The neurochip TOTEM: a case study in HEP.

S. Dusini [a,b], F. Ferrari [a,b], I. Lazzizzera [a,b,2], P. Lee [d],
A. Sartori [a,c], A. Sidoti [a,b], G. Tecchiolli [a,c] A. Zorat [a,b]

[a] *INFN - Sezione di Padova, Gruppo Collegato di Trento - Trento - Italy*

[b] *Università di Trento - Trento - Italy*

[c] *Istituto per la Ricerca Scientifica e Tecnologica - Trento - Italy*

[d] *University of Kent - Canterbury - United Kingdom*

Abstract

It is being proved that the neurochip TOTEM is a viable solution for high quality and real time computational tasks in HEP, including event classification, triggering and signal processing. The architecture of the chip is based on a "derivative free" algorithm called Reactive Tabu Search (RTS), highly performing even for low precision weights. ISA, VME or PCI boards integrate the chip as a coprocessor in a host computer. This paper presents: 1) the state of the art and the next evolution of the design of TOTEM; 2) its ability in the Higgs search at LHC as an example. Reference number: 303

Keywords: Neural networks, Processors, Computer Arithmetic.
PACS: 07.05.Mh 55.40.-e 84.35

1 Introduction

Neural networks implemented as VLSI hardware are being considered as good candidates to solve problems of time-critical and high quality performance pattern recognition in High Energy Physics (HEP) [1–3]. The main benefit is speed, because of the massive parallel architecture. A cost is usually a very complex architectural structure, since common algorithms such as backpropagation, being derivative-based, require high precision computation[4].
To gain significant improvement in this respect, Battiti and Tecchiolli devised a "derivative-free" algorithm in the context of a novel approach to the training problem, which is first transformed into a *combinatorial optimization* task, then solved by means of the heuristic method called *Reactive Tabu Search* (RTS)[7,8].

[1] Work supported by Istituto Nazionale di Fisica Nucleare (INFN)
[2] Corresponding author: Dipartimento di Fisica, Univ. Trento, I-38050 Povo (TN)
Tel:+39-461-881551, fax:+39-461-882014, e-mail: lazi@abacus.science.unitn.it

RTS is based on the construction of *search trajectories* in the space of the binary strings of length $L = N * B$, into which N weights, needed to configure a neural network, are suitably coded using B bits per weight. The search is intended to locate the best "suboptimal" minimum on a *cost surface* by means of a sequence of elementary moves, each consisting of a single bit-flip in the string of weights. When a move is done, its inverse is forbidden for a *prohibition period* of T successive steps (the Glover's *Tabu Search* method[10]), allowing some amount of *diversification* in the search process. RTS remarkably enhances such diversification by dynamically adjusting the parameter T through a simple mechanism that evaluates and *reacts to* the current local shape of the cost surface. This way it escapes rapidly from local minima and cyclings and finds solutions even for low precision weights, moreover quite independently from any starting point.
Sect. 2 is devoted to a description of the neurochip TOTEM; Sect. 3 presents a new architectural design; Sect. 4 gives the results of a sample application, namely the extraction of the Higgs events from background in simulated data at LHC energies.

2 The TOTEM chip

TOTEM is a full-custom chip designed to operate as a co-processor in a host system, carrying out the most compute-intensive operations for RTS[9]. ISA and high performance PCI and VME boards have been developed to set the coupling. The chip includes an array of 32 parallel processors with associated on-chip weight memory and control logic with broadcast and output buses.
Pipelinings are included to speed up operations. A 32-bit static storage register on the output of the MACs allows data transfer from the neurons of a layer in a MLP net to occur concurrently with a parallel input-multiply-accumulate operation on all the processors. The memory depth of 128 8-bit words allows neurons with up to 128 inputs to be implemented. Because of the sequential access to the weights, the chip can realize different MLP topologies with a high degree of flexibility: the memory bank can either be assigned to a single neuron or be partitioned among neurons on different layers. The sigmoid function is implemented off-chip by a RAM-based look-up table. Up to four chips can be paralleled in each layer of a network.
With 250,000 transistors on a 70 mm^2 die manifactured in a 1.2 μm CMOS technology, TOTEM performs 1 GMAC/sec when clocked at 30 MHz.
A doubling in the processor density and higher operational speed will be obtained by the transition to a 0.8 μm CMOS technology currently in progress.

