ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 349-356

Hardware solutions for implementation of neural networks in High Energy Physics triggers

J.-C. Prévotet[1], B.Denby[1], P.Garda[1], B.Granado[1], C. Kiesling[2]

[1]Laboratoire des Instruments et Systèmes d'Ile de France, Paris, France
[2]Max Planck Institut für Physik, München, Germany

Abstract. Neural networks have been used as triggers in HEP for more than ten years, and continue to deliver promising results. In this article, we will give an overview of the triggering problem and present general neural online solutions retained by physicists to process data in High Energy Physics triggers. We will finally describe an FPGA implemented architecture dedicated to fast neural computations, taking advantage of massive parallelism in order to meet the tight timing constraints imposed by Level 1 neural triggers.

1 Triggering problem in HEP

In modern accelerators, particle bunches collide with very high frequency, typically of order of 100 ns (bunch crossing time), resulting in possible interactions to be recorded by the data acquisition systems. The products resulting from such interactions are measured in complex electronic detector systems (divided into subdetectors realizing a variety of detection techniques) where their associated circuitry translates physical quantities into a data flow. Trigger systems aim at reducing the data flow by filtering interesting physical events and, as a consequence reject background coming from the noisy environment.

A trigger scheme generally consists of several triggering levels. The data flow coming from the subdetectors is continously reduced to a degree that can be handled by the subsequent trigger level: Level i trigger collects data from level i-1 and rejects most of the unwanted events. Such triggering schemes are described in figure 1.

In level 1, the incoming information is available only to the individual subdetector with very limited access to information of the other subdetectors. Because of tight timing constraints only simple processing is performed. For example, thresholds on energies in a specific subdetector may be applied. Such triggers have a typical processing time inferior to 1 μs.

In level 2, more sophisticated algorithms are generally performed. These algorithms aim at grouping data coming from various subdetectors by exploiting

ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 349-356

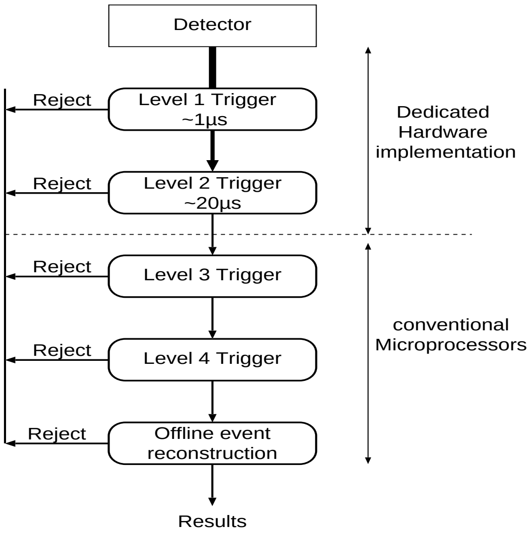

Figure 1: A multistep triggering scheme

correlations in order to build topological structures. Typical processing time for this level is 10-20 μs or more.

In levels 2, processing is generally distributed on processor farms running in parallel and destined to reconstruct the physical event, within a time of the order of 1s.

Generally, the first two levels are hardware triggers based on subdetector information available shortly after the interaction. Higher trigger levels are implemented in software since the processing time is less demanding.

2 Adopted solutions for triggering

2.1 Hardware technology

2.1.1 Programmable logic versus specific custom integrated circuits

There are currently many manufacturers offering a wide range of FPGA products, the architecture of which can differ substantially. Nevertheless, some points may be generalized : First, most of FPGAs based on memories consist of a matrix core of logic cells implementing both combinatorial and sequential logic. Second, all these cells are interconnected via a programmable bitstream according to the desired configuration.

FPGA devices are continuing to advance rapidly in both density and performances and seem efficient in a number of fields such as digital signal processing, communication, multimedia applications. Today's FPGAs offer the equivalent of 10 million system gates.

ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 349-356

Depending on the type and family of the FPGA device, other resources are added to the existing ones, targetting specific applications. For example, the VIRTEXII family of Xilinx contains embedded multipliers, which are appropriate to perform billion of MACs (Multiplication Accumulation) per second[1].

