregation layer, the powerful FPGAs can be used for local processing using the information provided by the 40 links from the sub-detectors. Intrinsically local clustering algorithms should be possible to implement in the FPGAs for all sub-detectors. Beyond this, it will be crucial to explore which other high-level physics primitives can be computed, without communication between readout cards, in order to reduce the processing load of the high level software stages.

In particular, the most efficient processor from a cost / energy point of view should be chosen for individual tasks, also taking into account an efficient data flow. To evaluate the best technology for a given task, a demonstrator test bed is being developed within the RTA project, such that existing and emerging architectures can be assessed and compared in terms of their performance for different tasks and dataflow bottlenecks can be identified.

Real-time reconstruction
In Upgrade II, the main computing challenge will come from HLT2 since the problem scales quadratically with luminosity. On one hand, the signal volume selected by HLT1 increases linearly with luminosity, while on the other hand, the size of every selected event also scales almost linearly with the luminosity. Therefore, HLT2 reconstruction algorithms will require the most computing resources. They shall be processed on GPUs for cost- and energy-efficiency.

A large fraction of Upgrade II sub-detectors will also provide timing information. So one can explore the option of performing pileup suppression already at the level of HLT1. Measurements which are clearly not associated to the primary interaction of interest are suppressed, reducing the event size.

As energy efficiency becomes increasingly important, it is also crucial to investigate the feasibility of almost real-time processing on data centers in France, as compared to running the full reconstruction and selection chain on a “small” computing farms located at the LHCb

LPNHE	Physicists Postdoctoral researchers Software engineers Doctoral students	V. Gligorov, F. Polci M. Fontana, C. Agapopulou A. Bailly-Reyre, N. Garroum A. Scarabotto, T. Fulghesu
IJCLab	Physicists Doctoral students	Y. Amhis, F. Machefert, P. Robbe F. Volle, V. Vayeroshenko
CPPM	Physicists Doctoral students	A. Poluektov, R. Le Gac, D. vom Bruch V. Dedu

Tab. 3: People involved in the LHCb Real Time Analysis project within IN2P3.

site. Once the data volume has been sufficiently reduced by HLT1 and possibly parts of HLT2 reconstruction, the remaining data stream can be sent to HPC facilities. There, information about the signal candidates would be built and the exclusive HLT2 selections performed. The optimal configuration will be a balance of financial and energy cost implications when using different resources.

4.4 Visibility, strengths, weaknesses

The PCIe40 project has good visibility in the HEP community, as it provides high bandwidth data acquisition cards. In particular, four different experiments use it: LHCb, ALICE, Belle-II and Mu3e. The only other comparable card is the FELIX board developed by the ATLAS collaboration, which mainly differs by the FPGA family: Xilinx instead of Intel. The successor of the PCIe40 card, developed in close collaboration with CERN, might be used in the foreseen upgrade of Alice and Belle 2. Although the CPPM team is the natural candidate to lead this project, it suffers from retirement of its key engineers. Reinforcing the team, and knowledge transfer to the next generation, are mandatory if we would like to keep this know-how at IN2P3.

LHCb's innovative real-time reconstruction has received wide recognition in the community. First, the offline quality reconstruction and online calibration and alignment for Run 2. Then the implementation of HLT1 on GPUs for Run 3. These developments have been recognized with prizes, two ERC grants for French researchers (RECEPT and ALPaCA) and two ANR grants (ANN4EUROPE and BACH).

Allen provides a natively cross-architecture framework for heterogeneous real-time data processing, but is in competition with other heterogeneous frameworks. Here, long-term maintenance support will be crucial for the continuation of Allen and its application to other areas (such as heterogeneous simulation) and experiments. Continuous support by engineers or applied physicists will be needed to support a diverse range of use-cases.

CPPM	Physicists Research engineers	R. Le Gac K. Arnaud, P. Bibron, J.-P. Cachemiche, J. Langouet
LAPP	Physicists Research engineers	S. T'Jampens G. Vouters

Tab. 4: People from LHCb involved in PCIe40/400 projects within IN2P3.