3 Advances in the design of TOTEM and the *plog* encoding

Considerable percentage of the silicon real estate of the TOTEM chip had to be devoted to the multiplication units and to the memories where the weights are stored. Both areas will be reduced by means of the already mentioned technological migration. However in the case of the multipliers an alternative and complementary approach can be considered: going to a *logarithmic representation* of the feature inputs and the axon weights, so that multiplication is replaced by addition, which is cheaper in terms of silicon area requirements and allows both lower power consumption and faster clock rate. Some of the authors (P. Lee, I. Lazzizzera, A. Sartori, G. Tecchiolli and A. Zorat) are exploring this way indeed[5]. The first problem is to find a reasonable approximation to the *bin-to-log* and *log-to-bin* conversion, since they are quite expensive[6]. If one defines for any positive real number x the functions $\eta(x) = n \in \mathbf{N}$ such that $x \in [2^n, 2^{n+1}[$ and $plog(x) = \eta(x) + x/2^{\eta(x)} - 1$, one gets an approximation of $log_2(x)$ that has a maximum error of only 0.0861. When $x = \sum_{i=0}^{W-1} b_i 2^i$ is a binary encoded positive number with $W = 2^w$ bits, then evidently $b_i = 0$ for $W - 1 \geq i > \eta(x)$ and $b_{\eta(x)} = 1$. Writing $plog(x) = \sum_{i=-f}^{w-1} p_i 2^i$, it follows that the bits $p_{w-1}, p_{w-2}, \ldots, p_0$ are the binary encoding of $\eta(x)$ and the bits $p_{-1}, p_{-2}, \ldots, p_{-f}$ are given by $b_{\eta(x)-1}, b_{\eta(x)-2}, \ldots b_{\eta(x)-f}$ respectively. Clearly f is an integer parameter stating a truncation (quantization error).

This way one gets the basis of a sign-magnitude, fixed-point *plog-encoding* of an integer $x \in [-2^W - 1, 2^W - 1]$ with $1 + w + f$ bits: 1 bit (given by b_W) for the sign; w bits encoding $\eta(x)$ (the integer part); f bits (given by $b_{\eta(x)-1} b_{\eta(x)-2} \ldots b_{\eta(x)-f}$) for the fractional part ($x = 0$ is coded in a particular way). The total error includes the quantization ($\sim 2^{-f}$) and the approximation to log_2 (≤ 0.0861): it amounts at most to a 10%. When applying the *plog* encoding to neural nets, the multiplier stage of a neuron is replaced by an adder and a plog-to-bin unit. In such a *plog* based architecture the RTS training method for a Multi Layer Perceptron (MLP) is applied without modifications. The above estimated error of at most 10% turns out to be the same that TOTEM owes to a 4-bit weight setup in the conventional multiplier architecture: with such low precision weights, however, adequate solutions for many problems[8] are still obtained. The point is that, assuming the same fabrication technology, the area of the multiplication blocks are reduced by a factor 10, with a reduction in power consumption by a factor 12 and an increase in computational speed by a factor 3. The reduction in power consumption can be exploited to increase both the number of processors per die and the operating speed. These figures pave the way to new implementations with high performance factors at the same fabrication costs. As an example the fabrication of a neural processor hosting hundreds of neurons running at a reasonable 100 MHz clock rate is feasible within a couple of years. With such a processor, triggering tasks requiring neural nets of approximately 100 neurons can easily substain input rates of the order of 10^7 events per second, thus making its use possible even in the most time critical experiments, such as LHC.

4 Higgs search: observables and performance of TOTEM

TOTEM has been tested in the discrimination of Higgs events from background at LHC energies using simulation data obtained by the PYTHIA/JETSET Monte Carlo code.