The flexibility linked to the ease of reprogrammability is of major interest since it allows a certain independance towards the application, compared to VLSI typical devices which impose important changes in their conception flow if a minor change occurs in the application. Another advantage of programmable logic compared to ASIC is its lower cost in terms of money and development time.

2.1.2 FPGAs and DSP processors

Despite their intrinsic technological difference, FPGA's and DSP processors target the same signal processing applications. DSP processors are widely employed because of their software oriented approach, which does not require specific knowledge in electronics engineering. Moreover, their internal speed make them potential candidates to many applications requiring massive iterative computations. Today's DSPs are very cheap and constitute a good answer to many problems in the signal processing world.

For several years, FPGAs have been a satisfactory alternative to DSP processors in signal processing. Despite their lower internal resources in terms of clock speed frequency notably, they have closed the gap by massively parallelizing the computations. At the same time, the implementation of applications in the circuit demands good knowledge of synthesis flow and of the technology of the device. This point is currently circumvented by utilizing predefined IP (intellectual property) modules performing algorithms and DSP functions. These configurable cores may be very easily integrated within a design.

2.2 Hardware in Neural networks triggers

2.2.1 Level 2 neural trigger

One example of the use of a neural trigger at level 2 is well illustrated in the H1 experiment at DESY which has been running for several years. More details about the trigger concepts of this experiment are given in [2]. The Level 2 total processing time is of 20 μs, half of it being dedicated to the neural network preprocessing, which consists of an intelligent way of grouping data from Level 1 [3]. $64\times64\times1$ MLP nets are implemented using CNAPS boards [4] whereas the preprocessing algorithms are distributed among several FPGAs on separate data distribution boards (DDB).

Another example of neural network implementation at Level 2 is described in [5]. It has been tested on simulated data from the LHC environment and consists of implementing a principal component analysis that performs dimensionality reduction of the input data space. The neural network is mapped onto a multiprocessor parallel machine, based on DSP boards and offers high processing speed compatible with the Level 2 timing constraints.

Basically, it can be assumed that today's programmable devices such as FPGAs or DSPs may be used since they can meet the timing requirements of this level. Moreover, the trend is for engineers to integrate more and more "intelligence" at this level, developing more complex processing such as powerful pattern recognition algorithms.

2.2.2 Level 1 neural trigger

With an incoming data flow 2 orders of magnitude greater than that of Level 2, specifications are very different for this level. For example, in the calorimeter trigger of ATLAS experiment, a decision has to be provided with a latency of about 500 ns, with a data flow coming into the trigger at every bunch crossing (25 ns). Until now, implementing digital circuits at this level was not considered since the constraints were too strong for the technology and only analog designs were studied [6]. Nevertheless, some digital solutions have been envisaged in [7] and consist in using RAM memories with preloaded contents to implement the neural net. This approach, implemented in the DIRAC experiment allows very fast processing (˜65 ns) but to the detriment of the precision and size which might be insufficient in other noisy environments.

Recent progress in digital technology allows to envisage utilizing digital processors at Level 1 and thus implement more powerful tasks with even more precision. The problem resides in the amount of data to be processed and imposes a massively parallel distribution of the computations.

3 A neural architecture for Level 1 triggering

Promising experimental results in Level 2 have naturally brought physicists to the idea of transposing Level 2 trigger concepts into Level 1. For example, architectures like MLPs (multi-layer perceptrons) have shown their robustness in Level 2 and seem appropriate to perform the same tasks at Level 1. A circuit is currently being designed at LISIF (Laboratoire des Instruments et Systèmes d'Ile de France), to match these expectations.

Although the vocation of the circuit is to be as general as possible, it aims at fulfiling the constraints imposed by the Level 1 trigger scheme in the ATLAS experiment. It would then integrate the actual Level 1 trigger [8],[9], enabling a conjugation of different processing methods.

As reported above, the main point consists of data arriving every 25 ns, corresponding to a bunch crossing in the LHC detector. Data are first filled into a pipeline and a decision is taken regarding the event in order to provide pipelined data to Level 2.

