NEURAL NETWORK TRIGGERS FOR GLOBAL EVENT DECISION AT THE LHC

C. Kiesling, Max-Planck-Institut für Physik
Föhringer Ring 6, D-80805 München, Germany (email: cmk@mppmu.mpg.de)

Abstract
A global event decision hardware, suitable for the planned LHC experiments, is presented, based on the neural network architecture to be implemented typically at the second trigger level. A prototype for such a system is successfully operating in the H1 experiment at the HERA *ep* collider. The latency of the network triggers, which are of the feed-forward type, is about 20 microseconds. The inputs to the networks are suitably preprocessed quantities available at level 2. We describe the neural hardware and its use within the overall trigger strategy, exemplified with the H1 experiment. We discuss an interesting application of the network hardware for fast secondary vertex finding at the trigger level, useful for studies of b and top physics and searches for hadronic Higgs final states.

1 Introduction

It is expected that the extremely high interaction rates at the LHC (about 20 events on average per bunch collision, at 40 MHz) will pose a serious challenge to the trigger systems. High selectivity of the interesting physics processes, which are usually rare, is of utmost importance. Typically, the trigger systems have to reduce the rates of order 40 MHz down to a manageable figure of around 10-100 Hz, suitable for permanent logging to tape. Multi-level trigger systems for the large colliding-beam experiments ATLAS [1] and CMS [2] have been proposed, reducing the rate step by step with increasingly complex decision machines, to cope with this formidable task. In Atlas, the first two trigger levels will be realized in dedicated hardware, from the third level onwards general processors (software) will manage the data reduction, while the CMS collaboration plan to go directly to a general purpose processor farm after the first hardware level. We present here a trigger concept, designed for the second level, which is particularly well suited for the typical trigger task of recognizing and discriminating complicated event patterns in a multi-component detector system. The concept is based on the neural network technology, realized in dedicated hardware, which has convincingly demonstrated its potential in a concrete application within the H1 experiment at the HERA *ep* collider [3].

We summarize here the concepts and the technical realization of the H1 neural network trigger, details on the system can be found elsewhere (see, e.g. [3, 4]). Besides the superior global event decision capabilities à la H1, which can be naturally extended to the LHC experiments, we present a recent pilot study for a fast secondary vertex finder, based on the hit information from the Silicon trackers available at level 2.

2 Prototyping: The H1 Detector and its Trigger Scheme

Modern, large particle detector systems such as ATLAS or CMS at the LHC are designed with the intention to serve as general purpose facilities, pushing into a new kinematic frontier, and be prepared to be sensitive for the expected physics as well as potentially new phenomena. In this spirit, a large variety of detection principles have been foreseen, allowing for efficient detection and measurement of hadronic particles (jets) as well as of photons and leptons (elec-

trons and muons). This is no different for the running experiments at the HERA collider, which are presently probing a dramatically enlarged kinematic domain of deep-inelastic scattering.

For triggering the apparatus, H1 has installed a scheme of three levels, two hardware levels and one software level ("level 4"). An intermediate software level ("level 3") is provided, but not used at present. At level 1, each of the detector components (subdetectors) provides a set of triggers to a central trigger box, where they can be subjected to simple coincidence logic. Details on the H1 detector are given elsewhere[5]. The first level trigger is pipelined and does not generate deadtime (the same strategy is applied for the LHC experiments). This architecture implies that the level 1 trigger processors are able to provide a trigger decision for each bunch crossing (BC, a 96 ns interval for HERA, 25 ns for the LHC). For H1, the pipeline memory is 30 BC's long. After about 2.3 μs (24 BC's) a level 1 trigger decision is formed ("L1-keep") and the trigger information from all subdetectors is transferred to the level 2 systems. At this point the primary deadtime starts, no further triggers can be accepted until the event buffers are fully read out or a "fast clear" from the level 2 trigger system has been issued, rejecting the event. When the event is accepted by the level 2, the detector readout is initiated and the entire event information is sent to the level 4 processor farm, where a full event reconstruction is performed and the final event decision, using standard C-code programming, is taken.