Arbitrarily we assume the Higgs mass to be $M_H = 200\ GeV/c^2$, just above the threshold for the creation of two real Z's[12]. In this case the dominant production mechanism is the gluon–gluon fusion and the best decay channel for its identification is the so–called *gold plated channel*:

$pp \to HX \to ZZ \to 4\mu X.$

whose cross section is $2.84 \times 10^{-12}\ mb$ as computed by the Pythia MC code. We provide the two expected main backgrounds according to the actual top quark mass ($M_t = 175$ GeV[14]):

$p\,p \to t\,\bar{t}\,X \to \mu^+\mu^-\mu^+\mu^- X'$

with 4 muons produced by semileptonic decays of the top and antitop;

$p\,p \to Z^0\,b\,\bar{b}\,X \to \mu^+\mu^-\mu^+\mu^- X'$

with a muon pair produced by Z^0 decay and the other one by semileptonic b and $\bar{b}$ decays. These two backgrounds have cross-sections respectively of $7.84 \times 10^{-9}\ mb$ and $6 \times 10^{-9}\ mb$ as computed by the Pythia MC code.
We order the final muons according to the magnitude of their transverse momenta and use the following ten variables as physical observables :
($X_1 - X_4$) the transverse momentum of the four muons;
($X_5 - X_8$) the invariant masses of the four $\mu^+\mu^-$ pairs;
(X_9) the four muons invariant mass;
(X_{10}) the hadron multiplicity of the hard jets, discriminated according to the $K_\perp$ *Clustering algorithm* for hadron-hadron collisions [13].

TOTEM has been trained using a sample of 4000 Higgs events, mixed with 2000 of each of the backgrounds. The test set, totally different from the training one, consisted of $N_H = 2000$ Higgs events mixed with about 360,000 $t\bar{t}$ and 270,000 $Zb\bar{b}$ event samples (thus respecting only the ratio between the cross sections of the two backgrounds). Some results are listed in Table 1, where N_H^c is the number the events correctly classified as Higgs, N_B^c is the number of the events wrongly classified as Higgs and δ is the interval amplitude within which the classification of an Higgs is assumed certainly correct, in units corresponding to 1/8192 of the gap between the *truth value* of an Higgs event and the *truth value* of a background event. Efficiency (N_H^c/N_H) and purity ($N_H^c/(N_H^c + N_B^c)$) are also shown, linearly extrapolated to a number of background events in a ratio with 2000 Higgs events as required by the exitimated cross sections above.

δ	N_H^c	N_B^c	$eff.$	$extrap.\ pur.$
1	753	31	0.37	0.61
2	1102	55	0.55	0.56
5	1228	95	0.61	0.45
10	1498	172	0.74	0.36
100	1863	848	0.93	0.12

Table 1

References

[1] B. Denby, *The use of Neural Networks in High Energy Physics*, in Neural Computation **4**(5) 1976

[2] R.K. Boch, I. Carter and L.C. Legrand, ATLAS/DAQ-No-11 EAST 94-08, CERN (1994)

[3] Th. Linblad et al., *Nucl. Instrum. Methods*, **356** (1995) 498.

[4] R. Battiti and G. Tecchiolli, Learning with first, second, and no derivatives: A case study in high energy physics. *Neurocomputing* **6** (1994) 181–206.

[5] P. Lee, I. Lazzizzera, A. Sartori, G. Tecchiolli and A. Zorat *Nuclear Instruments & Methods in Physics Research A* in print.

[6] R. De Mori, R. Cardin, A new design approach to binary logarithm computation, *Signal Processing*, **13**(2), Sept. 1987, pp. 177–195.

[7] R. Battiti and G. Tecchiolli, The Reactive Tabu Search, *ORSA Journal on Computing*, **6** (2) (1994) 126-140.

[8] R. Battiti and G. Tecchiolli, Training Neural Nets with the Reactive Tabu Search, *Tech. Rep. UTM 421*, Dip. di Matematica, Univ. di Trento - Italy. Shorter version to appear in *IEEE Transactions on Neural Networks*

[9] G.Anzellotti et al., *J. of Mod. Phys.C*, **6** (1995) 555-560

[10] F. Glover, *ORSA Journal on Computing*, **1**(3), pp.190-206 (1989)

[11] C. R. Baugh and B. A. Wooley, *A two's Complement Parallel Array Multiplication Algorithm* IEEE Transactions on Computers **C-22** (12) 1045-1047.

[12] M. Lüscher, P. Weisz *Nucl. Phys.* **B290** (1987) 5; *ibid.* **B295** (1988) 65; *ibid.* **B318** (1989) 705

[13] S.Catani, Y.L.Dokshitzer, M.H.Seymour, B.R.Weber, *Nucl. Phys.* **B406** (1993) 187

[14] M. Mangano and T. Trippe (of the Particle Data Group), in *Review of Particle Poperties*, Phys. Rev. D54 (1996) 309.