The circuit is designed to classify showers in a calorimeter and then discriminate electrons from hadrons and jets. A preprocessor sums analog signals from the calorimeter and performs analog to digital conversion to limit the number

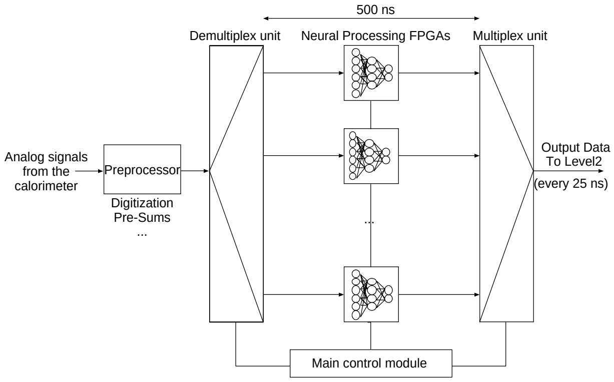

Figure 2: Level 1 trigger hardware

of inputs to the neural network. Signals above a certain threshold constitute towers and a local area around a maximum energy is furnished to the neural network whose goal is to identify the signature of a particle.

An estimated processing time of 500 ns has been adopted assuming that the total level 1 triggering time is 2.0 μs and that 1.5 μs are reserved for data transfer and preprocessing. Since the data arrive every 25 ns, a demultiplex unit distributes them sequentially to several identical circuits allowing then to parallelize the calculations thus reducing the processing time. A multiplex unit then distributes the processed data, delivering outputs to the level 2 trigger every 25 ns (Figure 2). Typical outputs of the neural processor consist of indicating whether an electron, a jet or a hadron has been found.

In the following, it is assumed that the neural network is a MLP with N_i inputs, N_h hidden units and N_o outputs units. The general idea of the architecture is based on a distributed computation of all neurons. N PEs (Processing Elements) work in parallel, each of them performing a part of the neurons' potentials computation. The PEs organization consists of a matrix of $p{\times}q$ elements and is summarized in Figure 3.

Each PE contains a MAC (Multiply Accumulate) unit, an address generation unit and its associated internal memory for weight storage. A control module generates different signals via a control bus and manages the entire PE. Additionnal registers are foreseen in order to store intermediate results. The weights are coded into 16 bits and activations into 8 bits, but these are generic values that are easily configurable. An overview of the PE architecture is given in Figure 4.

The neural network computation is performed in two steps. First, q out of the N_i inputs are provided at every clock cycle on the q input parallel buses. This enables to minimize the number of input data to be provided at each cycle, and simplifies considerably the circuit layout by reducing the number of I/O ports of the device. After N_i/q cycles, all inputs are then processed. Neuron parallelism is retained for the hidden layer computation: in this configuration,

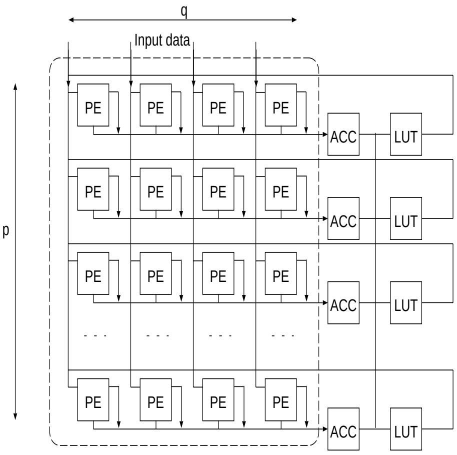

Figure 3: Distribution of the processing elements (PE)

the network slicing is equivalent to storing one row of the weight matrix within a row of PEs. All sums of products are then accumulated within a row to constitute the potential of the corresponding neuron. All potentials address in parallel p LUTs (Look Up Tables) in which the activation function is stored. The size of the table is configurable but a value of 256 is generally retained in most of applications.