At the second hardware trigger level in H1, the decision time is limited to 20 μs in order to digest a maximum of 1-2 kHz from level 1 while keeping the deadtime below 2 %. At level 2, the information from all level 1 trigger processors is available, so that "intelligent" use of this information is possible, exploiting the correlations among the various trigger quantities. The output of the level 2 trigger is around 50 Hz presently, which is the maximum input rate for the level 4 RISC processor farm (30 CPUs). The output rate of level 4 which is limited to about 10 Hz, is dumped to disk and then stored permanently to tape.

3 Principles of the H1 Neural Network Trigger

The H1 design of the level 2 neural network trigger is based on the empirical observation [6], that *small* nets trained for *specific* physics reactions, working all in parallel, are the most efficient and flexible way to use neural nets at the trigger level (compared to a single large net trained on all physics reactions simultaneously). Most importantly, putting these nets to a real trigger application, the degree of modularity is extremely helpful when a new trigger for a new kind of physics reaction is to be implemented: there is no need to retrain the other nets, the new physics net is simply added to the group of the others.

The present strategy of using the networks is the following: Each of the networks is trained for a specific physics channel and is coupled to a set of level 1 subtriggers, particularly efficient for that physics channel. Because some of the level 1 subtriggers need to be made sufficiently relaxed to be efficient, their rate is usually unacceptably high. The level 2 trigger therefore has the task to reduce the excess background rate in these subtrigger sets while keeping the efficiency for the chosen physics channel high. At present, 12 networks are running in parallel, mostly optimized for electro- and photoproduction of vector mesons, which are difficult to separate from the background at level 1.

4 L2 Trigger Hardware

According to the principles described above, the hardware realization for the neural network trigger is modularized as follows (see [3]): Receiver cards collect the incoming trigger information of the various subdetectors and distribute them via a 128 *bit* wide L2 bus to preprocessing units, called *Data Distribution Boards* (DDB). Each DDB is able to pick up a freely chosable set of items from the L2 input data stream. The DDB can perform some basic operations on the items (e.g bit summing) and provides an input vector of maximally 64 8*bit* words for its companion CNAPS/VME board. Controlling and configuring of the complete system is done by a THEMIS

VME SPARCstation, which is located in the CNAPS crate (see below).

4.1 The CNAPS board

The algorithms calculating the trigger decision are implemented on VME boards housing the CNAPS 1064 chip [7] (see fig. 1). It is a parallel fixed-point arithmetic computer in SIMD architecture. The CNAPS-1064 chip (also called *array*) houses 64 processor nodes (PN). Up to eight chips (512 PNs) can be combined on one VME board. A PN is a processor for itself except that it shares the instruction unit and I/O busses with all the other PNs. An on-chip instruction unit handles the command and data flow. The commands are distributed via a 32 *bit* PN command bus. The 8 *bit* wide input and output busses are used for the data transfer to and from the CNAPS array. A direct access to these I/O busses is realized with a mezzanine board developed at MPI Munich. Through the mezzanine board the input vector is loaded into the CNAPS chip and the trigger result is sent back to the DDB. For synchronization reasons the CNAPS boards are driven with an external clock at 20.8 MHz (2 times the HERA clock frequency of 10.4 MHz).

The main internal parts of the PNs are arithmetic units like adder(32 *bit*) and multiplier(24 *bit*), logic unit, register unit, 4K memory and a buffer unit. Calculations are done in fixed-point arithmetic with chosable precision. The sigmoidal transfer function is implemented via a 10 *bit* look-up table (*LUT*) on chip. A full net with 64 inputs, 64 hidden nodes and 1 output node can be computed in 8 μs at 20.8 MHz, or in 166 clock cycles. To get the same speed with a single conventional CPU one would have to clock it at several GHz.

4.2 The Data Preprocessing

The Data Distribution Board (DDB) resides in a special *"L2 VME crate"* equipped with the L2 Bus, an 8 times 16 bit parallel data bus running with the HERA clock speed in an interleaved mode, yielding an effective 20 MHz transfer rate. For each subdetector the level 1 data, which are a quite heterogeneous such as