4.5 Resources

The French community is among the leading national contributors to real-time processing in LHCb, at a similar level as the UK, CERN, Germany and the Netherlands. This will continue for the foreseeable future. However, the contributions in reconstruction have been largely auto-financed in the previous years by ERC and ANR projects. For R&D concerning processing in near-real time on HPC centers, we have submitted one proposal within the French Exascale initiative and one within the PEPR NumPEX.

Table 3 shows that the Real Time Analysis (RTA) project has been mainly supported by physicists, postdoctoral researchers and doctoral students, with small contributions from engineers. The PCIe project on the other hand is based on strong support from engineers (see table 4) working closely together with physicists. Only this close collaboration enabled the successful development, production and deployment of the PCIe40 cards. The success of the PCIe400 cards will depend on continuous support from both engineers and physicists. A similar collaboration is required for the RTA developments to make the contributions truly sustainable and maximize the scientific output we can get from our work to date. Therefore, we need to reinforce the software engineering teams, a group of people which IN2P3 is uniquely positioned to provide.

5 OWEN project: Optimized Waveform for Electronic Nodes

5.1 Timeline

OWEN stands for "Optimal Waveform recognition Electronic Node". The project consists of developing a readout system with on-line signal waveform analysis of data originating from a spherical high pressure gaseous TPC. Such a detector is used for example in direct dark matter detection and neutrinoless double beta decay observation. 2022 will be the end of the proof of concept from both OWEN and the R2D2 RT project, aimed at R&D for a TPC detector. Thanks to the low data rate of R2D2 [13], we could test a new approach of detector signal processing. The goal is to invert what has been mixed in the detector from the incident particles (temperature, pressure, gas mixture, nature of track and signal vs background). Contrary to other approaches processed after the analog signal-to-noise (SNR) optimization, we want to place the classifier just after the charge preamplifier for the raw signal. One of the various possible architectures of neural networks could classify interesting signals (double beta decay) versus background. As the NN is near the detector, the system needs to control directly the analog to digital converter (ADC) without latency. So, GPUs and CPUs are out of the race because of the need for memory buffering and therefore tight latency constraints. FPGAs can handle the DAQ functionality and NN inference with an expected response lower than 10's of µs. As a next step, a decision has to be taken from IN2P3 or agencies to fund the construction of the complete experiment in phase 1. In case the project is not continued, the components developed within OWEN will remain available for others projects.

5.2 Ongoing developments, major roles of IN2P3 people

The main features of the R2D2 detector are its simplicity with a single or a few readout channels, the low mass material budget resulting in a low radioactive background, and its excellent energy resolution and low threshold. Industrial applications of such a TPC include neutron measurements and radioprotection. The on-line signal analysis relies on determining the signal waveform parameters and the energy without degradation in precision, as well as performing particle identification. This implies the use of artificial intelligence (AI) on specifically developed electronics, which is the core of the proposed project. Innovations occur in three sectors: 1) Developing a versatile charge amplifier allowing to read every kind of low capacitance detector with different gain and a high resolution ($<$0.3% rms). 2) Creating an algorithm in an embedded system, estimating the best intrinsic signal inside the detector using the inverse problem method with AI. The signal processing will select data depending on the waveform based on Deep Learning methods. It will be a new way to filter signal in embedded systems and open a wide range of applications like image processing for scientific image application. 3) Changing the way to build the control & command system, based on the users' experience. The developments in these three sectors would open a wide range of applications such as industrial test benches for aeronautic and automotive companies, reducing development cost.

Since the starting of the project, a custom made low noise charge amplifier was developed and an embedded DAQ has been built, aimed at tagging events to label them in order to test AI algorithms. A DAQ software architecture, based on OSGI standard on windows, has been put in place to prepare the control and command of the system.

The learning phase and technical discussion is strongly related to the R&T THINK project described in section 6. There are also links with strategic initiatives (GPR, LabEx, cluster of excellence, EUR) : SFRI Infinity2, project GPR ORIGINES, AAP Intelligence artificielle 2021)

5.3 Future plans

One aim is to scale the number of channels from the R2D2 spherical gaseous detector and to build a DAQ board with a more powerful FPGA to hold all digital data processing in line with a single board. The expected outcome of the OWEN project is a complete open source framework to create embedded neural network models, a publication on the methodology of embedded AI in Zynq FPGAs and a publication on the working example with R2D2.