In a second step, the activations of all hidden neurons are broadcasted horizontally to all PEs within a row. Each PE within a column computes a partial sum of an output neuron's potential according to a synapse parallelism and subresults are accumulated before addressing once again the LUTs in which the activation function is stored. The circuit is currently designed to implement a 128×64×4 MLP in 500ns. In this configuration $N_i = 128$, $N_h = 64$, $N_o = 4$, $p = 64$, $q = 4$. The number of 128 inputs has been chosen as an example and corresponds to a local area with a 8×8 granularity around a tower in both hadronic and electromagnetic part of the calorimeter. The 4 outputs may correspond respectively to the identification of the four possible particle flavors: electrons (photons),, tau-leptons, hadrons and jets.

The clock frequency of the circuit has been chosen to be 160 MHz since it corresponds to a multiple of the LHC clock frequency of 40 MHz. States are coded in 8 bits and weights in 16 bits which provides sufficient precision for many applications. These values are nevertheless configurable.

The major advantage of this architecture is its processing speed. It is obtained by distributing the tasks over many simple processing elements working in parallel on the same data. Another advantage of this architecture is its relative independance to the number of hidden units, since it is sufficient to add a new array of line processors to compute a hidden layer neuron. Finally, the architecture targets implementation in an FPGA device. XILINX VIRTEXII

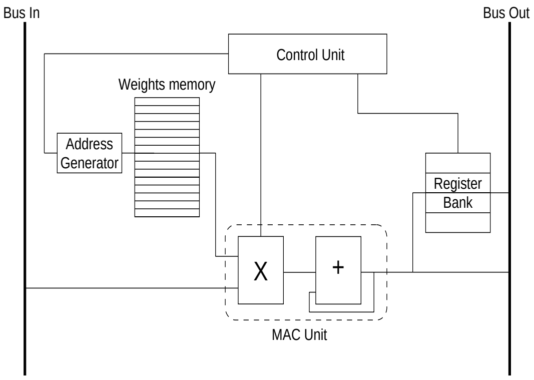

Figure 4: Overview of a Processing Element

devices have been chosen since they combine both speed and internal resources. Moreover the flexibility of this circuit makes it possible to adopt many configurations of neural networks with a variety of size and data precisions.

4 Conclusion

Neural Networks have proved their efficiency in HEP triggers as discriminators between background processes and the interesting signal. The online utilization of such networks have been implemented in hardware at different levels in the triggering process. Until now, digital technology did not seem to be able of processing the data flow within the time allowed for Level 1 and typical systems were mostly analog.

Today, recent progress in digital signal processing electronics have opened the way to extend Level 2 concepts in which sophisticated algorithms are already implemented into Level 1. This implies transfering more complexity near the subdetector, and consequently performing better event filtering. In order to illustrate this idea, we have presented an original architecture the goal of which is to provide high speed neural processing, that also takes advantage of the FPGAs' intrinsic flexibility.

References

[1] Xilinx Virtex II Databook, Xilinx Corporation, San Jose, CA., http://www.xilinx.com.

[2] J.H Koehne et al., "Realization of a second level neural network trigger for the H1 experiment," *Nucl. Inst. Meth.* A389, pp. 128, 1997.

[3] J.-C Prévotet et al. "Intelligent preprocessing for neural networks in the H1 experiment,"in *Proc. ACAT*, 2000, pp. 73-75.

[4] CNAPS system, Adaptative Solutions Company, Beaverton Oregon, USA, commercialized by Cromemco, CH 6340 Baar, Switzerland.

[5] E.S Caner et al. "Developing fast neural classifiers on a parallel processing environment,"in *Proc. New Computing Techniques in Physics Research VI*, 1999, pp. 40-44.

[6] Peter Masa et al., "70 Input, 20 nanosecond Pattern Classifier,"in *Proc. ICNN*, 1994.

[7] M.Steinacher "Hardware implementation of a fast neural network," *in Proc. New Computing Techniques in Physics Research VI*, 1999, pp. 156-165 .

[8] ATLAS Level-1 trigger Group, "ATLAS first Level trigger technical Design Report, ATLAS TDR-12, CERN/LHCC/98-14, CERN, Geneva(1998)

[9] G.Anagnostou et al. "The final multi-chip module of the ATLAS level-1 calorimeter trigger preprocessor" in *Proc. LEB2001*, to be published.