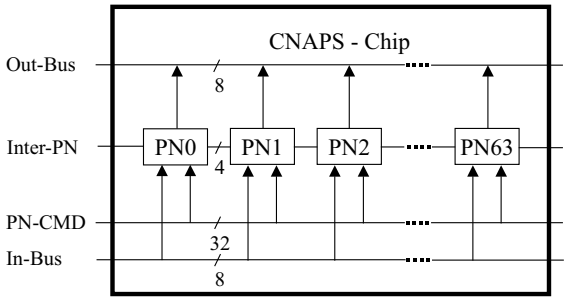

Figure 1: *Architecture of the CNAPS 1064 chip: It realizes an array of 64 fixed-point arithmetic processors, operating in the Single Instruction Multiple Data (SIMD) mode. Data are clocked in serially with 8 bit width ("in-bus") and distributed to the processing nodes. The weights for the multiplication step are stored in on-chip memory. The "out-bus" transfers the results for optional further processing (e.g. for a look-up table to emulate a sigmoid function).*

calorimetric energy sums, tracker vertex histograms, tracker rays (bits in the $\theta - \phi$ plane), bit-coded muon hit maps etc., are sent serially onto one of the eight subbusses of the L2 backplane. For system control purposes, a special monitor board ("spy") with an independent readout of the data transmitted over the L2 bus is residing in the same crate.

On the DDB, the L2 data received are passed through a data type selection where they can be transformed (e. g. split into bytes or single bits) using a look-up table. After bit splitting, several preprocessing algorithms like summing of bits and bytes, bit selections or general functions (look-up) can be applied. The data may, of course, also be sent unchanged to the selection RAM, where the input vector for the neural network computer is stored. Through the use of XILINX 40XX chips the hardware can be flexibly adapted to changes, e.g. for new data formats in the received input. Using selection masks the data are transmitted via a parallel data bus from the RAM to a mezzanine receiver card directly connected to the local data bus on the CNAPS board. Figure 2 shows the hardware of the neural network trigger operating in H1-experiment. The upper crate

houses the CNAPS neuro-computer boards and the VME control computer (at the left). The lower (9U) crate houses the receiver units for the L1 information (at the right) and the data distribution boards. Each board is associated with its neuro-computer, connected via cable between the backplanes, carrying the preprocessed network input to the CNAPS board.

The system described above can be translated into an LHC environment without any principal problems: Also there, partial (coarse granularity) detector data are available for the level 2 and these informations can be used in full correlation to arrive at highly selective triggers. The main advantage of using neural nets as opposed to standard multi-purpose processors (using a high level computer language) is their inherent speed. It seems also that complex high-dimensional correlations - in absence of known apriori algorithms, which usually is the case at the trigger level - can be exploited in an optimal way with the adaptive methods provided by neural networks.

5 Vertex Finding with Neural Nets

An interesting field not yet fully exploited by neural networks is the tracking area. Here, the basic event pattern is provided by an ensemble of hits from the tracking detectors (two- or three-dimensional information), which give already a pretty clear "view" of the event origin and basic kinematic features. Of particular interest is the information from high precision silicon trackers, which are used offline to find secondary vertices in the events, i.e. to tag heavy flavor decays. Bottom-quarks, e.g., provide a unique signature for many known and new phenomena expected at the TeV scale such as top quark physics, Higgs searches and supersymmetry, in addition to the physics of B- hadrons themselves. By finding the secondary vertex from the B hadron decay at the trigger level, using the fine granular silicon strip information available at level 2, one is in principle sensitive to all B hadron decays, not only to the semileptonic decays considered in present trigger schemes [1].