LP2I Bordeaux	Physicists	A. Meregaglia, C. Jollet, F. Piquemal
	Research engineers	F. Duillole, P. Hellmuth, A. Rebii
	Electrical engineers	R. Bouet
	Doctoral students	P. Charpentier
Subatech	Doctoral students	V. Cecchini

Tab. 5: People involved in the OWEN project.

5.4 Visibility, strengths, weaknesses

Embedded AI is one of the foreseen technologies to handle the high data stream from particles physics experiments. OWEN will build a methodology to develop embedded neural networks on Xilinx FPGAs to process detector waveforms. OWEN's approach is one option to add AI algorithms in instruments near the detector signal. It is also the possibility to learn how to build AI neural networks to process data from the R2D2 gaseous detector. The current version could be scaled to a greater number of electronic channels in the future.

OWEN is the first project with the objectives to develop AI embedded algorithms at the nearest place from the detector, compared to others approach which use NNs after the analog SNR optimization. If successful, it could demonstrate a new approach of numerical signal processing to measure physics quantities such as the energy and track of particles. Weaknesses are mostly related to the AI implementation. We run AI deep neural network algorithms to define the process to translate standard AI algorithms written in python to firmware algorithms for FPGAs. For this, we had to define a scoring mechanism and to analyze the performances depending on FPGA resources. The challenge is to understand what parameters, FPGA ressources we have to play with to manage to implement AI DNNs in FPGA. We have to conceive an FPGA function around the algorithm to be able to control and run it. We also have to test and find good neural network architectures in the zoology of them. This process is time consuming.

5.5 Resources

Table 5 shows the people involved in the OWEN project at LP2I Bordeaux. The project is financed by the IdEx Programme Emergence 2019 for two years (2019 - 2022) and supported by the R&T IN2P3 THINK project. After the demonstration of the concept with the Idex program, we are planning to ask for a follow-up budget from Idex to finalize the complete OWEN development for the complete R2D2 instrument. An ANR demand has been submitted and we are waiting for the phase 2 result.

6 THINK

The effectiveness of deep learning methods and neural networks in many fields is well known. These techniques are now beginning to be implemented in the field of particle physics with remarkable efficiency. They are a relevant answer to the limits encountered by classical algorithms in the face of the ever increasing luminosity in detectors such as the LHC. The increase in luminosity in physics detectors, due to background noise and pile-up phenomena, makes the task of recognition algorithms very complex. So more intelligence is required to discriminate real data from noise.

While DAQ systems without hardware-level trigger can process sophisticated algorithms on heterogeneous computing farms, it is difficult to adopt the same approach for experiments like ATLAS or CMS. One reasong being the closed geometry of the detector which makes it impossible to extract all the data to the outside at a reasonable cost. The notion of hardware triggering must therefore remain. Interesting perspectives emerge with more intelligent hardware triggers implemented without increasing the amount of hardware logic cells. This approach is based on neural networks. Until now, few developments of this type have been made. However, it should be noted that some teams are quite active in the CMS experiment and have obtained encouraging results [16]. These teams have successfully realized real time jet extraction by implementing neural networks on FPGA.

If FPGAs appear as a natural candidate for implementing a neural engine, some very interesting other solutions exist, deserving to be closely studied. We can mention neuromorphic chips from Brainchip including up to 1.2 million neurons connected by 10 billions synapses, or Massively Parallel Processor Array (MPPA) from Kalray embedding several hundreds of CPU cores in a single IC.

It is however difficult to understand at first how all these solutions compare with each other or even with traditional solutions used so far such as the implementation of neural networks on CPU or GPU. This is the goal of the THINK project : exploring the respective performances of a few physics cases on several hardware engines, respectively : FPGA, MPPA, neuromorphic chips and GPU. The comparison will be made on many aspects: performance but also costs, access to manufacturer information, learning curve, speed of implementation. The objective is to give a project manager the tools allowing him/her to reasonably select the best solution for his/her problem in the early stage of the project without conducting costly studies.

The project is supported by participants already involved in the implementation of neural networks on a hardware target. The results of the following projects contribute to the knowledge sharing: the OWEN project (see section 5), the AIDAQ project (see section 2) and the HGCNN project (see section 3). It has to be noted that neural implementations of AIDAQ on FPGA already outperform classical algorithms.