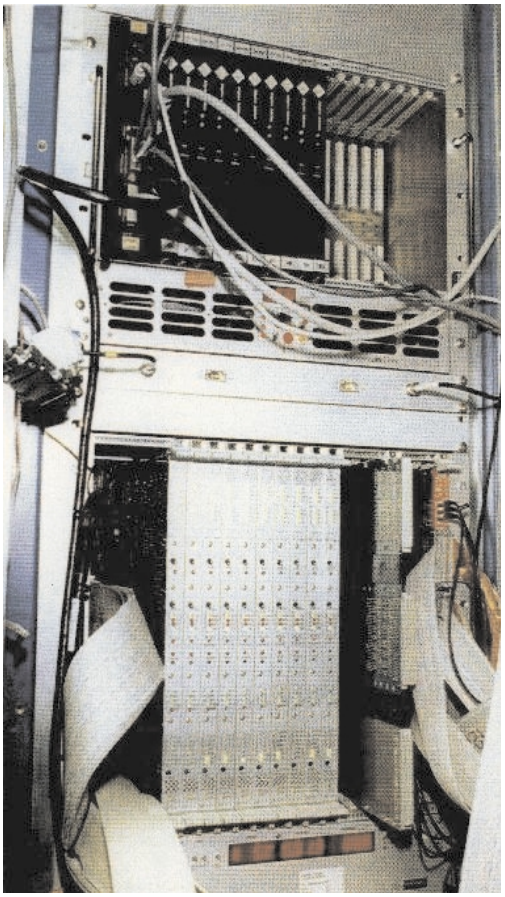

Figure 2: *View of the hardware of the H1 neural network trigger (see text).*

As a pilot study we investigated the potential to "reconstruct", at the trigger level, a possible secondary vertex associated with a certain set of tracks. In order to facilitate the network training in this initial study, the silicon strip information to be used was limited to the regions of interest (ROI) defined in the first level calorimeter trigger (the ROI strategy will be employed for the second level triggers at ATLAS). Typical decay lengths for charmed mesons expected in the LHC energy range are several centimeters, whereas the precision of space points for a modern Si strip detector is around 20 microns. This

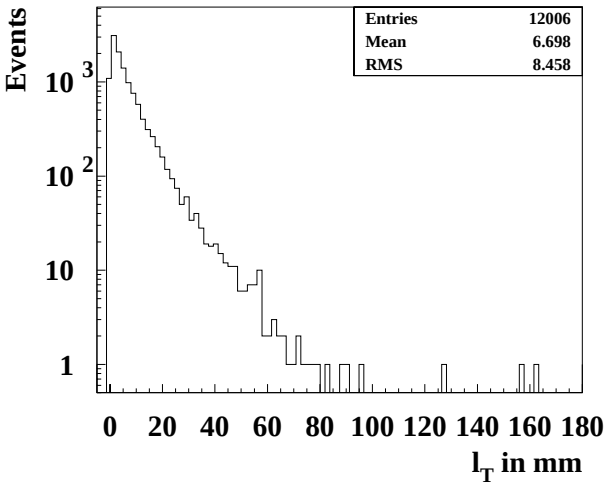

Figure 3: *Distribution of the decay length of D mesons originating in the decay chain of Top and Higgs ($m = 200$ GeV/c^2 decays) from proton-proton interactions at LHC energies.*

gives a ratio for vertex distance to space point precision of roughly 100:1. This ratio seems sufficiently large to try fast secondary vertex recognition with the space hits alone as inputs to a neural net. No attempt was made to form tracks from the space points.

Using Monte-Carlo simulation, hits were generated in the various layers of the Si-detector of ATLAS, originating from D^* decays which where given distributions in momentum space according to the expectation from top and Higgs ($m_{\mathrm{Higgs}} = 200$ GeV) production. A typical decay length distribution for the D mesons (coming from the prompt D^* decays) is shown in fig. 3. Without attempting any conventional track reconstruction method, which would be forbiddingly long for a hardware trigger application, the hit pattern originating from the tracks within a cone around the ROI was given to a feed-forward network with 2 hidden layers (typically 20 nodes in each layer) and one output node. The task of the linear output node was to estimate the decay length associated with the hit pattern. Each input to the network is a 1 bit number, either 0 (no hit) or 1 (hit), ordered in rising ϕ coordinate, mimicking a realistic readout scenario. The nets were trained with a backpropagation algorithm with a target output value equal to the decay length of the simulated $D \to K\pi$ decay. The training was controlled by an independent sample of events, yielding the result shown in fig. 4, where the difference between the target value and the decay length estimated by the network is plotted. A resolution of $\sigma \sim 1.6$ mm is achieved in the simple case of only two tracks (originating from the $D \to K\pi$ decay).

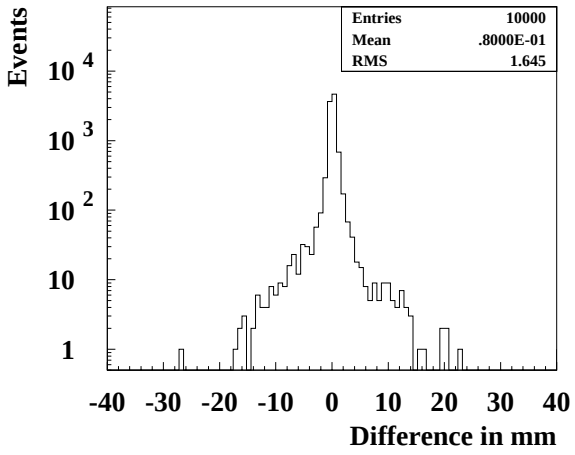

Figure 4: *Difference between the true decay length and the value estimated by the neural network for $D \to K\pi$ decays. The results shown satisfy the condition that no other track except the two decay products from the D meson are present in the search sector.*

In order to investigate the stability of the nets against noise, we added additional tracks coming from the primary vertex with corresponding hits to the neural input and repeated the training. As expected, the performance of the networks, keeping the network architecture the same, deteriorated progressively with increasing number of additional tracks. Doubling the number of tracks from the primary vertex resulted in an increase of the resolution for the decay length by about a factor of two ($\sigma \approx 3$ mm). Further details can be found elsewhere[8].

Further studies are required with increased complexity of the network architecture (more nodes in the hidden layers) in order to improve the perfor-

mance in presence of many fake tracks close to the two tracks originating from the secondary vertex. In addition, the possibility of more than two tracks coming from the secondary vertex have not been attacked yet. The preliminary results, however, are quite encouraging and demonstrate the ability of the neural networks to extract the relevant information (presence of a secondary vertex) from a complex hit pattern without explicit track reconstruction. Such an algorithm would be extremely fast and well suited for an application at an early trigger level.

6 Conclusions

We have presented the principles and the hardware realization for the second level neural network trigger, operational in the H1 experiment at HERA since 1996 as a global event decision machine. Based on commercially available, massively parallel digital ULSI neural network chips, a 20 μs decision time is achieved for the network trigger. The network inputs are derived from the trigger information provided by the various level 1 trigger systems and are preprocessed by custom-designed hardware.

For the physics data taking at the LHC, e.g. in ATLAS or CMS, such a trigger system could be readily adapted. Depending on the field of application (level 1 or level 2) the preparation of the network inputs needs special attention, both on the electronics level as well as in the successful training of the networks. Besides the proven strength of neural networks in pattern recognition tasks, one of the main motivations of using the neural approach in a trigger application at the LHC is the speed provided by dedicated neuromorphic hardware executing the inherently parallel computations of the neural algorithms.

As a promising physics application we have studied the neural networks ability to recognize secondary vertices in an ensemble of tracks, based on the information from the simulated ATLAS Si-detector. Feedforward networks were trained to estimate the decay lengths of D mesons from top and Higgs decays, given solely the information from the Si-hits in the various layers of the detector. It is found that the resolution for the decay length is adequate (order of a few mm) to recognize a secondary vertex, thus being able to tag charmed and beauty decays with high efficiency.

References

[1] ATLAS-Collaboration, Technical Proposal, CERN/LHCC/94-42, LHCC/P2, December 1994

[2] CMS-Collaboration, Technical Proposal, CERN/LHCC/94-38, LHCC/P1, December 1994

[3] J.H. Köhne et al., Nuclear Instruments and Methods in Physics Research A **389**(1997)128

[4] C. Kiesling et al., *The H1 Neural Network Trigger*, to appear in the Proceedings of the VI International Workshop on Artificial Intelligence in High Energy and Nuclear Physics, Herakleion, Crete, April 1999
S. Udluft et al., *The H1 Neural Network Trigger - Training and Monitoring*, ibid.
L. Janauschek et al., *Artificial Neural Networks as a Level 2 Trigger at the H1 Experiment - Performance Analysis and Physics Results*, ibid.

[5] I. Abt et al. *The H1 detector at HERA*, Nuclear Instruments and Methods in Physics Research A **386**(1997)310

[6] C. Kiesling et al., *A Neural Network Second Level Trigger for the H1-Experiment at HERA*, Contributed paper to the Lepton - Photon Conference, Beijing, August 1995. See http://wwwh1.mppmu.mpg.de/ projects/ neuro/ publications.html

[7] CNAPS Release Notes 2.0, Adaptive Solutions, Inc. , Beaverton Or. (1993)

[8] M. Schied, *Untersuchungen zur Erkennung sekundärer Vertices in hochenergetischen Teilchenreaktionen mit Hilfe Neuronaler Netze*, diploma thesis, Ludwig-Maximilians-Universität München, July 1998