7 Summary

The French groups of the ATLAS, CMS and LHCb experiments as well as of the OWEN project historically have a strong involvement within real-time data processing. Both within the low level triggers implemented in hardware and within high level software triggers they have held various roles of responsibility and are major contributors to current upgrades (Run 3) and future upgrades (Run 4 and beyond) and proposed experiments. In particular, the novel developments of a high-level trigger with GPUs (LHCb) and processing AI algorithms on FPGAs (ATLAS, CMS, OWEN) is in line with the goal for moving towards heterogeneous computing solutions called for within both the European strategy for particle physics [3] and the Roadmap of the HEP Software Foundation [7]. In view of the computing challenges ahead, incremental changes and a simple extrapolation of methods used until now are not sufficient.

It is therefore crucial to strengthen and maintain the novel approaches developed within the French HEP community, as highlighted as major goal within working group 9 "Computing, algorithms and data" of the prospectives of IN2P3 [6]. Two of the priorities are the efficient usage of acclerators like GPUs and FPGAs and the continuous effort to lead in the domain of AI in science domains of IN2P3. Both of these are combined in the developments described in this document. To continue the success story of computing and data science at IN2P3, we rely on the ability to attract and train experts in the field to build long-term teams of both physicists and engineers.

Acknowledgements

Special thanks for helping with preparing this document goes to: G. Aad, J.-P. Cachemiche, F. Druillole, R. Le Gac, V. Gligorov, J.-F. Marchand, J.-B. Sauvan

References

[1] Framework TDR for the LHCb Upgrade II Opportunities in flavour physics, and beyond, in the HL-LHC era, LHCB-TDR-023-001 (2021). https://cds.cern.ch/record/2776420

[2] Intérêts technologiques de LHCb Upgrade II pour les laboratoires et pour l'IN2P3, Prepared for submission to IN2P3 (2022).

[3] The European Strategy Group, Document explicatif relatif à la mise à jour 2020 de la stratégie européenne pour la physique des particules, CERN-ESU-014, 2020. https://cds.cern.ch/record/2721050

[4] Physics case for an LHCb Upgrade II - Opportunities in flavour physics, and beyond, in the HL-LHC era (2019) arXiv:1808.08865

[5] LHCb Trigger and Online Upgrade Technical Design Report. https://cds.cern.ch/record/1701361/files/LHCB-TDR-016.pdf

[6] Exercice de prospective nationale en physique nucléaire, physique des particules et astroparticules, rapports des groupes thématiques. https://prospectives2020.in2p3.fr/?page_id=313

[7] A Roadmap for HEP Software and Computing R&D for the 2020s HEP Software Foundation. https://arxiv.org/pdf/1712.06982.pdf

[8] J.P. Cachemiche *et al*, The PCIe-based readout system for the LHCb experiment, JINST 11 (2016) P02013

[9] Cleland W E, Stern E G (1994) Signal processing considerations for liquid ionization calorimeters in a high rate environment, NIM A Volume 338:467-497

[10] ATLAS Collaboration (2017) Technical Design Report for the Phase-II Upgrade of the ATLAS LAr Calorimeter, CERN-LHCC-2017-018, ATLAS-TDR-027. https://cds.cern.ch/record/2285582

[11] Aad, G., Berthold, AS., Calvet, T. et al. Artificial Neural Networks on FPGAs for Real-Time Energy Reconstruction of the ATLAS LAr Calorimeters. Comput Softw Big Sci 5, 19 (2021)

[12] Duarte et al., Fast inference of deep neural networks in FPGAs for particle physics, JINST 13 P07027 (2018), arXiv:1804.06913

[13] Study of a spherical Xenon gas TPC for neutrinoless double beta detection https://arxiv.org/pdf/1710.04536.pdf

[14] Allen: A High-Level Trigger on GPUs for LHCb 10.1007/s41781-020-00039-7

[15] LHCb Upgrade GPU High Level Trigger Technical Design Report CERN-LHCC-2020-006. LHCB-TDR-021

[16] Fast inference of deep neural networks in FPGAs for particle physics arXiv1804.06913